WASP WINTER CONFERENCE 2022

Poster Catalogue

AI MLX

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

WASP WINTER CONFERENCE 2022
POSTER CATALOGUE 2/4

AI MLX

WASP WINTER CONFERENCE 2022
POSTER CATALOGUE 2/4

AI MLX

Ahmadian, Amirhossein
Linköping University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Likelihood-free Out-of-Distribution Detection with Invertible Generative Models

Likelihood of generative models has been used traditionally as a score to detect atypical (Out-of-Distribution, OOD) inputs. However, several recent studies have found this approach to be highly unreliable, even with invertible generative models, where computing the likelihood is feasible. In this paper, we present a different framework for generative model--based OOD detection that employs the model in constructing a new representation space, instead of using it directly in computing typicality scores, where it is emphasized that the score function should be interpretable as the similarity between the input and training data in the new space. In practice, with a focus on invertible models, we propose to extract low-dimensional features (statistics) based on the model encoder and complexity of input images, and then use a One-Class SVM to score the data. Contrary to recently proposed OOD detection methods for generative models, our method does not require computing likelihood values. Consequently, it is much faster when using invertible models with iteratively approximated likelihood (e.g. iResNet), while it still has a performance competitive with other related methods.

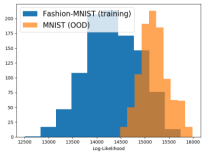
Ahmadian, Amirhossein
Linköping University

Likelihood-free Out-of-Distribution Detection with Invertible Generative Models

Amirhossein Ahmadian and Fredrik Lindsten
Department of Computer and Information Science, Linköping University

WASP Winter Conference 2022 / Published at IJCAI 2021

Introduction



iResNet loglikelihoods on two datasets
In-distribution: FashionMNIST (blue)
OOD: MNIST (orange)

- *Out-of-Distribution (OOD) Detection* : **detecting inputs that are considerably different from the data on which a machine learning model is trained** (anomalies, novelties). Generative models can be used to solve this in a completely unsupervised setting.
- **Generative models may assign much higher likelihoods to OOD test data than in-distribution data**. Also, **likelihood can be too costly to compute.**
- We propose a method to detect OOD inputs using **pretrained invertible generative** models that **does not need to compute the model likelihood** (the log-determinant term). It is based on fast computation of very **low-dimensional statistics** (features) using the invertible model and an **image complexity measure**, and employing a **One-class SVM (OSVM)** in this new feature space.

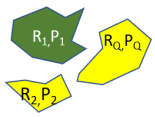
OOD Detection: Similarity Score and Null Hypothesis Testing

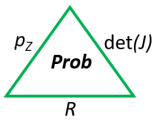
$$u(x) = \sum_{x_i \in D_{in}} \phi_i \exp\left(\frac{-||T(x) - T(x_i)||^2}{\sigma_0^2}\right) - \rho$$

OSVM is trained on in-distribution data in the space of statistics T, to obtain {Φ,ρ}

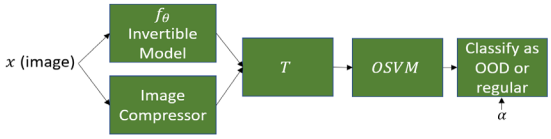
- Idea: OOD Detection should be based on measuring a **similarity/distance between the input and in-distribution (training) data**. We use a one-class SVM (**OSVM**) to define **a score function** measuring this similarity of interest.
- The representation space is particularly important. **Features called 'statistics' are extracted from the input (image) and the generative model state**. Hence, **the generative model is involved indirectly through the statistics**, and not its likelihood.
- Classical **null hypothesis testing** is used to adjust the threshold on similarity, given the training data and a significance level.
- The idea of using statistics with OSVM is also in **Density of States Estimation (DoS)** [Morningstar et al,2021]. However, we use different statistics that are faster to compute, while having a competitive performance. Our theoretical motivation is also different.

OOD Statistics (Features) for Invertible Models





- *Simple but neglected fact*: the **probability** of a class under a generative model depends both on the model **likelihood** and the **region** (particularly the volume) corresponding to the class.
- With **invertible models**, the probability is a function of: **prior density, determinant of encoder Jacobian, integration region**. Some are costly or impossible to obtain, but we may use *any function related to them* as our *statistics*.
- We replace the determinant term, which can be costly to compute, by the **average of column-sum of the encoder Jacobian matrix**, that is another function of Jacobian matrix.
- Heuristic: more **complex inputs** are likely to come from a region with high volume. Assuming image data, we use the **compression size of image using FLIF format** as a proxy for the **region volume**.



Performance Results and Comparison

- Evaluated on some popular OOD detection tasks such as *FashionMNIST vs. MNIST* and *CIFAR-10 vs. SVHN* with the models *Glow*, *iResNet*, and *ResFlow*, and compared to 4 other methods in terms of **Area Under ROC curve (AUROC)**
- Two combinations of the following statistics:

$$T_1 = log\ p_Z(f(x)), T_2 = C_{FLIF}(x),\ T_3 = \text{mean}|\textstyle\sum_i J_{ij}|$$

where *f(.)* is the encoder of invertible model, and *J* is its Jacobian matrix.

- Our method is **competitive with the other ones overall**. It **ranks first in 5 cases**, although it is outperformed by another method in 2 cases. It is better than S-Score [Serra et al,2020] in most cases, which also uses FLIF image compression size complexity measure.
- Most near OODs, e.g. CIFAR-10 vs. CIFAR-100 are not detected well by any of the methods

Model/Data	T_1,T_2	T_1,T_3	DoS	S-Score	Simple LL
iResNet trained on MNIST					
OOD: FashionMNIST	0.99	0.99	0.99	0.97	0.99
iResNet trained on FashionMNIST					
OOD: MNIST	**0.96**	0.89	0.88	0.95	0.07
OOD: Vertical Flipped	0.62	0.60	0.63	0.66	0.55
ResFlow trained on CIFAR10					
OOD: SVHN	**0.96**	0.92	0.94	0.89	0.10
OOD: Vertical Flipped	0.50	0.54	0.52	0.54	0.51
Glow trained on CIFAR10					
OOD: SVHN	0.96	0.91	0.95	0.88	0.09
OOD: CIFAR100	0.57	0.56	0.57	0.49	0.52

Computational Cost Advantage

Model/Data	Log-likelihood Approximation	T_3 statistic
iResNet - MNIST	209ms	3ms
ResFlow – CIFAR10	378ms	20ms
GLOW-CIFAR10	-	18

T_2 statistic: 13ms per image
(Linux core-i7 machine)

- All the compared methods **are dependent on model likelihood** values. In some types of invertible models such as iResNet and ResFlow, likelihood is quite costly to compute since the Jacobian determinant term needs iterative approximation.
- **None of the statistics we suggest depend on explicit evaluation of likelihood** (Jacobian determinant). Consequently, our method can be about **10 times faster at test time**. T_2 comes from a fast image compression algorithm and T_3 from a typical automatic differentiation.
- Considering the close AUROC performance with the other methods, the primary advantage of our method is its speed.

Alkhatib, Amr
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Global Explanation by Characteristic Rules Extraction

Current techniques for explaining black-box predictions rarely produce explanations that generalize beyond the explained instances and hence do not allow for verification or prediction. A method for generating global explanatory rules by aggregating multiple local explanations is proposed. The generated rules can be used to understand how the black-box model operates in general and also emulate the model. The proposed method is applied to several different explanation techniques, individually and in combination. Experimental results show that the method produces high fidelity rules that are often as accurate as, and sometimes even more accurate than, the underlying black-box model.

Alkhatib, Amr
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Global Explanation by Characteristic Rules Extraction

Amr Alkhatib and Henrik Boström

KTH Royal Institute of Technology
School of Electrical Engineering and Computer Science (EECS)

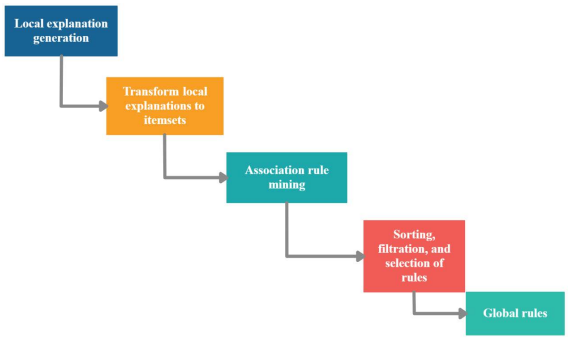
Abstract

Current techniques for explaining black-box predictions rarely produce explanations that generalize beyond the explained instances and hence do not allow for verification or prediction. A method for generating global explanatory rules by aggregating multiple local explanations is proposed. The generated rules can be used to understand how the black-box model operates in general and also emulate the model. The proposed method is applied to several different explanation techniques, individually and in combination. Experimental results show that the method produces high fidelity rules that are often as accurate as, and sometimes even more accurate than, the underlying black-box model.

Introduction

Techniques for explaining black box models can be classified into local and global methods. A local explanation method explains a single prediction made by a model, while global interpretations provide an understanding of how the model behaves in general and which features are globally important.
The local explanations are not empirically verifiable and cannot be easily used to understand the general behavior of the underlying model. We propose a method for generating explanations in the form of general rules, by aggregating multiple specific explanations. The rules can be combined to explain and emulate the underlying black-box model.

From Local Explanations to Global Rules



Discriminative vs. Characteristic Rules

The discriminative rules are a way to distinguish the given class from other classes using the minimum possible number of features, while a characteristic rule is the conjunction of all features that are common to all instances in the class.
Extracting the characteristic explanations to a representative subset of a class of instances provides a good approximation to that class's true set of global explanations.

The Accuracy of The Rules

Dataset	XGBoost Model*	Characteristic Explanatory Rules**	The Rules' Fidelity***
Adult	0.92	0.84	0.90
Compass	0.83	0.75	0.81
Spambase	0.99	0.95	0.95
Blood-transf.	0.71	0.69	0.93
German-credit	0.81	0.78	0.83

The explanation rules' performance in terms of area under ROC curve
*The accuracy of the black-box model.
**The accuracy of the rules.
***The accuracy of the rules in predicting the XGBoost model.

Conclusion and Future Work

This work proposes a method to aggregate local explanations and extract characteristic global explanations that we can test and measure their fidelity to the underlying model and apply to new data instances. We also show that the characteristic global explanations have high fidelity and can be more accurate than the black-box model. The method was used to compare different local explanation techniques and also to combine them for more accurate global explanations. An interesting direction for future work is to include the relative importance of each feature in the local explanations to compute a global explanation. It could also be interesting to use conformal prediction frameworks with explanation techniques to measure and quantify the provided explanations' good.

References

1. Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: Advances in Neural Information Processing Systems 30. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. url: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
2. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. 2016, pp. 1135–1144.
3. Rakesh Agrawal and Ramakrishnan Srikant. "Fast Algorithms for Mining Association Rules in Large Databases". In: Proceedings of the 20th International Conference on Very Large Data Bases. VLDB '94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499. isbn: 155860-153-8. url: http://dl.acm.org/citation.cfm?id=645920.672836.

Almeida, Tiago
Örebro University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Mental well-being awareness for Human-Robot Interaction

Psychological assistive robots are the next generation of service robots due to the continuous increase of mental illnesses in modern societies. In order to provide the best human-robot experience, the robot should be able to perceive and predict future human behaviors. Therefore, this research stands out for the study of Machine Learning methodologies for learning people's behavior and conduct during daily life. The hypothesis is that the human's state (a latent variable describing the particular human state) is a powerful feature for predicting future behaviors, and a central point for designing a better Human-Robot Interaction.
In this way, we intend to delve into this problem by finding behavioral patterns (ways of moving, daily life activities, physical reactions, etc.) induced by different human states and then learn the best robot interaction.

Almeida, Tiago
Örebro University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Mental well-being awareness for Human-Robot Interaction

Tiago Almeida, Örebro University
Center for Applied Autonomous Sensor Systems (AASS)

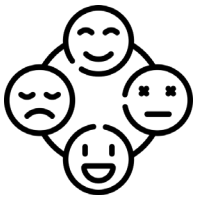
Motivation & Research Goals

Psychological assistive robots are the next generation of service robots due to the continuous increase of mental illnesses in modern societies. In order to provide the best human-robot experience, the robot should be able to **perceive** and **predict future human behaviors**. Therefore, this research stands out for the study of **Machine Learning** methodologies for learning people's behavior and conduct during daily life. The hypothesis is that the human's state (a latent variable describing the particular human state) is a powerful feature for predicting future behaviors, and a central point for designing a better **Human-Robot Interaction**. In this way, we intend to delve into this problem by finding behavioral patterns (ways of moving, daily life activities, physical reactions, etc.) induced by different human states and then learn the best robot interaction.

Overview

Mental well-being

- happiness
- sadness
- neutral
- angriness
- relaxation
- stress

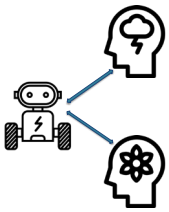


Awareness

- perceive the mental well-being [1]
- find behavioral patterns that indicate the mental well-being
- predict behaviors conditioned on different mental states
- use evidences from Psychology

Human-Robot Interaction

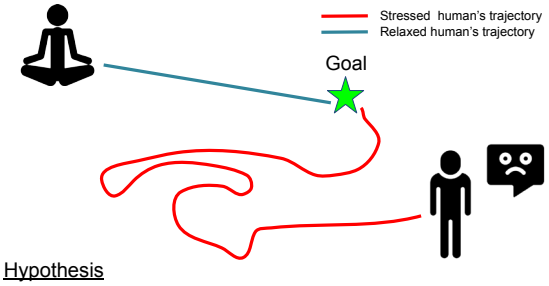
- how the robot should navigate in a human-centred environment [2, 3]
- how the robot should approach the human [2]
- adaptive behaviors [3]



First Ideas

Trajectories conditioned on different mental states

1. Design mental well-being constrained experiments
2. Collect different trajectories of different people under different mental states
3. Assess the mental state of people during the experiments

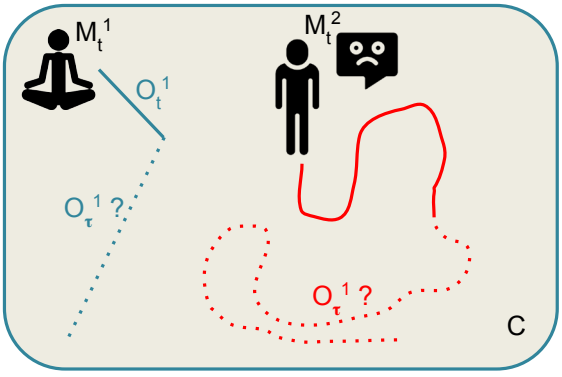


Hypothesis

- There is a correlation between behavioral patterns (motion trajectories) and mental states

Mental well-being aware Human Motion Prediction

1. Observe *(x, y)* position for person *i* during period *t*, so O_t^i in context *C*
2. Estimate mental-well being, M_t^i based on O_t^i, *and C*
3. Predict the next coordinates *(x', y')* conditioned on M^i over a period τ



References

1. T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera and D. Manocha, "Learning Perceived Emotion Using Affective and Deep Features for Mental Health Applications," *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2019, pp. 395-399, doi: 10.1109/ISMAR-Adjunct.2019.000-2.
2. V. Tolani, S. Bansal, A. Faust and C. Tomlin, "Visual Navigation Among Humans With Optimal Control as a Supervisor," in IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 2288-2295, April 2021, doi: 10.1109/LRA.2021.3060638.
3. P. Scales, O. Aycard and V. Aubergé, "Studying Navigation as a Form of Interaction: a Design Approach for Social Robot Navigation Methods," 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 6965-6972, doi: 10.1109/ICRA40945.2020.9197037.
4. V. Narayanan, B. M. Manoghar, V. Sashank Dorbala, D. Manocha and A. Bera, "ProxEmo: Gait-based Emotion Learning and Multi-view Proxemic Fusion for Socially-Aware Robot Navigation," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8200-8207, doi: 10.1109/IROS45743.2020.9340710.

Banerjee, Sourasekhar
Umeå university

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Fed-FiS : A Novel Information-theoretic Federated Feature Selection for Learning Stability

In the era of big data and federated learning, traditional feature selection methods show unacceptable performance for handling heterogeneity when deployed in federated environments. We propose Fed-FiS, an information-theoretic federated feature selection approach to overcome the problem occur due to heterogeneity. Fed-FiS estimates feature-feature mutual information (FFMI) and feature-class mutual information (FCMI) to generate a local feature subset in each user device. Based on federated values across features and classes obtained from each device, the central server ranks each feature and generates a global dominant feature subset. We show that our approach can find stable features subset collaboratively from all local devices. Extensive experiments based on multiple benchmark iid (independent and identically distributed) and non-iid datasets demonstrate that Fed-FiS significantly improves overall performance in comparison to the state-of-the-art methods.

Banerjee, Sourasekhar
Umeå university

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Fed-FiS: A Novel Information-theoretic Federated Feature Selection for Learning Stability

UMEÅ UNIVERSITY

Sourasekhar Banerjee, Umeå University
Dept. of Computing Science
Supervisors: Monowar Bhuyan, Erik Elmroth
sourasb@cs.umu.se, monowar@cs.umu.se, elmroth@cs.umu.se

Autonomous Distributed Systems Lab

Abstract

In the era of big data and federated learning, traditional feature selection methods show unacceptable performance for handling heterogeneity when deployed in federated environments. We propose Fed-FiS, an information-theoretic federated feature selection approach to overcome the problem occur due to heterogeneity. Fed-FiS estimates feature-feature mutual information (FFMI) and feature-class mutual information (FCMI) to generate a local feature subset in each user device. Based on federated values across features and classes obtained from each device, the central server ranks each feature and generates a global dominant feature subset. We show that our approach can find stable features subset collaboratively from all local devices. Extensive experiments based on multiple benchmark iid (independent and identically distributed) and non-iid datasets demonstrate that Fed-FiS significantly improves overall performance in comparison to the state-of-the-art methods. This is the first work on feature selection in a federated learning system to the best of our knowledge.

Introduction

- **Privacy preserving**, **collaborative machine learning** technique.
- Trains **local models** on data samples of each edge device without exchanging raw data.
- Server receives local models from edge devices
- Aggregate the models and produce global model.
- This process continues until the global model converges.

Problems

- Edge devices are normally low performing devices with limited resources. Therefore, computing models from terabytes of data is difficult.
- Sending terabytes of unprocessed data to server is costly and also violate privacy of users.

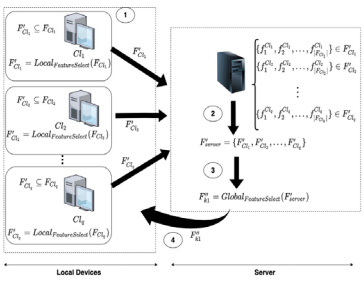
Objective

- Feature selection is essential and paramount to process such data and uncover useful knowledge for developing low-cost models.

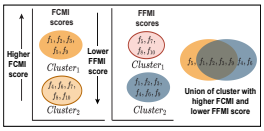
Contributions

- Fed-FiS introduces a local feature subset selection method by using mutual information and clustering.
- We develop a score function based on FCMI and aFFMI for global feature subset selection.
- Fed-FiS finds a most relevant features set from all devices where data is distributed in iid and non-iid manner.

Fed-FiS: Proposed Approach



Local feature selection



Global feature selection

- **Feature triplet**
  - feature_id,
  - fcmi score
  - affmi score

$$f_{kFCMI} = \frac{\sum_{j=1}^{q} f_{kFCMI}^{Cl_j}}{j}, where\, j \in [1,q]$$

$$f_{kaFFMI} = \frac{\sum_{j=1}^{q} f_{kFFMI}^{Cl_j}}{j}, where\, j \in [1,q]$$

- **Global score function**

$$S(f_k) = f_{kFCMI} - \frac{1}{(|F'_{server}|-1)} \times f_{kaFFMI}$$

- Where,

$$S(f_k) \in [1,-1] \quad |F'_{server}| > 1$$

$$S(f_k) = -1 \text{ iff } f_{kFCMI} = 0 \text{ and } f_{kaFFMI} = 1, \; |F_{server}| = 2$$

$$S(f_k) = 1 \text{ iff } f_{kFCMI} = 1 \text{ and } f_{kaFFMI} = 0$$

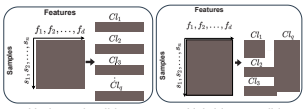
Simulation setup

Parameter(s)	Value	Description
local devices	5	Local feature selection
Server	1	Global feature selection
local device's participation	100%	Full participation of each local device
Cluster Validity Index (CVI)	Silhouette Coefficient	Identify the quality of Cluster
Clustering algorithm	k-means	Cluster with nearest mean
Learning algorithm	2	KNN-5, Federated forest
Performance metrics	1	Accuracy

Datasets

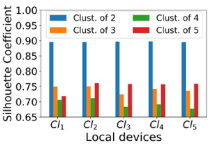
- NSL-KDD99
- Anonymized creditcard transaction dataset

Data division

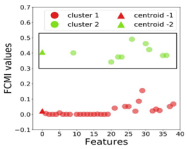


Results and discussions

- Cluster analysis and local feature subset (for Cl_1)

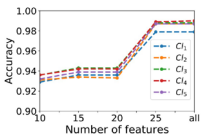


FCMI cluster analysis for NSL-KDD99 iid dataset

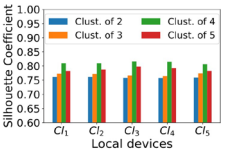


Clusters of features of NSL-KDD99 dataset with respect to the FCMI values.

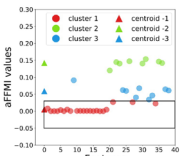
- Global feature subset selection



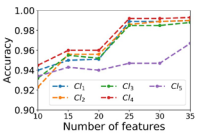
Feature subset for NSL-KDD99 iid dataset



aFFMI cluster analysis for NSL-KDD99 iid dataset



Clusters of features of NSL-KDD99 dataset with respect to the aFFMI values .



Feature subset for NSL-KDD99 non-iid dataset

- Steady learning ability

Datasets	Fed-FiS		without FS		classical MI based FS	
	Feature subset	Accuracy (%)	Feature set	Accuracy (%)	Feature subset	Accuracy (%)
NSL-KDD99 iid	23	99.3	41	99.5	23	99.24
NSL-KDD99 non-iid	9	89.33	15	54.07	7	92.75
ACC iid	10	99.95	29	99.96	6	99.3
ACC non-iid	9	99.83	14	99.81	10	99.83

- **Inference**
  - For iid datasets the performance of Fed-FiS is equivalent to classical MI based Feature selection method.

Conclusion

- We propose Fed-FiS, a mutual information-based federated feature selection method to select strongly relevant features set for stable and low-cost federated learning.
- Fed-FiS achieved expected model performance with lower number of features set, verified with federated forest algorithm.
- For IID dataset Fed-FiS gives expected performance but for non-IID dataset, size of overlapping feature set is important

Future Work

- More extensive experiments to check robustness of Fed-FiS for different type of dataset, such as, Feature distribution skew, Label distribution skew, Same label different features, Same features different labels, quantity skew or unbalancedness.
- Application specific implementation of Fed-FiS. Ex. Anomaly detection.

References

1. Manikandan, G., Abirami, S.: Feature selection is important: state-of-the-art methods and application domains of feature selection on high-dimensional data. In: Kumar, R., Paiva, S. (eds.) Applications in Ubiquitous Computing. EICC, pp. 177–196. Springer, Cham (2021).
2. Hoque, N., et al.: MIFS-ND: a mutual information-based feature selection method. Expert Syst. Appl. 41(14), 6371–6385 (2014)
3. Liu, G., et al.: Feature selection method based on mutual information and support vector machine. Int. J. Pattern Recogn. Artif. Intell. 35, 2150021 (2021)
4. Zheng, L., et al.: Feature grouping and selection: a graph-based approach. Inf. Sci. 546, 1256– 1272 (2021)
5. Gui, Y.: ADAGES: adaptive aggregation with stability for distributed feature selection. In: Proceedings of the ACM-IMS on Foundations of Data Science Conference, pp. 3–12 (2020)
6. Soheili, M., et al.: DQPFS: distributed quadratic programming based feature selection for big data. J. Parallel Distrib. Comput. 138, 1–14 (2020)
7. Moran-Fern ´ andez, L., et al.: Centralized vs. distributed feature selection methods based on ´ data complexity measures. Knowl.-Based Syst. 117, 27–45 (2017)
8. Tavallaee, M., et al.: A detailed analysis of the KDD CUP 99 data set. In: IEEE Symposium on Computational Intelligence for Security and Defense Applications, pp. 1–6 (2009)

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Bereza, Robert
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Parameter Estimation for Non-linear Differential-Algebraic Equation Models with Unknown Disturbances

Differential-algebraic equations (DAEs) arise naturally as a result of equation-based object-oriented modeling. Such models often contain unknown parameters that have to be identified using measured data. A challenge with the identification of physical systems is the effect of unknown disturbances. If such disturbances are ignored during the identification procedure, one can obtain poor parameter estimates. To the best of the authors' knowledge, there are no general methods successfully dealing with parameter estimation for this type of model. In this work, we propose a simulation-based prediction error method for non-linear DAEs where disturbances are modeled as continuous-time stochastic processes. We assume that the model can be simulated using available DAE solvers. Our method is tested on a simulated pendulum example, which suggests that our method provides consistent parameter estimates.

Bereza, Robert
KTH

Parameter Estimation for Non-linear DAE Models with Unknown Disturbances

Robert Bereza[1]
Mohamed R.-H. Abdalmoaty[1], Oscar Eriksson[2], David Broman[2] and Håkan Hjalmarsson[1]
[1]KTH, Division of Decision and Control Systems
[2]KTH, Division of Software and Computer Systems

Motivation & Research Goals

Differential-algebraic equations (DAEs) appear naturally when modeling many physical systems, such as mechanical or electrical systems. They are also the underlying model in high-level modeling languages such as Modelica or Simscape. Such models can contain unknown parameters that have to be **estimated using experimental data**. However, disturbances present in real-life systems makes such estimation difficult, as neglecting the influence of such disturbances can lead to biased parameter estimates. Therefore, the goal of this research is to develop **computationally tractable** parameter estimation methods for **nonlinear DAEs**. These methods should model **process disturbances** and take their effect into account to obtain **consistent parameter estimates**.

Differential-Algebraic Equations

- Differential equations with algebraic constraints
- **Example:**

$$\ddot{x}_1 + \theta x_2 = w(t)$$
$$x_1^2 = u(t)$$

with disturbance $w(t)$, control input $u(t)$, and parameter θ

- **General form:**

$$F(\dot{x}(t), x(t), u(t), w(t); \theta) = 0$$
$$y(t;\theta) = q(x(t), u(t); \theta) + noise \quad (*)$$

- Appear naturally when modeling many physical systems
- Underlying model in high level component based modeling languages (e.g. Modelica and Simscape)
- Non-trivial to solve and/or re write as ODEs

Problem and Method

- Identify unknown parameters θ using only knowledge of input $u(t)$ and samples of output $\{y(t_1), y(t_2), \dots, y(t_N)\}$
- **Challenge:** System influenced by unknown disturbances, which can lead to biased parameter estimates
- **Solution:** Minimize difference between measured output and expected value of model output w.r.t. disturbances [2]

$$\theta_\circ = \underset{\theta}{\operatorname{argmin}} J_N(\theta) = \underset{\theta}{\operatorname{argmin}} \frac{1}{N}\sum_{k=1}^{N} \| y(t_k) - \mathrm{E}_w[y(t_k;\theta)] \|^2$$

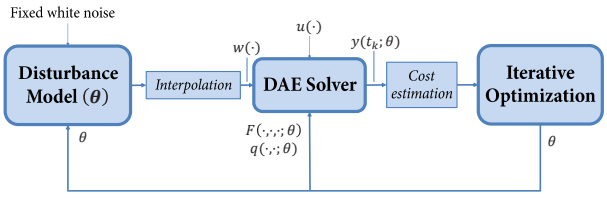


Figure 1. Parameter estimation process

References

1. M. R.-H. Abdalmoaty, O. Eriksson, R. Bereza, D. Broman, and H. Hjalmarsson. (2021). Identification of Non-Linear Differential-Algebraic Equation Models with Process Disturbances. Proceedings The 60th IEEE Conference on Decision and Control (CDC)
2. M. R.-H. Abdalmoaty, "Identification of stochastic nonlinear dynamical models using estimating functions," Ph.D. dissertation, KTH Royal Institute of Technology, 2019.

Simulation Results

Method tested on simulated example of a pendulum

$$m\ddot{x}_1(t) = x_3(t)x_1(t) - k|\dot{x}_1(t)|\dot{x}_1(t) + u(t) + w^2(t)$$
$$m\ddot{x}_2(t) = x_3(t)x_2(t) - k|\dot{x}_2(t)|\dot{x}_2(t) - mg$$
$$L^2 = x_1^2(t) + x_2^2(t)$$

Results when identifying parameter k [1]

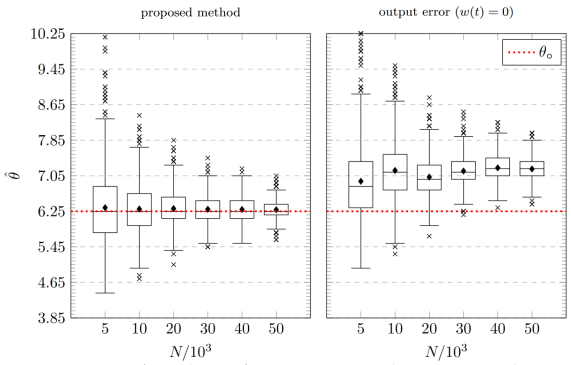


Figure 2. Left: Distribution of parameter estimates when using proposed method. Right: Distribution of parameter estimates when neglecting the effects of the process disturbance during parameter estimation

Stochastic Approximation

- For above results, cost function was approximated using Monte Carlo simulations
- This is time-consuming for large models

$$\nabla_\theta J_N(\theta) = \frac{2}{N}\sum_{k=1}^{N} (y(t_k) - \mathrm{E}_w[y(t_k;\theta)])(-\nabla_\theta \mathrm{E}_w[y(t_k;\theta)])$$

- For **stochastic gradient descent**, we need an **unbiased estimate** of $\nabla_\theta J_N(\theta)$

$$E_w[y(t_k;\theta)] \approx y^{(1)}(t_k;\theta)$$
$$\nabla_\theta E_w[y(t_k;\theta)] \approx \nabla_\theta y^{(2)}(t_k;\theta)$$

where $y^{(1)}(t_k;\theta)$ and $y^{(2)}(t_k;\theta)$ are two independent solutions to $(*)$

- $\nabla_\theta y^{(2)}(t_k;\theta)$ can be obtained while solving DAEs through **forward sensitivity analysis**

Blöcker, Christopher
Umeå University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Map Equation Centrality

To measure node importance, network scientists employ centrality scores that typically take a microscopic or macroscopic perspective, relying on node features or global network structure. However, traditional centrality measures, such as degree centrality and PageRank, neglect the community structure found in real-world networks. To study node importance based on network flows from a mesoscopic perspective, we exploit the coding principles behind the map equation framework, and derive a community-aware information-theoretic centrality score analytically. Applied to artificial and real-world networks, we demonstrate that our approach enables a more fine-grained differentiation between nodes than node-local or network-global measures, and highlight the role that local network context plays in determining node importance.

Blöcker, Christopher
Umeå University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Map Equation Centrality: A Map Equation-based Community-Aware Centrality Score

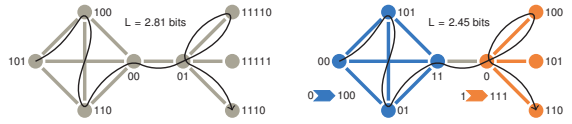
Christopher Blöcker[1], Juan Carlos Nieves Sanchez[2], and Martin Rosvall[1]

[1]Integrated Science Lab, Department of Physics, Umeå University
[2]Department of Computing Science, Umeå University

Abstract

To measure node importance, network scientists employ centrality scores that typically take a microscopic or macroscopic perspective, relying on node features or global network structure. However, traditional centrality measures, such as degree centrality and PageRank, neglect the community structure found in real-world networks. To study node importance based on network flows from a mesoscopic perspective, we exploit the coding principles behind the map equation framework, and derive a community-aware information-theoretic centrality score analytically. Applied to artificial and real-world networks, we demonstrate that our approach enables a more fine-grained differentiation between nodes than node-local or network-global measures, and highlight the role that local network context plays in determining node importance.

The Map Equation Framework



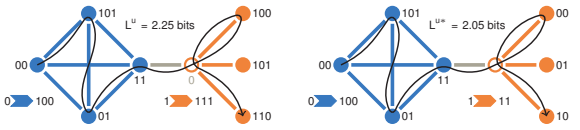
(a) A one-level partition with unique codewords for each node.

(b) A modular partition with unique codewords for nodes within modules.

Figure 1: Coding example for a small network with eight nodes and two communities. (a) In the one-level partition where each node has a unique codeword, the codelength is 2.81 bits. (b) In a modular partition, the codelength is reduced because codewords can be re-used between modules, reducing the codelength to 2.45 bits.

- The map equation is an information-theoretic objective function for community detection
- It measures the quality of a network partition by relating it to the lower bound of the average per-step description length for a random walk
- A network partition corresponds to a modular coding scheme, based on a Huffman code

Map Equation Centrality



(a) Using the same coding scheme as before, but without using the codeword for the node shown as a circle.

(b) Designing a new coding scheme where the node shown as a circle has no codeword.

Figure 2: Effect of silencing a node, that is, not encoding for random-walker steps to the respective node. (a) Using the same coding scheme, but simply omitting the node shown as a circle from the random walk's description. This is inefficient because the codeword assigned to the node is never used. (b) Designing a new coding scheme that does not assign a codeword to the node shown as a circle. The rest of the nodes receive shorter codewords.

- To measure node importance, we consider by how much more the random walk's description can be compressed if the node was not present
- By not assigning a codeword to a node, on average, the rest of the nodes can receive shorter codewords
- Essentially, a node's importance is the effective marginal harm it causes to other nodes by its existence

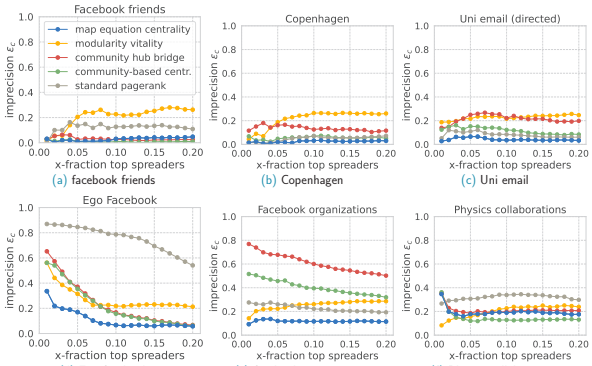
Results

- We have tested Map Equation Centrality on a set of online social networks and compared it with other community-aware centrality measures
- To measure the performance of centrality scores, we test how well they identify nodes with high spreading power
- Spreading power is the expected number of nodes that get infected in an SIR disease spreading model where the disease starts at the node in question, and nodes recover after 1 time step

Results

Table 1: List of networks and their properties: number of nodes (N), number of links ($|E|$), number of communities as detected with infomap (M), mixing (μ), and epidemic threshold (λ).

Network	N	$\|E\|$	M	μ	λ
facebook friends	329	1,954	20	0.127	0.048
Copenhagen	800	6,429	36	0.502	0.038
Uni email	1,133	5,452	52	0.402	0.057
Ego facebook	4,039	88,234	74	0.082	0.009
facebook org.	5,524	94,219	48	0.352	0.016
Physics collab.	8,798	27,416	863	0.279	0.066



(a) facebook friends (b) Copenhagen (c) Uni email

(d) Ego facebook (e) facebook organizations (f) Physics collaborations

Figure 3: Results for three community-aware centrality scores as well as standard PageRank on six online social networks.

- In general, Map Equation Centrality outperforms the other community-aware centrality measures
- However, it is no silver bullet: on the physics collaborations network, the other measures, including PageRank perform better

Conclusion

- We derived a community-aware centrality score from the map equation to measure node importance in modular networks
- In most cases we tested, Map Equation Centrality outperforms other centrality measures

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

ArXiv

http://arxiv.org/abs/2201.12590

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

UMEÅ UNIVERSITY

icelab INTEGRATED SCIENCE LAB

christopher.blocker@umu.se

Brunnsåker, Daniel
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

High-throughput metabolomics for the characterization and validation of a regulatory diauxic shift model

Biological systems are very complex due to the nature of their regulation and the sheer amount of interconnected components present in even the simplest of organisms. To properly understand and fully quantify these systems, thousands and thousands of experiments have to be made; A herculean effort that would take decades of work to achieve with conventional methods. Previously a robot scientist was employed in implementing iterative closed cycles of experimentation and model improvement to semi-autonomously generate new knowledge about a biological phenomenon in S. cerevisiae, the diauxic shift.
This project aims to extend the original work by adding additional robustness and biological context to the existing framework by introducing the integration of contextual biological data in the form of untargeted metabolomics to further characterize the shift.

Brunnsåker, Daniel
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

High-throughput metabolomics for the characterization and validation of a regulatory diauxic shift model

Brunnsåker D[1], Reder GK[1], Tiukova IA[1], Soni NK[1], Gower A[1] & King RD[123]

[1] Chalmers University of Technology, Gothenburg, Sweden
[2] Cambridge University, Cambridge, United Kingdom
[3] Alan Turing Institute, London, United Kingdom

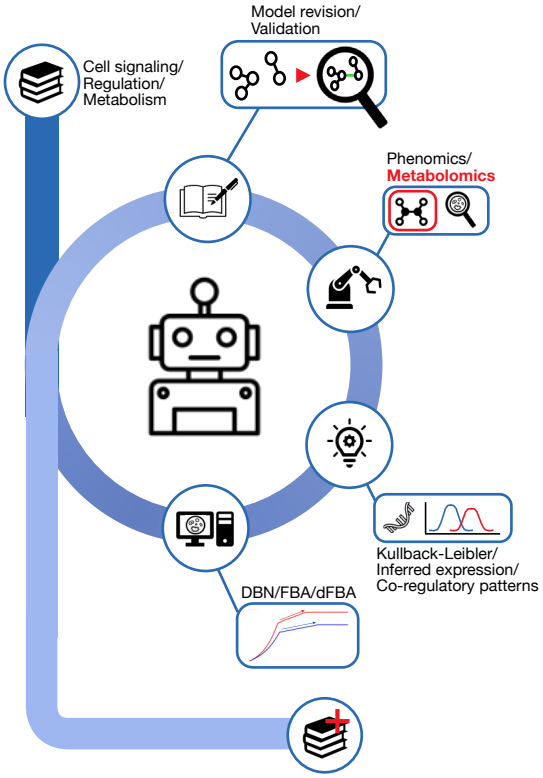
Background

Biological systems are very complex due to the nature of their regulation and the sheer amount of interconnected components present in even the simplest of organisms. To properly understand and fully quantify these systems, thousands and thousands of experiments have to be made; A herculean effort that would take decades of work to achieve with conventional methods. Previously a **robot scientist** was employed in implementing iterative **closed cycles of experimentation** and **model improvement** to semi-autonomously generate new knowledge about a biological phenomenon in *S. cerevisiae*, **the diauxic shift.** This project aims to extend the original work by adding additional robustness and biological context to the existing framework by introducing the integration of contextual biological data in the form of untargeted metabolomics to further characterize the shift.

Closed-cycle model improvement

The complexity of biology necessitates a vast amount of explorative experiments to map a eukaryotic system, a number even conventional automated systems are incapable of handling. In order to meet this challenge, the system needs **autonomous** control and the ability to form low-level **hypotheses** and independently perform model revision.

This is done through stages of **guided experiment selection** and **model revision** using both simulated and experimental data generated by a **automatic laboratory**. The basis of experiment selection is through partial evidence inference of gene expression using **phenotypical** measurements along with co-regulatory patterns from regulatory networks acquired from literature.



The regulatory model is then refined through iterative edge removal using minimized growth-curve differences between simulated and experimental measurements as the objective function. Likewise, additional validation metrics for model revision will help guide the process further.

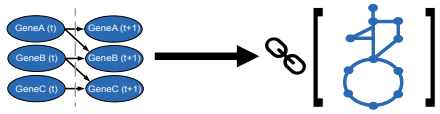
High throughput metabolomics

The project makes use of an untargeted metabolomics platform in the form of **FIA-MS** (flow injection analysis mass-spectrometry) to acquire biological measurements from the cultivations generated by the automated laboratory cell. In order to achieve a sufficiently high-throughput analysis, **automated pipelines** for metabolomics data processing are used.

Model validation

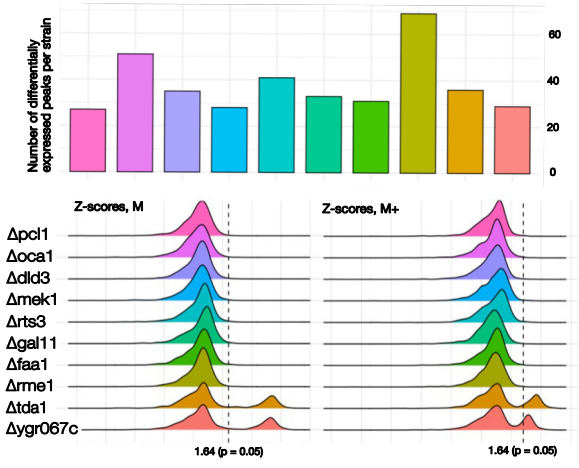
As the model was improved using only partial evidence on physiology and prior knowledge of regulatory interactions, the biological relevance of the proposed changes need to be validated with **contextual biological data.**

Reference dFBA simulations for use with model validation were performed using the framework previously developed by Coutant et al. in which a model of control of metabolism (**iMM904**) was constrained by simulated gene expression from a regulatory model represented by a dynamic Bayesian network (**DBN**).



To validate the model batch cultivations were performed in the automated laboratory cell Eve on multiple regulatory deletion mutants from the EUROSCARF collection along with their related reference strain (BY4741).

Identifying **regulatory hotspots of metabolism** by observing patterns in differentially expressed metabolites allow for some elucidation regarding the metabolic perturbation of regulatory gene deletion. Likewise, observing patterns and zones of differential fluxes using the model allows for identification of affected parts of metabolism. Establishing the similarities and dissimilarities between these can then **guide model revision**, allowing for higher biological relevance and **higher model fidelity.**



References

1. Anthony Coutant et al. **Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast.** Proceedings of the National Academy of Sciences **Sep 2019**.

2. Svetlana Kutuzova et al. **SmartPeak Automates Targeted and Quantitative Metabolomics Data Processing.** Analytical Chemistry, **Dec 2020** (92)

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

sys² bio www.sysbio.se

Bökman, Georg
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

ZZ-net: A Universal Rotation Equivariant Architecture for 2D Point Clouds

In this work, we investigate a novel neural network architecture for 2D point clouds, guaranteeing rotation equivariance and invariance to permutations of the points.

Bökman, Georg
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

CHALMERS UNIVERSITY OF TECHNOLOGY

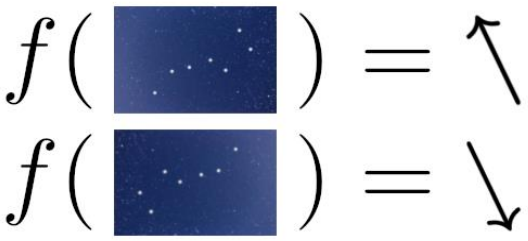
ZZ-net: A Universal Rotation Equivariant Architecture for 2D Point Clouds

There has been much recent work on group equivariance and group invariance of neural networks. Refer to the figure on the right for an example where the group is the rotation group in 2D. Equivariance means that when the input of the network is acted on by a group element, then the output is acted on by the same group element. Invariance means that the output of the network stays constant when the input is acted on by the group.

In this work, we investigate a novel neural network architecture for 2D point clouds, guaranteeing rotation equivariance and invariance to permutations of the points. Investigations have already been done on the 3D case and the general nD case. We find here that the ability to use complex numbers as representations for both the points and the rotations, makes the 2D case stand out and renders it possible for us to find a universality result for neural network architectures that we can't easily generalize to higher dimensions.

There are three main ideas in the paper. First, we discuss how rotation equivariant, permutation invariant functions can be decomposed into a sum of evaluations of a rotation invariant function – see the "Flavour of the results" to the right. Second, we explain a way to approximate this decomposition with neural networks and prove the approximation to be universal. Third, we describe how to apply our framework to the case of an input of correspondences between two point clouds. To illustrate the last idea, we perform experiments on the estimation of essential matrices in stereo vision and find that our framework outperforms state of the art methods when the test data contains rotations unseen in the training data.

Reference:
Bökman, Georg, Fredrik Kahl, and Axel Flinth. November 2021.
'ZZ-Net: A Universal Rotation Equivariant Architecture for 2D Point Clouds'
http://arxiv.org/abs/2111.15341.



A simple example of rotation equivariance. The illustration shows the task of determining the direction to the North Star given other stars in the night sky. The input is the set of locations of the visible stars in some 2D coordinate frame. Rotation equivariance of the point cloud processor f means here that the determined direction should rotate if the night sky (or the observer) rotates.

Image modified from https://commons.wikimedia.org/wiki/File:Big_Dipper_20210116.jpg, used under Creative Commons Attribution-ShareAlike 4.0 International license. Stars in main constellation brightened.

Flavour of the results

Let Z denote a 2D point cloud with m points, viewed as a vector in $\mathbb{C}^m$. Denote the index set $\{0, 1, 2, \dots, m-1\}$ by $[m]$ and let τ_i be the transposition of i and 0. Further, let $\mathrm{Stab}(0)$ be the stabilizer of 0, i.e. the group of all permutations of the m indices that fix 0.

Theorem 2

The set of functions on the form

$$f(Z) = \sum_{i \in [m]} \gamma(\tau_i(Z)) Z_i, \tag{1}$$

where γ is an arbitrary continuous, rotation invariant and $\mathrm{Stab}(0)$-invariant function, is dense in the set of continuous rotation equivariant and permutation invariant functions from $\mathbb{C}^m$ to $\mathbb{C}$.

Georg Bökman
bokman@chalmers.se
Computer Vision
Department of Electrical Engineering
Chalmers University of Technology

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Caylak, Gizem
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Automatic Static transformation of Probabilistic Programs

Probabilistic program languages (PPLs) provide tools to write a model and do statistical inference on the model ideally without considering the internals of the inference algorithm. While simulation is the strong aspect of the Monte Carlo methods in PPLs to generate samples, it needs to be done in a smart way such that the variance of the sampler reduces in a computationally feasible time. Especially, the models, having high number of random variables, require manual transformation of the model to get an efficient sampler. The aim of this research is to automatically analyse and transform the model given by the user to a more efficient equivalent representation using analytical relations between random variables in compile time.

Caylak, Gizem
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Automatic Static Transformations of Probabilistic Programs

Gizem Caylak, KTH Royal Institute of Technology
Department of Computer Science
Main Advisor: David Broman

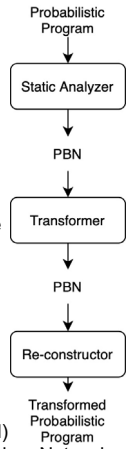
Abstract

Probabilistic program languages (PPLs) provide tools to write models and do statistical inference on the model ideally without considering the internals of the inference algorithm. While simulation is an important aspect of Monte Carlo methods in PPLs to generate samples, it needs to be done in a effective way that the variance of the sampler reduces in a computationally feasible time. Especially, the problem is that models, which have high number of random variables, require manual transformations to get an efficient sampler. The aim of this research is to **automatically analyze and transform** the model given by the user to a more **efficient** equivalent representation. We use analytical relations between random variables **at compile-time** in transformations.

Research Problem

PPLs provide an expressive interface to represent probabilistic models ideally without dealing with the inference. However, making the sampler efficient requires manual transformation of the model. Run-time algorithms, such as Delayed Sampling [1], are suggested to do the transformation automatically. To reduce the run-time overhead, we propose an algorithm to do the transformation **automatically at compile time.**

Method



Static Analyzer
Creates a Programmatic Bayesian Network (PBN) from the probabilistic program. Vertices are either random variables or code blocks and edges represent the dependency between vertices.

Transformer
Takes the PBN as input and uses conjugate prior relationships between random variables to derive the posterior. We guarantee that the transformation is conservative, always producing a correct program.

Re-constructor
Takes the transformed PBN and reconstructs the probabilistic program.

Main Contributions

- We extend the concept of Bayesian Networks (BN) new kind of graph called Programmatic Bayesian Networks (PBN). A PBN encapsulates random variables and the code structures in a probabilistic program as well as reconstructing a probabilistic program from a PBN.
- We transform the PBN based on analytical relations, such as conjugate prior relations, between the model parameters.
- All steps are done **automatically at compile time.**
- We implement this method in the meta language framework Miking and demonstrate the efficiency of our algorithm on non-trivial models such as LDA

References

1. Murray, L., Lundén, D., Kudlicka, J., Broman, D. & Schön, T. (2018). Delayed Sampling and Automatic Rao-Blackwellization of Probabilistic Programs. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, in *Proceedings of Machine Learning Research* 84:1037-1046

Transformations

We use conjugate prior relations between random variables to derive the posterior hyperparameters. This gives a closed-form expression for the posterior.

$$\underbrace{P(\theta \mid Y)}_{\text{Posterior}} = \frac{\overbrace{P(Y \mid \theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{Prior}}}{P(Y)}$$

Example: A simple coin toss mode

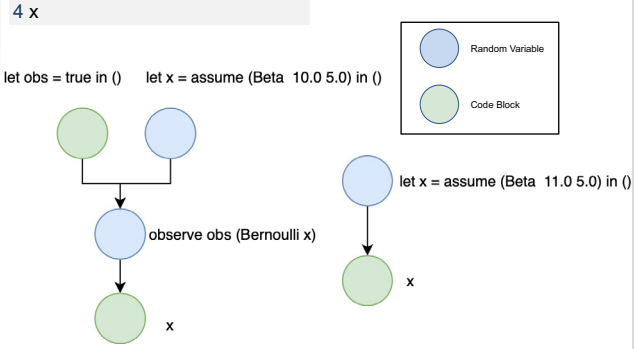
$$P(x) \sim Beta(5,5)$$

$$P(Y \mid x) \sim Bernoulli(x)$$

$$P(x \mid Y = true) \sim ?$$

```
1 let x = assume (Beta 5.0 5.0) in
2 let obs = true in
3 observe obs (Bernoulli x);
4 x
```

```
1 let x = assume (Beta 6.0 5.0) in
2 x
```



Limitations and work in progress

- Since the transformation is conservative, it may not output a more efficient representation, e.g. in **stochastic branching,** it may not capture different relations discovered between random variables.

Cornell, Filip
Gavagai AB

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Dimensionality reduction for Attributed Graphs (on-going work)

Embedding nodes in a large graph using vectors but most solutions rely on optimizing embedded representations. This yields problems; adding new nodes require re-training the embeddings, and training is costly and time-consuming. To mitigate this, we propose a simple, static and lossy way of compressing and representing higher-order connections in a network. We make use of Random Indexing, embedding the nodes through static vectors in a euclidean vector space using aggregation methods, showing a performance increase in comparison to baselines, needing less dimensions for more expressivity.

Cornell, Filip
Gavagai AB

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Dimensionality reduction for Attributed Graphs (on-going work)

Filip Cornell
KTH Royal Institute of Technology & Gavagai

Introduction

Embedding nodes in a large graph using vectors but most solutions rely on **optimizing** embedded representations. This yields problems:

- Adding new nodes require re-training the embeddings
- Training is costly and time-consuming

We propose:

- A simple, static and lossy way of compressing and representing higher-order connections in a network.
- We make use of **Random Indexing,** embedding the nodes through static vectors in a euclidean vector space

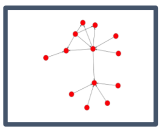
Two methods:

- Linear Graph Neural Network aggregation (**RIA**)
- Build up vectors from random walks (**RI + Random Walks**)

Data

Five benchmark datasets; citation & social networks

Dataset	$\|V\|$	$\|E\|$	$\|F\|$	$\|C\|$
Cora	2,708	5,429	1,433	7
Citeseer	3,312	4,715	3,703	6
Pubmed	19,717	44,338	500	3
Wiki	2,405	17,981	4,973	17
BlogCatalog	5,196	343,486	8,189	6



Datasets: Citation & social networks

Method

Empirical investigation[2]

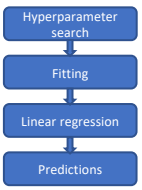
1. Embed the graphs using our methods
2. Evaluate using linear regression on the embeddings using **node** and **link** prediction

Random hyperparameter search
For all methods, perform a hyperparameter search

Baselines:

- QUINT[1] – recent hashing method baselines
- Attributes only (binary one-hot encodings)

Experiment pipeline



Results

Link prediction

- We beat our baselines in all cases
- Embedding higher order neighborhood is **crucial**

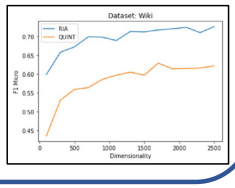
Method/Dataset	Cora		Citeseer		Pubmed		Wiki		BlogCatalog	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
Feature baseline	.768	.726	.756	.707	.8	.796	.706	.659	.925	.923
QUINT	.804	.768	.766	.713	.876	.875	.797	**.69**	**.968**	**.967**
RIW (ours)	.862	.841	.746	.705	.762	.735	.815	.712	.749	.749
RIA (ours)	**.89**	**.88**	**.798**	**.723**	**.89**	**.887**	**.8**	.669	.966	.966

Node classification experiments. Our method is superior to our baselines in most cases

Method/Dataset	Cora	Citeseer	Pubmed
	with/out	with/out	with/out
Features only	.747	.826	.808
QUINT	.865/.722	.92/.669	.914/.892
RI + Random walks	.72/.696	.719/.628	.891/.868
Aggregation, RI	**.895/.814**	**.968/.809**	**.975/.891**

Link prediction. We beat our baselines in all cases

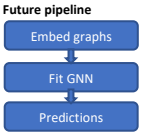
- Our method needs **less dimensions** to produce better results than other hashing methods -> higher expressiveness



Future work

- Test features' quality when combined with a Graph Neural Network (GNN)
- Can be useful to initialize representations when graphs have **binary** or **no attributes**
- Initial experiments show up to an 80 % reduction of parameters with **slight** performance decrease

Future pipeline



Acknowledgements

This work was partially supported by Wallenberg Autonomous Systems Program (WASP).

1 QUINT: Node embedding using network hashing. arXiv preprint arxiv:2109.04206, 2021.

Doostmohammadi, Ehsan
Linköping University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Verb Understanding in Video Grounded Language Models

Training a language model only on textual data does not give the model a comprehensive understanding of the world. One way to amend this and make more capable models is to train them on more modalities, for example, images. There is also a vast literature on video grounding, which usually means further training (most of the times transformer-based) language models on videos and their captions. Training on videos is usually more expensive than images and requires 3D convnet encoders, which raises the question, how does it benefit a language model? In this work, we try to investigate this further and see if grounding in videos improves model's understanding of actions (verbs) and objects (nouns), over only text, text and images, and possibly sparsely sampled video frames.

Doostmohammadi, Ehsan
Linköping University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Does Grounding in Videos Help with Verb Understanding?

Ehsan Doostmohammadi, Linköping University
Department of Computer and Information Science
Supervisors: Marco Kuhlmann (LiU) and Richard Johansson (Chalmers)

What is Grounding?

Grounding is defined as "Learning language representations from explicit visual associations" (Sileo, 2021), which basically means using other knowledge sources (e.g., images or knowledge graphs) in addition to text for training language models. We expect such models to perform better in some aspects and have a better understanding (?) of the world.

Grounding in Videos; How Does it Help?

Have you seen those perfect pictures in image-caption datasets that are self-contained and obvious? Unfortunately, the world is not that perfect. Can you guess what is being added in the left picture and what they are doing in the right one?

Or what are they doing in this one? Should you plug in before doing a certain task or unplug? In this case, the information is available in text, but that is not always the case.

Ideally, one would create a dataset (maybe similar to Hendricks et al., 2021) and be done with it. But creating datasets is expensive and even more complicated for videos.

Imageability

As it is not currently feasible to create a dataset, we rely on a body of work in psychology about an easy-to-grasp concept called imageability (Bird et al., 2001). Words like "(to) walk" and "(a) wasp" are more imageable than "(to) think" and "jealousy".
We hypothesize that a language model that is grounded in videos will perdict highly imageable words more accurately than less imageable ones and also compared to a model that is not grounded.

References

- Bird, Helen, Sue Franklin, and David Howard. "Age of acquisition and imageability ratings for a large set of words, including verbs and function words." Behavior Research Methods, Instruments, & Computers 33.1 (2001): 73-79.
- Hendricks, Lisa Anne and Aida Nematzadeh. "Probing Image-Language Transformers for Verb Understanding." FINDINGS (2021).
- Miech, Antoine, et al. "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
- Sileo, Damien. "Visual Grounding Strategies for Text-Only Natural Language Processing." Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN). 2021.
- Sun, Chen, et al. "Videobert: A joint model for video and language representation learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
- A. -M. Oncescu, et al., "QUERYD: A Video Dataset with High-Quality Text and Audio Narrations," ICASSP 2021 - 2021 ICASSP, 2021, pp. 2265-2269.

Data

We need video data paralleled with text, such as video captions. But such datasets are expensive to create and small in size. Therefore we use **Howto100m** dataset, which is a large-scale dataset of narrated instructional videos (Miech et al., 2019). As this dataset is automatically collected, there is a considerable amount of noise in it, which is partly compensated by the sheer size of it, ∼12 TB.

How to train such a model?

We used a pretrained distil-BERT language model and fine-tuned on parallel video-text data. Random words and video inputs are masked and the model learns to predict them. Additionally, the text and video in some random training samples are not aligned and the model is trained to distinguish them. For more details see **VideoBERT** (Sun et al., 2019).

Preliminary Results

We train the language model as described above and test it on a held-out test set of 5 thousand videos. The words that are available in the Bird dataset are masked for the model to predict them. Only words with high and low imageablity are kept, as interpreting the results on medium imageability would be difficult.

Train/Test	Img.	Acc. (Δ)	V. Acc. (Δ)	N. Acc. (Δ)
T/T	Low	34.3	35.7	24.0
	High	16.8	17.5	16.6
TV/TV	Low	33.7 (-0.6)	35.0 (-0.7)	23.7 (-0.3)
	High	17.7 (0.9)	18.9 (1.4)	17.2 (0.6)
TV/T	Low	34.1 (-0.2)	35.5 (0.2)	24.0 (0.0)
	High	17.1 (0.3)	18.0 (0.5)	16.7 (0.1)

In the first column, T stands for text and V for video. In the first row, Acc. is accuracy, V. is verb, and N. stands for noun. Delta is the difference between that cell and the corresponding cell in the T/T row.

The results show 1.4% increase in the accuracy when we train on videos and text, and test on both. When only testing on text, we see a 0.5% increase for the highly imageable words. The results also show that there is a higher increase for verbs compared to nouns.

What's next?

Train a better model, test on high quality data, such as **QueryD** (Oncescu et al., 2021) (as only 50% of the text is refering to an object/action in the scene in Howto100m), and analyze and interpret the results.
Have any qustion? Don't hesitate to contact me:
ehsan.doostmohammadi@liu.se :)

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Drexler, Dominik
Linköping University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Learning Language-Based Representations for Efficient and Intelligent Acting

We learn compact and reusable control knowledge as representations of a target language. The control knowledge captures the subgoal structure in a planning domain. We can formally prove that the learned representations can be used to solve any instance from the domain efficiently. In this setting, we are restricted to tractable domains. Those are domains, where it is easy to find a suboptimal solution. We address the learning problem using answer set programming. We present some results of the learning and directions for future work.

Drexler, Dominik
Linköping University

Learning Language-Based Representations for Efficient and Intelligent Acting

Dominik Drexler, PhD Student, Linköping University

Supervisors: Hector Geffner and Jendrik Seipp

Motivation & Research Goals

- Learn reusable control knowledge for intelligent acting
- Learned target is language representation that
  - is compact and simple
  - represents the subgoal structure
  - generalizes over target class $\mathcal{Q}$ of planning problems
  - can be used to solve any problem $P \in \mathcal{Q}$ efficiently
- Goals in future work: scalability, learn hierarchies, learn from arbitrary inputs

Methods

Classical Planning Problem.

- **Given:** A **planning problem** P models a world and consists of
  - **Domain** D, i.e., **Predicates** over **objects**, **Actions** over predicates
  - **Instance information**, i.e, sets of literals made from ground atoms over predicates and objects:
    - Fully observable **initial state** I
    - **Goal** description G
- **Objective:** find action sequence that leads from I to G

Width.

- **Width** $w(P)$ measures difficulty of problem P
- **Theorem:** If $w(P) \leq k$ then $IW(k)$ algorithm solves P in time $\exp(k)$

Serialization & Policy Sketches.

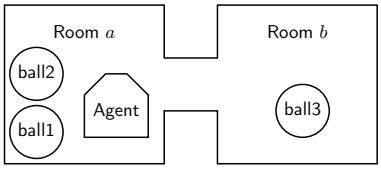
- **Policy sketch** R_Φ is set of rules of form $C \mapsto E$ over state features Φ
- R_Φ defines subgoals in $P \in \mathcal{Q}$ over common domain
- R_Φ decomposes $P \in \mathcal{Q}$ into subproblems
- **Sketch width** $w_{R_\Phi}(\mathcal{Q})$ is largest width of any subproblem in all $P \in \mathcal{Q}$
- **Serialization** SIW_{R_Φ}: search greedily to closest subgoal and repeat
- **Theorem:** if $w_{R_\Phi}(\mathcal{Q}) \leq k$ then SIW_{R_Φ} solves $P \in \mathcal{Q}$ in time $\exp(k)$

Learning Sketches.

- As combinatorial optimization problem: answer set programming
- Unsupervised from example problems $P_1, \ldots, P_n$
- Input parameter $k \in \mathbb{N}_0$ controls difficulty of subproblems
- Current limitations: tractable domains, scalability, requires PDDL like inputs

Selected Results

Gripper. An agent must move all balls from room a to room b.



Consider **features** $\Phi = \{g_a, g_b\}$ where

- g_a is number of balls in room a
- g_b is number of balls in room b

The **learned width-2 sketch** R^2 over features $\{g_b\}$ is

$$r = \{\} \mapsto \{g_b\uparrow\}$$

- r: increasing number of balls in room b ($g_b\uparrow$) is good

The **learned width-1 sketch** R^1 over features $\{g_a, g_b\}$ is

$$r_1 = \{\} \mapsto \{g_a\downarrow, g_b?\}$$
$$r_2 = \{\} \mapsto \{g_b\uparrow\}$$

- r_1: decreasing number of balls in room a ($g_a\downarrow$) and arbitrarily changing number of balls in room b ($g_b?$) is good
- r_2: increasing number of balls in room b ($g_b\uparrow$) and keeping number of balls in room a the same (no effect) is good

Experimental Results.

	$w=1$				LAMA		BFWS	
Domain	S	T	AW	MW	S	T	S	T
Blocks-clear (30)	30	3	0.78	1	30	4	30	8
Blocks-on (30)	30	36	1.00	1	30	4	30	52
Childsnack (30)	30	4	0.11	1	9	3	4	10
Delivery (30)	30	1	1.00	1	30	3	30	2
Gripper (30)	30	5	0.50	1	30	3	30	7
Miconic (30)	30	140	0.54	1	30	7	30	28
Reward (30)	30	3	1.00	1	30	3	30	4
Spanner (30)	30	1	0.25	1	0	–	0	–
Visitall (30)	30	1687	0.01	1	29	507	18	113

References

[1] Expressing and Exploiting the Common Subgoal Structure of Classical Planning Domains Using Sketches
Dominik Drexler, Jendrik Seipp and Hector Geffner
KR2021, 3-12 November, 2021, Hanoi, Vietnam

[2] Learning Sketches for Decomposing Planning Problems into Subproblems of Bounded Width
Dominik Drexler, Jendrik Seipp and Hector Geffner
submitted to ICAPS2022, 19-24 June, 2022, Singapore, Singapore

Dürr, Alexander
Lund University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Robot Skill Learning based on Interactively Acquired Knowledge-based Models

For small and medium-sized enterprises collaborative robots represent a possible way to improve their manufacturing practices. Yet, programming such a robot to perform many different and complicated tasks poses a problem even for a skilled operator - this is a hurdle for enterprises. We present our approach to tackling this problem with machine learning and classic AI. We showcase contributions and results from one published paper, which improves robot simulations. Last, we present current work in progress.

Dürr, Alexander
Lund University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Robot simulations can have large **errors in force and moment** output but these errors **can be corrected** with a **neural network architecture** based on **LSTM**

Robot Skill Learning based on Interactively Acquired Knowledge-based Models

Alexander Dürr, Lund University - LTH
Robotics and Semantic Systems
Main advisor: Elin Anna Topp

LUND UNIVERSITY | LTH FACULTY OF ENGINEERING

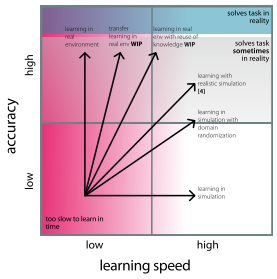
Background

For small and medium-sized enterprises **collaborative robots** represent a possible way to improve their **manufacturing** practices. Yet, programming such a robot to perform many different and complicated tasks poses a problem even for a skilled operator - this is a hurdle for enterprises.

Methods

In general, we seek new approaches to improve the robot's autonomy. For this, we try to increase the **training speed, accuracy, and success rate** of policies for **robots interacting with the real environment.**

We look at **combinations** of modern **machine learning** [1] and **classic AI for robotics.** For example, deep neural networks to analyze images for task-relevant content, knowledge bases [2,3], and **reasoners and planners** to build an executable program for the robot.



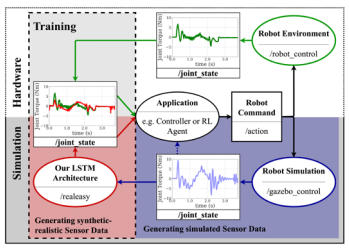
Selected Results

Traning in simulation is a popular approach to scale policy learning. Simulations can run in parallel, but this creates a new problem as the learned policies can often not be transferred directly to reality due to:

- Inaccurate numerical solvers
- Inaccurate model of the task
- Low tolerances allowed in the execution

Contributions (see [4]):

- Two **systematically developed neural network** architectures to improve simulated robot state data, easily integrated into the **ROS** framework.
- Shared data set description, data set, trained model, network parameters, a real-time graphics-enabled docker image for robot control.

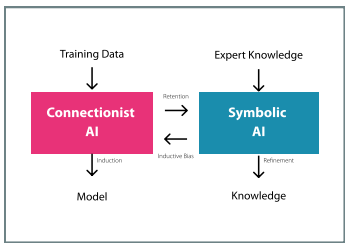


Current work in progress

We try to bring the ideas from Neuro Symbolic Visual Question Answering [5] and the Neuro Symbolic Concept Learner [6] to our environment where interaction with the environment is required.

For this, we explore:

- **Reuse of knowledge**: Where and how can knowledge be integrated into the learning algorithm?
- **Knowledge representation**: How to represent knowledge for skills and behaviors? Which representations help with learning? How do they help?
- **Connectionist ideas and symbolic ideas**: How to refine knowledge? How to guide learning with knowledge? How to extract symbolic knowledge from a neural network?



References

[1] Ghadirzadeh, A., Bütepage, J., Maki, A., Kragic, D., Björkman, M.: A sensorimotor reinforcement learning framework for physical human-robot interaction. In: IEEE International Conference on Intelligent Robots and Systems 2016, pp. 2682--2688.
[2] Topp, E. A., Stenmark, M., Ganslandt, A., Svensson, A., Haage, M., Malec, J.: Ontology-Based Knowledge Representation for Increased Skill Reusability in Industrial Robots. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5672-567, IEEE--Institute of Electrical and Electronics Engineers Inc.
[3] Stenmark, M., Haage, M., Topp, E. A., Malec, J.: Supporting Semantic Capture during Kinesthetic Teaching of Collaborative Industrial Robots. In: International Journal of Semantic Computing 2018, vol. 12, no. 1, pp. 167--186.
[4] Dürr, A., Neric, L., Krueger, V., Topp, E. A.: Realeasy: Real-Time capable Simulation to Reality Domain Adaptation. In: 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE) (pp. 2009-2016). IEEE.
[5] Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. B. (2018): Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 1039-1050).
[6] Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019): The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. International Conference on Learning Representations.

Take a picture to download the full paper

{alexander.durr,elin_anna.topp}@cs.lth.se
Department of Computer Science
Lund University, LTH
Sweden

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Ek, Sofia
Uppsala University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Learning Language-Based Representations for Efficient and Intelligent Acting

We learn compact and reusable control knowledge as representations of a target language. The control knowledge captures the subgoal structure in a planning domain. We can formally prove that the learned representations can be used to solve any instance from the domain efficiently. In this setting, we are restricted to tractable domains. Those are domains, where it is easy to find a suboptimal solution. We address the learning problem using answer set programming. We present some results of the learning and directions for future work.

Ek, Sofia
Uppsala University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Learning Pareto-Efficient Decisions with Confidence

UPPSALA UNIVERSITET

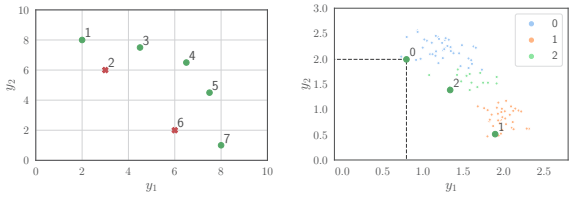
Sofia Ek, Dave Zachariah, Petre Stoica

Uppsala University, Sweden

Introduction

- Interested in problems where the objective functions are unknown but we have access to data from past outcomes of decisions. The goal is to balance trade-offs between several outcome objectives using Pareto-efficiency.
  - Decision $\mathbf{x}$ is *Pareto-efficient* if it is not strictly dominated by any alternative decision $\mathbf{x}'$.
  - Efficient decisions $\mathcal{X}^\alpha$ form a 'Pareto-frontier' $\mathcal{Y}^\alpha$ in reward space.

Main features of proposed method:
- Considers the lower tail rewards at any specified level.
- Provides finite-sample coverage guarantees for the rewards even when the data-generating process is unknown.
- Enables the study of decision trade-offs using a Pareto-frontier that takes reward uncertainties into account.



(a) (b)

Figure: (a) Illustration of efficient decisions $\mathcal{X}^\alpha = \{1, 3, 4, 5, 7\}$ shown as solid green dots. Decisions 2 and 6 (red crosses) are thus strictly dominated. (b) Decision $\mathbf{x} = 0$ and its corresponding outcome rewards $\mathbf{y}_1$ and $\mathbf{y}_2$.

Problem Formulation

We have observed data $\mathcal{D} = \{(\mathbf{x_i}, \mathbf{y_i}, \mathbf{z_i})\}_{i=1}^{n}$, where:
- $\mathbf{x} \in \{0, 1, 2, \dots\}$ are discrete decisions (treatments) in a space $\mathcal{X}$.
- $\mathbf{y}$ are reward vectors (utility/negative loss) in a reward space $\mathcal{Y} \subset \mathbb{R}^{\mathbf{m}}$. The number $\mathbf{m}$ is typically small, the examples here illustrate cases with $\mathbf{m} = 2$ rewards.
- $\mathbf{z}$ are $\mathbf{d}$-dimensional covariate vectors representing decisions taken in different contexts.

Goal: Learn covariate-specific efficient decisions $\mathcal{X}^\alpha(\mathbf{z})$ and the corresponding 'Pareto-frontier' $\mathcal{Y}^\alpha(\mathbf{z})$ in reward space that *lower bounds* random rewards with a high probability of at least $1 - \alpha$:

$$\mathbb{P}\big(\mathbf{y}^\alpha(\mathbf{x}, \mathbf{z}) \preceq \mathbf{y} \mid \mathbf{x}\big) \geq 1 - \alpha, \qquad \forall \mathbf{x} \in \mathcal{X}. \tag{1}$$

If a nontrivial bound can be found, we use it to define efficient decisions that take into account the uncertainty of outcome. We also consider a covariate shift, i.e. the training distribution $\tilde{\mathbf{p}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is not the same as the testing distribution $\mathbf{p}(\mathbf{x}, \mathbf{y}, \mathbf{z})$.

Method

We want to ensure (1) . Using the union-bound, we have that:

$$\mathbb{P}\big(\cup_{\mathbf{k}=1}^{\mathbf{m}} \mathbf{y_k} < \mathbf{y}_\mathbf{k}^\alpha(\mathbf{x}, \mathbf{z}) \mid \mathbf{x}\big) \leq \sum_{\mathbf{k}=1}^{\mathbf{m}} \mathbb{P}(\mathbf{y_k} < \mathbf{y}_\mathbf{k}^\alpha(\mathbf{x}, \mathbf{z}) \mid \mathbf{x}).$$

Building upon the 'weighted conformal quantile regression' method [4, 3, 2], we construct $\mathbf{y}_\mathbf{k}^\alpha$ such that:

$$\mathbb{P}(\mathbf{y_k} < \mathbf{y}_\mathbf{k}^\alpha(\mathbf{x}, \mathbf{z}) | \mathbf{x}) \leq \frac{\alpha}{\mathbf{m}} \qquad \forall \mathbf{k},$$

given $\mathbf{p}(\mathbf{z}|\mathbf{x})$. This ensures that (1) is satisfied. The conditional *quantile function* of an individual continuous reward $\mathbf{y}$ can be expressed as:

$$\mathbf{q}^\alpha(\mathbf{x}, \mathbf{z}) = \inf\left\{\tilde{\mathbf{y}} : \frac{\alpha}{\mathbf{m}} = \mathbb{P}(\mathbf{y} < \tilde{\mathbf{y}} \mid \mathbf{x}, \mathbf{z})\right\}.$$

This quantity is unknown and $\hat{\mathbf{q}}^\alpha(\mathbf{x}, \mathbf{z})$ is estimated from data using *any* flexible machine learning method. Determine an *adjustment* variable $\kappa^\alpha(\mathbf{x})$ and form:

$$\mathbf{y}^\alpha(\mathbf{x}, \mathbf{z}) := \hat{\mathbf{q}}^\alpha(\mathbf{x}, \mathbf{z}) - \kappa^\alpha(\mathbf{x}).$$

The adjustment variable is needed to guarantee coverage in finite samples, but should also be weighted to take distributional shift into account

$$\tilde{\mathbf{p}}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \mathbf{p}(\mathbf{z})\mathbf{p}(\mathbf{x}|\mathbf{z})\mathbf{p}(\mathbf{y}|\mathbf{x}, \mathbf{z}), \quad \mathbf{p}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \mathbf{p}(\mathbf{z})\mathbb{1}(\mathbf{x} = \mathbf{x}^*)\mathbf{p}(\mathbf{y}|\mathbf{x}, \mathbf{z}),$$

using weights $\mathbf{w}(\mathbf{x}, \mathbf{z})$ that must be estimated when the past decision policy is unknown.

Synthetic Data

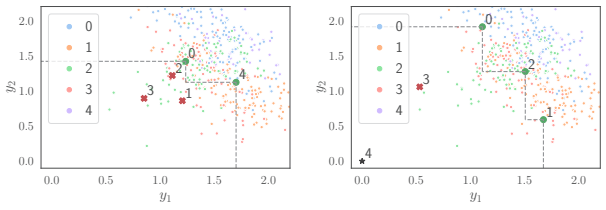
One-dimensional covariate $\mathbf{z} \sim \mathcal{N}(60, 100)$. Five decision alternatives $\mathcal{X} = \{0, 1, 2, 3, 4\}$. The rewards $\mathbf{y}$ are drawn as:

$$\begin{cases} \mathbf{y}_1 &= \mathbf{a_x} + \mathbf{b_x}\mathbf{f}\left(\frac{\mathbf{z}-55}{9}\right) + \mathbf{u}_0, \\ \mathbf{y}_2 &= \mathbf{c_x} + \mathbf{d_x}\mathbf{f}\left(\frac{\mathbf{z}-50}{8}\right) + \mathbf{u}_1, \end{cases}$$

where $\mathbf{f}(\cdot)$ is the sigmoid function, $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ are constants and $\mathbf{u}_0, \mathbf{u}_1$ are noise terms. We create an unbalanced scenario where some of the decisions $\mathbf{x}$ are rare and learn a (two-component) Gaussian mixture model of $\mathbf{p}(\mathbf{z}|\mathbf{x})$ in:

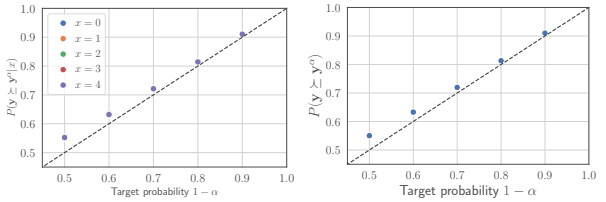
$$\mathbf{w}(\mathbf{x}, \mathbf{z}) = \frac{\mathbb{1}(\mathbf{x} = \mathbf{x}^*)}{\mathbf{p}(\mathbf{x}|\mathbf{z})} = \frac{\mathbb{1}(\mathbf{x} = \mathbf{x}^*)}{\mathbf{p}(\mathbf{z}|\mathbf{x})\mathbf{p}(\mathbf{x})} \sum_{\mathbf{x}' \in \mathcal{X}} \mathbf{p}(\mathbf{z}|\mathbf{x}')\mathbf{p}(\mathbf{x}').$$

Results: Synthetic Data



(a) $\mathbf{z} = 56$ (b) $\mathbf{z} = 68$

Figure: Efficient decisions with confidence 80% for two different contextual covariates (a) and (b). In (b) there was no data for decision $\mathbf{x} = 4$ and the method adapts by setting the minimum possible reward as a boundary (a black star). The dashed lines illustrate the frontiers $\mathcal{Y}^\alpha(\mathbf{z})$ for each context.

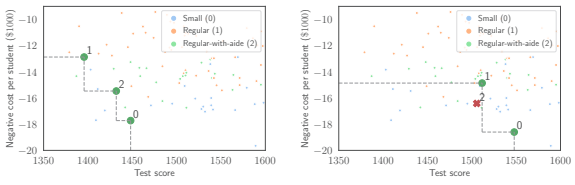


(a) (b)

Figure: Evaluation of reward boundary $\mathbf{y}^\alpha(\mathbf{x}, \mathbf{z})$ using 500 Monte Carlo simulations. (a) Experimental data scenario with known random assignment probabilities $\mathbf{p}(\mathbf{x})$. (b) Observational data scenario where $\mathbf{p}(\mathbf{z}|\mathbf{x})$ is estimated.

Semi-real Data

- Teachers and students randomly assigned into:
  - Small class size: 13 to 17 students per teacher,
  - Regular class size: 22 to 25 students per teacher,
  - Regular-with-aide class size: 22 to 25 students per teacher plus a full-time teacher assistant.
- The study was a *randomized controlled trial*, with interventions assigned by a given random policy $\mathbf{p}(\mathbf{x})$.
- Outcome $\mathbf{y}_1$ is an achievement test score (the sum of standardized math, reading and listening test scores).
- Outcome $\mathbf{y}_2$, is a non-linear generated value [1] and could correspond to a (negative) cost per student.



(a) $\mathbf{z}'$ (b) $\mathbf{z}''$

Figure: Frontiers $\mathcal{Y}^\alpha(\mathbf{z})$ for student groups with two different sets of covariates (a) and (b) and learned from real data with three different class size interventions. Confidence level 80%.

Conclusions

In this work we extended the concept of Pareto-efficiency to include statistical confidence. We propose a method that learns efficient decisions and a frontier that takes into account the lower tail rewards at any specified confidence level. Such a frontier can quantify trade-offs and provide valuable insight to decision makers due to its finite-sample statistical guarantees.

Acknowledgment

This research was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation.

References

[1] J. L. Hill.
Bayesian nonparametric modeling for causal inference.
Journal of Computational and Graphical Statistics, 20(1):217–240, 2011.

[2] L. Lei and E. J. Candès.
Conformal Inference of Counterfactuals and Individual Treatment Effects.
arXiv preprint arXiv:2006.06138., 2020.

[3] Y. Romano, E. Patterson, and E. Candès.
Conformalized quantile regression.
Advances in Neural Information Processing Systems, 32, 2019.

[4] R. J. Tibshirani, R. Foygel Barber, E. Candès, and A. Ramdas.
Conformal prediction under covariate shift.
Advances in Neural Information Processing Systems, 32, 2019.

sofia.ek@it.uu.se

Englesson, Erik
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Generalized Jensen-Shannon Divergence Loss for Learning with Noisy Labels

Prior works have found it beneficial to combine provably noise-robust loss functions e.g., mean absolute error (MAE) with standard categorical loss function e.g. cross entropy (CE) to improve their learnability. Here, we propose to use Jensen-Shannon divergence as a noise-robust loss function and show that it interestingly interpolate between CE and MAE with a controllable mixing parameter. Furthermore, we make a crucial observation that CE exhibit lower consistency around noisy data points. Based on this observation, we adopt a generalized version of the Jensen-Shannon divergence for multiple distributions to encourage consistency around data points. Using this loss function, we show state-of-the-art results on both synthetic (CIFAR), and real-world (e.g., WebVision) noise with varying noise rates.

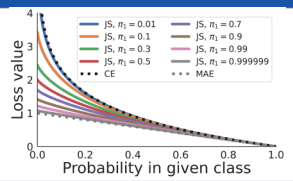
Englesson, Erik
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

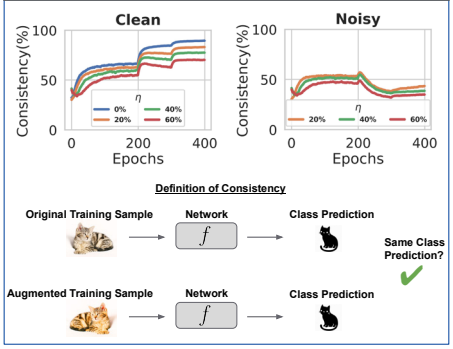
Generalized Jensen-Shannon Divergence Loss for Learning with Noisy Labels

Erik Englesson, Hossein Azizpour

1. Jensen-Shannon Loss Generalizes CE and MAE



2. Learning with Label-Noise Reduces Consistency (CE)



3. Generalized JS loss to Improve Consistency

y = One-hot label

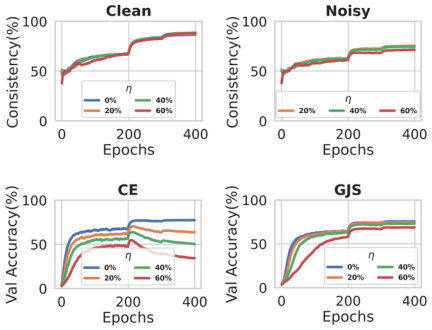
p_1, p_2 = Predictive distributions of augmented versions of the sample

$$\mathrm{GJS}(y, p_1, p_2) := \overbrace{\mathrm{JS}_\pi(y, \frac{p_1+p_2}{2})}^{\text{Label-dependent term}} + \overbrace{(1-\pi)\mathrm{JS}_{\frac{1}{2}}(p_1, p_2)}^{\text{Consistency term}}$$

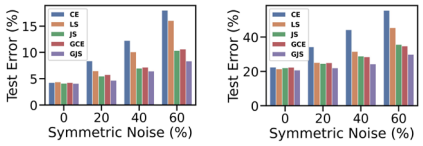
$$\mathrm{JS}_\pi(p, q) := \pi \mathrm{KL}(p \| m) + (1-\pi)\mathrm{KL}(q \| m)$$

$$m := \pi p + (1-\pi) q$$

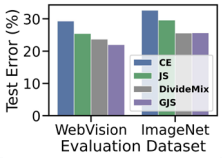
4. GJS Improves Consistency and Generalization



5. Results: Synthetic Noise - CIFAR-10 & CIFAR-100



6. Results: Real-World Noise - WebVision



References

- Wei, Jiaheng, and Yang Liu. "When Optimizing f-divergence is Robust with Label Noise." ICLR 2021.
- Zhang et al. "Generalized cross entropy loss for training deep neural networks with noisy labels." *NeurIPS* 2018.
- Ghosh et al. "Robust loss functions under label noise for deep neural networks." *AAAI* 2017.
- Hendrycks, et al. "AugMix: A simple data processing method to improve robustness and uncertainty." ICLR 2020.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Eriksson, Hannes
Zenseact AB

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Safe Decision-Making for Autonomous Driving

For sequential decision-making problems such as autonomous driving it is imperative to consider the full range of outcomes as they might range from arriving to a location faster than expected, to being part of a catastrophic crash. In particular, we are mostly concerned with the left-tail properties of the distribution of outcomes.
By devising risk-aware agents that focus on performance in the worst outcomes, we can arrive at safer decision-makers.

Eriksson, Hannes
Zenseact AB

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Safe Decision-Making for Autonomous Driving

Hannes Eriksson
hannese@chalmers.se, hannes.eriksson@zenseact.com
Zenseact AB, Chalmers University of Technology

Abstract

For sequential decision-making problems such as autonomous driving it is imperative to consider the full range of outcomes as they might range from arriving to a location faster than expected, to being part of a catastrophic crash. In particular, we are mostly concerned with the left-tail properties of the distribution of outcomes.
By devising risk-aware agents that focus on performance in the worst outcomes, we can arrive at *safer* decision-makers.

Paper I: Epistemic Risk-sensitive Reinforcement Learning

In this work we studied the concept of *epistemic risk*, that is the risk that arises due to uncertainty about the model parameters μ. Typically this situation occurs when we have a belief over MDPs $\xi(\mu)$ and we want to optimize for a risk-sensitive objective w.r.t. the uncertainty due to ξ.
The main contributions were defining an entropic risk measure for epistemic risk and the delivery of two algorithms, one based on **Approximate Dynamic Programming** and one based on **Bayesian policy gradient**.

$$\pi^E(U) = \arg\max_\pi \frac{1}{\beta} \log \mathbb{E}\Big(\exp(\beta R)\Big). \quad (1)$$

Eq. 1. defines the objective over the utility function mirroring an entropic risk measure with some nice properties. By also considering the uncertainty induced by ξ we can arrive at the full objective in Eq. 2., by replacing U with the considered utility function in Eq. 1.

$$\pi^E(U, \xi) \triangleq \arg\max_\pi \int_{\mathcal{M}} U(\mathbb{E}^\pi_\mu[R])\, \mathrm{d}\xi(\mu). \quad (2)$$

SENTINEL Schematic

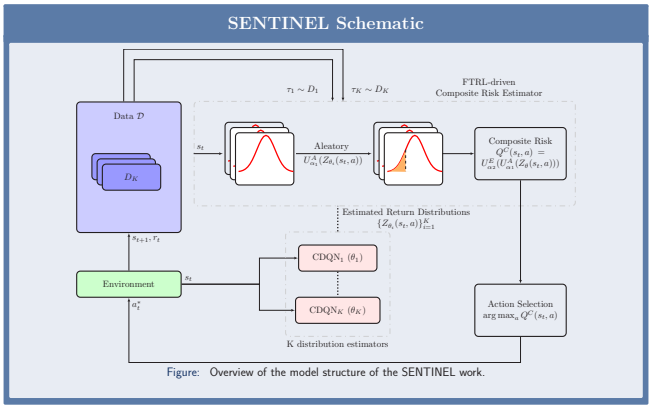


Figure: Overview of the model structure of the SENTINEL work.

Paper II: Inferential Induction: A Novel Framework for Bayesian Reinforcement Learning

In this work we introduced a novel Bayesian Reinforcement Learning framework that correctly infers value function distributions from data. From this framework, depending on what you marginalize over, gives arise to a whole new class of BRL algorithms. In particular, we develop and demonstrate comparable to state-of-the-art performance of **Bayesian Backwards Induction**.

$$\mathbb{P}^\pi_\beta(V_i \,|\, D) = \int_{\mathcal{V}} \mathbb{P}^\pi_\beta(V_i \,|\, V_{i+1}, D)\, \mathrm{d}\mathbb{P}^\pi_\beta(V_{i+1} \,|\, D). \quad (3)$$

Paper III: SENTINEL: Taming Uncertainty with Ensemble-based Distributional Reinforcement Learning

In *SENTINEL*, we study a novel kind of risk measure, in this work termed *composite risk*, which combines both the risk due to aleatory uncertainty and the risk due to epistemic uncertainty into one risk measure. We prove that this new risk measure better estimates the total risk than one that considers both risks separately.

$$\text{Comp. Risk} \triangleq \int_\Theta \int_Z Z\, \mathrm{d}(U^A_{\alpha_1} \circ \Pr)(Z|\theta)\, \mathrm{d}(U^E_{\alpha_2} \circ \beta)(\theta) \quad (4)$$

Further, we demonstrate how to design an agent that optimizes for this risk, using distribution estimators, as seen in the above schematic.

Paper IV: *In progress* Transfer Reinforcement Learning with Risk

In an ongoing work we are considering techniques that leverage knowledge transfer from a set of *source* domains to a *target* domain. This setting is interesting when you for instance have a task that you know how to solve and you now want to solve a different task but with similar structure. For instance, knowing how to drive in **Europe** should inform you to some extent how to drive in the **US**, but there are some differences, namely traffic rules, road signs and road behavior.

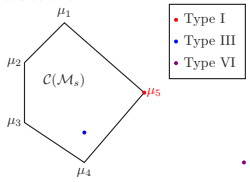


Figure: Overview some of the types of settings that may arise in Transfer Reinforcement Learning.

In particular, we consider three kinds of structures over model space. The first, which gives arise to *Type I*, assumes the target MDP μ_t is part of the finite set of source MDPs $\mathcal{M}_s$. The second, *Type III*, which considers the convex set of source MDPs $\mathcal{C}(\mathcal{M}_s)$, searches for the best representative $\hat{\mu}_t \in \mathcal{C}(\mathcal{M}_s)$. Finally, the last type of structure is the general case where the target MDP can be arbitrarily different from the source MDPs. In that case, you are studying problems of *Type V*.

Acknowledgements

I want to thank all my co-authors and the people I have worked with, some of these are, Emilio Jorge, Divya Grover, Debabrota Basu, Tommy Tram, Christos Dimitrakakis, Mina Ali-Beigi Nabi, Aristide Tossou.

Faridghasemnia, Mohamadreza
Örebro University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Robot Learning of Symbol Grounding in Multiple Contexts Through Dialog

When a person talks to a robot about an object in the environment, the robot has to find which object the person is talking about. Symbol grounding is the task of finding a link between a word and one of the observed objects. It is a complex task: What information the robot can perceive (language and vision)? What links exist between language and vision? What links exist between the objects that can help symbol grounding? In (1) we build a platform with a Pepper robot that can see, listen, and talk about objects that it can see, and save this information. In (2) we extended (1) and assumed that the user wants to ground objects implicitly. In (3) we discuss a better vision model, that is not limited to recognizing the category of objects.

Faridghasemnia, Mohamadreza
Örebro University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

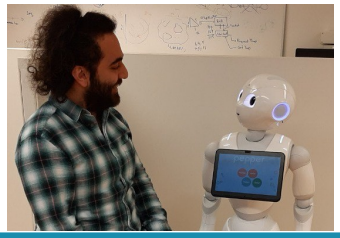
Robot Learning of Symbol Grounding in Multiple Contexts Through Dialog

Mohamadreza Faridghasemnia, Örebro University
Centre for Applied Autonomous Sensor Systems (AASS)
Supervisors: Prof. Alessandro Saffiotti, Prof. Lars Karlsson
Project partner: Bram Willemsen (KTH)

ÖREBRO UNIVERSITET

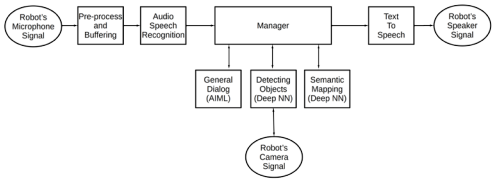
A brief outline

When a person talks to a robot about an object in the environment, the robot has to find which object the person is talking about. **Symbol grounding**: The task of finding a link between a word and one of the observed objects. It is a complex task: What information the robot can perceive (language and vision)? What links exist between language and vision? What links exist between the objects that can help symbol grounding? In 1. we build a platform with a Pepper robot that can see, listen, and talk about objects that it can see, and save this information. In 2. we extended (1) and assumed that the user wants to ground to objects implicitly. In (3) we discuss a better vision model, that is not limited to recognizing the category of objects.



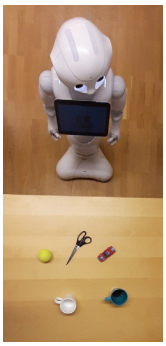
(1). Extract information from language and vision

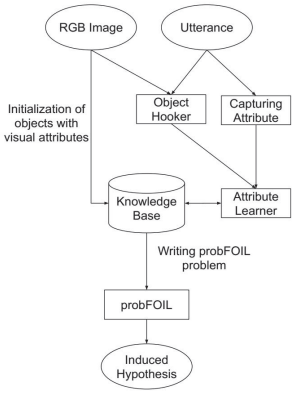
Audio Speech Recognition: Audio signal to phrase.
Semantic mapping: Extract frame and frame element of language.
Vision: Extract object categories in the image.
AIML: Handling general dialog.
Manager: Holds detected object, Holds information from Semantic mapping module, Generate an utterance to confirm the user input
Manager works with symbols, while symbols are generated from applying deep neural networks on sensory inputs.

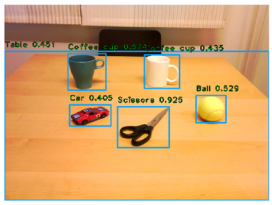


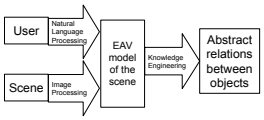
(2). Relationships that help symbol grounding

Ground to an object that is green and belongs to Mary. We have a mug that is green and kitchenware, a mug that is white and kitchenware, a scissor that is black, kitchenware, and belongs to Mary, and a ball and a car that are toys and are for Tom.









Robot

Hi there!

I see, the white mug is on the table

Got it! it belongs to mary

Ok, I save it as kitchenware

Human

Hello.

the white mug on the table

i guess it is for mary

its label is kitchenware

EAV data:

$mug(object001)$
$white(color001)$
$on-table(position001)$
$kitchenware(label001)$
$1.0 :: mary(object001, color001, position001, label001)$
$0.0 :: john(object001, color001, position001, label001)$
To: $mary(A, B, C, D, ...) : -kitchenware(C)$

(3). Extracting rich information from vision

The vision model was a deep Neural network object detector, and colors were recognized from RGB values. Why not a smarter vision model?

References

[1] Faridghasemnia, M. , et al., (2020). Towards Abstract Relational Learning in Human Robot Interaction.
[2] Faridghasemnia, M. (2020). Reducing Uncertainty in Contextualized Dialog.
[3] Faridghasemnia, M. , et al., (2019). Capturing Frame-Like Object Descriptors in Human Augmented Mapping.

Fay, Dominik
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Federated Learning for Smart Radiotherapy Systems

Today, treatment planning for radiotherapy involves several tedious and time-consuming tasks such as the segmentation of tumors in an MR image. Automating these tasks using machine learning could make radiotherapy substantially more cost-effective. Here, complications arise in cases where patient data must be stored locally at the clinic for privacy reasons. We investigate federated learning as a solution to learn a global privacy-preserving model. Aside from data storage requirements, we also aim for learning algorithms that satisfy statistical notions of privacy, adapt to local differences in clinical practice and continuously update the global model efficiently.

Fay, Dominik
KTH

Fredin Haslum, Johan
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Exploring the use of Phenotypic Screening data for Bioactivity Prediction

Developing new drugs is a long and costly process, both in terms of time and resources. Early stages of the process include High Throughput Screenings (HTS), probing the effects of hundreds of thousand or millions of compounds, on the target of interest. Predicting the outcome of such experiments could speed-up the drug discovery process and provide new insight into the underlying biological processes. Phenotypic screening data could potentially provide a high-resolution method for characterizing such compounds. In this work we explore the use of phenotypic screening data for predicting compound bioactivity. Our results indicate that the data captures information that can be linked to activity in a wide variety of assay targets, as well as information directly related to the underlying biology by confirming predictive performance across assay and target types as well as validating the predictive performance across assays.

Deep Learning for Drug Discovery

Johan Fredin Haslum[1,2,3]

Supervisors: Kevin Smith[1,2], Erik Müllers[3], Karl-Johan Leuchowius[4], Riku Turkki[5]

[1] Science for Life Laboratory, Stockholm, Sweden

[2] School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

[3] Bioscience Cardiovascular, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

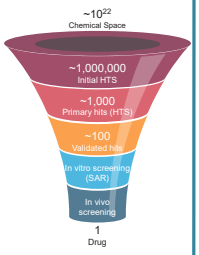
[4] Mechanistic Biology, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden , [5] Data Sciences & Quantitative Biology, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden

Exploring the use of Phenotypic Screening data for Bioactivity Prediction

Developing new drugs is a long and costly process, both in terms of time and resources. Early stages of the process include High Throughput Screenings (HTS), probing the effects of hundreds of thousand or even millions of compounds, on the target of interest. Predicting the outcome of such experiments could speed-up the drug discovery process and provide new insight into the underlying biological processes. Phenotypic screening data could potentially provide a high-resolution method for characterizing such compounds. In this work we explore the use of phenotypic screening data for predicting compound bioactivity. Our results indicate that the data captures information that can be linked to activity in a wide variety of assay targets, as well as information directly related to the underlying biology by confirming predictive performance across assay and target types as well as validating the predictive performance across assays focused on the same target.

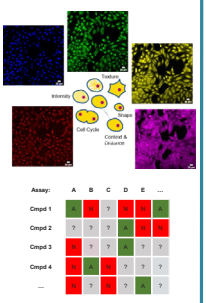
Problem

Identifying compounds active towards a target of interest is an important step of early drug discovery. High Throughput Screening (HTS) Assays, are experiments focused on identifying compounds with biological activity towards a particular target and are designed in order to efficiently screen millions of compounds. Bioactivity Prediction is an attractive solution to limit the number of experiments that needs to be performed. Enabling the prioritization of compounds with higher likeliness of activity. Potentially reducing the size of such HTS Screens.



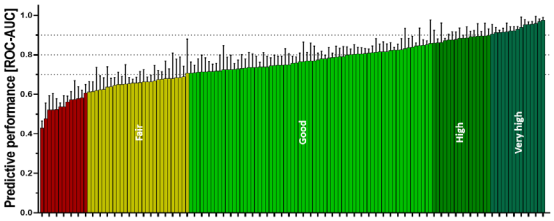
Method

Predicting compound bioactivity, based on phenotypic screening data is an attractive alternative to compound structure-based approaches[4], due to it's potential of capturing a compound's effect on a biologically system. Previous work have shown that such data contain information relevant for predicting bioactivity in orthogonal assays [1,2]. In this work we built a phenotypic screening dataset, using the Cell Painting protocol [3] to generate high resolution fluorescent microscopy images. In effect characterizing compound by their biological effect in a cell line and the morphological changes it induces. Combined with repurposed bioactivity readouts for a wide range of compounds and targets. A CNN is trained in a supervised multi-task manner to predict bioactivity in multiple assay at once.



Results

Predicting Bioactivity over 140 different assays using Cell Painting images as input. 6-fold cross validation gives an average ROC-AUC prediction score of 0.744. With a majority (71%) of assays being predicted with a ROC-AUC above 0.7. Indicating that information relevant for assay activity prediction is captured in the Cell Painting images and that our network is able to recognize it.



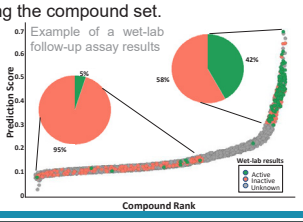
Screening compounds according to the models ranking, significantly enriches the number of actives identified. Yielding a much higher HIT-rate than expected by randomly sampling the compounds, but also better than competing methods using compound structure (ECFP4 [4]) or Cell Profiler features [5] as input.

Input modality	Phenotypic Deep Learning	Phenotypic Cell Profiler	Structure Based, ECFP4
ROC-AUC	0.744	0.714	0.687

Beyond higher ROC-AUC performance, we show that Cell Painting based predictions results in a more diverse set of compounds, compared to structure based. With a Wilcoxon two-sided test of structure diversity between the top ranked compounds in the test set and the known hits from the training set, we find that there is a significant difference between the two input modalities.

To further validate the performance of our model we also probe the activity of the top ranked compounds in several follow-up assays, meant to confirm the activity of each active compound under different experimental conditions. Five different assays with varying targets and model performances were selected and followed up in wet-lab experiments. Showing similar results in HTS and wet-lab assays, confirming that the network is capturing target relevant features when predicting bioactivity and enriching the compound set.

Assay	HTS Assay ROC-AUC	Enrichment at X percentile	
		90%	98%
A	0.95	6.3	30
B	0.68	3.6	15
C	0.94	3.8	16
D	0.63	4.2	7
E	0.81	3.5	3.9



References

1. Simm, Jaak, et al. "Repurposing high-throughput image assays enables biological activity prediction for drug discovery." Cell chemical biology 25.5 (2018): 611-618.
2. Hofmarcher, Markus, et al. "Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks." Journal of chemical information and modeling 59.3 (2019): 1163-1171.
3. Bray, Mark-Anthony, et al. "Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes." *Nature protocols* 11.9 (2016): 1757-1774.
4. Rogers, David, and Mathew Hahn. "Extended-connectivity fingerprints." *Journal of chemical information and modeling* 50.5 (2010): 742-754.
5. Carpenter, Anne E., et al. "CellProfiler: image analysis software for identifying and quantifying cell phenotypes." Genome biology 7.10 (2006): 1-11.

Fu, Jingru
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Generative Aging of Brain Images with Diffeomorphic Registration

Predicting subject-specific brain aging can be used for improving the diagnosis and prognosis in neurodegenerative diseases. Previous approaches have been restricted to group-level predictions, or yielded unreal results. This study addresses these issues by proposing a novel method that generates synthetic MRI images of the brain to simulate its changes due to aging. The method is based on diffeomorphic image registration, which can provide more accurate and fidelity-controllable subject-level predictions.

Fu, Jingru
KTH

Generative Aging of Brain Images with Diffeomorphic Registration

Jingru Fu[1], Daniel Ferreira[2], Rodrigo Moreno[1]
1 KTH Royal Institute of Technology
2 Karolinska Institute

Research goals

Predicting subject-specific brain aging can be used for improving the diagnosis and prognosis in neurodegenerative diseases. Previous approaches have been restricted to group-level predictions [1], or yielded unreal results [4]. This study addresses these issues by proposing a novel method that **generates synthetic MRI images** of the brain to simulate its changes due to **aging.** The method is based on **diffeomorphic image registration**, which can provide more ***accurate*** and ***fidelity-controllable subject-level*** predictions.

Method

Diffeomorphic Registration:

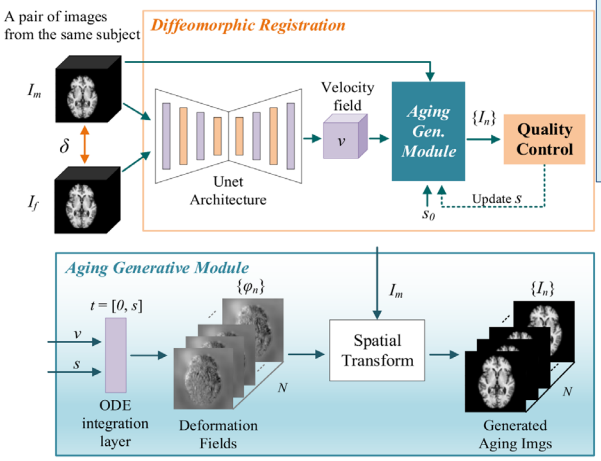
- **Diffeomorphic registration (*DR*)** enjoys many advantages [2], such as the learned deformation field are differentiable and invertible, and thus preserve topology. According to [3], the contraction of the brain structures shows a linear change in old age. This gives us theoretical support to use the registration of two images at different ages to simulate the changes in brain aging.

Architecture:

- As shown in the figure below, we used images acquired at different time points from the same subject. The objective is to generate images in between these two time points using the velocity field estimated through deep learning-based diffeomorphic registration.

Two modules are introduced within the *DR* framework:

- ***Aging Generative Module (AGM)*** integrates the velocity fields at different time steps to generate several deformation fields that are used to generate the synthetic images.
- ***Quality Control Module (QCM)*** takes synthetic images quality into consideration, leading to adjusting the stopping point of the integration layer (s) in the *AGM*. It is possible to extrapolate images by using s > 1.

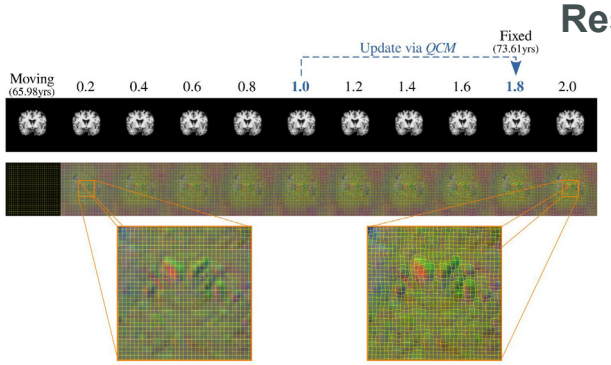


I_f: fixed image of subject
I_m: moving image of subject
$\{I_n\}$: set of generated imags
δ: age difference
N: number of generated images
$\{\varphi_n\}$: set of deformation fields
t: integration interval
s: stopping point of integration layer

References

1. Sharmin Pathan and Yi Hong. *Predictive image regression for longitudinal studies with missing data.* arXiv preprint arXiv:1808.07553, 2018.
2. Adrian V Dalca et al. *Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces.* **Medical image analysis**, 57:226–236, 2019.
3. Kristine B. Walhovd et al. *Effects of age on volumes of cortex, white matter and subcortical structures.* **Neurobiology of Aging**, 26(9):1261–1270, October 2005. ISSN 0197-4580.
4. Viktor et al. *Generative Aging of Brain MR-Images and Prediction of Alzheimer Progression.***GCPR**, 2019.

Results



Generative aging of one sample subject from the non-demented subset of the OASIS-3 dataset. The first row represents the generated images from 0 to 2; the second row represents the corresponding deformation.

Results show in the left figure:

- Twenty images were generated for this subject with values of *t* between 0 and 2 (the moving image is at *t*=0). Although t=1 should theoretically match the fixed image, in practice it is not (*t*=1.8 in the example). The *QCM* adjusts *s* with quality measurements to make the generated images better correspond to chronological age.

Conclusions:

- As shown in the second row of the image, brain deformations are increasing with time.
- The entire generation for an individual can be completed in about five minutes using only CPU.
- We have generated synthetic images for three large-scale longitudinal datasets so far.

Gedon, Daniel
Uppsala University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

ResNet-based ECG Diagnosis of Myocardial Infarction in the Emergency Department

Myocardial infarctions (MIs) are often missed in the emergency department. In managed settings deep learning models have shown promise in electrocardiogram (ECG) classification. However, in a real-world scenario there is a lack of high performing models for classification of MIs. We developed a ResNet-based deep neural network to classify the ECG between non-ST-elevation MI (NSTEMI), ST-elevation MI (STEMI), and control status in the more challenging real-world setting. In a test set, our model discriminates STEMIs/NSTEMIs with an AUROC of 0.99/0.83 and a Brier score of 0.05/0.05. The model also generalizes well and obtains a similar performance on an additional test set collected in the months following the initial collection and that does not overlap temporally with the set used for developing the model. Our results are above human-level performance reported in previous studies for STEMIs and NSTEMIs.

Gedon, Daniel
Uppsala University

Gillsjö, David
Lund University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

In Depth Semantic Scene Completion

This work studies Semantic Scene Completion which aims to predict a 3D semantic segmentation of our surroundings, even though some areas are occluded.
For this we construct a Bayesian Convolutional Neural Network (BCNN), which is not only able to perform the segmentation, but also predict model uncertainty.
This is an important feature not present in standard CNNs.
We show results for the Semantic Scene Completion task where a category is introduced at test time on the SUNCG dataset.
In this complex task the Bayesian approach outperforms the standard CNN, showing better Intersection over Union score and excels in mean Average Precision.
With the added benefit of having better calibrated scores and the ability to express model uncertainty.

Gillsjö, David
Lund University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

In Depth Semantic Scene Completion

David Gillsjö, Kalle Åström

Centre for Mathematical Sciences, Lund University, Sweden

Introduction

Semantic Scene Completion (SSC) is a challenging task in which both visible and occluded surfaces are labeled semantically in 3D. In Figure 1 we see an illustration of the problem where a UAV would benefit from knowing what to expect in occluded areas.

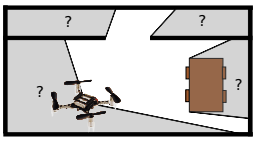


Figure 1: An UAV has some occluded areas in its surrounding and would like to have an idea about what to expect.

Our contributions include:

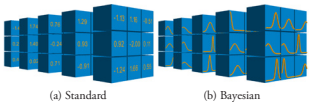
- An open source system for BSSC using Variational Inference released on `https://github.com/DavidGillsjo/bssc-net`.
- An extended SSC task on the SUNCG dataset with more occluded space.
- Experiments showing that the Bayesian approach is more robust to unseen data in the SSC task.
- Parameter studies on both MNIST and SUNCG.

Bayes by backprop

This method introduced by [1] is based on Variational Inference. Each weight in the network is sampled from a normal distribution, as illustrated in Figure 2.
We estimate the posterior $P(w|\mathcal{D})$ using a simpler model $q(w|\theta)$ with learnable parameters θ, which minimizes the approximate Kullback-Leibler (KL) divergence to the true posterior.

$$\theta^* = \arg\min_\theta \sum_{i=1}^{n} \frac{\beta}{n} \underbrace{\left[\log q(w^{(i)}|\theta) - \log P(w^{(i)})\right]}_{\text{Complexity}} \underbrace{- \log P(\mathcal{D}|w^{(i)})}_{\text{Likelihood}}.$$

where $w^{(i)}$ is a sample from the variational posterior $q(w^{(i)}|\theta)$. The scale factor $\frac{\beta}{n}$ with β as design parameter is introduced to tune the regularization.



(a) Standard (b) Bayesian

Figure 2: In 2a we see a filter bank from a standard 2D CNN, each weight is a scalar. In 2b we see a filter bank in a Bayesian Variational Inference 2D CNN, here each weight represented as a distribution which is sampled from at inference time.

Prediction & Uncertainty

An unbiased estimation of the expectation is given [2] by

$$\mathbb{E}_{q(w|\theta)}[P(\hat{y}|\hat{x}, w)] = \int q(w|\theta) p_t \, dw \approx \frac{1}{T} \sum_{t=1}^{T} p_t,$$

where $p_t := P(\hat{y}|\hat{x}, w^{(t)})$ is the softmax output from forward pass t.
For uncertainty we use Predictive Entropy,

$$H = -\sum_{t=1}^{T} p_t \log p_t.$$

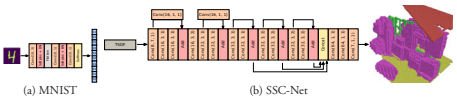
For metrics we use mean Average Precision (mAP), Intersection over Union (IoU) for performance. For separation metric we use the Bhattacharyya coefficient (BC)

$$BC(p, q) = \frac{1}{N} \sum_{i=1}^{N} \sqrt{p_i q_i},$$

where N is the number of categories, q_i and p_i are the number of TP and FN. Lower score indicates better separation.

Model

We have explored two network architectures. The first network architecture is inspired by the original SUNCG article [3]. We call it **SSC-Net**. The second architecture is a **UNet**. We chose softplus as activation functions instead of relu to have more active weights in the network [2]. The architecture is displayed in Figure 3.



(a) MNIST (b) SSC-Net

Figure 3: Architecture of SSC-Net used for MNIST and SUNCG experiments. Conv(d, k, l) stands for a 3D convolution filter stack of depth d and kernel size k and dilation l. Batch normalization and softplus activation is performed after every Conv layer. Softmax in the final layer.

Selected References

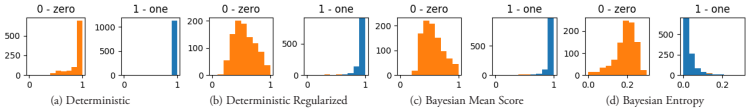
[1] C. Blundell, J. Cornebise, K. Kavukcuoglu and D. Wierstra, 'Weight uncertainty in neural networks,' *arXiv preprint arXiv:1505.05424*, 2015.

[2] K. Shridhar, F. Laumann and M. Liwicki, *A comprehensive guide to bayesian convolutional neural network with variational inference*, 2019. arXiv: 1901.02731 [cs.LG].

[3] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva and T. Funkhouser, 'Semantic scene completion from a single depth image,' *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

MNIST Experiment

In Figure 4 we see output distributions from MNIST test set for digits 0 and 1 when 0 is removed from the training data. The Bayesian Score is more better calibrated and the Entropy is higher for 0.



(a) Deterministic (b) Deterministic Regularized (c) Bayesian Mean Score (d) Bayesian Entropy

Figure 4: Here we see true (blue) and false (orange) predictions for 0 and 1.

SUNCG Experiment

SUNCG [3] is a large dataset with manually created and labeled synthetical indoor scenes. We've used a subset of 2000 training and 1000 testing scenes for the experiments. In Figure 5 we show a parameter study on β.Figure 6 shows example output.
We also conducted an experiment when category *bed* was removed from training, the result is presented in Table 1.

Table 1: BC, mAP and mIoU for different network architectures when the *bed* class is removed from training. S=Score, E=Entropy. We observe that Bayesian SSC-Net has the best score in all metrics.

CNN	mIoU	mAP: S	mAP: E	BC: S	BC: E
SSC-Net ω=0	0.19	0.2		0.31	
SSC-Net ω=0.01	0.14	0.23		0.29	
B-SSC-Net	**0.21**	**0.26**	**0.19**	**0.27**	**0.28**

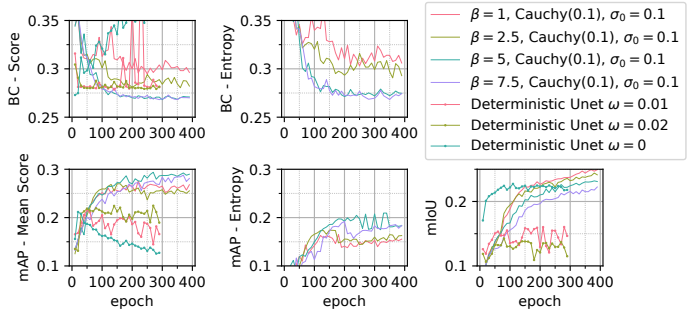


Figure 5: BC, mAP and mIoU for the Bayesian UNet with different weights β and ω for the SUNCG mini dataset. We observe that $\beta = 5$ is better in all metrics but mIoU, where $\beta = 1$ is best.

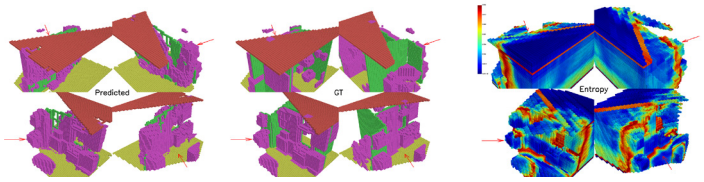


Figure 6: Example from the SUNCG test set. From the left we have predicted, true labels and entropy.

Govindarajan, Hariprasath
Linköping University / Arriver Sweden AB

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Self-Supervised Representation Learning for Content Based Image Retrieval of Complex Scenes

Although Content Based Image Retrieval (CBIR) is an active research field, application to images simultaneously containing multiple objects has received limited research interest. For such complex images, it is difficult to precisely convey the query intention, to encode all the image aspects into one compact global feature representation and to unambiguously define label similarity or dissimilarity. Motivated by the recent success on many visual benchmark tasks, we propose a self supervised method to train a feature representation learning model. We propose usage of multiple query images and use an attention based architecture to extract features from diverse image aspects that benefits from this.

Govindarajan, Hariprasath
Linköping University / Arriver Sweden AB

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Self-Supervised Representation Learning for Content Based Image Retrieval of Complex Scenes

Hariprasath Govindarajan, Peter Lindskog, Dennis Lundström,
Amanda Olmin, Jacob Roll, and Fredrik Lindsten
Department of Computer and Information Science (IDA)

LINKÖPING UNIVERSITY

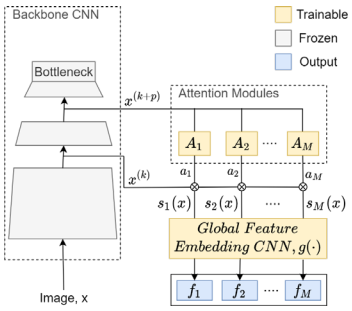
Arriver

Introduction

Although **Content Based Image Retrieval (CBIR)** is an active research field, application to images simultaneously containing multiple objects has received limited research interest. For such complex images, it is difficult to precisely convey the query intention, to encode all the image aspects into one compact global feature representation and to unambiguously define label similarity or dissimilarity. Motivated by the recent success on many visual benchmark tasks, we propose a **self- supervised method** to train a feature representation learning model. We propose usage of **multiple query images** and use an **attention-based architecture** to extract features from diverse image aspects that benefits from this.

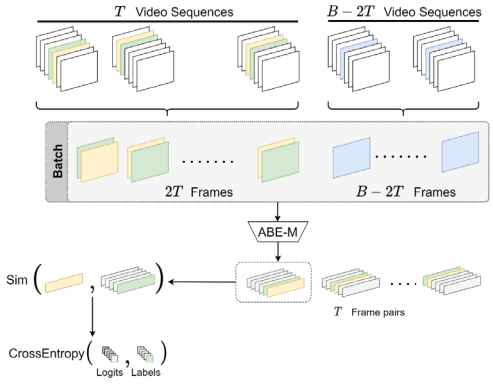
Method

Attention Based Ensemble architecture (ABE-M) [1]



Self-supervised Training

We propose a self-supervised training method based on a contrastive pretext task using frames from videos. We expect a frame to be more similar to an adjacent frame in the video than to frames from other videos.



Encourages similar frames to have similar feature representations.

NCE Loss

$$\mathcal{L}_{\text{NCE},t} = -\log \frac{\exp(f(x_t)\cdot f(x_{t'})/\tau)}{\sum_{x\in X_{-t}} \exp(f(x_t)\cdot f(x)/\tau)}$$

$$\mathcal{L}_{\text{NCE}} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_{\text{NCE},t}$$

Divergence Loss

$$\mathcal{L}_{\text{div}}^{(b)} = \frac{1}{M(M-1)} \sum_{1\le p,q\le M} \max\left(0, \delta - \mathrm{d}\left(f_p^{(b)}, f_q^{(b)}\right)\right)$$

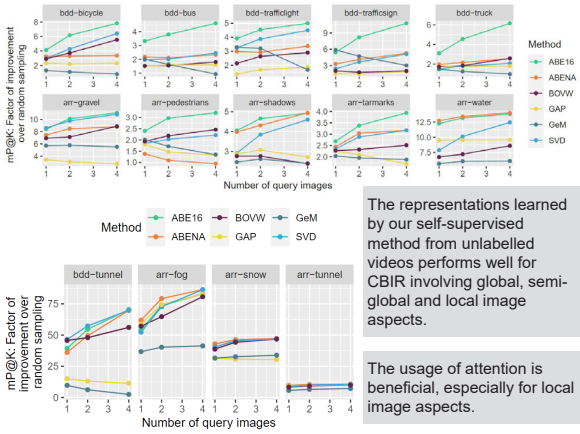
$$\mathcal{L}_{\text{div}} = \frac{1}{B}\sum_{b=1}^{B} \mathcal{L}_{\text{div}}^{(b)}$$

Encourages different attention heads to focus on different spatial regions.

Overall Loss

$$\mathcal{L} = \mathcal{L}_{\text{NCE}} + \lambda \mathcal{L}_{\text{div}}$$

Results



The representations learned by our self-supervised method from unlabelled videos performs well for CBIR involving global, semi-global and local image aspects.

The usage of attention is beneficial, especially for local image aspects.

The usage of multiple query images is useful for conveying query intention for CBIR involving images with multiple image aspects. It also consistently improves the performance of our method on all image aspects.

References

1. W. Kim, B. Goyal, K. Chawla, J. Lee, and K. Kwon, “Attention-based ensemble for deep metric learning,” in ECCV 2018.

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Gower, Alexander
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Automation of scientific discovery in systems biology using active learning

In our research, we are seeking to automate the discovery of scientific knowledge by building a robot scientist – a combination of laboratory automation hardware and artificial intelligence (AI) capable of closed-loop cycles of experimentation. This means the robot scientist will design experiments, execute them, analyse results and generate new scientific knowledge without human intervention. We direct our robot scientists, Eve and Genesis, toward generating new scientific knowledge about the metabolism of the yeast Saccharomyces cerevisiae. Two crucial parts of the scientific process are: hypothesis generation and applying results of experiments to refine theory. We implement experiment selection algorithms to decide gene knockouts based on influence profiles of genes over certain sections of metabolism. The aim is to select experiments that will provide maximum information gain for the area of metabolism we seek to understand.

Gower, Alexander
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Automation of scientific discovery in yeast systems biology using active learning

Alexander Gower[a], Daniel Brunnsåker[a], Ievgeniia Tiukova[a] and Ross King[abc]

[a] Chalmers University of Technology, Gothenburg, Sweden
[b] Cambridge University, Cambridge, United Kingdom
[c] Alan Turing Institute, London, United Kingdom

CHALMERS UNIVERSITY OF TECHNOLOGY

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

The case for a robot scientist

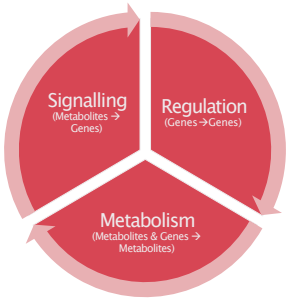
Our research aims to automate the discovery of scientific knowledge by building a **robot scientist** – a combination of laboratory automation hardware and artificial intelligence (AI) capable of closed-loop cycles of experimentation. This means the robot scientist will design experiments, execute them, analyse results and **generate new scientific knowledge without human intervention**. We direct our robot scientists, Eve and Genesis, toward generating new scientific knowledge about the metabolism of the yeast *Saccharomyces cerevisiae*.

A common approach to generate new knowledge about cell metabolism is through **factorial experimentation**, where the scientist creates a mutant strain by removing one or more genes and cultivates the yeast in various conditions. The potential number of experiments is vast–*Saccharomyces cerevisiae* has approximately 6000 genes. As such, heuristics are employed to select experiments that will generate the most knowledge. A robot scientist has many advantages: **selecting experiments likely to yield the most information for the least cost**; greater precision and reproducibility; greater capacity; and high-throughput data analysis.

An integrated model of a yeast cell

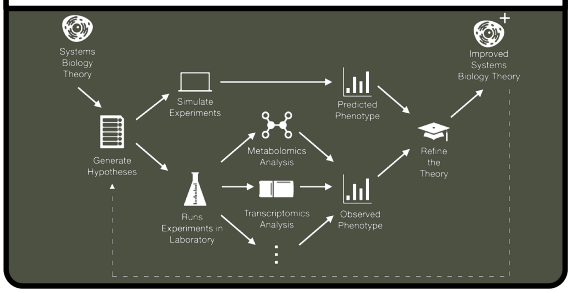
The background knowledge of yeast cell biology that our robot scientists use is in the form of an integrated computational model. Here we use a **modular framework**–we model cell signalling, gene regulation and metabolism separately and integrate them during simulation calculations. This means we can choose model structures that suit the underlying biological processes.

The model for our robot scientists is comprised of: a boolean cell-signaling network; a directed hypergraph of co-regulatory relationships between genes; and a genome-scale metabolic model (Yeast8). The regulatory model is reduced to a dynamic Bayesian network for simulation. As all these processes happen on different timescales and with different magnitudes, we are experimenting with several simulation protocols.



A modular model has other advantages for a robot scientist: machine learning algorithms can be selected to match the model structure and best make use of experimental data; the robot scientist can isolate changes a specific section of the model; and we can select a mathematical formulation that is suited to representing the biological processes (e.g. stoichiometric matrix for metabolic reactions, Boolean rules for signaling).

The robot scientist's research pipeline



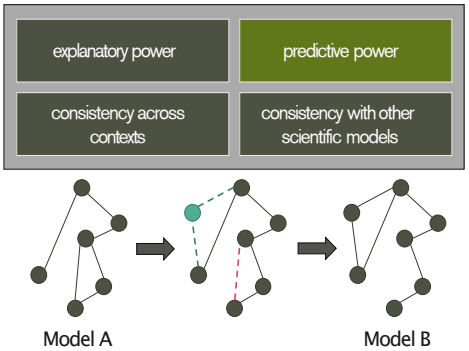
Automated model improvement

A model of yeast cell metabolism can be improved in many different ways. For a robot scientist to be able to make improvements algorithmically, a specific procedure to reach a goal must be specified. We define these goals as **units of discovery**, for example:

- a new regulatory interaction;
- updating an existing regulatory interaction;
- a regulatory effect between groups of genes; and
- a new signaling interaction (link between some condition and some response).

Algorithms can be defined within the structure of our integrated yeast model in order that the **robot scientist can suggest revisions to the model**. These algorithms exploit an optimisation objective that represents the **quality of the model**. There are many such objectives one could use, however we focus on a model's predictive power. In practice this means the robot scientist compares simulations based on different prospective models against experimentally obtained transcriptomic and metabolomic data, along with fermentation profiles.

Metrics of quality for scientific models



[1] A. Coutant et al., 'Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast', *PNAS*, vol. 116, no. 36, pp. 18142–18147, Sep. 2019, doi: 10.1073/pnas.1900548116.

[2] D. T. Banos, P. Trébulle, and M. Elati, 'Integrating transcriptional activity in genome-scale models of metabolism', *BMC Systems Biology*, vol. 11, no. 7, p. 134, Dec. 2017, doi: 10.1186/s12918-017-0507-0.

This work is partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Grönqvist, Johan
Lund University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Closed Loop Guarantees with Neural Networks in Control

We use methods from Robust Control to obtain guarantees of closed loop stability for a dynamical system controlled using a neural network.

Grönqvist, Johan
Lund University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

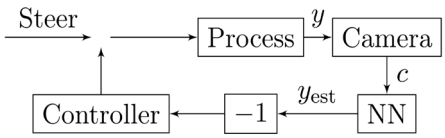
Closed Loop Guarantees with Neural Networks in Control

Johan Grönqvist, Lund University
Department of Automatic Control

LUNDS UNIVERSITET

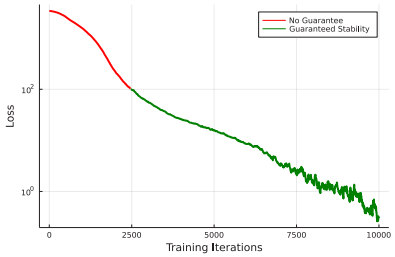
Charge while Driving

- Pick-Up Contact under vehicle
- PID controller to track Charging Rail
- Camera and Neural Network Estimate State
- Master Thesis at Automatic Control LTH



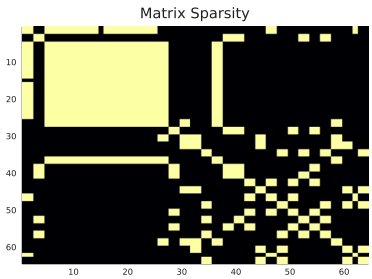
Closed Loop System

- Linear Time-Invariant Process
- Camera and Neural Network Estimate State
- Inaccurate Estimate: Nonlinear Disturbance
- Nonlinear Blocks: Quadratic Forms



Neural Network and Stability Guarantee

- Prediction Accuracy as Loss Function
- NN challenges: Overfitting and Adversaries
- Neurons described by Quadratic Forms
- IQC Proof of Global Stability
- Several Kinds of Guarantees Possible



Sparse Semi-definite Programming

- Problem Dependent Structure
- Depends on Choices in IQC Description
- Challenge: Deep and Large Networks
- Sparsity to the Rescue

Gugliermo, Simona
Örebro University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Can Industrial Transport Applications be improved using AI planning?

AI planning methods are fundamental in industrial transport applications. These methods typically rely on manually-specified knowledge to derive plans. The goal of my PhD is to use Machine Learning to enhance AI planning methods by learning from human planning experts. The purpose is to develop learning algorithms to help humans build domains, aiming at

1. reducing the knowledge engineering effort for humans
2. providing better quality plans in many domains
3. allowing automated planning and scheduling systems to learn how to execute plans or policies from previous experience.

In this poster I present an approach to address this problem with its the main challenges and I highlight the industrial need.

Gugliermo, Simona
Örebro University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Can Industrial Transport Applications be improved using AI planning?

Simona Gugliermo, Örebro University
Center of Applied and Autonomous Sensors (AASS)
Advisors: Federico Pecora, Christos Koniaris

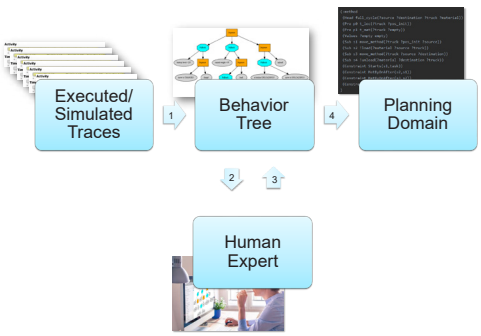
Motivation

- AI planning methods are useful for industrial transport applications.
- Current AI planning methods rely on manually-specified knowledge, encoded in the *planning domain*.
- AI planning domains are difficult to write for humans.
- Aim: use Machine Learning to
  1. Reduce the knowledge engineering effort for humans
  2. Provide better quality plans in many domains
  3. Allow AI planning systems to learn how to execute plans or policies from previous experience.

Proposed Approach

Main Features

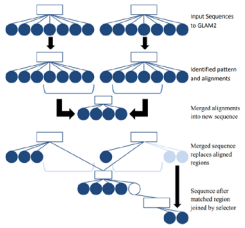
- Involve fleet transport managers in the learning loop
- Make the learning process more transparent and tractable
- Increase the generality step by step
- Create an application-independent method, usable with different scenarios.



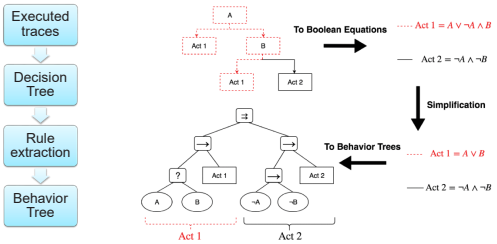
1. Learn a Behavior Tree (BT) from executed traces (data mining techniques).
2. Revise and correct (if needed) the learned Behavior Tree by humans to fully control the resulting behavior.
3. Approval of the final Behavior Tree.
4. Learn a domain description from the Behavior Tree

Research Questions

1. **How to learn behaviors from execution traces?**
   Related work
   - Learn Behavior Trees from traces in the real-time strategy gaming context [1].
     1. A maximally specific BT is created from the given trace
     2. The BT is iteratively reduced in size by finding and combining common patterns of actions.
     3. If no new satisfactory patterns are found, the process stops.



   - Learn Behavior Trees and transfer new skiils using programming by demonstration (PbD) [2,3]



2. **How to facilitate human modelling of BTs?**

3. **How to obtain a domain theory from BTs?**
   Research studies [4,5] have been conducted to extract BTs from planning domains. The idea is instead to extract a planning domain from a BT.

References

1. Robertson and Watson, (2015). Building behavior trees from observations in real-time strategy games. In *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*.
2. Sagredo-Olivenza et al., (2019). Trained Behavior Trees: Programming by Demonstration to Support AI Game Designers. *IEEE Transactions on Games*.
3. French et al., (2019). Learning Behavior Trees From Demonstration. In *2019 International Conference on Robotics and Automation (ICRA)*.
4. Colledanchise et al. (2019a). Towards Blended Reactive Planning and Acting using Behavior Trees. *In IEEE International Conference on Robotics and Automation*, Montreal, Canada.
5. Martín et at., (2021). Optimized Execution of PDDL Plans using Behavior Trees. In *International Foundation for Autonomous Agents and Multiagent Systems (AAMAS)*

Industrial need

AI planning can introduce disruptive changes in industrial transport applications.

A decision support tools can increase the capabilities of fleet transport managers.

Better company performance and better satisfaction of their customers.

Overcome the challenge of transferring the fleet transport managers knowledge

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Gummesson Svensson, Hampus
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Sequential Decision-making in Drug Discovery

Two new paradigms have emerged in the pharmaceutical industry to increase the productivity in drug design: (1) AI-augmented molecular design that utilizes generative models for sampling the chemical space; (2) Automated laboratories together with machine learning to make, test and analyze potential drug candidates without human intervention. This work focuses on how to select what molecules to make in order to explore and exploit the chemical space in an efficient way. The results can enable a faster and better way to find drug candidates, ultimately providing new drugs for unmet medical needs faster to the benefit of patients worldwide.

Sequential Decision-making in Drug Discovery

Hampus Gummesson Svensson[†,‡],

Esben Jannik Bjerrum[‡], Christian Tyrchan[‖], Alexander Schliep[♯],

Ola Engkvist[‡], Morteza Haghir Chehreghani[†]

UNIVERSITY OF GOTHENBURG

CHALMERS UNIVERSITY OF TECHNOLOGY

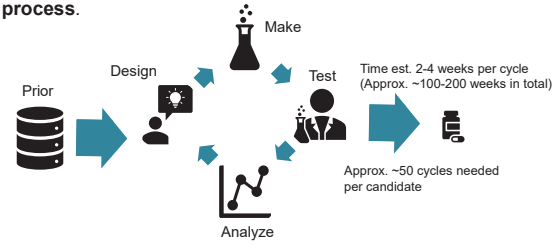
AstraZeneca

Motivation & Research goals

Two new paradigms have emerged in the pharmaceutical industry to increase the productivity in drug design: (1) AI-augmented molecular design that utilizes generative models for sampling the chemical space; (2) Automated laboratories together with machine learning to make, test and analyze potential drug candidates without human intervention. This work focuses on how to select what molecules to make in order to explore and exploit the chemical space in an efficient way. The results can enable a faster and better way to find drug candidates, ultimately providing new drugs for unmet medical needs faster to the benefit of patients worldwide.

Background

Drug optimization is a **complex**, **multiparameter optimization** with dozens of non-correlating and even **opposing parameters**. Therefore, the process of finding drug candidates is an **iterative process**.



Generative models such as REINVENT [1] can be used to **design** thousands of molecules by sampling the chemical space, which is estimated to consist of up to 10^{60} drug-like molecules. Subsequently, to understand the true properties of a molecule, we need to **make** it. If a molecule is successfully made, we can **test** and **analyze** its properties to better steer the generation of molecules. On the other hand, it is only possible to make a few molecules since each experiment is **costly** and **time-consuming**, which limits the amount of information that can be acquired in each iteration.

Goal: Optimize the selection of molecules to test, to explore and exploit desired areas of the chemical space in an optimal way.

Previous contributions (see [2]):

- **Active learning** to help to decrease the amount of data needed to **develop robust models for reaction yield predictions**, helping to successfully make molecules.

References

1. Blaschke, Thomas, et al. "REINVENT 2.0: an AI tool for de novo drug design." *Journal of Chemical Information and Modeling* 60.12 (2020): 5918-5922.
2. Johansson, Simon Viet, et al. "Using Active Learning to Develop Machine Learning Models for Reaction Yield Prediction." *ChemRxiv* (2021).
3. Slivkins, Aleksandrs. "Contextual bandits with similarity information." *Proceedings of the 24th annual Conference On Learning Theory*. JMLR Workshop and Conference Proceedings, 2011.
4. Chen, Lixing, Jie Xu, and Zhuo Lu. "Contextual combinatorial multi-armed bandits with volatile arms and submodular reward." *Advances in Neural Information Processing Systems* 31 (2018): 3247-3256.
5. Nika, Andi, Sepehr Elahi, and Cem Tekin. "Contextual combinatorial volatile multi-armed bandit with adaptive discretization." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

Approach

Challenges:

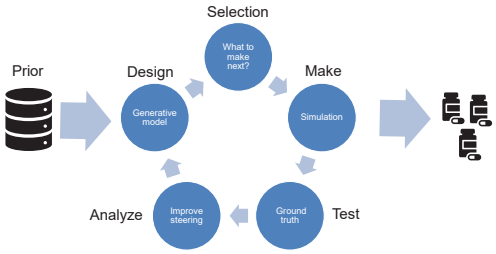
1. "Infinite" space (generated)
2. New molecules in each cycle
3. Select several molecules to make and test

Proposed solution:
To study this problem in the context of **sequential decision-making**, where the goal is to **adaptively** compute the most informative decisions.

Formerly, we currently seek a multi-armed bandit that can handle the following settings:

- **Contextual**
- **Infinite** action and/or context space
- **Volatile** arms
- **Combinatorial**

Previous works show approaches for several or all these settings using similarity information [3,4,5]. We have a **high-dimensional** and complicated context and large **unfixed** dataset (generated). Hence, we need to adapt and extend these methods to our problem.



A framework for **simulating** the different steps in the drug discovery process is being developed. This will help fast **evaluation and comparison** of different sequential decision-making strategies for selecting what molecules to make.

† Department of Computer Science and Engineering, Chalmers University of Technology
‡ Molecular AI, Discovery Sciences, R&D, AstraZeneca
‖ Medicinal Chemistry, Research and Early Development, Respiratory and Immunology (R&I), BioPharmaceuticals R&D, AstraZeneca
♯ Department of Computer Science and Engineering, University of Gothenburg

Gutierrez Maestro, Eduardo
Örebro University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

AI-based Mental Well-being monitoring during activities of daily life

The way technology has evolved in the last decades has changed our vision on how to face the problems of our daily life. Artificial Intelligence (AI) is taking an important role, being present in many fields, as for example Health Care. This research aims to fuse the different types of laboratory measures (including behavioral, peripheral, and central nervous systems) with cutting-edge technology based on predictive algorithms. The main objective is to discover patterns in the data that give us cues of how a person is feeling with the final goal of designing models that are able to monitor mental well-being, avoiding depressive, anxious, or stressful states during daily life activities.

Gutierrez Maestro, Eduardo
Örebro University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

AI-based Mental Well-being monitoring during activities of daily life

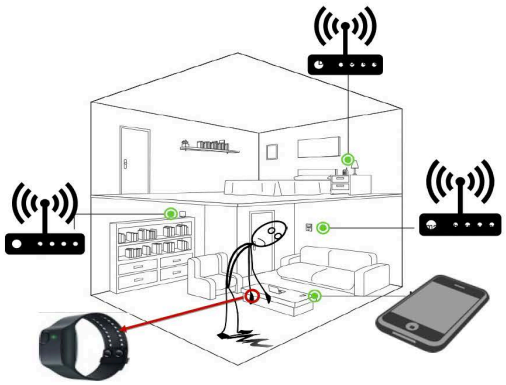
Eduardo Gutierrez-Maestro, Örebro University
Center for Applied Autonomous Sensor Systems (AASS)

Motivation & Research goals

The way technology has evolved in the last decades has changed our vision on how to face the problems of our daily life. Artificial Intelligence (AI) is taking an important role, being present in many fields, as for example Health Care. This research aims to fuse the different types of laboratory measures (including behavioral, peripheral, and central nervous systems) with cutting-edge technology based on predictive algorithms. The main objective is to discover patterns in the data that give us cues of **how a person is feeling** with the final goal of designing models that are able **to monitor mental well-being**, avoiding **depressive, anxious, or stressful states** during daily life activities.

Approach

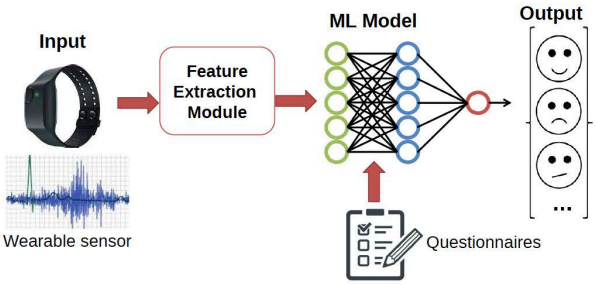
- **Well-being:** individuals' own abilities to cope with normal stresses of life, to work productively and to be able to contribute his or her community (World Health Organization)
  - Examples: stress, happiness, calm, sad, depressed, etc.



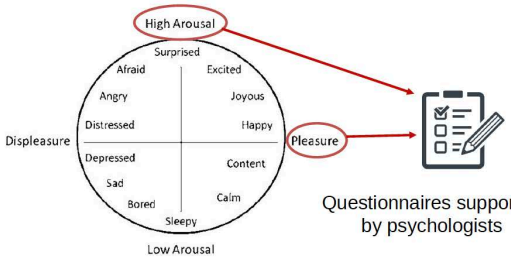
- **Multi-modal** scenario:
  - Smartphones, wearables, presence sensors.
- Hypothesis: context information can help to understand well-being states → **Activities of Daily Living (ADL)**
- **ADL**: non-controlled environment
  - Routine analysis
  - Real-world solution

Current work

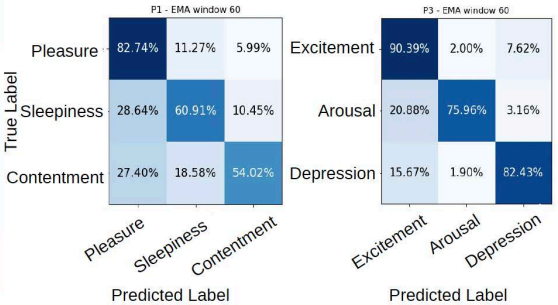
- Well-being **prediction** based on **physiological** data [1] [2]



- How can we **label** physiological data with **emotions**? [3]



- First results on **mood prediction** [1]



References

1. Calatrava-Nicolás, F.M.; Gutiérrez-Maestro, E.; Bautista-Salinas, D.; Ortiz, F.J.; González, J.R.; Vera-Repullo, J.A.; Jiménez-Buendía, M.; Méndez, I.; Ruiz-Esteban, C.; Mozos, O.M. Robotic-Based Well-Being Monitoring and Coaching System for the Elderly in Their Daily Activities. Sensors 2021
2. Bautista-Salinas, D.; Gonzalez, J.R.; Mendez, I.; Mozos, O.M. Monitoring and Prediction of Mood in Elderly People during Daily Life Activities. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, 23–27 July 2019
3. Russell et al., "A circumplex model of affect.", 1980

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Hagström, Lovisa
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?

Large language models are known to suffer from the hallucination problem in that they are prone to output statements that are false or inconsistent, indicating a lack of knowledge. A proposed solution to this is to provide the model with additional data modalities that complements the knowledge obtained through text. We investigate the use of visual data to complement the knowledge of large language models by proposing a method for evaluating visual knowledge transfer to text for uni- or multimodal language models. The method is based on two steps, 1) a novel task querying for knowledge of memory colors, i.e. typical colors of well-known objects, and 2) filtering of model training data to clearly separate knowledge contributions. Additionally, we introduce a model architecture that involves a visual imagination step and evaluate it with our proposed method. We find that our method can successfully be used to measure visual knowledge transfer capabilities in models and that our novel model architecture shows promising results for leveraging multimodal knowledge in a unimodal setting.

Hagström, Lovisa
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?

Tobias Norlund*, Lovisa Hagström*, Richard Johansson
Chalmers University of Technology

Motivation & Summary

Despite the ability of language models to learn and hold large quantities of structural knowledge [1], LMs are are also known to suffer from the **hallucination problem** in that they are prone to output statements that are false or inconsistent, indicating significant knowledge gaps [2]. A hypothesis is that this knowledge in many cases simply is missing in the large text corpora typically used for training the models, due to e.g reporting bias [3]. In such cases certain types of knowledge might also be more readily available in a different data modality.

In this work, we investigate **visual knowledge transfer**, i.e. to which extent language models can incorporate and *textually* express knowledge originating from a visual modality.

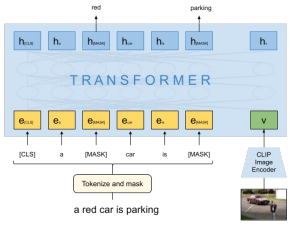
We investigate this by constructing a novel cloze-style task testing knowledge of **memory colors** for common objects (such as blood is red, a lemon is yellow etc). We also build a large vision-and-language dataset used for self-supervised training, and by careful filtering we make sure the color information is only available through the images and not the text.

Finally, we compare two strategies for how to most effectively query the trained multimodal language model for this visual knowledge.

We show that language models are able to texually express knowledge obtained from a visual modality, as a result from multimodal self-supervised training.

References
[1] Fabio Petroni et al. Language Models as Knowledge Bases? EMNLP (2019).
[2] Robert Logan et al. 2019. Barack's wife Hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.
[3] Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*.
[4] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision.

Multimodal Self-supervised Training: CLIP-BERT



We propose a novel visual-and-language model denoted **CLIP-BERT**, where the image encoder of the pre-trained CLIP [4] model is used to represent the image before appended to the input of a BERT-base model. We train this model using MLM on our vision-and-language dataset, and seek to evaluate how this affects performance on the memory colors task.

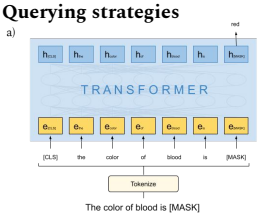
Task: Memory colors

With the help of 11 human annotators, we have created a dataset of 109 common objects and their memory colors, with high annotator agreement.

a	*lemon*	*yellow*	
descriptor	item	color	picture

* Equal contribution

Querying strategies



We compare two strategies for querying the language models for this visual knowledge, through [MASK] token prediction.

a) Implicit: The visual knowledge is retrieved from the trained parameters of the language model.

b) Explicit: We use the text encoder of the CLIP model to "imagine" a visual representation of the query text and append to the transformer input. This way the visual knowledge can partly be retrieved from this input, as well as from the trained model parameters.

Results

- ①: The original (text-only) BERT-base performs poorly on this task, close to majority baseline
- ① vs ⑥: Continue MLM training of BERT-base on the text part of our multimodal dataset improves performance slightly despite filtering
- ⑥ vs ⑧: Adding images to training improves performance significantly, showing effective visual knowledge transfer!
- ③ vs ④ and ⑦ vs ⑧: The explicit querying strategy performs better than the implicit

Training	Model	Accuracy	
	Random baseline	0.091 ± 0.026	
	Majority baseline	0.229 ± 0.000	
	Human baseline	0.937 ± 0.051	
None	BERT-base	0.252 ± 0.102	①
Unfiltered	BERT-base	0.724 ± 0.112	②
	CLIP-BERT		
	implicit	0.744 ± 0.080	③
	explicit	0.870 ± 0.086	④
	images	0.876 ± 0.063	⑤
Filtered	BERT-base	0.460 ± 0.083	⑥
	CLIP-BERT		
	implicit	0.541 ± 0.060	⑦
	explicit	0.733 ± 0.098	⑧
	images	0.785 ± 0.055	⑨

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Recorded Future

Heimerson, Albin
Lund University

Adaptive Control of Data Center Cooling using Reinforcement Learning

Data centers (DCs) have complex thermal environments which traditional cooling controllers are not able to fully capture.
These controllers are tuned using simple heuristics, which can result in inefficient operation and suboptimal cooling.
A Reinforcement Learning (RL) agent is developed for controlling the cooling fans, featuring system awareness from sensors on servers, in the room, and outdoors.
To realistically run the training and evaluate agent effectiveness, a model of a DC environment is developed.
The model features a CFD simulation of the DC room, heat-generating servers, and fans as well as heat exchangers, compressors, and dry coolers.
Experiments show that the RL agent can outperform baseline agents modeling current best practices in a simple setting where some external disturbances act on the system. Additionally, the
RL agent can adapt to larger changes in the environment, such as systems breaking down.

Heimerson, Albin
Lund University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Adaptive Control of Data Center Cooling using Reinforcement Learning

Albin Heimerson, Johannes Sjölund, Rickard Brännvall, Jonas Gustafsson, Johan Eker

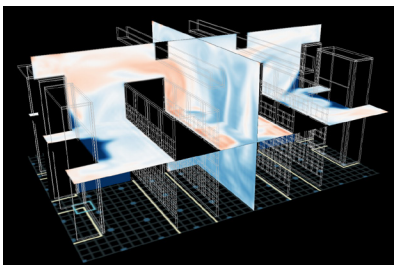
Collaborations

Ericsson Reserach Data Center and RICE SICS North are research data centers collaborating with us to develop better methods for data center control. The work was supported by Vinnova grant ITEA3-17002 (AutoDC).

Problem

The total energy consumption by data centers is expected to grow from 1.15% of global energy consumption in 2016 to around 1.86% in 2030 [a] . The International Energy Agency states that one of the innovation gaps in the IT-sector that needs to be filled is applying AI to data centers.

We want to create smarter control algorithms for the cooling systems that are both adaptive and can reason based on more information available in the datacenter.

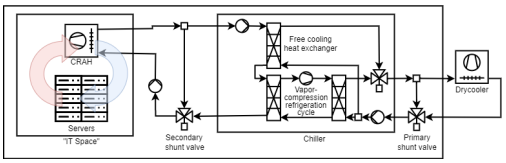


CFD Modelling of data center heat flow

[a] M. Koot and F. Wijnhoven, "Usage impact on data center electricity needs: A system dynamic forecasting model, 2021"

Modelling

A conceptual schematic of the physical model is shown below where the different parts are modeled using different strategies. The "IT space" is modeled using an CFD method called Lattice Boltzmann Method, using an algorithm called Single Relaxation Time Bhatnagar-Gross-Krook, and the model was first presented by Sjölund[a]. Boundary conditions of the servers and cooling systems are based on mathematical modeling of hardware such as the power used by the IT load, the fan speeds, the vapor compression, and the heat exchangers, which all affect the temperatures and air velocities. In previous work[b] a similar model was used with a simpler room simulation.



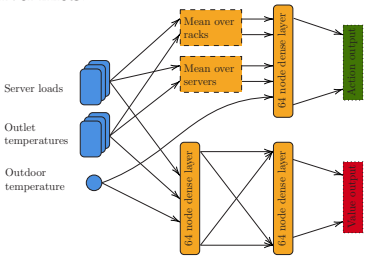
A typical cooling loop for a data center. It can utilize a drycooler to unload the compressor and reduce energy consumption.

[a] J. Sjölund, "Real-time thermal flow predictions for data centers: Using the lattice boltzmann method on graphics processing units for predicting thermal flow in data centers," 2018
[b] A. Heimerson, R. Brännvall, J. Sjölund, J. Eker, and J. Gustafsson, "Towards a Holistic Controller: Reinforcement Learning for Data Center Control", 2021

Reinforcement Learning

The states used is server loads, server outlet temperatures and outdoor temperature. The action is temperature and flow setpoints for the cooling units.

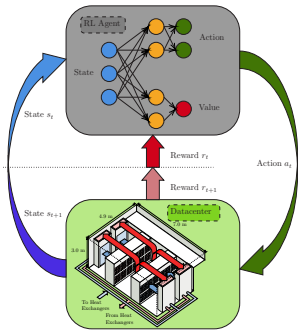
The reward was a weighted sum of two objectives, the energy consumption in the cooling system and a penalty on breaking a temperature threshold of 27° C for the server inlets.



The algorithm needed to be a bit stable while learning, so PPO[a] was used throughout this work.

PPO is an actor-critic algorithm, where the actor maps state to action and the critic maps state to expected reward. The critic tries to learn the bellman equation, $V(s_t) = r_{t+1} + \gamma V(s_{t+1})$, while the actor tries to increase the probability of actions that will have a good future reward according to the critic.

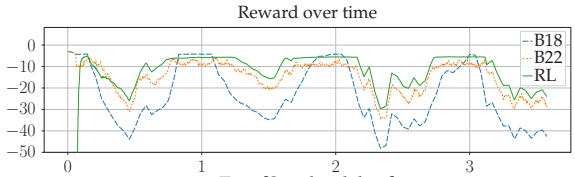
The special thing about PPO is that it doesn't allow the actor to change too much during each update, before it has been evaluated on new data.

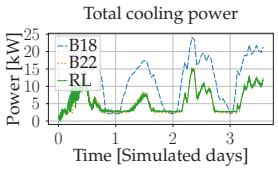


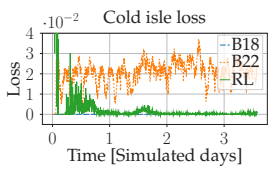
[a] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms", 2017

Results

We compare the RL agent to two baselines, one that is very good at the energy objective by keeping a higher outlet temperature of 22° C, and another that fixes the temperature threshold by keeping the minimal temperature of 18° C.

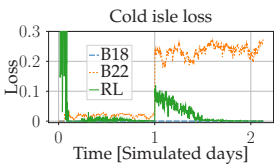


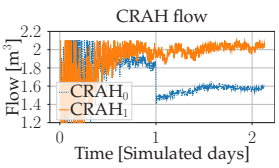




We see that the RL agent matches the energy efficient baseline quite well in energy, while still managing to keep the cold isle loss to a minimum, and achieves the highest reward.

Next we want to see how the agent can adapt to changing conditions such as a cooling unit breaking down $CRAH_0$ loses efficiency, and the cold isle loss will be harder to uphold, but the RL agent will adapt with time to remove the loss (left). The RL agent increases the flow of $CRAH_0$ after the inefficiency is introduced, as well as the flow of $CRAH_1$ which share the same cold aisle.





Hvarfner, Carl
Lund University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization

Bayesian optimization (BO) has become an established framework and popular tool for hyperparameter optimization (HPO) of machine learning (ML) algorithms. While known for its sample-efficiency, vanilla BO can not utilize readily available prior beliefs the practitioner has on the potential location of the optimum. Thus, BO disregards a valuable source of information, reducing its appeal to ML practitioners. To address this issue, we propose \method, an acquisition function generalization which incorporates prior beliefs about the location of the optimum in the form of a probability distribution, provided by the user. In contrast to previous approaches, πBO is conceptually simple and can easily be integrated with existing libraries and many acquisition functions. We provide regret bounds when \method is applied to the common Expected Improvement acquisition function and prove convergence at regular rates independently of the prior. Further, our experiments show that \method outperforms competing approaches across a wide suite of benchmarks and prior characteristics. We also demonstrate that πBO improves on the state-of-the-art performance for a popular deep learning task, with a $12.5\times$ time-to-accuracy speedup over prominent BO approaches.

Hvarfner, Carl
Lund University

Jin, Yifei
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Open World Learning Graph Convolution for Latency Estimation in Routing Networks

Accurate routing network status estimation is a key component in Software Defined Networking. However, existing deep-learning-based methods for modeling network routing are not able to extrapolate towards unseen feature distributions. Nor are they able to handle scaled and drifted network attributes in test sets that include open-world inputs. We propose a novel approach for modeling network routing, using Graph Neural Networks. Our method can also be used for network latency estimation. Supported by a domain-knowledge-assisted graph formulation, our model shares a stable performance across different network sizes and configurations of routing networks, while at the same time being able to extrapolate towards unseen sizes, configurations, and user behavior. We show that our model outperforms most conventional deep-learning-based models,

Jin, Yifei

KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Open World Learning Graph Convolution for Latency Estimation in Routing Networks

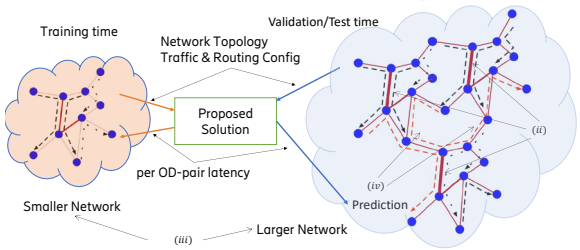
Yifei Jin, Marios Daoutis, Sarunas Girdzijauskas, Aristides Gionis

KTH Royal Institute of Technology, Ericsson Research

Routing Network Status Estimation

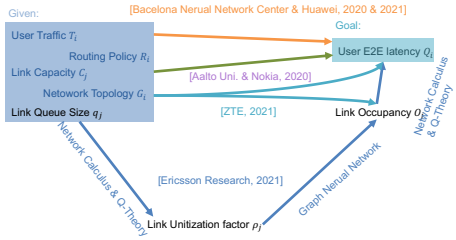
Accurate routing network status estimation is a key component in Software Defined Networking. However, existing deep-learning-based methods for modeling network routing are not able to extrapolate towards unseen feature distributions. Nor are they able to handle scaled and drifted network attributes in test sets that include open-world inputs. We propose a novel approach for modeling network routing, using Graph Neural Networks. Our method can also be used for network-latency estimation. Supported by a domain-knowledge-assisted graph formulation, our model shares a stable performance across different network sizes and configurations of routing networks, while at the same time being able to extrapolate towards unseen sizes, configurations, and user behavior. We show that our model outperforms most conventional deep-learning-based models,

Extrapolation in Routing Network



Expansion of the network --- An Open World Learning Problem

- More users (i) & Drifted Traffic Distribution
- Larger link capacity (ii)
- Larger network topology (iii)
- Per OD-pair have longer Trajectories (iv)



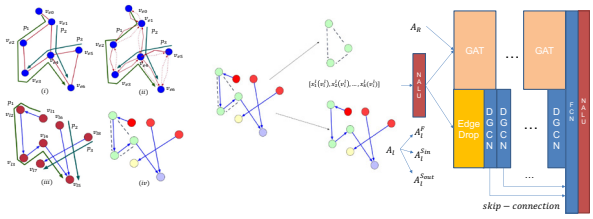
Unsolved problem in state of art models:

- **Structual Feature** of a link is underrepresented.
- There is no structure to gurantees the **extrapolation ability** on:
  - Attribute value (larger link capacity & drifted traffic throughput)
  - Larger sized graph G_i.
- There is no **theoratical guarentee** of the model validity

References

1. Suárez-Varela J, Ferriol-Galmés M, López A, et al. The graph neural networking challenge: a worldwide competition for education in AI/ML for networks[J]. ACM SIGCOMM Computer Communication Review, 2021, 51(3): 9-16.
2. Ferriol-Galm´es et al. Scaling graph-based deep learning models to larger networks.arXiv preprint arXiv:2110.01261, 2021.
3. Happ, Martin, et al. "Graph-neural-network-based delay estimation for communication networks with heterogeneous scheduling policies." *ITU Journal on Future and Evolving Technologies* 2.4 (2021).
4. Kong, Yangzhe, et al. "Path-Link Graph Neural Network for IP Network Performance Prediction." (2020).
5. Rusek K, Suárez-Varela J, Almasan P, et al. RouteNet: Leveraging Graph Neural Networks for network modeling and optimization in SDN[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(10): 2260-2270.

Proposed Solution



We proposed a Graph Convolution based approach for route state estimation that:

- Works in a **Directed Graph Formulation**.
- Introduce Node **Role adjacency** Information in pair-wise attention, i.e., improves GNN expressivity, identifies similar (and **key**!) hops in a path.
- Does **Arthmetic Extrapolation** towards unseen node/edge attrbutes.
- Our model **can be mathematically proven**, on the basis of network calculus and queuing theory, to be expressive enough for routing network staute estimation task.

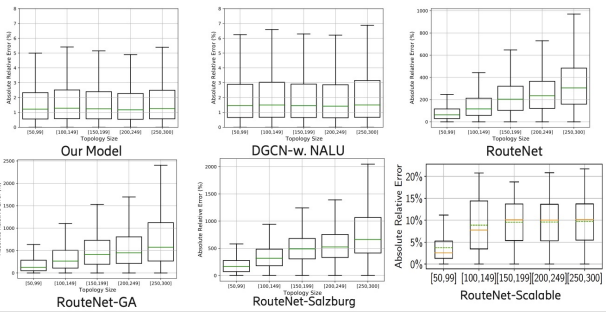
Experimental Result

Inference Accuracy

Model	MAPE
Ours	2.2
RouteNet	274.5
RouteNet-Salzburg	596.8
RouteNet-GA	557.8
DGCN.w.NALU	2.9
RouteNet-Scalable	10

Inference Speed

Topology Size	50	100	200	300
Ours	161.5	373	1115	2718
RouteNet	485.3	1972	6084	>10s
RouteNet-Salzburg	>10s	>10s	>10s	>10s
RouteNet-GA	1363	4786	>10s	>10s
DGCN.w.NALU	105.5	220.3	861.9	1770



- Our proposed solution requires No GPU to train, requires only 6-25% embedding size comparing with the known published benchmarks.

Jorge, Emilio
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Inferential Induction: A Novel Framework for Bayesian Reinforcement Learning

Bayesian Reinforcement Learning (BRL) offers a decision-theoretic solution to the reinforcement learning problem. While “model-based” BRL algorithms have focused either on maintaining a posterior distribution on models, BRL “model-free” methods try to estimate value function distributions but make strong implicit assumptions or approximations. We describe a novel Bayesian framework,\emph {inferential induction}, for correctly inferring value function distributions from data, which leads to a new family of BRL algorithms. We design an algorithm, Bayesian Backwards Induction (BBI), with this framework. We experimentally demonstrate that BBI is competitive with the state of the art. However, its advantage relative to existing BRL model-free methods is not as great as we have expected, particularly when the additional computational burden is taken into account.

Inferential Induction: A Novel Framework for Bayesian Reinforcement Learning

Emilio Jorge*,1 Hannes Eriksson*,1,2 Christos Dimitrakakis*,1,3 Debabrota Basu1,4 Divya Grover1

*Equal contribution [1]Chalmers university of Technology [2]Zenseact [3]University of Oslo [4]Scool, Inria Lille-Nord Europe

The problem

- **Setting**: Bayesian reinforcement learning (BRL).
- **Model-based BRL**: Straightforward formalisation by model distributions.
- **Model-free BRL**: Value function distributions via implicit approximations.

This work

- **Solution**: Derive correct value function distributions directly.
- **Expectation**: Improved modeling could lead to better performance
- **Reality**: BBI is competitive, but more complicated and slower.

Reinforcement learning

An unknown Markov Decision Process (MDP) μ with state s_t, action a_t, reward $r_t \sim P_\mu(r_t \mid s_t, a_t)$, next state $s_{t+1} \sim P_\mu(s_{t+1} \mid s_{t+1}, a_t)$.

Objective: Maximize utility $u_t = \sum_{k=t}^{T} \gamma^t r_t$

The value function V^π of a policy π is

$$V_\mu^\pi(s) \triangleq E_\mu^\pi[u_t \mid s_t = s_0], \qquad a_t \sim P^\pi(a \mid s_t)$$

Bayesian reinforcement learning

The Bayes-optimal solution is

$$\max_\pi E^\pi(u|D)$$

Two main Bayesian approaches

- **Model based:** Belief $\beta \triangleq P(\mu \mid \mathrm{D})$. We can then obtain
$$V_\beta^\pi(s) = \int_\mu V_\mu^\pi(s) dP(\mu \mid \mathrm{D})$$
- **Model free:** Estimate $P(V \mid D)$ directly.

Bayesian value function estimates

Existing model-free BRL algorithms follow the GPTD[1] framework.

- Gaussian process prior over the $P(V)$
- Likelihood function
$$P(D \mid V) \approx \prod_{i=1}^{t} \exp\{-|V(s_i) - r_i - \gamma V(s_{i+1})|^2\}, \; s_i \in D.$$
- At a high level, the inference is :
$$P(V \mid D) = \frac{P(V)P(D \mid V, \hat{\mu}(D))}{P(D)}$$
- Implicitly assumes the empirical MDP $\hat{\mu}(D)$ is correct
⇒ ignores model uncertainty.

Inferential induction

We propose a framework, **Inferential Induction**, to calculate the value function distribution $P^\pi(V \mid D_t)$ for policy π, correctly.

Data $D_t = s_1, a_1, r_1, \ldots, s_t, a_t, r_t$
⇒VF posterior$P(V_T|D_t), \ldots, P(V_i|D_t), \ldots, P(V_t|D_t)$.

Calculate the value functions with the inductive integral

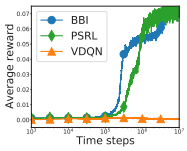
$$P^\pi(V_i \mid D_t) = \int_{\mathcal{V}} P^\pi(V_i \mid V_{i+1}, D_t) \, \mathrm{d}P^\pi(V_{i+1} \mid D_t) \qquad \text{(induction)}$$

$$P^\pi(V_i \mid V_{i+1}, D_t) = \int_{\mathcal{M}} \underbrace{P^\pi(V_i \mid \mu, V_{i+1})}_{\text{Bellman operator}} \, \mathrm{d}\overbrace{P^\pi(\mu \mid V_{i+1}, D_t)}^{\text{link distribution}}. \qquad \text{(marginalisation)}$$

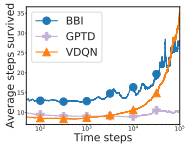
We introduce **Bayesian Backwards Induction** for calculating $P^\pi(V \mid D_t)$.

- Calculate integral through Monte Carlo sampling of V_{i+1} and μ.
- Define Gaussian kernel relating V_i and utility samples from μ to calculate link distribution $P^\pi(\mu \mid V_{i+1}, D_t)$.
- Importance sampling weights on $P(V_i \mid \mu, V_{i+1})$
- Utilising link distribution may above all be useful when true μ not in model class.

Experiments



(a) Maze



(b) Inverted Pendulum

CPU-time (in seconds) used for each algorithm with 100 policy updates for Chain and 5 updates for Maze.

	BBI	PSRL
Chain	14	5
Maze	921	6

Conclusion

- New framework for Bayesian RL.
- Beats GPTD, but results are not as impressive as we hoped.
- BBI uses $P(\mu \mid D)$ to obtain $P(\mu|V, D)$. Additional methods suggested in the work avoid this.
- It does not appear to be possible to do purely "model-free" Bayesian value function estimation.

References

[1] Yaakov Engel, Shie Mannor, and Ron Meir.
Bayes meets Bellman: The Gaussian process approach to temporal difference learning.
In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 154–161, 2003.

Karlsson, Alexander
KTH / SAAB

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Model-Aided Drone Classification Using Convolutional Neural Networks

Abstract - Classifiers using convolutional neural networks (CNNs) often yield high accuracies on samples that come from the same distribution as the training data. In this study we evaluate a CNN classifier's ability to discriminate drones from non-drone targets, such as birds, when they are not represented in the training data. We found that the mean accuracy on such out-of-distribution drones was 78%. By introducing a synthetic drone class, generated from a mathematical model, the out-of-distribution drone accuracy was improved to 86%. When trained on all drone types the mean accuracy over all classes was 90%, and greater than 95% for signal to noise ratios of at least 17.5 dB. The data was collected with a 77 GHz mechanically scanning radar with only 9 ms dwell time.

Karlsson, Alexander
KTH / SAAB

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Model-Aided Drone Classification Using Convolutional Neural Networks

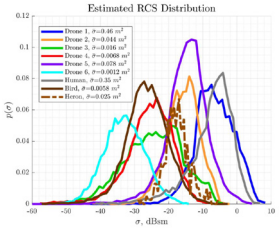
Alexander Karlsson*†, Magnus Jansson*, Mikael Hämäläinen†
*Division of Information Science and Engineering KTH,
†Electronic Warfare Systems, Saab.

Abstract - Classifiers using convolutional neural networks (CNNs) often yield high accuracies on samples that come from the same distribution as the training data. In this study we evaluate a CNN classifier's ability to discriminate drones from non-drone targets, such as birds, when they are not represented in the training data. We found that the mean accuracy on such out-of-distribution drones was 78%. By introducing a synthetic drone class, generated from a mathematical model, the out-of-distribution drone accuracy was improved to 86%. When trained on all drone types the mean accuracy over all classes was 90%, and greater than 95% for signal to noise ratios of at least 17.5 dB. The data was collected with a 77 GHz mechanically scanning radar with only 9 ms dwell time

Background

A drone, or unmanned aerial vehicle (UAV), can be used to deliver harmful payloads such as bombs, cause disturbances at e.g. airports, and collect footage of sensitive sites to name just a few potential threats. The ability to accurately detect and classify drones is therefore of great importance in present and future radar surveillance systems. By classification we refer to the process of distinguishing drones from non-drone targets.

In the literature, the non-drone class is mainly constituted by birds, but depending on what area is under surveillance humans, animals, flying frisbees etc. may also become relevant non-drone targets. Airplanes and other aerial and ground vehicles can often be discarded as false alarms by their radar cross section (RCS) values alone. Birds, humans and drones however may very well have comparable RCS.



Our main objective is to evaluate a neural network based classifier on drone types not present in the training data. To make the classifier more robust to such out-of-distribution drones we introduce training on both synthetic and real data. The radar we use in this study is a SAAB SIRS 1600 frequency modulated continuous wave (FMCW) radar operating at 77 GHz. This is a mechanically scanning radar and the dwell time is limited to the scan rate and beamwidth, in this case 9 ms.

Method

Data was collected from six different drones, birds (mostly seagulls) and humans. We trained the neural network on each drone, yielding six different scenarios. For reference we also trained on all six drones. The number of classes is N_c.

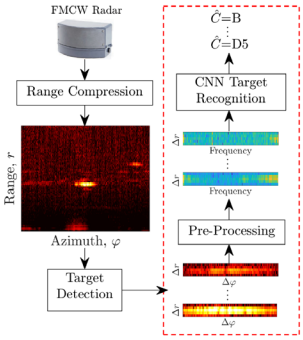
Layer type	Weight Shape	Strides	Output Shape	activation
Convolutional	F: 10, K: 5×8	[1,1]	1 × 143 × 10	ReLu
Convolutional	F: 10, K: 1×8	[1,2]	1 × 68 × 10	ReLu
Convolutional	F: 10, K: 1×8	[1,3]	1 × 21 × 10	ReLu
Fully Connected	32 × 210	-	32×1	ReLu
Fully Connected	N_c × 32	-	N_c×1	Softmax

F: Number of filters/kernels. K: Kernel size. Total number of parameters: 8782 + N_c(32 + 1).

The input data is a segment from a scan corresponding to 5 m in range and 2.3° in azimuth.

Each segment is then preprocessed by taking the discrete Fourier transform over azimuth (yielding a Doppler spectrum) and then normalized. Synthetic drone data at range r and FM-sweep p was generated as

$$\tilde{s}_{r,p} = G_{r,p}\left(1 + \rho z_p/\bar{z}\right) + w_{r,p}$$



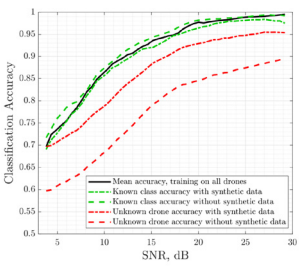
where z_p is the combined propeller return at pulse/FM-sweep p, $w_{r,p}$ is complex white Gaussian noise, $G_{r,p}$ is the combined pulse gain at range index r and azimuth beam gain at index p, and $\bar{z}$ is the rms value of z over all p pulses. For each pulse

$$z_p = \sum_{n=1}^{N_p}\sum_{i=1}^{N_s}\sum_{k=1}^{2}\alpha_k \exp\left(j(-1)^k I(\theta_{n,i})\theta_{n,i}\right)\operatorname{sinc}(\theta_{n,i})$$

$$\theta_{n,i} = 2\pi d_i \sin\left(2\pi\nu p + \phi_n\right)/\lambda$$
$$N_s = \lceil 10d/\lambda \rceil$$
$$d_i = di/N_s$$
$$I(\theta) = \begin{cases} 1, & \text{if } \cos(\theta) \geq 0 \\ -1, & \text{otherwise} \end{cases}$$

where N_p is the number of propellers, λ is the carrier wavelength, ν is the propeller's rotation rate in rounds per second divided by the pulse repetition frequency, φ_n is the initial phase of propeller n, d is the blade length and α_1 and α_2 are weights that determine the symmetry of the Doppler spectrum.

Results



We see that for SNR> 17.5 dB the mean accuracy is 95% or more when all drones are used in training. When only one drone class is used, the mean accuracy on the known classes is also > 95% for SNR> 17.5 dB whether synthetic data is used or not. The accuracy on the out-of-distribution drones is only > 82% without the synthetic data class and > 91% with synthetic data when the SNR> 17.5 dB.

Kidane, Lidia
Umeå University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Novel Data Selection Strategies and Associated Machine Learning Algorithms for Cloud Management

Cloud management systems are increasingly using machine learning models for autonomous resource provisioning. These systems need to be frequently calibrated and their models re trained to capture and understand the changing behaviors in the cloud system.
Fundamental assumptions include: workload volume may drastically increase from initial deployment to normal use years later, and that major changes to the machine learning models are expected at times of software and hard ware upgrades and with changing user trends. Models need also be able to deal with periods of volatility in various metrics. A notable problem that arises during those conditions is the change in the statistical properties of the monitoring data. This condition known as concept drift results in incorrect predictions and reduced efficiency of the models. Focus of this research is investigate and propose efficient method that adapt the dynamic behavior of cloud systems. Moreover, since the amount of monitoring data available in these systems is virtually unlimited, our second challenge includes developing new methods and algorithms for selecting important subsets of data for efficient training while being in control of prediction uncertainty.

Kidane, Lidia
Umeå University

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

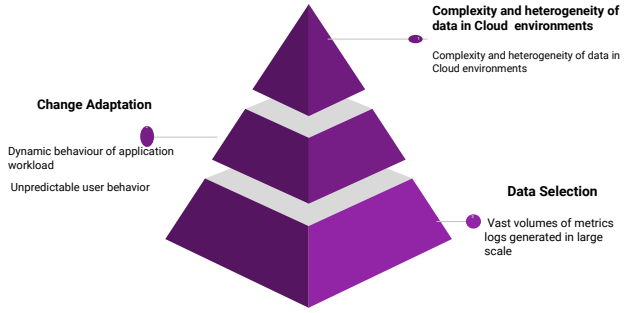
Novel Data Selection Strategies and Associated Machine Learning Algorithms for Cloud Management

Lidia Kidane, Umeå University
Computing Science
Main Supervisor: Erik Elmroth
Co-Supervisor: Paul Townend

Motivation and Research Goals

Cloud management systems are increasingly using machine learning models for autonomous resource provisioning. These systems need to be frequently calibrated and their models retrained to capture and understand the changing behaviors in the cloud system. Fundamental assumptions include: workload volume may drastically increase from initial deployment to normal use years later, and that major changes to the machine learning models are expected at times of software and hardware upgrades and with changing user trends. Models need also be able to deal with periods of volatility in various metrics. A notable problem that arises during those conditions is the change in the statistical properties of the monitoring data. This condition known as concept drift results in incorrect predictions and reduced efficiency of the models. Focus of this research is investigate and propose efficient method that adapt the dynamic behaviour of cloud systems. Moreover, since the amount of monitoring data available in these systems is virtually unlimited, our second challenge includes developing new methods and algorithms for selecting important subsets of data for efficient training while being in control of prediction uncertainty.

Challenges



Methods

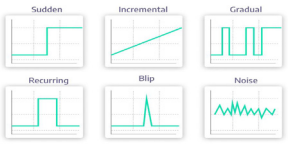
Workload Predictions based on only historical logs does not consider changes in usage patterns or resources.

Concept Drift:

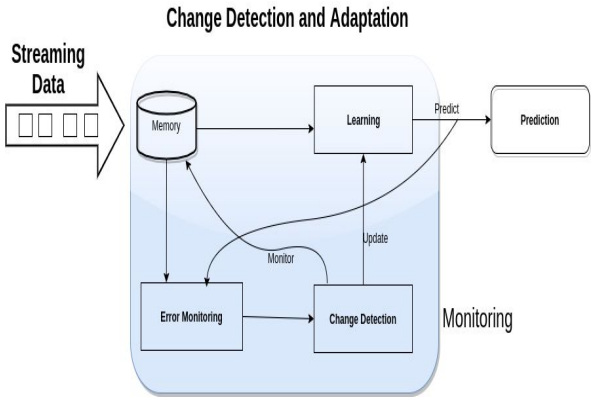
$$p(c_i|X) = \frac{P(c_i)p(X|c_i)}{p(X)}$$

Class definition change p(c|X) while p(X) remains the same. (real drift / concept drift / functional relation change)

- Sudden shifts in workload
- Change in user usage pattern
- Software or hardware upgrade



Methods



Change Adaptation:

- We implement State-of-the-art concept drift detection algorithms for time series analysis for workload prediction within cloud environment.
- We utilize both **machine learning** time series prediction techniques incorporated with **stream processing** algorithms to update workload prediction models on the fly.

Future Work

Smart data selection strategies

- **Uncertainty Sampling** targets data that is obviously confusing to your model in its current state
- **Diversity sampling** targets data that are gaps in your model's knowledge.
  - Eliminating overlapping of information
- Find metrics that collectively provide accurate, contextual, and insightful information on various aspects of model performance.

Konuk, Emir
KTH

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

An empirical study of the relation between network architecture and complexity

In this project we propose an empirical study of how networks handle changes in complexity of the data. We investigate the effect of network capacity on generalization performance in the face of increasing data complexity. For this, we measure the generalization error for an image classification task where the number of classes steadily increases. We compare a number of modern architectures at different scales in this setting.

Konuk, Emir
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

KTH ROYAL INSTITUTE OF TECHNOLOGY

An empirical study of the relation between network architecture and complexity

EMIR KONUK, KEVIN SMITH

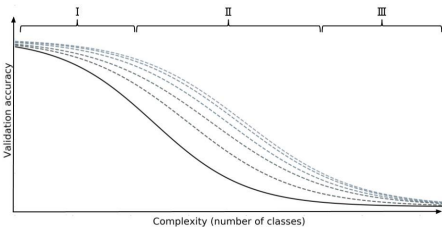
Abstract

We propose an empirical study of how networks handle changes in complexity of the data. We investigate the effect of network capacity on generalization performance in the face of increasing data complexity. For this, we measure the generalization error for an image classification task where the number of classes steadily increases. We compare a number of modern architectures at different scales in this setting.

Questions

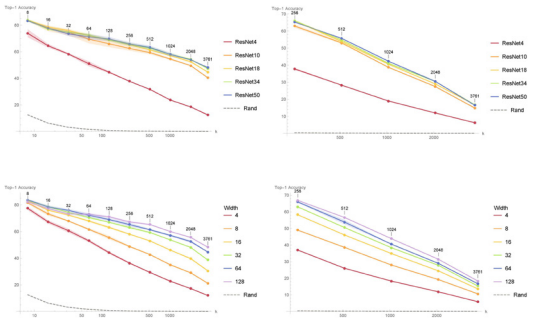
- How can we characterize the changes in generalization error as complexity is increased?
- How does capacity relate to generalization error as complexity increases? Does scaling the network by a particular dimension (e.g. depth) offer an advantage?
- How do architectural innovations such as group convolutions affect the capacity/generalization/complexity relationship?

Hypothesis



Ensure monotonic increase in complexity by introducing new classes but not necessarily linear increase.

Results and Conclusion



Effects of increasing data complexity on architectures with different capacity. Top-1 image classification validation accuracy for networks with various depths (a,b) and widths (c,d) as the number of classes k in the task increases (x-axis is scaled in log10). In (a,c), the number of training examples N grows with the number of classes, in (b,d) N=100,000. Image data was sampled from the broader ImageNet repository, not the ILSVRC subset, and includes up to 3,761 classes with ≈400 training samples per class. Results shown are the mean top-1 accuracy for basic ResNet blocks with up to 5 repetitions (some repetitions were not completed). Shaded areas indicate 1 s.d. The dashed lines indicate performance of a random classifier for reference.

Our empirical findings roughly align with our hypothesis as it was outlined in the Hypothesis.

For a limited but practical setting, we show that a deep network's performance decreases log-linearly with problem complexity as indicated by the number of categories.

Scaling the model size has diminishing returns for problems of different complexities - the diminishing returns effect is more prominent for more complex datasets but still observable for simpler ones.

SciLifeLab WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Acknowlegements: This work was supported by WASP-AI and the Swedish Research Council (VR) 2017-04609.

Email: ekonuk@kth.se

Kwatra, Saloni
Umeå University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

A Privacy Preserving Federated Learning Framework with Decision Trees

The aim is to build models for decentralized data following the Federated Learning (FL) approach. In FL, a general approach for learning consists of an iterative process in which (i) a set of agents is selected and they download the current model; (ii) the agents compute an updated model based on their data; (iii) the model updates are sent to the server; and (iv) the server aggregates these models to construct an improved general model. Privacy issues arise concerning the updated model sent to the server and how the models are aggregated. We plan to work on non-numerical (e.g., decision trees) and unsupervised learning models. This study presents an approach for implementing FL with decision trees on horizontally partitioned data.

Kwatra, Saloni
Umeå University

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

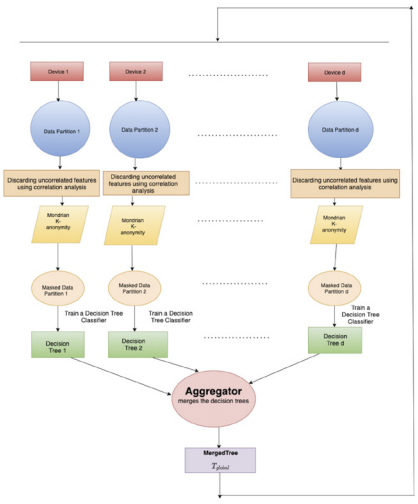
A Privacy Preserving Federated Learning Framework with Decision Trees

Saloni Kwatra, PhD at Umeå University
Department of Computing Science
Main Supervisor: Prof. Vicenç Torra
Co-Supervisor: Associate Prof. Lili Jiang

Motivation & Research Goals

The aim is to build models for decentralized data following the Federated Learning (FL) approach. In FL, a general approach for learning consists of an iterative process in which (i) a set of agents is selected and they download the current model; (ii) the agents compute an updated model based on their data; (iii) the model updates are sent to the server; and (iv) the server aggregates these models to construct an improved general model. Privacy issues arise concerning the updated model sent to the server and how the models are aggregated. We plan to work on non-numerical (e.g., decision trees) and unsupervised learning models. This study presents an approach for implementing FL with decision trees on a horizontally partitioned data.

Methods



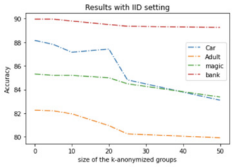
Creating Non-Independent and Identically Distributed Partitions

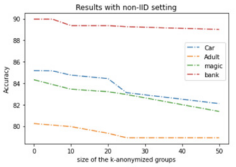
$$\begin{aligned} \text{Minimize} \quad & p^T Q p + p^T L \\ \text{subject to} \quad & \\ & \textstyle\sum_{i=1}^{d} p_{ij} = d n_j / n \quad \text{for each} \quad j = 1, \ldots, l \\ & \textstyle\sum_{j=1}^{l} p_{ij} = 1 \quad \text{for each} \quad i = 1, \ldots, d \\ & p_{ij} \geq 0 \quad \text{for each} \quad i = 1, \ldots, d \text{ and } j = 1, \ldots, l \\ & p_{ij} = 0 \quad \text{for each} \quad p_{ij} \in \mathcal{N} \end{aligned} \tag{1}$$

where, n is number of records, l is number of classes, p_{ij} denotes probability for i^{th} device and j^{th} class, $\sum_{j=1}^{l} p_{ij} = 1$ for all devices $i = 1, \ldots, d$, N denotes the set of probabilities that needs to be set to zero, square matrix $Q = Id$ (i.e., the identity matrix of size $d \cdot l$) and the vector $L = -2\mathrm{A}$ where $\mathrm{A} = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{dl})$. We consider a quadratic objective function of the form $OF(p; \alpha_{ij}) = (p_{ij} - \alpha_{ij})^2 = (p_{ij}^2 - 2\alpha_{ij} p_{ij} + \alpha_{ij}^2)$, where α_{ij} is a random number taken from a uniform distribution in [0,1].

Selected Results

Firstly, each device protects its raw data using Mondrian, and then trains a decision tree classifier on its protected data. Devices share the root node of their trees with the aggregator. The aggregator merges the trees by choosing the most common split attribute and grows the branches based on the split values of the chosen split attribute. This recursive process stops when all the nodes to be merged are leaf nodes. After the merging operation, the aggregator sends the merged decision tree to the distributed devices. Hence, we build a joint machine learning model based on the data from multiple devices while offering k-anonymity to the participants.





We show the results when the devices use data from the UCI datasets and where the databases follow non-IID data partitions of the whole dataset. We observe a drop of 2% in accuracy in the non-IID setting compared with the IID setting, when k=0 (k=0 means without anonymisation). This shows that the case of non-IID partitioning of data has some effect on the performance of our framework. From our perspective, this behavior is due to the fact that different devices have data with different probability distributions. The classification accuracy remains acceptable in the non-IID case, even when the value of k is as big as 50.
From the privacy point of view, using Mondrian k- anonymity for converting raw dataset into anonymised dataset is a good choice, as it is robust to various kinds of attacks and also it takes into account the multivariate distribution of the data.

Possible Future Improvements of our work

- In our implementation, all devices have the policy of sharing all their nodes with the aggregator. It would be possible that different devices have different policies about sharing (or not) their nodes.
- In our approach, node sharing is based on the decision tree learned with the original data. If partial decision trees are shared by the aggregator, devices can recompute their trees at each iteration.

References

[1] C. Fan, P. Li, Classification acceleration via merging decision trees, in: Pro- ceedings of the 2020 ACM-IMS on Foundations of Data Science Conference 2020, pp. 13–22

[2] LeFevre, K., DeWitt, D. J., & Ramakrishnan, R.(2006, April). Mondrian multidimensional k-anonymity. In 22nd International conference on data engineering (ICDE'06) (pp. 25-25). IEEE.

Källström, Johan
Linköping University / Saab Aeronautics

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Utility-Based Reinforcement Learning in Support of Simulation-Based Training

Team training in complex domains often requires a substantial amount of resources, e.g., instructors, role-players and vehicles. For this reason, it may be difficult to realize efficient and effective training scenarios in a real-world setting. Instead, intelligent agents can be used to construct synthetic, simulation-based training environments. However, building behavior models for such agents is challenging, especially for the users of the training systems, who typically do not have expertise in artificial intelligence. In this project, we study how reinforcement learning can be used to simplify the process of constructing agents for simulation-based training. By constructing smarter synthetic agents the dependency on human training providers can be reduced, while the availability and quality of training is improved.

Källström, Johan
Linköping University / Saab Aeronautics

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Utility-Based Reinforcement Learning in Support of Simulation-Based Training

Johan Källström, Linköping University, johan.kallstrom@liu.se
Department of Computer and Information Science

Abstract

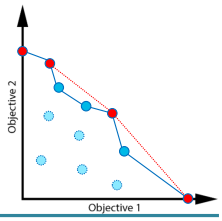
Team training in complex domains often requires a substantial amount of resources, e.g., instructors, role-players and vehicles. For this reason, it may be difficult to realize efficient and effective training scenarios in a real-world setting. Instead, intelligent agents can be used to construct synthetic, simulation-based training environments. However, building behavior models for such agents is challenging, especially for the users of the training systems, who typically do not have expertise in artificial intelligence. In this project, we study how reinforcement learning can be used to simplify the process of constructing agents for simulation-based training. By constructing smarter synthetic agents the dependency on human training providers can be reduced, while the availability and quality of training is improved.

The Utility-Based Approach

In this work, we use Multi-Objective Markov Decision Processes (MOMDPs) to model problems with multiple objectives, represented by multiple reward signals. We develop Multi-Objective Reinforcement Learning (MORL)[1] algorithms to solve these problems. MORL allows synthetic agents to learn how to prioritize among multiple, possibly conflicting objectives. The priorities among the objectives of the learning agent are defined by a utility function. Some advantages of MORL compared to standard single-objective reinforcement learning algorithms is that complex non-linear utility functions can be used, a greater degree of flexibility in adapting to changes in goals or utility is achieved, and a more diverse set of solutions can be found[2]. In MORL, there are two major types of optimization criteria: scalarized expected returns (SER) and expected scalarized returns (ESR).

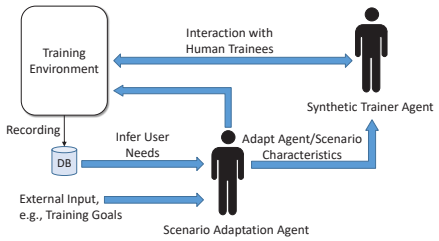
$$V_u^{\pi}(s) = u\left(E\left[\sum_{t=0}^{\infty}\gamma^t \mathbf{r}_t | s_0 = s\right]\right), for\ SER\ criterion$$

$$V_u^{\pi}(s) = E\left[u\left(\sum_{t=0}^{\infty}\gamma^t \mathbf{r}_t\right) | s_0 = s\right], for\ ESR\ criterion$$



Automating Training Systems

Synthetic, learning agents can be used to automate simulation-based training by replacing or assisting human instructors and role-players. In this work, we use two types of learning agents: A Synthetic Trainer Agent and a Scenario Adaptation Agent[3]. The Synthetic Trainer Agent participates as an actor in training scenarios. By adjusting the utility function of the agent the dynamics of the simulation can be adapted to current training needs, The Scenario Adaptation Agent considers trainees' proficiency in relation to training goals, and then populates training scenarios with human and synthetic agents to maximize improvement in trainees' performance.

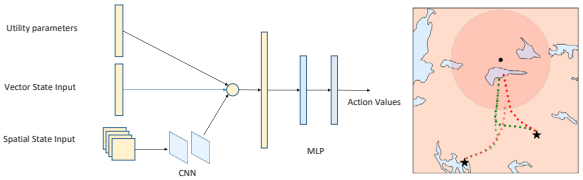


References

1. Hayes, C.F., Rădulescu, R., Bargiacchi, E., Källström, J., et al. (2021). A practical guide to multi-objective reinforcement learning and planning. Submitted to the Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS).
2. Vamplew, P., Smith, B.J., Källström, J., et al. (2021). Scalar reward is not enough. Submitted to the Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS).
3. Källström, J. (2020). Adaptive Agent-Based Simulation for Individualized Training. In Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS).
4. Källström, J., Granlund, R., & Heintz, F. (2021). Design of Simulation-Based Pilot Training Systems using Machine Learning Agents. Submitted to the Aeronautical Journal (AER).
5. Källström, J., & Heintz, F. (2020). Agent Coordination in Air Combat Simulation using Multi-Agent Deep Reinforcement Learning. In proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC).

Case Study

As a case study, we use a simulation-based air combat training system[4]. In this system, in addition to dealing with multiple training goals, learning agents also need to consider multiple conflicting objectives related to the simulated scenario, e.g., tactical mission goals, resource consumption, and safety. To learn adaptable policies, we use deep neural networks that are conditioned on parameters of the agent's utility function[5]. By letting the agent explore with different utility functions, the agent can learn a set of Pareto optimal policies. Then, after learning, the policy that maximizes the user's current utility can be selected.



Lindqvist, Jakob
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Posterior linearisation smoothing with robust iterations

This work considers the problem of robust iterative Bayesian smoothing in nonlinear state-space models with additive noise using Gaussian approximations. Iterative methods are known to improve smoothed estimates but are not guaranteed to converge, motivating the development of more robust versions of the algorithms. The poster presents Levenberg-Marquardt (LM) and line-search extensions of the classical iterated extended Kalman smoother (IEKS) as well as the iterated posterior linearisation smoother (IPLS). The IEKS has previously been shown to be equivalent to the Gauss-Newton (GN) method. We derive a similar GN interpretation for the IPLS.
Furthermore, we show that an LM extension for both iterative methods can be achieved with a simple modification of the smoothing iterations, enabling algorithms with efficient implementations.

Lindqvist, Jakob
Chalmers

Posterior linearisation smoothing with robust iterations

Jakob Lindqvist

Iterative smoothing

The work presented here is based on an article submitted to Transactions on Signal Processing and is also available as a pre-print [3]

Smoothing is a form of state estimation where a sequence of latent states in a Markov process $x_{1:K} \coloneqq (x_1, x_2, \ldots, x_K)$ are estimated from noisy measurements $y_{1:K} \coloneqq (y_1, y_2, \ldots, y_K)$.

The system is described by a state-space model with additive Gaussian noise

$$\begin{aligned} x_{k+1} &= f_k(x_k) + q_k & q_k &\sim \mathcal{N}(0, Q_k) \\ y_k &= h_k(x_k) + r_k & r_k &\sim \mathcal{N}(0, R_k). \end{aligned} \tag{1}$$

- Rauch-Tung-Stribel (RTS) smoothing computes a closed form solution for linear (affine) motion $f(\cdot)$ and measurement $h(\cdot)$ models.
- General Gaussian RTS smoothers are a family of methods which linearise the state-space models and then perform exact RTS smoothing of the approximate system. Members of this family are extended Kalman smoothing (EKS) and SLR-smoothers.
- Linearisation is done around the best available estimate of the state, commonly the predicted estimates $\hat{x}_{k|k-1}$.
- Iterative smoothers repeatedly perform smoothing and use the posterior estimates of the previous iteration for linearisation points. General Gaussian smoothers have natural iterative extensions, e.g. the IEKS and the IPLS.
- The smoothing problem can be viewed as optimisation problems. The IEKS is equivalent to Gauss-Newton (GN) opt. of the neg. log. likelihood of the state-space in eq. (1) [1].

Posterior linearisation

The choice of linearisation method defines a particular smoother with different properties.

- The EKS and IEKS use first-order Taylor approximation around the estimated mean.
- The IPLS use SLR based on the full distribution [2]:

$$F_k(\hat{x}_k, \hat{P}_k) = \Psi_{f_k}^\top \hat{P}_k^{-1} \tag{4a}$$
$$b_k(\hat{x}_k, \hat{P}_k) = \bar{x}_k - F_k \hat{x}_k \tag{4b}$$
$$\Omega_k(\hat{x}_k, \hat{P}_k) = \Phi_{f_k} - A\hat{P}_k A^{\mathrm{T}}, \tag{4c}$$

$$\begin{aligned} \bar{x}_k &= \int f_k(x_k) p(x_k) \mathrm{d}x_k \\ \Psi_{f_k} &= \int (x_k - \hat{x}_k)(f_k(x_k) - \bar{x}_k)^\top p(x_k) \mathrm{d}x_k \\ \Phi_{f_k} &= \int (f_k(x_k) - \bar{x}_k)(f_k(x_k) - \bar{x}_k)^\top p(x_k) \mathrm{d}x_k. \end{aligned} \tag{5}$$

Iterative smoothers can select a better linearisation point.

- Non-iterative methods use the predicted state $(\hat{x}_{k|k-1}, \hat{P}_{k|k-1})$, which do not take the measurements $y_{k:K}$ into account.
- Ideally, linearisation would be done w.r.t. the posterior distribution of the state.
- Iterative smoothers repeatedly refine the estimates until the linearisation point is approximately chosen w.r.t. the posterior.

Smoothers with robust iterations

Iterative smoothers might diverge and more robust versions are necessary. We take inspiration from general optimisation and use the connection to GN optimisation to propose such modified smoothers.

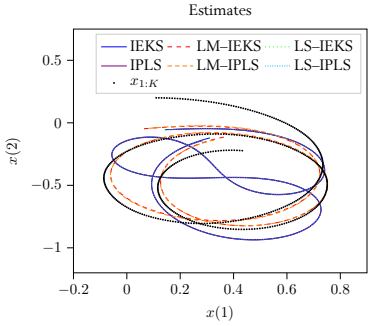
- We show that the iterations of IPLS of eq. (1) is equivalent to the iterations of GN optimisation of the cost function

$$\begin{aligned} L^{(i)}_{IPLS}(x_{1:K}) = \frac{1}{2}\Big(&(x_1 - \hat{x}_{1|0})^\top \hat{P}_{1|0}^{-1}(x_1 - \hat{x}_{1|0}) \\ &+ \sum_{k=1}^{K-1} (x_{k+1} - \bar{x}_k(x_k))^\top \left(Q_k + \Omega_k^{(i)}\right)^{-1} (x_{k+1} - \bar{x}_k(x_k)) \\ &+ \sum_{k=1}^{K} (y_k - \bar{y}_k(x_k))^\top \left(R_k + \Gamma_k^{(i)}\right)^{-1} (y_k - \bar{y}_k(x_k))\Big), \end{aligned} \tag{2}$$

- We extend the results of [4] and show that Levenberg–Marquardt (LM) regularised versions of the IEKS and IPLS can be achieved by extending eq. (1) with a pseudo-measurement of the state:

$$\hat{x}_k^{(i)} = x_k + e_k, \qquad e_k \sim \mathcal{N}(0, (\lambda^{(i)})^{-1} S_k^{(i)})$$

- We propose line-search versions of the IEKS and IPLS.



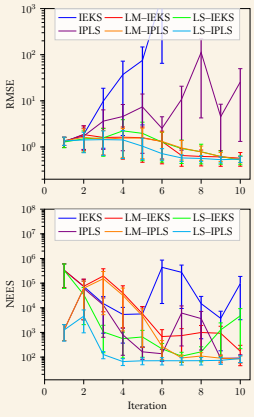
(3)Visualisation of a single realisation of a coordinated turn (CT) experiment with varying bearings only measurements. The measurements at $k = 50, 100, \ldots, 500$ are low noise bearings measurements from a single sensor at $(1, 1)^\top$ Note that for this particular realisation, it is only the LM-regularised smoothers, LM–IEKS and LM–IPLS that estimate the general shape of the true trajectory.

Experimental results

We propose the robust smoothers LM–IEKS and LM–IPLS as well as the line-search based LS–IEKS and LS–IPLS.

We report root mean square error (RMSE) and normalised estimation error squared (NEES) for the CT experiment above. The plots show the metrics averaged over 100 independent simulations with the standard error in the error bars.

- The IPLS based smoothers perform better overall.
- The large spread in the IEKS metrics are caused by some divergent realisations.
- The regularised version perform better than the original methods. In particular, they exhibit faster convergence.



Discussion

- We present robust versions of the IEKS and IPLS smoothers. LM regularisation is achieved by a simple modification of the state-space model and the line-search versions are conceptually simple.
- The IPLS is robust in its original form since it linearises w.r.t. to the full distribution and not just a point estimate, like the IEKS, and benefits less from the robustness modifications.
- Making the linearisation based on the full distribution is more computationally expensive but the added complexity is ameliorated somewhat by the observation that this group of smoothers seems to require fewer iterations to converge to an acceptable trajectory.

References

[1] B. M. Bell. The iterated Kalman smoother as a Gauss–Newton method. *SIAM Journal on Optimization*, 4(3):626 – 636, 1994.

[2] Á F. García-Fernández, L. Svensson, and S. Särkkä. Iterated posterior linearization smoother. *IEEE Transactions on Automatic Control*, 62(4):2056–2063, 2017.

[3] Jakob Lindqvist, Simo Särkkä, Ángel F. García-Fernández, Matti Raitoharju, and Lennart Svensson. Posterior linearisation smoothing with robust iterations. *arXiv e-prints*, page arXiv:2112.03969, December 2021.

[4] S. Särkkä and L. Svensson. Levenberg-Marquardt and line-search extended Kalman smoothers. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5875–5879, 2020.

Lourenço, Inês
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

A teacher-student framework for online correctional learning

A classical learning setting is one in which a student collects data, or observations, about a system, and estimates a certain quantity of interest about it. Correctional learning is a type of cooperative teacher-student framework where a teacher, who has knowledge about the system, has the possibility to observe and alter (correct) the observations received by the student in order to improve its estimation. In this poster, we present our formulation of both the batch and online correctional learning problem - while the former is approached as an optimisation problem, for the latter we formulate the MDP and solve it using dynamic programming. Our results show that in both cases the variance of the estimate of the student is reduced with the help of the teacher.
An the end of the poster there is a short summary of our other research areas that we would be very glad to share with you. These go beyond the cooperative setting to adversarial, biologically-inspired, and medical decision-making scenarios.

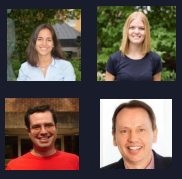
Lourenço, Inês
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

A TEACHER-STUDENT FRAMEWORK FOR ONLINE CORRECTIONAL LEARNING

Correctional Learning for Cooperative System Identification

Inês Lourenço ineslo@kth.se • Rebecka Winqvist rebwin@kth.se
KTH Royal Institute of Technology, Sweden



Background

Setup:

The student samples observations from the system

$$y_k \sim p(y|y_{k-1}, \dots, y_1; \theta_0)$$

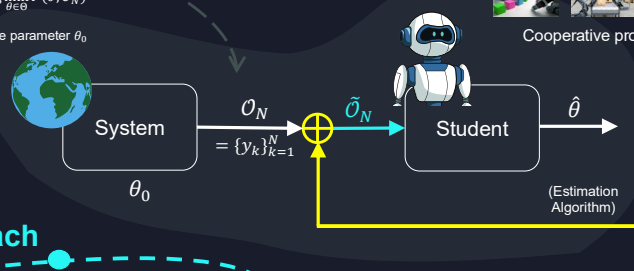
And uses them to estimate a model of the system

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} F(\theta, \mathcal{O}_N)$$

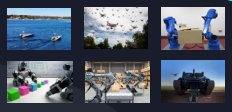
Goal: Find the true parameter θ_0

Problem: Costly and time consuming to go from experimental design to resulting model

How can we improve the learning process of the student?



Motivation



Cooperative problems

- Sparse rewards
- Difficult reward function

These problem can be overcome using advice from an expert agent, such as in

- Transfer learning
- Learning from demonstration

Teacher-student framework

Research Goal & Questions

Cooperative System Identification

Teacher

It is difficult to transmit knowledge:

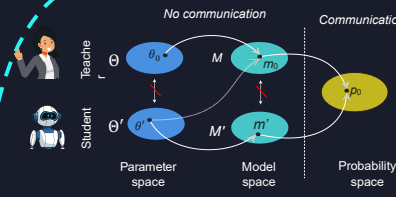
- Abstract knowledge - e.g., driving a car
- Too complex - e.g., GTP-3 model
- Privacy constraints
- **Different model classes or parametrizations**

For a consistent estimator, "improving" can mean:

$\hat{\theta} \to \theta_0$ "fast" $||\hat{\theta}_N - \theta_0||$ "small"

How can a teacher improve the learning process of the student?

Approach



Approach :
Operate in the space of induced probability distributions instead of the parameter space

True distribution p_0

Observations collected → $\mathcal{O}_N = \{y_k\}_{k=1}^N$ → Estimated distribution with the original observations: $\hat{p}$

Teacher modifies → $\tilde{\mathcal{O}}_N = \{\tilde{y}_k\}_{k=1}^N$ → Estimated distribution with the modified observations: $\tilde{p}$

By altering the observations $\tilde{\mathcal{O}}_N$ received by the student

How should the teacher alter the observations?

Methods & Preliminary Results — Correctional Learning

BATCH SETTING [2]

Answer:
The teacher's goal is to minimize the distance between the true distribution p_0 and the student's received distribution $\tilde{p}$

$$\min_{\tilde{\mathcal{O}}} \; V(p_0, \tilde{p}) \quad \text{(distance)} \quad \text{E.g., l2-norm } \|p_0 - \tilde{p}\|_2$$

$$\text{s.t.} \quad \tilde{y}_k \in Y, \quad \text{for all } \tilde{y}_k \in \tilde{\mathcal{O}},$$

$$B(\mathcal{O}, \tilde{\mathcal{O}}) \le b \quad \text{(budget)} \quad \text{E.g., l1-norm } \sum_{k=1}^{N} |y_k - \tilde{y}_k| < b$$

ONLINE SETTING [1]

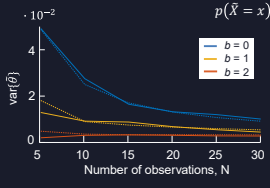
Answer:
The teacher's optimal policy is to decide at each moment whether or not to change the current observations. Using dynamic programming to solve the MDP.

States : $s = (x_{1:k}, b_k, y_k)$
Actions : a = { keep y_k, change to $\tilde{y}_k$}
Time-horizon : N
Reward function : M - $\|p_0 - \tilde{p}_N\|_1$
Constraint : number of times the action "change to $\tilde{y}_k$" is taken $\le c_b$
Transition probabilities : see [1]

Example: Binomial systems

$$p(\tilde{X} = x) = \begin{cases} p(X = \bar{x} - b), & if \ \bar{x} < \mathbb{E}[X] \\ p(X = \bar{x} + b), & if \ \bar{x} > \mathbb{E}[X] \\ \sum_{-b}^{b} p(X = \mathbb{E}[X] + b'), & if \ \bar{x} = \mathbb{E}[X] \end{cases}$$

$$\mu^* = \begin{cases} a_k = \text{keep } y_k, & if \ [x]_{y_k} \le [\theta_0]_{y_k} \\ a_k = \text{change to } \tilde{y}_k = 1 - y_k & \text{otherwise} \end{cases}$$



$\text{var}\{\tilde{\theta}\} \le \text{var}\{\hat{\theta}\}$ — The teacher is helping the student by reducing the variance of its estimator

$\frac{\text{var}\{\tilde{\theta}\}}{\text{var}\{\hat{\theta}\}} \to 0$ $_{b \to N}$ — As the budget increases, the teacher is able to increasingly accurately transfer its knowledge to the student

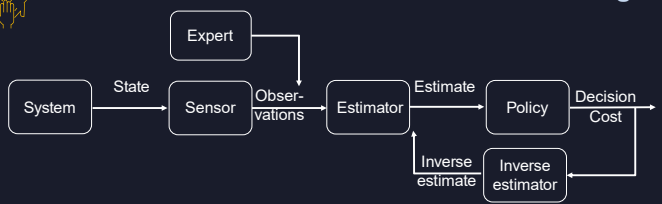
We can quantify by how much, for discrete systems

ML estimator

Roadmap & Milestones

- Noisy expert
- Generalizations to e.g., filtering (state estimation) and **reinforcement learning** (policy learning).
- Connections to the optimal mass transport problem

Forward and Inverse Decision-Making



Cooperative

- Can a teacher help a student perform its task?

Adversarial

- Can we estimate characteristics about an agent by observing its decisions?

Biologically-inspired

- Can we enable intelligent agents to perceive time similarly to humans?
- Can we estimate the internal parameters of the agents' frameworks from their behavior?

Medical

- Can we estimate biases in health care (e.g. doctors' biases)?
- Can we use estimates to make medical suggestions/interventions?
- Not assuming perfect policy: is it bias or simply bad policy?

[1] I. Lourenço, R. Winqvist, C. R. Rojas, and B. Wahlberg, "A teacher-student framework for online correctional learning," under review, 2021
[2] I. Lourenço, R. Mattila, C. R. Rojas, and B. Wahlberg, "Cooperative System Identification via Correctional Learning," 19th IFAC Symposium on System Identification (SYSID), 2021

Marti, Miquel
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

An analysis of over-sampling labeled data in semi-supervised learning with FixMatch

Most semi-supervised learning methods over-sample labeled data when constructing training mini-batches. This paper studies whether this common practice improves learning and how. We compare it to an alternative setting where each mini-batch is uniformly sampled from all the training data, labeled or not, which greatly reduces direct supervision from true labels in typical low-label regimes. However, this simpler setting can also be seen as more general and even necessary in multi-task problems where over-sampling labeled data would become intractable. Our experiments on semi-supervised CIFAR-10 image classification using FixMatch show a performance drop when using the uniform sampling approach which diminishes when the amount of labeled data or the training time increases. Further, we analyse the training dynamics to understand how over-sampling of labeled data compares to uniform sampling. Our main finding is that over-sampling is especially beneficial early in training but gets less important in the later stages when more pseudo-labels become correct. Nevertheless, we also find that keeping some true labels remains important to avoid the accumulation of confirmation errors from incorrect pseudo-labels.

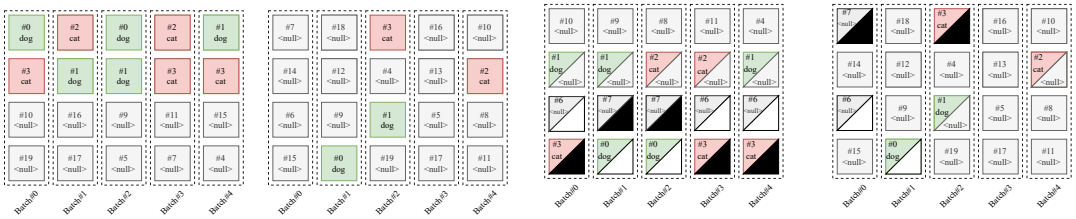
Marti, Miquel
KTH

An analysis of over-sampling labeled data in semi-supervised learning with FixMatch

Miquel Martí i Rabadán[1,2], Sebastian Bujwid[1], Alessandro Pieropan[2], Hossein Azizpour[1] and Atsuto Maki[1]
[1]KTH Royal Institute of Technology, Division of Robotics, Perception and Learning, Sweden
[2]Univrses AB, Sweden

Motivation & Goals

Most semi-supervised learning methods construct mini-batches sampling its parts from a set of labeled samples and a set of unlabeled ones, then repeating the first set that gets depleted [1,2,3,4], effectively over-sampling labeled data. We aim at studying how this common practice affects learning and compare it to the simplest alternative: uniform sampling of all training data, labeled or not. Alternative approaches to over-sampling can be a requirement in certain learning problems, e.g. in multi-task learning (MTL) samples are said to be labeled with respect to each of the tasks considered. In such case, a sample can not be thought as being "labeled" in general. Thus there is no single, direct approach to over-sample "labeled" data in MTL.



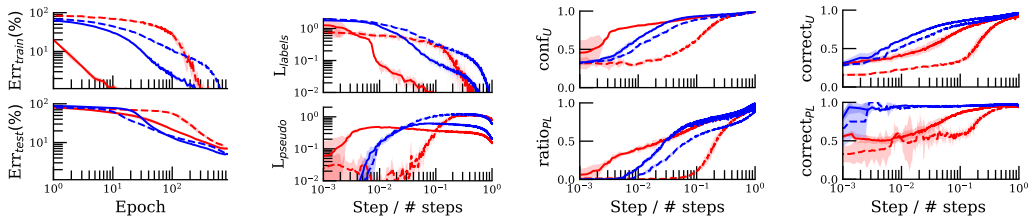
(a) Over-sampling labeled data (b) Uniform sampling (c) MTL ext. of over-sampling (d) MTL uniform sampling

Experiments

We compare both sampling approaches in semi-supervised image classification with CIFAR-10/100 using FixMatch [4] and analyze their training dynamics. (O)ver-sampling - (U)niform sampling - 6x is longer training. Mean/std.dev. of top-1 test accuracy (%) over 5 runs.

CIFAR-10	40 labels	250 labels	4000 labels	All labels
Supervised	36.77 ± 4.48	$59.88 \pm .73$	$87.39 \pm .20$	$96.54 \pm .11$
FixMatch(O)	83.44 ± 6.68	$93.01 \pm .58$	$94.89 \pm .16$	-
FixMatch(U)	73.90 ± 8.04	$91.42 \pm .98$	$94.84 \pm .05$	-
FixMatch(O) 6x	85.40 ± 2.83	$94.40 \pm .78$	-	-
FixMatch(U) 6x	86.59 ± 4.14	$94.00 \pm .54$	-	-
CIFAR-100		2500 labels	10000 labels	All labels
Supervised		47.26	67.58	82.38
FixMatch(O)		62.86	74.61	-
FixMatch(U)		55.53	71.83	-

- **Obs. 1** Both approaches outperform strong supervised baselines.
- **Obs. 2** Over-sampling outperforms uniform sampling. More labels or longer training improve both and reduce the performance gap.
- **Obs. 3** With few labels, uniform sampling is largely inefficient early in training - does not use unlabeled samples, which are the large majority.
- **Obs. 4** Over-sampling is especially beneficial early in training. High-confidence predictions to be used as pseudo-labels are available earlier.



Mean/std.dev. over 5 runs for CIFAR-10: 250 labels, 4000 labels, solid line for over-sampling, dashed for uniform sampling.

Training without true labels in models already producing high rates of correct pseudo-labels eventually leads to collapse to random accuracy, likely due to the accumulation of errors. This hints to the presence of *some* samples with true labels remaining important throughout training to anchor learning to the true task.

Conclusions

The relative success and generality of the simple uniform sampling approach enables its direct use in learning problems where over-sampling labeled data is not possible, such as semi-supervised multi-task learning. Future research should be directed at 1) validating these results for methods substantially different to FixMatch, 2) exploring other sampling approaches for semi-supervised learning and especially in MTL.

References

[1] Semi-supervised Learning with Ladder Networks
A. Rasmus et al. Conference on Neural Information Processing Systems 2015

[2] Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning
T. Miyato et al. IEEE Transactions on Pattern Analysis and Machine Intelligence 2019

[3] MixMatch: A Holistic Approach to Semi-Supervised Learning
D. Berthelot et al. Conference on Neural Information Processing Systems 2019

[4] FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence
K. Sohn et al. Conference on Neural Information Processing Systems 2020

Link to pre-print

With the funding of Swedish Research Council

Matsoukas, Christos
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Should we Replace CNNs with Transformers for Medical Images?

Convolutional Neural Networks (CNNs) have reigned for a decade as the de facto approach to automated medical image diagnosis, pushing the state-of-the-art in classification, detection and segmentation tasks. Recently, vision transformers (ViTs) have appeared as a competitive alternative to CNNs, yielding impressive levels of performance in the natural image domain, while possessing several interesting properties that could prove beneficial for medical imaging tasks. In this work, we explore whether it is feasible to switch to transformer-based models for medical image classification as well, or if we should keep working with CNNs - can we trivially replace CNNs with transformers? We consider this question in a series of experiments on several standard medical image benchmark datasets and tasks. Our findings show that, while CNNs perform better if trained from scratch, off-the-shelf vision transformers can perform on par with CNNs when pretrained on ImageNet, both in a supervised and self-supervised setting.

Should we Replace CNNs with Transformers for Medical Images?

Christos Matsoukas[1,2,3], Johan Fredin Haslum[1,2,4], Moein Sorkhei[1,2], Magnus Söderberg[3], Kevin Smith[1,2]

[1] Science for Life Laboratory, Stockholm, Sweden
[2] School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden
[3] Cardiovascular, Renal and Metabolism Safety, Clinical Pharmacology & Safety Sciences, R&D AstraZeneca, Gothenburg, Sweden
[4] Bioscience Cardiovascular, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D AstraZeneca, Gothenburg, Sweden

Convolutional Neural Networks (CNNs) have reigned for a decade as the de facto approach to automated medical image diagnosis, pushing the state-of-the-art in classification, detection and segmentation tasks. Recently, vision transformers (ViTs) have appeared as a competitive alternative to CNNs, yielding impressive levels of performance in the natural image domain, while possessing several interesting properties that could prove beneficial for medical imaging tasks. In this work, we explore whether it is feasible to switch to transformer-based models for medical image classification as well, or if we should keep working with CNNs - *can we trivially replace CNNs with transformers?* We consider this question in a series of experiments on several standard medical image benchmark datasets and tasks. Our findings show that, while CNNs perform better if trained from scratch, off-the-shelf vision transformers can perform on par with CNNs when pretrained on ImageNet, both in a supervised and self-supervised setting.

Methods

We compare two mainstream models for classification:
- ResNet50 [2], as CNN representative.
- DeiT-S [3], for ViTs.

We consider three initialization strategies:
- Randomly initialized weights [4].
- Transfer learning using supervised ImageNet [1] pretrained weights.
- Self-supervised pretraining using DINO [5] on the target dataset, after initialization as in (2).

To asses whether vision transformers are able to produce high quality embeddings for segmentation we consider DeepLabV3 [14] and we simply replace its ResNet50 encoder with DeiT-S.

Results

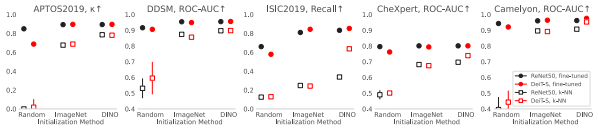


Figure 1: Performance comparison of RESNET50 and DEIT-S, two commonly used CNN-based and ViT-based architectures. The comparison covers several standard medical image classification datasets and different types of initialization including random init, IMAGENET pretraining, and self-supervision using DINO. Performance is measured after fine-tuning on the dataset, as well as using k-NN evaluation without fine-tuning. We report the median over 5 repetitions, error bars represent standard deviation.

Model	ISIC2018, IoU ↑	CSAW-S, IoU ↑
DEEPLAB3-RESNET50	0.802 ± 0.012	0.320 ± 0.008
DEEPLAB3-DEIT-S	0.845 ± 0.014	0.322 ± 0.028

Table 1: *Medical image segmentation with* DEEPLAB3 *comparing CNN vs. ViT encoders.*

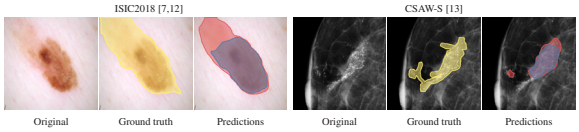


Figure 2: Medical image segmentation results comparing DEEPLAB3-RESNET50 (blue), DEEPLAB3-DEIT-S (red). Ground truth mask appears in yellow. Note that the ViT segmentations tend to do a better job of segmenting distant regions.

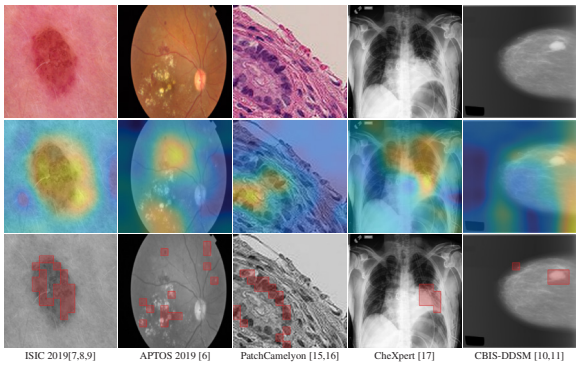


Figure 3: Comparing saliency for RESNET50 (2^{nd} row) and DEIT-S (3^{rd} row) on medical classification. Each column contains the original, a Grad-CAM visualization visualisation for ResNet50 and the top-50% attention map of the CLS token of DEIT-S.

Conclusions

- ViTs can reach the same level of performance as CNNs in small medical datasets, but require transfer learning in order to do so.
- ViTs can outperform CNNs using SSL pre-training when working with limited number of samples, but only marginally.
- ViTs offer built in high-resolution saliency maps that can be used to better understand the model's decisions.

References

[1] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee. 2009.
[2] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[3] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." International Conference on Machine Learning. PMLR. 2021.
[4] He, Kaiming, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." Proceedings of the IEEE international conference on computer vision. 2015.
[5] Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." arXiv preprint arXiv:2104.14294. 2021.
[6] Kaggle. "Aptos 2019 blindness detection". 2019
[7] Tschandl, Philipp , et al. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions." Scientific data 5.1. 2018.
[8] Codella, Noel CF, et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)." 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE. 2018.
[9] Combalia, Marc, et al. "BCN20000: Dermoscopic lesions in the wild." arXiv preprint arXiv:1908.02288. 2019.
[10] Sawyer Lee, R., et al. "Curated Breast Imaging Subset of DDSM. The Cancer Imaging Archive.". 2016.
[11] Clark, Kenneth, et al. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository." Journal of digital imaging 26.6. 2.013
[12] Codella, Noel, et al. "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)." arXiv preprint arXiv:1902.03368 (2019).
[13] Matsoukas, Christos, et al. "Adding seemingly uninformative labels helps in low data regimes." International Conference on Machine Learning. PMLR, 2020.
[14] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).
[15] Bejnordi, Babak Ehteshami, et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer." Jama 318.22 (2017): 2199-2210.
[16] Veeling, Bastiaan S., et al. "Rotation equivariant cnns for digital pathology." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2018.
[17] Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

Matsson, Anton
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Prototype-Based Off-Policy Evaluation

Before applying a new decision-making policy in safety-critical domains, e.g., in healthcare, we need a reliable estimate of the policy's value. While sampling from this target policy is not possible, we have access to samples from an unknown behavior policy which represents current practice. The problem of evaluating a target policy using data gathered under a behavior policy is known as off-policy evaluation (OPE). Importance sampling (IS) is often used to perform OPE but can provide uncertain value estimates when there are significant differences between the policies. To better diagnose potential problems, we propose estimating the unknown behavior policy for IS using prototype learning. We apply this approach in the evaluation of policies for sepsis treatment, demonstrating that the learned prototypes give a condensed summary of differences between the policies.

Matsson, Anton
Chalmers

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Prototype-Based Off-Policy Evaluation

Anton Matsson, Chalmers University of Technology

CHALMERS UNIVERSITY OF TECHNOLOGY

Computer Science and Engineering

Main Advisor: Fredrik Johansson

Abstract

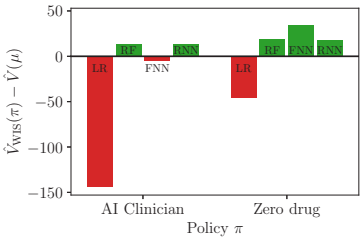
Before applying a new decision-making policy in safety-critical domains, e.g., in healthcare, we need a reliable estimate of the policy's value. While sampling from this *target policy* is not possible, we have access to samples from an unknown *behavior policy* which represents current practice. The problem of evaluating a target policy π using data gathered under a behavior policy μ is known as *off-policy evaluation* (OPE). Importance sampling (IS) is often used to perform OPE but can provide uncertain value estimates when there are significant differences between the policies. To better diagnose potential problems, we propose estimating the unknown behavior policy for IS using prototype learning. We apply this approach in the evaluation of policies for sepsis treatment, demonstrating that the learned prototypes give a condensed summary of differences between the policies.

Background

Importance sampling. Given an observational dataset of trajectories H (sequences of contexts X and actions A) and outcomes R, the standard IS estimator weights the outcomes by the density ratio of the target policy π and the behavior policy μ:

$$\hat{V}_{\text{IS}}(\pi) = \frac{1}{N}\sum_{i=1}^{N}\prod_{t=0}^{T}\frac{p_\pi\left(A_t=a_t^{(i)}\middle|H_t=h_t^{(i)}\right)}{\hat{p}_\mu\left(A_t=a_t^{(i)}\middle|H_t=h_t^{(i)}\right)}r^{(i)}.$$

What can go wrong? The figure shows an example of naïve OPE of two target policies for sepsis management: the AI Clinician [1] and a zero-drug policy. Weighted IS (WIS) was used with different models $\hat{p}_\mu$ of the unknown behavior policy. The zero-drug policy seems to be superior to the behavior policy, followed by physicians in data. But never treating patients with sepsis goes against intuition – can we trust these estimates?



Human evaluation. For $\hat{V}_{\text{IS}}(\pi)$ to be unbiased, overlap must be satisfied. That is, for all t, $p_\pi(A_t|H_t) > 0 \Rightarrow p_\mu(A_t|H_t) > 0$. Because the extent of overlap is unknown when μ is unknown, assessing the quality of $\hat{V}_{\text{IS}}(\pi)$ relies on evaluation by a domain expert. We identify three key questions in such an evaluation:

A. Which observations contribute to the IS estimate?
B. In which situations is overlap violated?
C. If $\hat{V}(\pi) > \hat{V}(\mu)$, what gives π the edge?

References

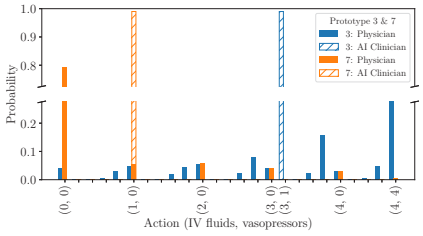
1. Komorowski, M. et al. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11): 1716–1720, 2018.
2. Matsson, A. and Johansson, F. D. Case-based off-policy policy evaluation using prototype learning. *arXiv preprint arXiv:2111:11113*, 2021.

Results

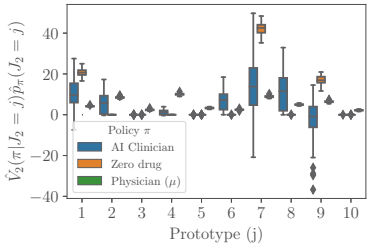
Using prototypes. When μ is estimated with black-box models, it can be difficult to detect problems with the IS estimate. Instead, we propose performing OPE using a prototype-based estimate of μ [2]. The overall idea is to compute the probability $p_\mu(A|H)$ by comparing the history H to a small set of prototype cases, which are

- representative trajectories from the training data
- automatically selected by the learning algorithm
- interpretable by a domain expert.

Answering A & B. By evaluating μ and π for each of the prototypes, we get a condensed summary of differences between the policies; see the figure for an example. We can identify groups of patients for which the ratio $p_\pi(A|H)/\hat{p}_\mu(A|H) \gg 1$ for certain actions (see prototype 7) and spot violations of overlap (see prototype 3). Note that the zero-drug policy always predicts action (0, 0) with probability 1.



Answering C. The prototypes allows for computing prototype-based contributions to the overall value $V(\pi)$ at each time step. In this way, we can see for which groups of patients a certain policy is most beneficial. The figure shows an example for $t = 2$. Here, $V_t(\pi|J_t = j)$ is the value of π for prototype j at time t and $p_\pi(J_t = j)$ is the probability of being assigned to prototype j at time t.



Mehta, Shivam
KTH

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Neural HMMs are all you need (for high-quality attention-free TTS)

Speech synthesis is an application of Generative modelling, where the output is generally conditioned on the input text. This is also referred as Text-To-Speech or TTS Systems. Current experiments on the state of the art speech synthesis systems takes days to realise if the energy hungry GPUs are chunking numbers properly to generate speech and can break down into gibberish. How can we make our current speech / audio experimentation iterations better and save time, effort and energy, without compromising with the quality of synthesised speech? We propose an autoregressive TTS system with a combination of Hidden Markov Models and Deep Neural Networks giving us a smaller size, comparable naturalness, faster iterations, control over speaking rate.

Mehta, Shivam
KTH

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Neural HMMs are all you need (for high-quality attention-free TTS)

Shivam Mehta – supervised by Gustav Eje Henter and Jonas Beskow

KTH ROYAL INSTITUTE OF TECHNOLOGY

Research Question

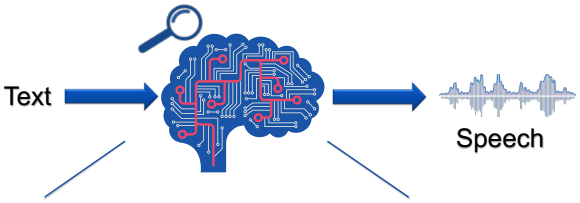
How can we make our current speech / audio experimentation iterations better and save time, effort and energy, without compromising with the quality of synthesised speech?

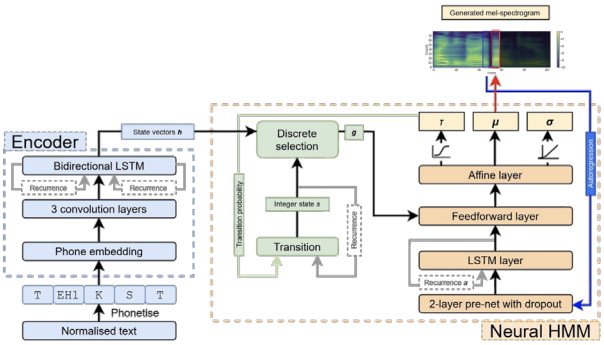
Background

Current state of the art speech synthesis systems (Tacotron 2):

- Takes days to realise if the energy hungry GPUs are chunking numbers properly enough to generate speech
- Can break down into gibberish
- Requires significant amount of data to work at all

Listen to synthesised audio samples





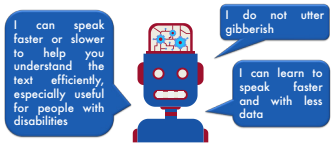
Approach

We propose an autoregressive text-to-speech (TTS) synthesis system that combines the best characteristics of Hidden Markov Models (HMMs) and Deep Neural Networks (Mehta, 2022*)

Key advantages:

- Smaller size than state-of-the-art models.
- Achieves comparable naturalness as state-of-the-art models
- Learns to speak and align much faster with significantly less data withing 2000 iterations (see fig. below)
- Allows for easier control over speaking-rate
- Fast training time allows for swift speech research

Type	Tacotron 2		Neural HMM	
Condition	T2+P	T2−P	NH2	NH1
Size	28.2M	23.8M	15.3M	12.7M
MOS	3.41±0.01	3.25±0.01	3.24±0.01	2.68±0.01



The research is supported by Wallenberg AI, Autonomous Systems and Software Program (WASP) at Department of Speech, Music and Hearing (TMH)

* Neural HMMs are all you need (for high-quality attention-free TTS) - **Shivam Mehta**, Éva Székely, Jonas Beskow, and Gustav Eje Henter – Submitted to ICASSP 2022.

Melnyk, Pavlo
Linköping University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Embed Me If You Can: A Geometric Perceptron

Solving geometric tasks involving point clouds by using machine learning is a challenging problem. Standard feed-forward neural networks combine linear or, if the bias parameter is included, affine layers and activation functions. Their geometric modeling is limited, which motivated the prior work introducing the multilayer hypersphere perceptron (MLHP). Its constituent part, i.e., the hypersphere neuron, is obtained by applying a conformal embedding of Euclidean space. By virtue of Clifford algebra, it can be implemented as the Cartesian dot product of inputs and weights. If the embedding is applied in a manner consistent with the dimensionality of the input space geometry, the decision surfaces of the model units become combinations of hyperspheres and make the decision-making process geometrically interpretable for humans. Our extension of the MLHP model, the multilayer geometric perceptron (MLGP), and its respective layer units, i.e., geometric neurons, are consistent with the 3D geometry and provide a geometric handle of the learned coefficients. In particular, the geometric neuron activations are isometric in 3D, which is necessary for rotation and translation equivariance. When classifying the 3D Tetris shapes, we quantitatively show that our model requires no activation function in the hidden layers other than the embedding to outperform the vanilla multilayer perceptron. In the presence of noise in the data, our model is also superior to the MLHP.

Melnyk, Pavlo
Linköping University

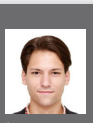
WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

LINKÖPING UNIVERSITY

ICCV 2021 OCTOBER 11-17 VIRTUAL

Embed Me If You Can: A Geometric Perceptron

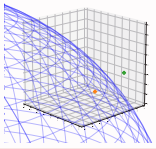
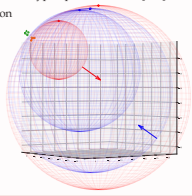
Pavlo Melnyk, Michael Felsberg, Mårten Wadenbäck
Computer Vision Laboratory, Linköping University, Sweden



{pavlo.melnyk, michael.felsberg, marten.wadenback}@liu.se

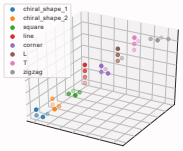
Contributions

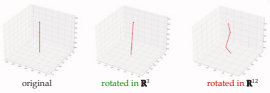
- We introduce a **conformal embedding approach** to problems in Euclidean 3D space and investigate its performance on **point clouds**
- Our embedding is based on **extended geometric modeling** of multiple points compared to standard MLPs and hypersphere neurons [1, 2]
- We obtain a geometric interpretation in terms of multiple 3D **spherical decision surfaces**

Problem

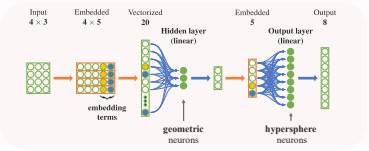
- We focus on **3D geometry**: important for pose estimation — a prerequisite for grasping, 3D inpainting, augmented reality
- Toy example:
  - classify the **3D Tetris** shapes [3]
  - each shape is $\mathbf{R}^{8\times3}$
- How to **represent such input** for the baseline MLHP [2]?
- Naïve approach:
  - vectorize to $\mathbf{R}^{12}$
  - then embed in $\mathbf{R}^{14} \cong \mathbf{ME}^{12}$
- Leads to **geometric inconsistency!**





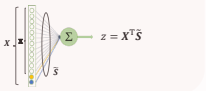
Geometric perceptron

- We **embed the input point-wise**: each point to be embedded in $\mathbf{R}^5 \cong \mathbf{ME}^3$
- **Consistent** with the geometry of the input
- Leads to the **change of the decision surface** geometry → **geometric neurons**
- Our extension of MLHP [2] – **multilayer geometric perceptron (MLGP)**

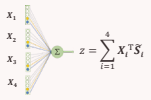


Hypersphere neuron [1] vs Geometric neuron (ours)

- One hypersphere in $\mathbf{R}^{12}$, learned as $\tilde{S} \in \mathbf{R}^{14}$
- Classifies a single input vector $\mathbf{x} \in \mathbf{R}^{12}$ embedded in $\mathbf{ME}^{12}$ as $X \in \mathbf{R}^{14}$



- Four spheres in $\mathbf{R}^3$, learned as $\tilde{S}_i \in \mathbf{R}^5$
- Each classifies the corresponding point $\mathbf{x}_i \in \mathbf{R}^3$ embedded in $\mathbf{ME}^3$ as $X_i \in \mathbf{R}^5$



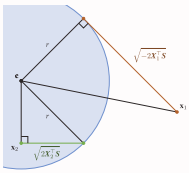
- Isomorphic implementation in the Euclidean space:

$$X = (x_1, \dots, x_n, -1, -\tfrac{1}{2}\mathbf{x}^2) \in \mathbf{R}^{n+2}$$

$$S = (c_1, \dots, c_n, \tfrac{1}{2}(\mathbf{c}^2 - r^2), 1) \in \mathbf{R}^{n+2}, \text{ learned } \textit{freely} \text{ as}$$

$$\tilde{S} = (s_1, \dots, s_{n+2}) \in \mathbf{R}^{n+2}$$

- Our graphical interpretation in 2D
- The scalar product $X^T S$ determines **the cathetus length!**
- Depends on the relative position of the point $\mathbf{x}$ wrt the learned sphere $(\mathbf{c}, r)$



Experiments

- **Classification test**:
  - apply random 3D rigid transformations and perturb transformed Tetris shapes
  - train and compare the performances of the proposed MLGP, baseline MLHP [2], and vanilla MLP

Table 1: Model accuracies on the test data (mean and std over 50 runs, %); values in parentheses represent the accuracy of the 10 best models selected based on the *validation* accuracy.

	Main dataset			Theta-split		
Noise	$\alpha = 0.0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.0$	$\alpha = 0.1$	$\alpha = 0.2$
MLP	71.7 ± 4.0 (78.0 ± 2.6)	65.3 ± 4.3 (71.7 ± 4.4)	60.7 ± 3.2 (65.9 ± 2.8)	55.3 ± 5.8 (63.8 ± 3.1)	51.8 ± 3.4 (56.5 ± 2.2)	48.4 ± 3.6 (53.3 ± 1.9)
Baseline	**92.5 ± 0.4** (**92.8 ± 0.2**)	81.3 ± 3.4 (86.9 ± 0.4)	68.2 ± 3.2 (72.2 ± 0.5)	87.6 ± 0.6 (88.3 ± 0.6)	78.4 ± 3.8 (83.8 ± 1.8)	65.7 ± 3.7 (70.4 ± 0.7)
Ours	91.8 ± 2.2 (92.5 ± 0.2)	**81.8 ± 5.9** (**89.3 ± 0.4**)	**69.7 ± 8.3** (**79.6 ± 0.2**)	**87.9 ± 0.8** (**89.0 ± 0.4**)	**80.3 ± 5.1** (**87.5 ± 0.5**)	**68.2 ± 8.4** (**79.4 ± 0.3**)

- **Isometry test**:
  - rigid transformations and the application of the learned model commute!

Table 2: Isometry test: the pretrained (original) and transformed MLGP accuracies on the original and transformed test split from the main dataset (mean and std over 10000 runs, %).

	MLGP	Transformed MLGP
Original data	**91.98 ± 0.00**	86.11 ± 3.46
Transformed data	86.09 ± 3.45	**91.98 ± 0.00**

Key features

- **No activation function necessary**: embedding is naturally non-linear
- **Superior performance**: to the baseline [2] (especially when the data are perturbed) and MLP
- **Geometric interpretability**: learned spheres each classifying the respective input point in the conformal space; possible *inversion* of the decision due to unembedding
- **Complexity**: typically, increased only with a factor of 1.6-2 compared to MLPs
- **Isometric activations**: crucial for rotational and translational equivariance

References

[1] Banarer, V., et al. "The hypersphere neuron." *ESANN*. 2003.

[2] Banarer, V., et al. "Design of a multilayered feed-forward neural network using hypersphere neurons." *International Conference on Computer Analysis of Images and Patterns*. Springer, Berlin, Heidelberg, 2003.

[3] Thomas, N., et al. "Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds." *arXiv preprint arXiv:1802.08219* (2018).

Mendez, Julian Alfredo
Umeå University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

A Language to Bridge the Ethical Gap Between Humans and Machines

AI has raised concerns on whether its use follows ethical principles. We present a new formal language to model ethical requirements. It is statically typed, object-oriented, purely functional, and can be used to create efficient prototypes.

Mendez, Julian Alfredo
Umeå University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

A new object-oriented purely **functional language** formalizes **ethical requirements** for **AI systems**.

A Language to Bridge the Ethical Gap Between Humans and Machines

Julian Alfredo Mendez — julian.mendez@umu.se
Umeå University

Introduction

AI has raised concerns on whether its use follows ethical principles. Our research questions are:

- Can we design a **readable language** to express **ethical requirements** to monitor AI systems?
- Can the specifications be efficiently **prototyped**?

Material and Methods

Why a new language?

The language needs to be not only **expressive** to model ethical requirements, but also **easier** and **clearer** than other similar languages, to **agree on ethical requirements**.

Language properties to model ethical requirements:

the language should ...	we make it ...
be easy to understand	**of simple syntax**
be applicable to many domains	**expressive**
prevent side effect errors	**purely functional**
have powerful design tools	**object-oriented**
detect errors in compile time	**statically typed**
be efficient and easy to integrate	**JVM**

Results and Discussion

We designed a new language that provides:

- **Classes** (to model entities)
- **Class extensions** (Open-Close Principle)
- **Constants and functions** (inside classes)
- **Standard arithmetic operations**
- **Standard basic types**
- **Lambda expressions**
- **Class constructors** (to instantiate objects)
- **Pattern matching** (for class deconstruction)

The descriptions are translated to Scala code, and then to Java Virtual Machine (JVM) bytecode.

Example

```
class RequirementMonitor = {

  has pricing_agent: PricingAgent

  get_price (customer: Customer, flight: Flight, date_in_days: Int): Int =
    pricing_agent.get_price (customer, flight, date_in_days)
}

class Report2 (compliant: Boolean, old_price: Int, new_price: Int)

class Requirement2Monitor
  extends RequirementMonitor = {

  acceptable_yearly_increase = 1.25

  get_report (customer: Customer, flight: Flight, date_in_days: Int): Report2 =
    let
      old_price = get_price (customer, flight, get_a_year_before (date_in_days) )
      new_price = get_price (customer, flight, date_in_days)
    in Report2 (new_price <= old_price * acceptable_yearly_increase,
        old_price, new_price)

  get_a_year_before (date_in_days: Int): Int =
    date_in_days - 365
}
```

State of the Art

We compared available programming languages:

language	version	A	B	C	D	E	F
Agda	2.6.2	Yes	Yes	No	Yes	No	10^9
Clojure	1.10.3	Yes	No	No	No	Yes	10^9
Coq	8.13.2	Yes	Yes	No	Yes	No	10^7
Haskell	8.6.5	Yes	Yes	No	Yes	No	10^8
Idris (2)	0.4.0	Yes	Yes	No	Yes	No	10^{10}
OCaml	4.08.1	Yes	No	Yes	Yes	No	10^9
Prolog	7.6.4	No	No	No	No	No	10^8
Python	3.8.10	No	No	Yes	No	No	10^8
Scala	3.0.2	Yes	No	Yes	Yes	Yes	10^{10}
our language	0.12.0	Yes	Yes	Yes	Yes	Yes	10^{10}

References: **A.** dominantly functional **B.** purely functional (no imperative features) **C.** object-oriented **D.** statically typed **E.** JVM integration **F.** repetitions in 30 s

Conclusion and Future Work

This language is very expressive and can model ethical requirements. We plan to include verification of pieces of code using Coq. Its **readability**, its **expressiveness**, and its **self-consistency** help to bridge the ethical gap between humans and machines.

umu.se/staff/julian-mendez

UMEÅ UNIVERSITY

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Mwai, Newton
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Machine learning for improved decision making based on historical data

Historical Data: Simulators make unique benchmarks for causal effect estimation since they do not rely on unverifiable assumptions or the ability to intervene on real-world systems, but are often too simple to capture important aspects of real applications. We propose a simulator of Alzheimer's disease aimed at modeling intricacies of healthcare data while enabling benchmarking of causal effect and policy estimators. We fit the system to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and ground hand-crafted components in results from comparative treatment trials and observational treatment patterns. The simulator includes parameters which alter the nature and difficulty of the causal inference tasks, such as latent variables, effect heterogeneity, length of observed history, behavior policy and sample size. We use the simulator to compare estimators of average and conditional treatment effects.

Decision Making: In personalized medicine, we can assume that there exists a patient latent state $Z\in{R^d}$ which we aim to learn from historical data, and once we learn it, we can recommend personalized near-optimal treatments. In searching for treatments, we usually are constrained in the number of trials that we can perform. We formulate a problem of near-optimal treatment search in a latent bandits setting where: $A\in \{1, ..., k\}, \mu_Z \in R^k$, and budget $T = \beta K$, where β is reasonably small. We apply pure exploration theory to investigate how to present a fixed confidence for near-optimal treatments with a budget in the order $T = \beta K$, where β is reasonably small. We run experiments and simulations with realistic Alzheimer's data from the ADCB environnment.

Mwai, Newton
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Machine learning for improved decision making in healthcare with historical data

Newton Mwai Kinyanjui, Fredrik D. Johansson,
Computer Science and Engineering Department, Chalmers

Motivation & Research goals

Healthcare organizations are eager to improve decision making using machine learning applied to records of past decisions and outcomes. Electronic healthcare records are constantly updated with decisions on tests, treatments, procedures and drug prescriptions. If used appropriately, machine learning has the potential to use this data to personalize and improve medicine. Key challenges are a) access to highly realistic observational simulation data in healthcare, and b) ensuring that machine learning systems do not pick up on associations that are not causally related to the results of decisions, to ensure robust decisions when the systems are applied to new problems or new domains.

Decision Making

Personalized medicine with observational data and latent bandits*

In personalized medicine, we can assume that there exists a patient latent state $Z \in \{R^d\}$ which we aim to learn from historical data, and once we learn it, we can recommend personalized near-optimal treatments. As an example of a latent state, it is believed that there are multiple subtypes of Alzheimer's disease which could respond differently to the same treatment.

In searching for treatments, we are usually constrained in the number of trials that we can perform. We formulate a problem of near-optimal treatment search in a latent bandit setting where: treatments (arms) $A \in \{1, ..., k\}$ and budget $T = \beta K$, where β is reasonably small.

We apply pure exploration theory to investigate how to present a fixed confidence for near-optimal treatments with a budget in the order $T = \beta K$, where β is reasonably small. We run experiments and simulations with realistic Alzheimer's data from the Alzheimer's Disease Causal estimation Benchmark (ADCB) environment.

Goal: Pure exploration with a very small budget

Identify an ϵ – optimal arm (treatment)

$$P(R_{AT} \leq R_A - \epsilon) \leq \delta$$

Subject to $T = \beta K$, where β is reasonably small

How does latent state information help us identify such an arm?

* *Project ongoing*

References

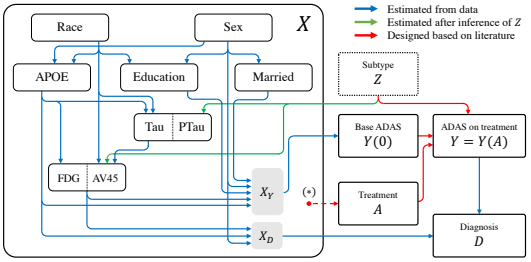
1. Kinyanjui, Newton Mwai, and Fredrik D. Johansson. "ADCB: An Alzheimer's disease benchmark for evaluating observational estimators of causal effects." ML4H 2021 - extended abstract *arXiv:2111.06811* (2021).
2. Meemansa Sood, et al. "Realistic simulation of virtual multi-scale, multi-modal patient trajectories using bayesian networks and sparse auto-encoders". Scientific reports, 10(1):1–14, 2020.
3. Hernandez, Santiago, et al. "Pharmacological treatment of Alzheimer's disease: effect of race and demographic variables." Journal of Alzheimer's Disease 19.2 (2010): 665-672.

Historical Data

Alzheimer's Disease Causal estimation Benchmark (ADCB)

Evaluating learned decision-making policies and observational estimates of causal effects is challenging, as it relies on strong assumptions and access to large samples, hence the research community often turns to simulators for benchmarking which often lack realism.

We propose a new benchmark for evaluating estimators of causal effects that combines the strengths of data-driven simulators with those of hand-crafted components by fitting a longitudinal causal model of patient variables to real data and providing tunable parameters for the generated data.



We assume the above causal graph based on previous literature (Sood et. al., 2020) and domain practitioner input. For each continuous(or discrete) attribute, a regression(or classification) model is fit with respect to its parents. We fit sequences autoregressively.

Tunable parameters

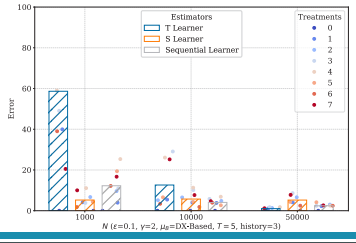
μ – Behavior policy ∈ {Random, Diagnosis-based, Hernandez-based (Hernandez et al.(2010))}

ϵ – Overlap parameter ∈ [0, 1]

γ – Treatment Effect Heterogeneity ≥ 0

N – Number of Samples ≥ 0

T – Sample trajectory length (history length) ∈ {0, 12, 24, …, 120}



Left: Results of using generated data from ADCB in comparing causal effect estimators based on Conditional Average Treatment Effects Error for Number of patient samples, N

Norlund, Tobias
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?

Large language models are known to suffer from the hallucination problem in that they are prone to output statements that are false or inconsistent, indicating a lack of knowledge. A proposed solution to this is to provide the model with additional data modalities that complements the knowledge obtained through text. We investigate the use of visual data to complement the knowledge of large language models by proposing a method for evaluating visual knowledge transfer to text for uni- or multimodal language models. The method is based on two steps, 1) a novel task querying for knowledge of memory colors, i.e. typical colors of well-known objects, and 2) filtering of model training data to clearly separate knowledge contributions. Additionally, we introduce a model architecture that involves a visual imagination step and evaluate it with our proposed method. We find that our method can successfully be used to measure visual knowledge transfer capabilities in models and that our novel model architecture shows promising results for leveraging multimodal knowledge in a unimodal setting.

Norlund, Tobias
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it?

Tobias Norlund*, Lovisa Hagström*, Richard Johansson
Chalmers University of Technology

Motivation & Summary

Despite the ability of language models to learn and hold large quantities of structural knowledge [1], LMs are are also known to suffer from the **hallucination problem** in that they are prone to output statements that are false or inconsistent, indicating significant knowledge gaps [2]. A hypothesis is that this knowledge in many cases simply is missing in the large text corpora typically used for training the models, due to e.g reporting bias [3]. In such cases certain types of knowledge might also be more readily available in a different data modality.

In this work, we investigate **visual knowledge transfer**, i.e. to which extent language models can incorporate and *textually* express knowledge originating from a visual modality.

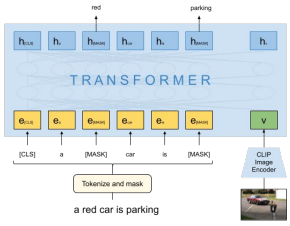
We investigate this by constructing a novel cloze-style task testing knowledge of **memory colors** for common objects (such as blood is red, a lemon is yellow etc). We also build a large vision-and-language dataset used for self-supervised training, and by careful filtering we make sure the color information is only available through the images and not the text.

Finally, we compare two strategies for how to most effectively query the trained multimodal language model for this visual knowledge.

We show that language models are able to texually express knowledge obtained from a visual modality, as a result from multimodal self-supervised training.

References
[1] Fabio Petroni et al. Language Models as Knowledge Bases? EMNLP (2019).
[2] Robert Logan et al. 2019. Barack's wife Hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.
[3] Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*.
[4] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision.

Multimodal Self-supervised Training: CLIP-BERT



We propose a novel visual-and-language model denoted **CLIP-BERT**, where the image encoder of the pre-trained CLIP [4] model is used to represent the image before appended to the input of a BERT-base model. We train this model using MLM on our vision-and-language dataset, and seek to evaluate how this affects performance on the memory colors task.

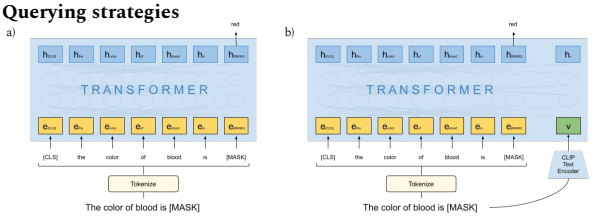
Task: Memory colors

With the help of 11 human annotators, we have created a dataset of 109 common objects and their memory colors, with high annotator agreement.

a	*lemon*	*yellow*	
descriptor	item	color	picture

* Equal contribution

Querying strategies



We compare two strategies for querying the language models for this visual knowledge, through [MASK] token prediction.

a) Implicit: The visual knowledge is retrieved from the trained parameters of the language model.

b) Explicit: We use the text encoder of the CLIP model to "imagine" a visual representation of the query text and append to the transformer input. This way the visual knowledge can partly be retrieved from this input, as well as from the trained model parameters.

Results

- ①: The original (text-only) BERT-base performs poorly on this task, close to majority baseline
- ① vs ⑥: Continue MLM training of BERT-base on the text part of our multimodal dataset improves performance slightly despite filtering
- ⑥ vs ⑧: Adding images to training improves performance significantly, showing effective visual knowledge transfer!
- ③ vs ④ and ⑦ vs ⑧: The explicit querying strategy performs better than the implicit

Training	Model	Accuracy	
	Random baseline	0.091 ± 0.026	
	Majority baseline	0.229 ± 0.000	
	Human baseline	0.937 ± 0.051	
None	BERT-base	0.252 ± 0.102	①
Unfiltered	BERT-base	0.724 ± 0.112	②
	CLIP-BERT		
	implicit	0.744 ± 0.080	③
	explicit	0.870 ± 0.086	④
	images	0.876 ± 0.063	⑤
Filtered	BERT-base	0.460 ± 0.083	⑥
	CLIP-BERT		
	implicit	0.541 ± 0.060	⑦
	explicit	0.733 ± 0.098	⑧
	images	0.785 ± 0.055	⑨

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Recorded Future

Olmin, Amanda
Linköpings universitet

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Robustness and reliability when training with noisy labels

Algorithms developed for the purpose of handling label noise in supervised training, are commonly evaluated in terms of accuracy. However, robustness in accuracy is not sufficient in applications where reliable uncertainty estimates are critical. For an input-dependent noise model, we investigate the effect of label noise on strictly proper loss functions as well as the set of robust loss functions characterised by noise-insensitive, asymptotic risk minimisers. We find that not only robust, but also strictly proper loss functions offer asymptotic robustness in accuracy. However, neither guarantee that the final model is calibrated. Moreover, both strictly proper and robust loss functions are susceptible to overfitting in practice.

Olmin, Amanda
Linköpings universitet

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Robustness and reliability when training with noisy labels

Amanda Olmin, Linköping University
Department of Computer and Information Science
Supervisors: Fredrik Lindsten (LiU) and Lennart Svensson (Chalmers)

LINKÖPINGS UNIVERSITET

Motivation

Supervised training of deep learning models is highly dependent upon labelled data. Curating a (possibly large) annotated training data set is costly and time-consuming and the risk of incorporating label noise is imminent. Label noise can hurt model performance by: (i) shifting the asymptotic risk minima towards the conditional distribution over noisy, instead of clean, labels, and (ii) increasing the risk of overfitting (Zhang et al., 2017). While algorithms robust to label noise are commonly evaluated in terms of accuracy, this is not enough if reliable uncertainty estimates are critical. We establish this idea and investigate the effect of label noise on model performance through a critical analysis of robust loss functions.

Simple non-uniform label noise

For noisy label $\tilde{Y} \in \mathcal{Y}$, true label $Y \in \mathcal{Y}$ and input variable $X \in \mathcal{X}$, simple non-uniform label noise (Ghosh et al., 2017) is defined by

$$\mathbb{P}(\tilde{Y} = \tilde{y} \mid Y = y, X = x) = \begin{cases} 1 - \omega(x), & \text{if } \tilde{y} = y \\ \frac{\omega(x)}{K-1}, & \text{otherwise} \end{cases}$$

with $\mathcal{Y} = \{1, \dots, K\}$ and noise parameter $0 \leq \omega(x) < \frac{K-1}{K}$.

Robust loss functions

A loss function ℓ is robust (Ghosh et al., 2017) to label noise if for all asymptotic minimisers f^* of the clean risk, $\mathcal{R}_\ell$, it holds that

$$\tilde{\mathcal{R}}_\ell(f^*) \leq \tilde{\mathcal{R}}_\ell(f), \quad \forall f \in \mathcal{F}$$

where $\tilde{\mathcal{R}}_\ell$ is the risk under the noisy data distribution.

Strictly proper loss functions recover the true conditional $f^*(x) = \mathbb{P}(Y \mid X = x)$ if labels are clean, but are not robust.

Symmetric loss functions are robust under simple non-uniform label noise (Ghosh et al., 2017). They satisfy

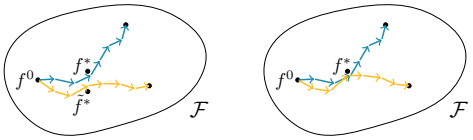
$$\sum_{k=1}^{K} \ell(q, k) = C, \quad \forall x \in \mathcal{X}, \ \forall q \in \Delta^{K-1}$$

for some constant C.

Training dynamics

We imagine two phases of training a model $f : \mathcal{X} \to \Delta^{K-1}$.

1. The training trajectory "aims" towards the true risk minimiser f^* (or $\tilde{f}^*$ if labels are noisy) and passes "close" to it.
2. The training trajectory diverges and the model overfits to the data.



(a) Strictly proper loss, $\tilde{f}^* \neq f^*$ (b) Robust loss, $\tilde{f}^* = f^*$

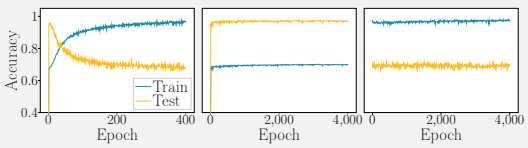
Robustness does not imply reliability

Under simple non-uniform label noise, we find that strictly proper and symmetric loss functions

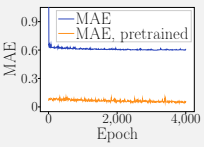
- are robust to label noise in accuracy,
- do not result in calibrated models (uncertainty estimates unreliable),

The robustness condition is not sufficient if reliable uncertainty estimates are critical.

- are both susceptible to overfitting in practice.



(a) CCE (b) MAE (c) MAE, pretrained



(d) MAE loss

Figure: Experiments on MNIST with simple non-uniform label noise, $\omega(x) = 0.3 \ \forall x$. The empirical robustness of Mean Absolute Error (b) has been attributed to its theoretical robustness (Ghosh et al., 2017). However, the theory does not account for overfitting. In practice, the overfitted model (c) achieves a smaller loss (d).

Calibration-based strictly properness

Let $\mathcal{F}_\mathcal{C} \subseteq \mathcal{F}$ be the set of calibrated models. The loss function ℓ, with asymptotic risk minimisers f^*, is calibration-based strictly proper if

$$f^* \in \mathcal{F}_\mathcal{C}, \quad \forall f^* \in \mathcal{F},$$

for all $\mathbb{P}(Y \mid X)$ and for all input distributions μ_X.

A loss function that is both robust and calibration-based strictly proper will preserve accuracy and ensure reliability.

References

Ghosh, A., Kumar, H., and Sastry, P. S. (2017). Robust loss functions under label noise for deep neural networks. In *AAAI Conference on Artificial Intelligence*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *ICLR*.

Oskarsson, Joel
Linköping University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Deep Gaussian Markov Random Fields for General Graphs

Machine learning methods on graphs have proven useful in many applications due to their ability to handle generally structured data. The framework of Gaussian Markov Random Fields (GMRFs) provides a principled way to define Gaussian models on graphs by utilizing the sparsity structure. We propose a highly scalable method for defining GMRFs on general graphs based on the layer structure of Deep GMRFs. The parameters of the resulting model can be trained efficiently using variational inference and existing software for Graph Neural Networks. For a Gaussian likelihood exact Bayesian inference is possible for predictions. The usefulness of the model and the multi-layer structure is verified by experiments on a number of synthetic and real world datasets.

Oskarsson, Joel
Linköping University

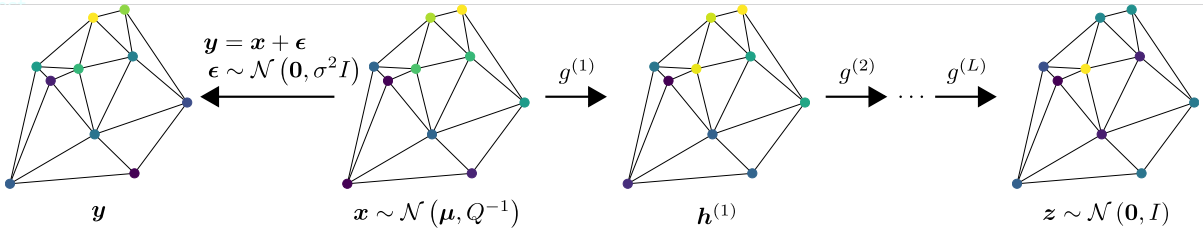
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Deep Gaussian Markov Random Fields for General Graphs[a]

Joel Oskarsson, joel.oskarsson@liu.se, Linköping University
Dept. of Computer and Information Science, Division of Statistics and Machine Learning
Supervisors: Fredrik Lindsten (LiU), Per Sidén (LiU, Arriver) and Jose M. Peña (LiU)

[a]Ongoing research

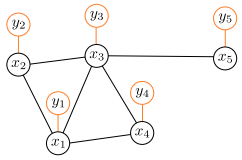
LINKÖPING UNIVERSITY



Background: Gaussian Markov Random Fields

Gaussian Markov Random Fields (GMRFs)

- Graphical models where the nodes $\boldsymbol{x}$ are jointly Gaussian [1]
- Conjugate prior to Gaussian likelihood $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{x}, \sigma^2 I)$
  - Posterior $p(\boldsymbol{x}|\boldsymbol{y})$ analytically tractable, also a GMRF!



Deep GMRFs (DGMRFs) [2]

- GMRF $\boldsymbol{x}$ defined implicitly by linear map g

$$\boldsymbol{z} = g(\boldsymbol{z}) = G\boldsymbol{x} + \boldsymbol{b}, \quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, I) \tag{1}$$

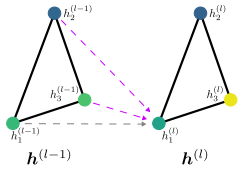
- g defined as a combination of L simpler layers $g = g^{(L)} \circ g^{(L-1)} \circ \cdots \circ g^{(1)}$
- Originally restricted to lattice-graphs (image-structured data)

Method: Graph DGMRF

- Consider graph $\mathcal{G}$ with adjacency matrix A and degree matrix $D = \mathrm{diag}([d_1, d_2, \ldots, d_N]^\intercal)$
- Define a layer construct $\boldsymbol{h}^{(l)} = g^{(l)}(\boldsymbol{h}^{(l-1)}) = G^{(l)}\boldsymbol{h}^{(l-1)} + \boldsymbol{b}^{(l)}$

$$G^{(l)} = \alpha_l D^{\gamma_l} + \beta_l D^{\gamma_l - 1} A \tag{2}$$
$$\boldsymbol{b}^{(l)} = b_l \mathbf{1} \tag{3}$$

  - b_l, α_l, β_l and γ_l are trainable parameters



- Closely corresponds to Graph Neural Network (GNN) formulation. Existing frameworks can be utilized for automatic differentiation and GPU acceleration.
- Variational training, maximizing the Evidence Lower Bound (ELBO)
- Training requires efficient evaluation of $\log|\det(G^{(l)})|$. We develop two solutions:
  - Exact method based on pre-computing the eigenvalues of $D^{-1}A$
  - Scalable approximate method based on reformulating log-determinant as a power series

References

[1] H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. Number 104 in Monographs on statistics and applied probability. Chapman & Hall/CRC, 2005.

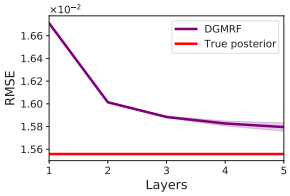
[2] P. Sidén and F. Lindsten. Deep gaussian markov random fields. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

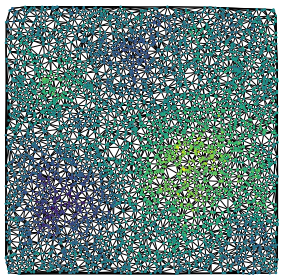
Experiments

- Experiments on inpainting problems where we train on 50% of nodes and evaluate on remaining 50%

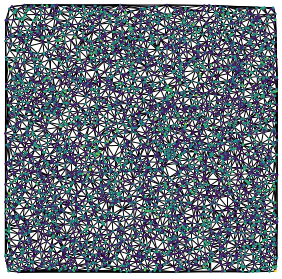
Synthetic GMRF Dataset

- Sample from known GMRF created as random mix of multiple DGMRFs
- 1-5 layer DGMRFs trained and evaluated





Posterior mean



Posterior marginal std.-dev.

Wikipedia Graph

- 11 631 Wikipedia pages as nodes
- Edges are links between pages
- Target $\boldsymbol{y}$ is log average monthly traffic
- CRPS, probabilistic metric taking uncertainty into account

	RMSE	CRPS
Baselines		
Graph GP	2.169	1.251
DGP (GNN)	1.293	0.769
I-GMRF	1.526	0.939
DGMRF		
1 Layer	1.312	0.704
3 Layer	1.230	0.652
5 Layer	**1.179**	**0.620**

$\Rightarrow$ Additional experiments on weighted graphs, use of node features and scaling to $> 10^5$ nodes

Summary

- Layered Gaussian model for graph-structured data
- Efficient variational training using Graph Neural Network framework
- Exact Bayesian inference gives principled uncertainty estimates

Patil, Minal Suresh
Umeå univeristet

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Towards Explainable Agency in Multi-Agent Systems Using Constraint Programming

Logical reasoning is a fundamental aspect of human behaviour, and this is an important criterion to build human-like reasoning in intelligent autonomous multi-agent systems. So far, the field of knowledge representation and reasoning has employed logic-based symbolic techniques to mimic the challenging task of incorporating human-like reasoning in multi-agent systems. However, the field of machine learning has shown increasing interest to take on this challenge. In this research, we describe a methodology that is based on Constraint Logic Programming that enables autonomous agents to generate explanations and logic-based reasoning from a knowledge base and monitor how explanations advance over time. Whilst this preliminary work addresses key limitations such as scalability and adaptability, we strongly emphasise the need for logic-based reasoning in multi-agent systems for interpretability and transparency in their behaviour.

Patil, Minal Suresh
Umeå univeristet

Towards Explainable Agency in Multi-Agent Systems Using Constraint Programming

Minal Suresh Patil, Umeå universitet
Main Supervisor: Kary Främling

Motivation & Research Goals

Logical reasoning is a fundamental aspect of human behaviour, and this is an important criterion to build human-like reasoning in intelligent autonomous multi-agent systems. So far, the field of knowledge representation and reasoning has employed logic-based symbolic techniques to mimic the challenging task of incorporating human-like reasoning in multi-agent systems. However, the field of machine learning has shown increasing interest to take on this challenge. In this research, we describe a methodology that is based on Constraint Logic Programming that enables autonomous agents to generate explanations and logic-based reasoning from a knowledge base and monitor how explanations advance over time. Whilst this preliminary work addresses key limitations such as scalability and adaptability, we strongly emphasise the need for logic-based reasoning in multi-agent systems for interpretability and transparency in their behaviour.

Method

The optimal linkage between explainability and causation, which is the cornerstone to effective human-agent explainability one of the most important ultimate goals of explainable AI systems. Explainability pertains to a system that has the ability to explain itself to others in a natural language. In other words, a system should be able to communicate the reasoning behind its decisions [1]. Cause and effect is crucial for making ethical decisions.

The logical way:

Logic can assist with technical aspects of the problems. As a consequence, we propose a Constraint Logic Programming (CLP) for machine explainability.. It's an approach that is indeed innately interpretable and it is easy to incorporate domain knowledge.

Results

Particularly compared to traditional techniques, CLP has a significant advantages:

- Because logic programs are expressive, CLP systems can understand intricate relational theories.
- Domain-Knowledge can also be used by CLP systems to learn domain specific constraints.
- Because domain knowledge is used as a type of inductive bias, CLP systems might generalize from a small number of examples.
- CLP systems are designed to promote continuous and transfer learning.

References

1. M.S. Patil. Explainability in Autonomous Pedagogically Structured Scenarios. In Proceedings of Workshop on Explainable Agency in Artificial Intelligence at 36th Association for the Advancement of Artificial Intelligence (AAAI), 2022

Paul, Sudipta
Umeå University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

A Subspace matched framework for Federated Learning

In the setup of Federated Learning[1], and more particularly, when using the FedAvg algorithm, all the weights of the locally trained model get averaged in each round. In this setup, the global model is being sent to the ready local devices and they, later, send back the locally trained model to the central server for the next averaging step. Though this setup claims to ensure the security and privacy of the local data at the devices, some research has shown that this algorithm is still prone to membership attack[2], model reconstruction attack and backdoor attack. Our approach introduces a new federated learning framework T\'opos-FL on the basis of subspace and correlation analysis upon the layers of the machine learning models. It mitigates several drawbacks of FedAvg. In particular, model and data reconstruction attacks, and membership attack. In our approach, a conjugated view of the layers is being transferred to the central server where the update is subject to maximizing the correlation between the global and the local models.

Paul, Sudipta
Umeå University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

A Subspace matched framework for Federated Learning

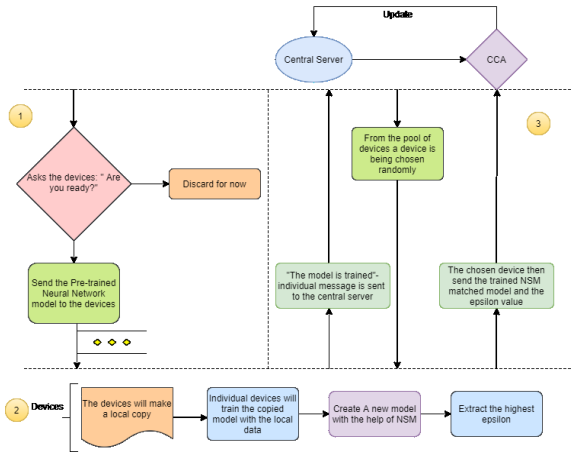
Sudipta Paul, PhD Student, Umeå University
Department of Computing Science
Supervisors: Prof. Vicenç Torra and Lili Jang (UMU)

UMEÅ UNIVERSITET

Motivation & Research Goals

In the setup of Federated Learning[1], and more particularly, when using the FedAvg algorithm, all the weights of the locally trained model get averaged in each round. In this setup, the global model is being sent to the ready local devices and they, later, send back the locally trained model to the central server for the next averaging step. Though this setup claims to ensure the security and privacy of the local data at the devices, some research has shown that this algorithm is still prone to membership attack[2], model reconstruction attack and backdoor attack. Our approach introduces a new federated learning framework Tópos-FL on the basis of subspace and correlation analysis upon the layers of the machine learning models. It mitigates several drawbacks of FedAvg. In particular, model and data reconstruction attacks, and membership attack. In our approach, a conjugated view of the layers is being transferred to the central server where the update is subject to maximizing the correlation between the global and the local models.

Methods



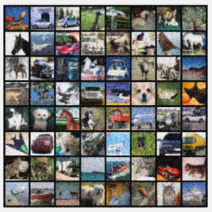
Algorithm 3: Tópos-FL.
The K clients are indexed by 1, 2, 3, 4, ..., k, ...
b is the batch size, η is the learning rate, e is the number of epoch, DB_k is the data at device k

At the Server:
initialization w_c;
This is the model at the central server
for *each round* t = 1,2,3, ... **do**
  $D_t \leftarrow$ set of ready devices;
  for *each client device* D_t **do**
    Receive $< w_n, \epsilon^* >$ from the client ;
    $w_c \leftarrow w_c + \eta SVCCA(w_c, < w_n, \epsilon^* >)$;
  end
  Send w_c to clients ;
end
At the Clients: *(Run on each client device in parallel)*
$w_c' \leftarrow w_c$ *(Make a copy of the central model)*;
for *each local epoch* $\in e$ **do**
  for *batch size* $\in b$ **do**
    w_c' is being trained by DB_k
  end
end
$< w_n, \epsilon^* > \leftarrow NSM(w_c, w_c')$;
Send to the server $< w_n, \epsilon^* >$;

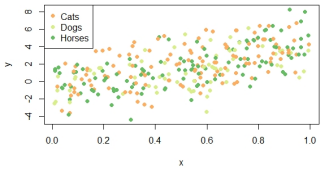
Selected Results

Result of the reconstruction attack in the Tópos-FL setup for CIFAR-10

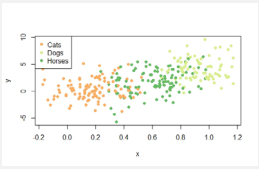
Result of the reconstruction attack in the FedAVG setup for CIFAR-10



Result of the membership attack in the Tópos-FL setup for CIFAR-10



Result of the membership attack in the FedAVG setup for CIFAR-10



References

[1] Advances and Open Problems in Federated Learning
Peter Kairouz et al.
arXiv preprint arXiv:1912.04977

[2] Membership Inference Attacks Against Machine Learning Models
Shokri, Reza ; Stronati, Marco; Song, Congzheng and Shmatikov, Vitaly
2017 IEEE Symposium on Security and Privacy (SP)

Pellaco, Lissy
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Deep Weighted MMSE Downlink Beamforming

The weighted minimum mean square error (WMMSE) algorithm was proposed to provide a locally optimum solution to the otherwise NP-hard weighted sum rate maximization beamforming problem, but it is still prohibitively complex. With the success of deep unfolding in trading off complexity and performance, we propose to apply deep unfolding to the WMMSE algorithm. With respect to traditional end-to-end learning, deep unfolding incorporates expert knowledge, with the benefits of well-grounded architecture selection, fewer trainable parameters, and better explainability. However, the classical formulation of the WMMSE algorithm given by Shi et al. is not amenable for deep unfolding due to matrix inversions, eigendecompositions, and bisection searches. Therefore, we present an alternative formulation that circumvents these operations. By means of simulations, we show that the deep unfolded WMMSE algorithm performs on par with the original WMMSE algorithm, at a lower computational load.

Pellaco, Lissy
KTH

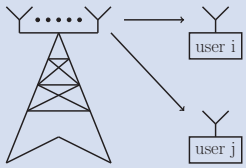
Deep Weighted MMSE Downlink Beamforming

Lissy Pellaco, Mats Bengtsson, Joakim Jaldén
pellaco@kth.se

https://github.com/lpkg/WMMSE-deep-unfolding/tree/ICASSP2021

Problem Formulation

- MU-MISO interference downlink channel
- Single base station with M transmit antennas
- N single-antenna users
- Linear beamforming



Received signal of user i

$$y_i = \boldsymbol{h}_i^H \boldsymbol{v}_i x_i + \sum_{j=1, j\neq i}^{N} \boldsymbol{h}_i^H \boldsymbol{v}_j x_j + n_i$$

$\boldsymbol{h}_i \in \mathbb{C}^M$: channel, n_i: noise
$\boldsymbol{v}_i \in \mathbb{C}^M$: beamformer vector
x_i: transmitted symbol
$\boldsymbol{V} \triangleq [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N]^T \in \mathbb{C}^{N\times M}$

- We address the **NP-hard** problem

$$\max_{\boldsymbol{V}} \quad \sum_{i=1}^{N} \alpha_i \log_2(1+\mathrm{SINR}_i) \tag{1a}$$
$$\text{s.t.} \quad \mathrm{Tr}(\boldsymbol{V}\boldsymbol{V}^H) \leq P \tag{1b}$$

  - $\log_2(1+\mathrm{SINR}_i)$ is the rate of user i
  - α_i is the priority of user i (assumed to be known)
- We want to comply with the **power consumption** and **latency constraints** at the base station

WMMSE algorithm

- It works on an **equivalent reformulation**

$$\min_{\boldsymbol{u},\boldsymbol{w},\boldsymbol{V}} \quad f(\boldsymbol{u},\boldsymbol{w},\boldsymbol{V}) \tag{2a}$$
$$\text{s.t.} \quad \mathrm{Tr}(\boldsymbol{V}\boldsymbol{V}^H) \leq P \tag{2b}$$

- f is jointly nonconvex over $(\boldsymbol{u},\boldsymbol{w},\boldsymbol{V})$
- f is **convex in each optimization variable**

ITERATIVE ALGORITHM PSEUDOCODE

repeat
$\boldsymbol{u} = \mathrm{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{\xi},\boldsymbol{w},\boldsymbol{V})$
$\boldsymbol{w} = \mathrm{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{u},\boldsymbol{\xi},\boldsymbol{V})$
$\boldsymbol{V} = \mathrm{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{u},\boldsymbol{w},\boldsymbol{\xi})$ s.t. $\mathrm{Tr}(\boldsymbol{\xi}\boldsymbol{\xi}^H) \leq P$
until convergence

- Guaranteed to **converge to a local optimum**
- **Relatively high computational complexity**

Deep unfolding

- **Goal**: trade off complexity and performance in presence of computational constraints and latency constraints for iterative algorithms
- **Key idea**: build and train a neural network whose structure is determined by the iterative algorithm
- It **incorporates domain knowledge**

Making the WMMSE unfoldable

- The update equations of $\boldsymbol{u}$ and $\boldsymbol{w}$ can be **easily mapped to neural network layers**
- Conversely, the update of $\boldsymbol{V}$, which is obtained by solving

$$\min_{\boldsymbol{\xi}} \quad f(\boldsymbol{u},\boldsymbol{w},\boldsymbol{\xi}) \tag{3a}$$
$$\text{s.t.} \quad \mathrm{Tr}(\boldsymbol{\xi}\boldsymbol{\xi}^H) \leq P, \tag{3b}$$

  requires **a matrix inversion, an eigendecomposition, and a bisection search** ☹
- We propose to solve (3) with the **projected gradient descent (PGD)** approach ☺
- We **truncate** the sequence of PGD steps to K

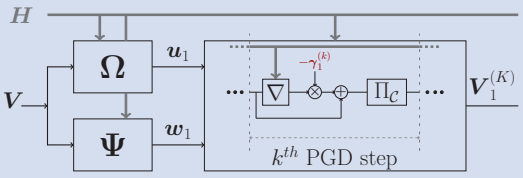
UNFOLDABLE WMMSE ALGORITHM PSEUDOCODE

for $l = 1, \ldots, L$
$\boldsymbol{u} = \mathrm{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{\xi},\boldsymbol{w},\boldsymbol{V})$
$\boldsymbol{w} = \mathrm{argmin}_{\boldsymbol{\xi}} f(\boldsymbol{u},\boldsymbol{\xi},\boldsymbol{V})$
for $k = 1, \ldots, K$
$\boldsymbol{V}^{(k)} = \Pi_{\mathcal{C}}\{\boldsymbol{V}^{(k-1)} - \gamma \nabla f(\boldsymbol{V}^{(k-1)})\}$

$$\Pi_{\mathcal{C}}\{\boldsymbol{V}\} = \begin{cases} \boldsymbol{V}, & \text{if } \mathrm{Tr}(\boldsymbol{V}\boldsymbol{V}^H) \leq P \\ \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|}\sqrt{P}, & \text{otherwise.} \end{cases} \tag{4}$$

We prove that the unfoldable WMMSE retains the **same convergence guarantees** of the WMMSE

Deep unfolded WMMSE

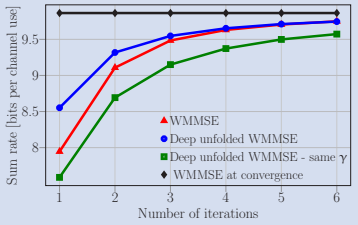


- The step sizes of the PGD ($\boldsymbol{\Gamma}$) are the trainable parameters, where $\boldsymbol{\Gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_L)$
- We minimize the **global objective** loss function

$$\mathcal{L}(\boldsymbol{\Gamma}) = -\frac{1}{N_s}\sum_{n=1}^{N_s}\sum_{l=1}^{L} f_{\mathrm{WSR}}(\boldsymbol{H}_n, \boldsymbol{V}_l; \boldsymbol{\Gamma}) \tag{5}$$

  where N_s is the size of the training set

Numerical Results



- $\boldsymbol{h}$ is drawn from a Rayleigh distribution
- We initialize $\boldsymbol{V}$ with matched filtering
- We initialize $\gamma_l^{(k)} = 1$ for $k = 1, \ldots, K$ and $l = 1, \ldots, L$
- We set $\alpha_i = 1$ for $i = 1, \ldots, N$
- For each combination of M, L, K, and SNR we train a different network
- We adopt Adam optimizer

Persson, Patrik
Lund University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Parameterization of Ambiguity in Monocular Depth Prediction

Monocular depth estimation is a highly challenging problem that is often addressed with deep neural networks.
While these use recognition of high level image features to predict reasonably looking depth maps the result often has poor metric accuracy.
Moreover, the standard feed forward architecture does not allow modification of the prediction based on cues other than the image.
In this paper we relax the monocular depth estimation task by proposing a network that allows us to complement image features with a set of auxiliary variables. These allow disambiguation when image features are not enough to accurately pinpoint the exact depth map and can be thought of as a low dimensional parameterization of the surfaces that are reasonable monocular predictions.
By searching the parameterization we can combine monocular estimation with traditional photo-consistency or geometry based methods to achieve both visually appealing and metrically accurate surface estimations.
Since we relax the problem we are able to work with smaller networks than current architectures.
In addition we design a self supervised training scheme, eliminating the need for ground truth image-depth-map pairs.
Our experimental evaluation shows that our method generates more accurate depth maps and generalizes better than competing state-of-the-art approaches.

Persson, Patrik
Lund University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Parameterization of Ambiguity in Monocular Depth Prediction

Patrik Perssion
Lund University

Introduction

Dense depth or disparity estimation is a classical problem in computer vision . Traditional methods use stereo (or multi-camera) setups and attempt to match every pixel in the reference image to a corresponding pixel in a neighbouring image using appearance cues. While the accuracy of the recovered depth is often very high for correctly matched pixels, ambiguous texture can degrade the matching and often leads to a noisy depth map. To stabilize the result a popular approach is to add geometric regularization terms such as derivative or curvature penalties. These can be realized as low order potentials in a conditional random field and efficient inference can be performed with move-making or message passing algorithms. While this kind of prior can drastically improve the estimation in ambiguous image regions, they lack any ability to recognize complex geometries and are basically limited to encouraging piece-wise planar or smooth surfaces.

A more recent approach is to use neural networks to directly infer depth or dense matching. An extreme case of this is monocular depth estimation where a neural network is used to estimate depth from a single image [3]. These networks typically require a huge amount of training data and may generalize poorly. In addition, while they achieve meaningful results with plausible object shapes, the resulting depth maps are often inaccurate because of the ambiguous nature of the problem. To resolve these ambiguities, CodeSlam [1] and DeepFactors [2] introduce an image dependent low-dimensional latent scene representation by training a conditional variational auto-encoder. Given an image, a depth basis is predicted and is linearly combined with the latent representation to form the depth. Here both methods are trained supervised using ground truth depth maps, limiting the datasets that can be used for training.

Our approach

Our goal is to use a neural network to extract a low dimensional shape parameterization from a single image that is flexible enough to allow depth fitting using traditional stereo cues such as photo consistency. To achieve this we add an additional input, the latent variable z, to a U-net architecture designed to complement the model with the information that is not directly observed in the image. Our latent variable model is not trained to recreate a depth map through an auto-encoder, as in CodeSlam or DeepFactors, but rather to complement image features with information needed to predict the depth map. A similar approach was introduced in DeepSDF [4] to learn signed distance functions.

Fig. 1: U-Net complemented with the latent vector z

During training and inference, we treat the latent vector z as a parameter to find during optimization instead of predicting it using an encoder. This allows us to train the network and performed inference in essentially the same way.

Training and Inference

During training and inference we use photometric and geometric consistency losses between overlapping views to constrain the depth instead of using ground truth depth maps. This makes the training self-supervised. For each view, we associate a latent vector z and mean depth α, and optimize these jointly for all images in the co-visible set giving us a depth in each view simultaneously. The losses are described below

$$X_i(p) = \pi^{-1}(D_i(p)) \tag{1}$$
$$q_{ij}(p) = \pi\left(T_{ij}X_i(p)\right) \tag{2}$$
$$\mathcal{L}_{\text{photo}_i} = \lambda_p \sum_{j \neq i} \|M_{ij} \odot (I_j \circ q_{ij} - I_i)\|_\delta \tag{3}$$
$$\mathcal{L}_{\text{depth}_i} = \lambda_d \sum_{j \neq i} \|\frac{1}{\alpha_j} M_{ij} \odot (D_j \circ q_{ij} - D_{ij})\|_\delta, \tag{4}$$

where π, π^{-1} are projection and un-projection operations, T_{ij} the relative transformation, M_{ij} a mask that removes invalid projections and $\|\cdot\|_\delta$ the huber norm. The only difference between training and inference is that during inference the network weights are not updated.

Additionally we explicitly handle occlusion by checking the difference between the depth D_j in view j with projected depth D_{ij} originating form D_i, using an adaptive threshold

$$\Delta_{ij}(p) \leq \textbf{Median}(\Delta_{ij}) - \tau \cdot \textbf{MAD}(\Delta_{ij}), \tag{5}$$

where $\Delta_{ij} = D_j - D_{ij}$. A point is classified as occluded if the condition is true.

Results

The Table below shows comparative results between DeepFactors, MegaDepth and Our method, where it can be seen that our method yield significantly better results.

		lower is better			higher is better		
scene	Method	Abs Rel	Sq Rel	RMSE	$\delta < 1.1$	$\delta < 1.25$	$\delta < 1.25^2$
scene0565_00	DeepFactors	0.1517	0.0693	0.3638	46.08%	76.68%	95.67%
	MegaDepth	0.2749	0.2879	0.6672	36.46%	62.34%	83.20%
	Ours	**0.0980**	**0.0492**	**0.3036**	**67.31%**	**88.70%**	**97.83%**
scene0606_02	DeepFactors	0.1736	0.1546	0.5799	44.96%	73.61%	91.01%
	MegaDepth	0.2312	0.1998	0.5959	35.70%	63.30%	87.39%
	Ours	**0.1232**	**0.0804**	**0.3989**	**61.62%**	**83.73%**	**95.32%**
scene0707_00	DeepFactors	0.1669	0.0913	0.3771	43.02%	73.62%	94.70%
	MegaDepth	0.2452	0.2226	0.5443	33.36%	61.90%	86.73%
	Ours	**0.0883**	**0.0365**	**0.2495**	**71.59%**	**91.31%**	**97.94%**
scene0715_00	DeepFactors	0.0959	0.0653	0.4599	64.16%	90.03%	97.74%
	MegaDepth	0.2291	0.4771	0.9298	45.54%	73.12%	89.03%
	Ours	**0.0672**	**0.0471**	**0.3724**	**80.20%**	**94.26%**	**98.20%**
scene0743_00	DeepFactors	0.1537	0.0549	0.23	45.61%	78.26%	95.70%
	MegaDepth	0.2111	0.1236	0.3570	39.86%	70.00%	89.74%
	Ours	**0.0779**	**0.0020**	**0.1709**	**74.52%**	**92.8%**	**98.70%**

One key difference between our method and CodeSlam and DeepFactors is that in our method, the depth is non-linear in z and each component of z has a local impact on the depth map in contrast to the aforementioned. We hypothesise that this is the reason for the significant performance increase since it may allow larger variations in depth and a slightly lower regularizing effect which may allow it to better adapt to new scenes while still regularizing local structure.

Examples

Conclusion

In this work, we have presented a learning approach for monocular depth estimation that takes ambiguities into account by providing a low dimensional parameterization of a family of feasible depth maps. We have shown that optimizing over this representation using photo-consistency losses yields accurate and realistic geometries. Our experimental results indicate, both qualitative and quantitative, that our approach performs better or on par with competing state-of-the-art methods.

Acknowledgements

This work was partially supported by the strategic research projects ELLIIT and eSSENCE, the Swedish Foundation for Strategic Research project, Semantic Mapping and Visual Navigation for Smart Robots (grant no. RIT15-0038) and Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation.

References

[1] M. Bloesch et al. "CodeSLAM - Learning a Compact, Optimisable Representation for Dense Visual SLAM". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 2560–2568. DOI: 10.1109/CVPR.2018.00271.

[2] J Czarnowski et al. "DeepFactors: Real-time probabilistic dense monocular SLAM". In: *IEEE Robotics and Automation Letters* 5 (2020), pp. 721–728. DOI: 10.1109/lra.2020.2965415. URL: http://dx.doi.org/10.1109/lra.2020.2965415.

[3] Zhengqi Li and Noah Snavely. "MegaDepth: Learning Single-View Depth Prediction from Internet Photos". In: *Computer Vision and Pattern Recognition (CVPR)*. 2018.

[4] Jeong Park et al. "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 165–174. DOI: 10.1109/CVPR.2019.00025.

Poceviciute, Milda
Linköping University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Unsupervised anomaly detection in digital pathology

Machine learning (ML) algorithms are optimized
for the distribution represented by the training data. For outlier data, they often deliver predictions with equal confidence, even though these should not be trusted. In order to deploy ML-based digital pathology solutions in clinical practice, effective methods for detecting anomalous data are crucial to avoid incorrect decisions in the outlier scenario. We propose a new unsupervised learning approach for anomaly detection in histopathology data based on generative adversarial networks (GANs). Compared to the existing GAN-based methods that have been used in medical imaging, the proposed approach improves significantly on performance for pathology data.
Our results indicate that histopathology imagery is substantially more complex than the data targeted by the previous methods. This complexity requires not only a more advanced GAN architecture but also an appropriate anomaly metric to capture the quality of the reconstructed images.

Poceviciute, Milda
Linköping University

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Unsupervised anomaly detection in digital pathology

Milda Pocevičiūtė, Gabriel Eilertsen, Claes Lundström
MIT, Linköpings Universitet

LINKÖPINGS UNIVERSITET

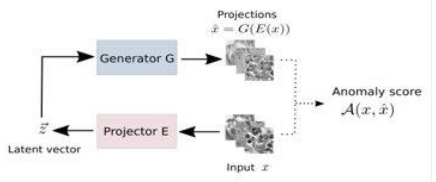
Introduction

- Anomaly detection is crucial for safe deployment of digital pathology methods to clinical practice.
- Outlier data is unknown or unavailable at the training time, hence unsupervised detection methods are required.
- GANs learn the distribution of training data, so they are expected to fail to reproduce realistic images of anomalous data
- In our experiments we use healthy patches for training and tumour patches as anomaly data to be detected.

Contributions

•**s2-AnoGAN**: unsupervised StyleGAN2 [1] based anomaly detection method tailored for digital pathology data
•Improved performance compared to two previous methods: f-AnoGAN [2] and pg-AnoGAN [3].

Method



General method for unsupervised anomaly detection with GANs:
•Projector (images -> latent representations)
•Generator (latent representations -> images)
•Anomaly score (comparison of image and its reconstruction)
The generator and gradient descent based projector from StyleGAN2 [1] are used in our method while anomaly score is based on Canny edges [4]: the difference between number of edges in the original images versus in the reconstructed image.
The lower the anomaly score, the better was the reconstruction which implies lower chance the image is an anomaly.
Area under ROC curve (AUC) is used to measure the success of detecting the anomalous images.

References

1 **Karras T et al.** Analyzingand improving the image quality of stylegan. *CVPR 2020*
2 **Schlegl T et al.** f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis 2019; 54.*
3 **Berg A et al.** Unsupervised adversarial learning of anomaly detection in the wild. *ECAI 2020*
4 **Canny J**. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence 1986; 6*

Results

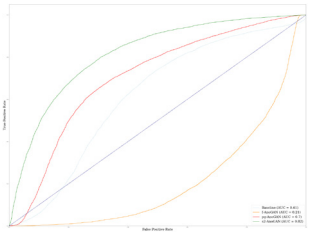


Fig. 1. ROC curves with the corresponding AUC achieved by the tested frameworks: s2-AnoGAN (our), f-AnoGAN, pg-AnoGAN and the baseline (for which only the number of edges is used as anomaly score). Our method had highest performance.

- **s2-AnoGAN** achieved the highest area under ROC curve (AUC) (see Fig. 1)
- As a baseline, we use the count of Canny edges within original images (no GANs' reconstructions involved)
- Poor f-AnoGAN performance is explained by MSE usage in the anomaly metric:
  - Pixel-wise comparison fails as the exact locations of the reconstructed features varies.
- In Fig. 2, we see that our method provides best visual reconstruction for healthy patches. All methods struggle to reconstruct the tumour patches:
  - Implies that more advanced GAN architecture is needed for digital pathology data.
- Table 1 shows AUC scores achieved by different combinations of the GAN + Projector and the anomaly scores

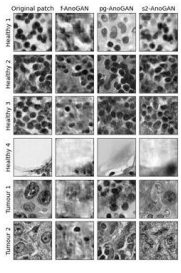


Fig. 2. Examples of test data and their projections by f-AnoGAN, pg-AnoGAN and s2-AnoGAN frameworks.

	f-AnoGAN	pg-AnoGAN	s2-AnoGAN
$\mathcal{A}_{f\text{-}AnoGAN}$	0.21	0.32	**0.37**
$\mathcal{A}_{pg\text{-}AnoGAN}$	**0.74**	0.70	0.71
$\mathcal{A}_{Canny}$	0.75	0.57	**0.82**
$LPIPS$	0.36	0.36	**0.64**
$\mathcal{A}_D$	0.24	0.35	**0.36**
$\mathcal{A}_{MSE}$	0.21	0.32	**0.37**
$\mathcal{A}_{res}$	0.19	**0.33**	0.27
$\mathcal{A}_0(z)$	**0.75**	0.70	0.71

Table 1. AUC values of anomaly detection when combined with different anomaly metrics.

Rahbar, Arman
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Do kernel and neural embeddings improve optimization and generalization?

We extend the recent results of (Arora et al. 2019) by spectral analysis of the representations corresponding to the kernel and neural embeddings. They showed that in a simple single-layer network, the alignment of the labels to the eigenvectors of the corresponding Gram matrix determines both the convergence of the optimization during training as well as the generalization properties. We generalize their result to the kernel and neural representations and show these extensions improve both optimization and generalization of the basic setup studied in (Arora et al. 2019). In particular, we first extend the setup with the Gaussian kernel and the approximations by random Fourier features as well as with the embeddings produced by two-layer networks trained on different tasks. We then study the use of more sophisticated kernels and embeddings, those designed optimally for deep neural networks and those developed for the classification task of interest given the data and the training labels, independent of any specific classification model.

Rahbar, Arman
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Do kernel and neural embeddings improve optimization and generalization?

Arman Rahbar, Emilio Jorge, Devdatt Dubhashi, Morteza Haghir Chehreghani
{armanr, emilio.jorge, dubhashi, morteza.chehreghani}@chalmers.se

CHALMERS UNIVERSITY OF TECHNOLOGY

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Description

The authors of [1] has showed that in a simple single-layer network, the alignment of the labels to the eigenvectors of the corresponding Gram matrix determines both the convergence of the optimization during training as well as the generalization properties. We thought about generalizing their results to other representations of the data. Specifically, we worked with representations induced by different Kernels as well as Neural Networks.
In particular, we extend the setup in [1] with Kernels and approximations of Kernel embeddings as well as with the embeddings produced by two-layer neural networks.

Background & Motivation

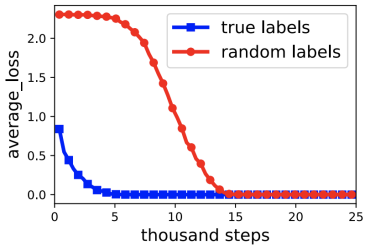


Figure 1: Rethinking Generalization Experiment. Source: ICML slides for [1]

Rethinking Generalization Experiment [3]: Gradient Descent for a neural network reaches nearly 0 training loss for both correct and random labels. **BUT** we see better generalization and faster convergence for correct labels. So we have two fundamental questions to answer:

Optimization. Why do true labels give faster convergence rate than random labels for gradient descent?

Generalization. What property of properly labeled data controls generalization?
In [1], the authors consider a specific simple two-layer network model:

$$f_{W,a}(x) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \max(0, w_r^T x_i), \quad (1)$$

with $x \in R^d$, $w_1, ..., w_m \in R^{d \times m}$ and $(a_1, ..., a_m)^T \in R^m$ (where m specifies the number of neurons in the hidden layer, i.e., its width). This network is trained on dataset of data points $\{x_i\}$ and their targets $\{y_i\}$.
They show that both training and generalization are better if the label vector y aligns with the eigenvectors corresponding to the top eigenvalues of $\mathbf{H}^\infty$ (Gram Matrix) where

$$\mathbf{H}^\infty_{i,j} := E_{\mathbf{W} \sim \mathcal{N}(0,\mathcal{I})} \left[\mathbf{x}_i^T \mathbf{x}_j 1[\mathbf{w}^T \mathbf{x}_i \geq 0, \mathbf{w}^T \mathbf{x}_j \geq 0] \right] \quad (2)$$

and

$$\mathbf{H}^\infty = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^T \quad (3)$$

is the orthonormal decomposition of $\mathbf{H}^\infty$.

Our Method

The simple two-layer network can be extended with adding different types of embeddings ϕ at the input layer corresponding to a kernel $\mathcal{K}$:

$$f_{\mathbf{W},\mathbf{a}}(\mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \max(0, \mathbf{w}_r^T \phi(\mathbf{x}_i)). \quad (4)$$

Then the Gram Matrix can be defined in the same way as before. Let its eigenvalues be ordered as $\lambda_0(\mathcal{K}) \geq \lambda_1(\mathcal{K}) \geq \cdots \geq \lambda_{n-1}(\mathcal{K})$ and let $\mathbf{v}_0(\mathcal{K}), \cdots, \mathbf{v}_{n-1}(\mathcal{K})$ be the corresponding eigenvectors. We can also view the representations generated by successive neural network layers as a type of embedding which helps us to understand what happens in deeper networks.

A kernel $\mathcal{K}$ such that the corresponding eigenvectors align well with the labels would be expected to perform well both for training optimization as well as generalization.
For our kernelized network the optimization and generalization are respectively controlled by:

$$\sqrt{\sum_i (1 - \eta \lambda_i(\mathcal{K}))^{2k} (\mathbf{v}(\mathcal{K})_i^T \mathbf{y})^2} \quad (5)$$

and

$$\mathbf{y}^T (\mathbf{H}(\mathcal{K})^\infty)^{-1} \mathbf{y} \quad (6)$$

Our Kernels

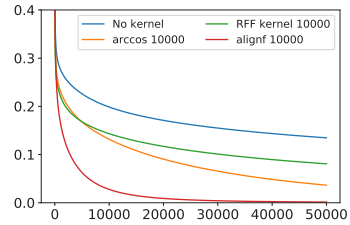
We have used the following kernels in our experiments:
1. Gaussian Kernel: The Gaussian kernel is given by $\mathcal{K}(x_i, x_j) := \exp\left(-\gamma \|x_i - x_j\|^2\right)$.
2. Neural Kernel: By adding another layer to the network and training it to convergence we can then use the weights of the first layer as a kernel embedding for training a new network.
2. Arc-cosine Kernel: This kernel mimics the computations in a neural network within an infinite dimensional feature space.
3. Optimized Kernel: We use the method proposed in [2] that suggests an algorithm to learn a new kernel from a group of kernels based on a similarity measure between the kernels, namely centered alignment. The learned kernel is expected to align well with the labels.

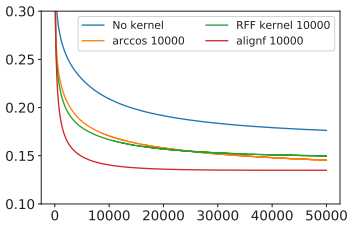
How To Find Kernel Embeddings?

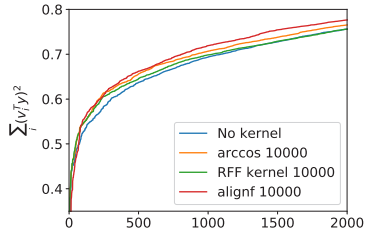
We can use existing methods for approximation of kernel embeddings. Specifically, we used *random Fourier features* (RFF) for Gaussian Kernel, and *Nyström method* for other Kernels.

Some of the Results

Below we can see the experimental train and test errors at the different steps of Gradient Descent as well as eigenvector projections for the CIFAR-10 dataset for various Kernels:







Bibliography

References

[1] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332, 2019.

[2] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *J. Mach. Learn. Res.*, 13:795–828, March 2012.

[3] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ArXiv*, abs/1611.03530, 2016.

Sanchez Aimar, Emanuel
Linköping University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Deep Expert Learning for Long-Tailed Recognition

Many real-world recognition problems present a highly imbalanced or long-tailed label distribution. This setting makes representation learning more challenging and tends to bias classifiers towards head classes, resulting in limited generalization for tail classes. By simultaneously addressing these issues, ensemble learning has shown promising results for long-tailed classification, exhibiting a good trade-off between head and tail performance. The aim of this research is to further study methods for learning and integration of multiple specialized models, called experts, to solve long-tailed visual problems. In addition, we investigate the calibration properties of the proposed expert-based model under long-tailed data regimes.

Sanchez Aimar, Emanuel
Linköping University

Deep Expert Learning for Long-Tailed Recognition

Emanuel Sanchez Aimar, Linköping University
Computer Vision Laboratory, ISY
Supervisors: Michael Felsberg, Marco Kuhlmann

Motivation & Research goals

Many real-world recognition problems present a highly imbalanced or long-tailed label distribution. This setting makes representation learning more challenging and tends to bias classifiers towards head classes, resulting in limited generalization for tail classes [1]. By simultaneously addressing these issues, ensemble learning has shown promising results for long-tailed classification, exhibiting a good trade-off between head and tail performance [2, 3]. The aim of this research is to further study methods for learning and integration of multiple specialized models, called *experts*, to solve long-tailed visual problems. In addition, we investigate the calibration properties of the proposed expert-based model under long-tailed data regimes.

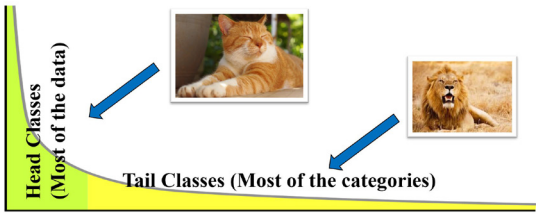


Illustration courtesy of [4].

Classical approaches & drawbacks

Data re-sampling
Loss re-weighting

→ Under-fitting to the head
Over-fitting to the tail
Sub-optimal for feature learning

(Class-dependent) margins

of ensemble learning

How to achieve diversity & specialization?
E.g., by data division

How to combine experts effectively?
E.g., by averaging

Framework

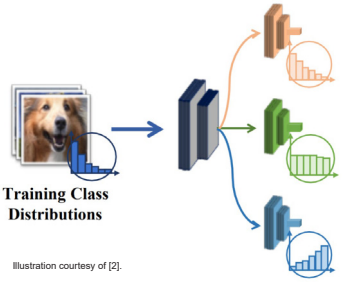


Illustration courtesy of [2].

- Stage 1: **train experts independently**, by simulating different class distributions
- Stage 2: **combine experts in an ensemble** (logit space)

1. *How to improve stage 1?* We propose a new parameterization for the expert losses inspired on [3]
2. *A more flexible fusion strategy for stage 2?* We investigate a learnable gating model
3. Interesting property to look into: *model calibration*

References

1. Kang, Bingyi, et al. "Decoupling Representation and Classifier for Long-Tailed Recognition." *International Conference on Learning Representations* (2019).
2. Zhang, Yifan, et al. "Deep long-tailed learning: A survey." *arXiv preprint arXiv:2110.04596* (2021).
3. Zhang, Yifan, et al. "Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision." *arXiv preprint arXiv:2107.09249* (2021).
4. Tang, Kaihua, et al. "Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect." *Advances in Neural Information Processing Systems* 33 (2020).

Shahriari-mehr, Firooz
Chalmers

Machine Learning over Networks: Optimization point of view

One reason for the spectacular success of machine learning models can be the appearance of large datasets. Large datasets cannot be processed on a single machine due to resource limitations or slow training. Moreover, some datasets are private, and data sharing is prohibited. Distributed optimization methods can help use much more sensitive data points to train a machine learning model quickly because the dataset is stored on different nodes, and each node does the computations associated with a collection of data points available to it.

Shahriari-mehr, Firooz
Chalmers

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Machine Learning over Networks: From an optimization point of view

Firooz Shahriari-mehr, PhD student at Chalmers University of Technology
Deptartment of Computer Science and Engineering
Supervisors: Ashkan Panahi

CHALMERS UNIVERSITY OF TECHNOLOGY

Motivation

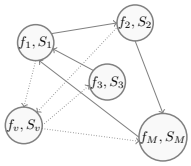
One reason for the spectacular success of machine learning models can be the appearance of large datasets. Large datasets cannot be processed on a single machine due to resource limitations or slow training. Moreover, some datasets are private, and data sharing is prohibited. Distributed optimization methods can help us to use much more sensitive data points to train a machine learning model quickly because the dataset is stored on different nodes, i.e., computational machines, each of them does the computations associated with a collection of data points available to it. In this poster, I will present a novel distributed optimization algorithm for the convex finite-sum minimization problem with explicit convex constraints over strongly-connected directed graphs.

Methods

Finite-sum minimization problem with explicit constraints:

$$\min_{\mathbf{x}\in\mathbb{R}^m} \frac{1}{M}\sum_{v=1}^{M} f_v(\mathbf{x}) \qquad \text{subject to} \quad \mathbf{x}\in\bigcap_{v=1}^{M} S_v$$

The setup that we have considered to solve this problem is a set of M nodes, each of them has access to its own local objective function and constraint set. The nodes communicate over a decentralized network represented by a directed graph.



Goal of the network: All nodes converge to a consensus solution which satisfies the following optimality condition:

$$\mathbf{0} \in \sum_{v=1}^{M} \left(\partial I_{S_v}(\mathbf{x}^*) + \nabla f_v(\mathbf{x}^*)\right)$$

Previous proposed algorithms[1] have assumed undirected graph, diminishing step-size, or identical constraints assumption for convergence analysis.

Double Averaging and Gradient Projection[2]**:** Each node v does the following updates in each iteration:

$$\begin{aligned}
\mathbf{z}^v &= \mathbf{x}_k^v - \sum_{u\in\mathcal{N}_v^{\text{in}}\cup\{v\}} w_{vu}\mathbf{x}_k^u - \mu\left(\nabla f_v(\mathbf{x}_k^v) - \mathbf{g}_k^v\right)\\
\mathbf{x}_{k+1}^v &= P_{S_v}(\mathbf{z}^v)\\
\mathbf{g}_{k+1}^v &= \mathbf{g}_k^v + \rho\left[\nabla f_v(\mathbf{x}_k^v) - \mathbf{g}_k^v + \frac{1}{\mu}\left(\mathbf{z}^v - \mathbf{x}_{k+1}^v\right)\right] + \alpha\left(\mathbf{h}_k^v - \mathbf{g}_k^v\right)\\
\mathbf{h}_{k+1}^v &= \mathbf{h}_k^v - \sum_{u\in\mathcal{N}_v^{\text{in}}\cup\{v\}} q_{vu}\left(\mathbf{h}_k^u - \mathbf{g}_k^u\right)
\end{aligned}$$

General concepts used in DAGP:

- Weighted averaging: each node receives some information from its neighbors and calculates a weighted average using gossip matrices.
- Projection: each node does a projection onto the local constraint set.
- Constrained gradient tracking: similar to the variance reduction techniques in stochastic optimization, the vectors called g^v store the past values of averaged (sub)gradients and feasible directions.
- Distributed null projection: to reach optimality, $\sum g^v = 0$ (null condition), which is achievable in a distributed way. To this end, we introduce the "distributed projection" of g^vs, called h^v, onto the space $\sum h^v = 0$.

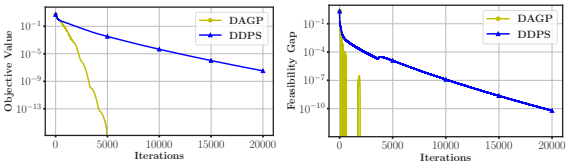
Selected Results

Theorem 1 (Consensus and Optimal Solution). *If the iterates of DAGP algorithm converge, any stopping point is an optimal and consensus solution, i.e., $\mathbf{x}^v = \mathbf{x}^*$, $\forall v \in \mathcal{V}$, and $\mathbf{x}^*$ satisfies the sufficient optimality conditions.*

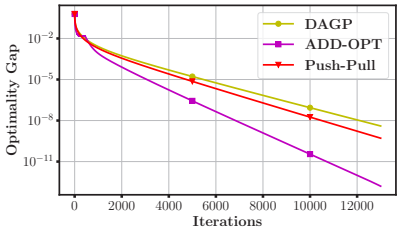
Theorem 2 (Convergence Rates). DAGP recovers the standard rates for convex and smooth objectives under some common assumptions without strong-convexity assumption. We showed feasibility and optimality gaps decay respectively with the rates of $1/K$ and $1/\sqrt{K}$, where K is the total number of iterations.

DAGP has superior empirical results in comparison to the DDPS[1] algorithm in the following constrained problem:

$$\begin{aligned}
f_v(\mathbf{x}) &= \log\left(\cosh(\mathbf{a}_v^T\mathbf{x} - b_v)\right), \qquad v = 1,\dots,20\\
S_v &= \{\mathbf{x}\in\mathbb{R}^{10} \mid \mathbf{c}_v^T\mathbf{x} - d_v \le 0\}. \qquad v = 1,\dots,20
\end{aligned}$$



Moreover, **DAGP** shows comparable performance with the state-of-the-art push-pull[3] algorithm in the logistic regression problem for two digits of the MNIST dataset, which is an unconstrained problem.



References

[1] Distributed Subgradient Projection Algorithm Over Directed Graphs, C. Xi and U. A. Khan, IEEE Transactions on Automatic Control, vol. 62, Aug. 2017.

[2] Decentralized Constrained Optimization: Double Averaging and Gradient Projection, F. Shahriari-Mehr, D. Bosch, A. Panahi, Conference on Decision and Control, Dec. 2021.

[3] Push-pull gradient methods for distributed optimization in networks, S. Pu, W. Shi, J. Xu, and A. Nedic, IEEE Transactions on Automatic Control, January, 2021.

Sidwall Thygesen, Signe
Linköping university

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Exploring Transition Ensembles using Hierarchical Clustering and Visual Representations

Ensemble analysis is a challenging problem, appearing in many scientific applications. This work introduces a pipeline to explore ensembles by combining automatic and interactive visual analysis, focusing on molecular electronic transition ensembles within the chemistry domain. An electronic transition describes the change in charge distribution between two molecular states. The goal of the pipeline is to characterize and compare the transitions, and how they correlate to physical properties. Each ensemble member is described with a quantitative feature vector, making it possible to utilize hierarchical clustering. A visual summarization for each cluster as well as the whole ensemble is proposed, building on the feature vector representation. Other interactive visual components are used, supporting both exploration of clusters and outlier detection, as well as investigation of correlation. The usefulness of the pipeline is shown by applying it to data from theoretical chemistry.

Sidwall Thygesen, Signe
Linköping university

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

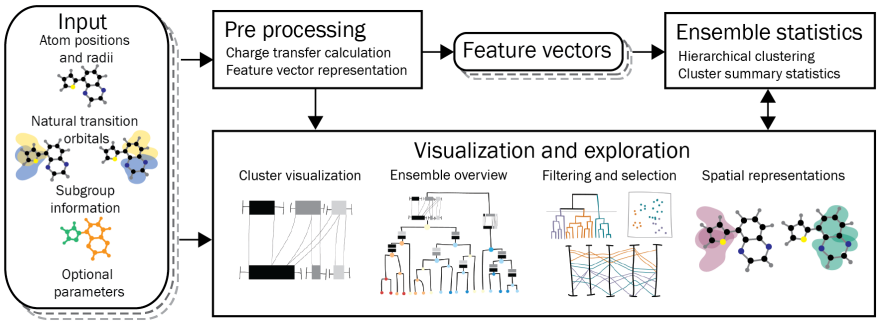
Exploring Transition Ensembles using Hierarchical Clustering and Visual Representations

LINKÖPING UNIVERSITY
Signe Sidwall Thygesen, Linköping University
Scientific Visualization, Media and Information Technology

Background & Motivation

Ensemble analysis is a challenging problem, appearing in many scientific applications. This work introduces a pipeline to explore ensembles by combining automatic and interactive visual analysis. It is applied it to molecular transition ensembles from theoretical chemistry, addressing the need to better explore such data. An electronic transition describes the change in charge distribution between two molecular states. The chemists are interested in finding out how much charge is transferred between different subgroups of the molecule. The purpose of the pipeline to characterize and compare transitions, and how they correlate to physical properties.

Pipeline



Each ensemble member is quantitatively described with a feature vector containing the charge in each subgroup of the molecule (building on a charge transfer matrix from previous work [1]). This quantitative description gives the possibility to utilize hierarchical clustering and statistical summarization methods. A measure of locality value (ML) is derived, capturing how much charge is transferred in the transition. The result is used in the visual representations. An automated dendrogram, augmented with cluster representations on different level of details gives an overview of the ensemble. Other interactive visual components are used, supporting both exploration of clusters and outlier detection, as well as investigation of correlation. Spatial representations of selected transitions links our abstract visualizations back to a view more familiar to the chemists.

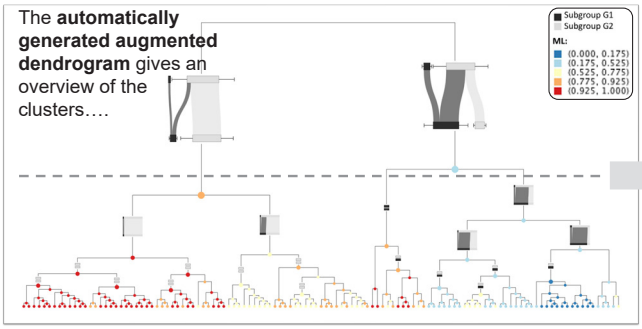
Contributions

- A novel **feature vector** to describe electronic transitions and a quantitative **measure of locality** for distinguishing between transitions of different nature
- A **visual pipeline** for ensemble analysis **combining automatic and explorative** methods
- A **level of detail representation** summarizing and conveying the mean behavior of a cluster
- Introduction of **augmented dendrograms** to provide a hierarchical visual representation of ensemble data

Example

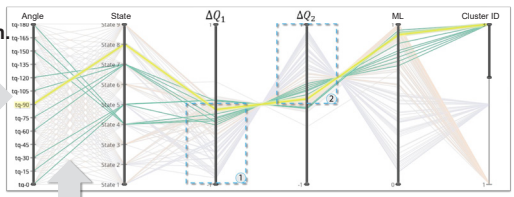
To illustrate the pipeline, it is applied to an electronic transition ensemble for a molecule consisting of two subgroups (G1 and G2).

The **automatically generated augmented dendrogram** gives an overview of the clusters….

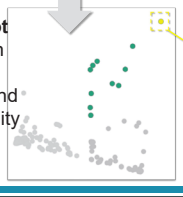


…which can be used to select a good level of detail in the **interactive exploration.**

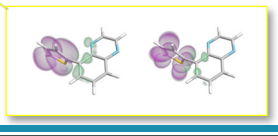
A **parallel coordinates plot** shows correlation between parameters, and makes it possible to filter and inspect ensemble members.



A linked **scatter plot** provides an additional overview and the possibility to select outliers.



The **spatial representations** for a selected transition helps the chemists to understand the data.



References

1. Bin Masood et al, *Visual Analysis of Electronic Densities and Transitions in Molecules,* Computer Graphics Forum, 2021.

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Stempfle, Lena
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Predicting progression & cognitive decline inamyloid-positive patients with Alzheimer's disease

In Alzheimer's disease, amyloid-β (Aβ) peptides aggregate in the brain forming CSF amyloid levels - a key pathological hallmark of the disease. However, CSF amyloid levels may also be present in cognitively unimpaired elderly individuals. We aim to explain the variance in disease progression among patients with Aβ-pathology. We perform prediction of a) the change of MMSE score using regression models for 2 and 4 years after follow-up and b) the change in diagnostic using classification model for 2 years after follow-up.
We show in our analysis that CSF levels of Aβ are not strong predictors of the rate of cognitive decline in Aβ-positive subjects when adjusting for other variables.

Stempfle, Lena
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Predicting progression & cognitive decline in amyloid-positive patients with Alzheimer's disease

Hákon Valur Dansson[1], Lena Stempfle[1]*, Hildur Egilsdóttir[1], Alexander Schliep[1], Erik Portelius[2,3], Kaj Blennow[2,3], Henrik Zetterberg[2,3,4,5], and Fredrik D. Johansson[1]

CHALMERS UNIVERSITY OF TECHNOLOGY

[1]CHALMERS, [2]GU, [3]SAHLGRENSKA UNIVERSITY HOSPITAL, [4]UCL NEURODEGENERATIVE DISEASES, [5]UCL DEMENTIA RESEARCH

GÖTEBORGS UNIVERSITET

Summary

In Alzheimer's disease, amyloid-β (Aβ) peptides aggregate in the brain forming CSF amyloid levels - a key pathological hallmark of the disease. However, CSF amyloid levels may also be present in cognitively unimpaired elderly individuals. We aim to explain the variance in disease progression among patients with Aβ-pathology. We perform prediction of a) the change of MMSE score using regression models for 2 and 4 year after follow-up and b) the change in diagnostic using classification model for 2 year after follow-up. We show in our analysis that CSF levels of Aβ are not strong predictors of the rate of cognitive decline in Aβ-positive subjects when adjusting for other variables.

Determination of amyloid-positive status

Even in individuals with Aβ pathology, there is substantial variation in symptoms, such as cognitive function, for this reason our work focuses on predicting progression in individuals with elevated Aβ CSF levels. Aβ pathology status was determined based on the Aβ42/Aβ40 ratio.

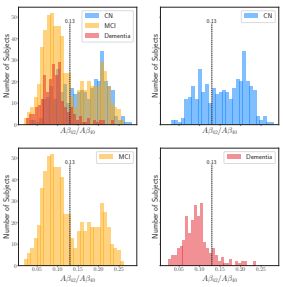


Figure 1: The full cohort was split into three groups: those who had a baseline Aβ ratio lower than 0.13 (Aβ-positive), those who had a higher ratio (Aβ-negative), and those with unknown status.

Potential predictors

Predictive models were built on *two* different sets of features. The first set of features *(all features)* was preselected following [48] and expanded to include key features from the ADNI TadPole competition [49] in addition to a few features that were available for over 90% of the ADNI cohort. This resulted in a set of 37 features including biomarkers tau, ptau and Aβ42 in CSF, and 15 different cognitive tests among others. The second feature set (cognitive tests only) consists only of the 15 cognitive tests.

Derivation and evaluation cohorts

Our experiments are performed on three different cohorts. To compensate for the small number of 1) only Aβ-positive subjects (*Aβ only*), were compared to training cohorts including 2) (*All Subjects*) combined Aβ-positive and Aβ-negative subjects and those without Aβ measurements into one derivation set and 3) (*All Subjects, Weighted*), with weighted samples with respect to the All Subjects cohort to mimic a larger sample of Aβ-positive subjects.

We let the latent state $C \in \{0,1\}$ of a Gaussian mixture model (GMM), fit to the Aβ-ratios of the All Subjects cohort, represent Aβ-positivity. The weight w_i was computed as:

$$w_i = \hat{p}(R = r_i | C = 1)/\hat{p}(\mathrm{R}= r_i)$$

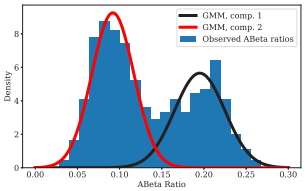


Figure 2: Each subject i was assigned a weight $w_i > 0$ based on the probability that their individual Aβ ratio r_i would be observed for an average hypothetical Aβ-positive subject

Prediction models and learning objectives

We studied the progression of Aβ-positive subjects with respect to two principal outcomes:

Task A: Predicting change in MMSE score for 2 (A1) and 4 (A2) years of follow-up

Change in MMSE score assessed as cognitive function [32], relative to baseline and 2 years. MMSE takes values on a scale from 0 to 30 where a lower score represents worse cognitive function [45].

We considered the performance of linear regression and gradient boosting models that predict the change in MMSE scores measured using the average cross-validated R^2 score and standard deviation.

Task B: Predicting change in diagnosis for 2 years (B1) follow-up

Changes in dementia diagnoses (CN/MCI/AD) were determined by comparing the disease status, indicating whether or not a subject's diagnosis had worsened in 2 year expressed by a binary variable.

For the classification task, a logistic regression model and the same tree-based gradient-boosting approach as for Task A was used and evaluated by cross-validated weighted F1 score.

Results

We perform a set of experiments using ADNI data to predict the change in MMSE after 2 and 4 year of follow-up and the change in diagnosis after 2 years.

The best predictive model of change in cognitive test scores for Aβ-positive subjects at the 2-year follow-up achieved an R^2 score of 0.388 while the best model predicting adverse changes in diagnosis achieved a weighted F1 score of 0.791. When predicting cognitive score change 4 years after baseline, the best model achieved an R^2 score of 0.325 and it was found that fitting models to the extended cohort improved performance. Moreover, using all clinical variables outperformed the best model based only on a suite of cognitive test scores which achieved an R^2 score of 0.228.

Table 1: Performance of the linear and gradient boosting regressions, predicting change in MMSE two and four years after baseline for three different cohort selections. We compare models trained on features a) the all features set from baseline and b) from baseline cognitive scores only.

R^2 (SD)	2-year follow-up	4-year follow-up
All features		
Linear regression, *Aβ Only*	0.372 (0.081)	0.205 (0.227)
Linear regression, *All Subjects*	0.354 (0.083)	0.325 (0.134)
Linear regression, *All Subjects, Weighted*	0.388 (0.073)	0.304 (0.152)
Gradient boosting, *Aβ Only*	0.287 (0.124)	0.156 (0.244)
Gradient boosting, *All Subjects*	0.356 (0.108)	0.252 (0.191)
Gradient boosting, *All Subjects, Weighted*	0.338 (0.950)	0.263 (0.192)
Cognitive tests only		
Linear regression, *Aβ Only*	0.343 (0.087)	0.178 (0.203)
Linear regression, *All Subjects*	0.333 (0.081)	0.228 (0.143)
Linear regression, *All Subjects, Weighted*	0.350 (0.079)	0.225 (0.160)
Gradient boosting, *Aβ Only*	0.272 (0.133)	-0.050 (0.358)
Gradient boosting, *All Subjects*	0.323 (0.118)	0.149 (0.224)
Gradient boosting, *All Subjects, Weighted*	0.293 (0.114)	0.118 (0.227)

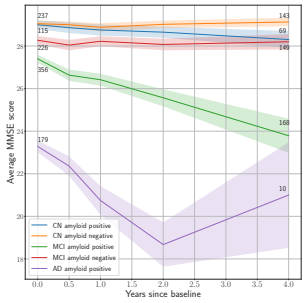


Figure 3: Aβ-positive subjects declined faster on average than those without Aβ pathology, but the specific level of CSF Aβ was not predictive of progression rate.

Conclusion

Our results illustrate high correlation between important predictors which offers future investigation to eventually handle the high missingness in the data. Baseline assessments of cognitive function accounts for the majority of variance explained in the prediction of 2-year decline but is insufficient for achieving optimal results in longer-term predictions.

Acknowledgements

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

References

[32] Galea, M., Woodward, M.: Mini-mental state examination (mmse).Australian Journal of Physiotherapy51(3), 198 (2005)

[45] Dick, J., Guiloff, R., Stewart, A., Blackstock, J., Bielawska, C., Paul,E., Marsden, C.: Mini-mental state examination in neurologicalpatients. Journal of Neurology, Neurosurgery & Psychiatry47(5),496–499 (1984)

[48] Nguyen, M., He, T., An, L., Alexander, D.C., Feng, J., Yeo, B.T.T.:Predicting alzheimer's disease progression using deep recurrent neuralnetworks. NeuroImage222, 117203 (2020).doi:10.1016/j.neuroimage.2020.11720349.

[49] Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W.,Weiner, M.W., Barkhof, F., Fox, N.C., Eshaghi, A., Toni, T., et al.:The alzheimer's disease prediction of longitudinal evolution (tadpole)challenge: Results after 1 year follow-up. arXiv preprintarXiv:2002.03419 (2020)

Sundqvist, Tobias
Umeå University / TietoEVRY

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Anomaly detection and root-cause analytics in 5G Radio Access Network

Our research objective is to improve the observability and speed up the fault finding process in the 5G Radio Access Network (RAN). We develop and utilize many different machine learning methods to analyze RAN system logs. From the logs we can learn and distinguish between the normal and abnormal behavior and aid the developers in locating the problems.

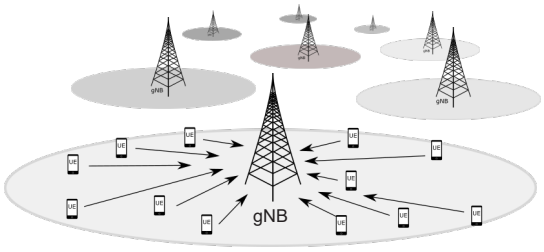
Sundqvist, Tobias
Umeå University / TietoEVRY

Anomaly detection and root-cause analytics in 5G Radio Access Network

We use AI to find faults quicker and more easily

Tobias Sundqvist, Monowar Bhuyan, Erik Elmroth, and Johan Forsman

5G RAN



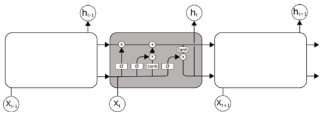
PROBLEM

It will be very challenging to make the 5G Radio Access Network (RAN) reliable due to its complexity:

- Increased network dynamics with different services
- Distributed system of radio units at antenna sites and applications in edge data center cites
- Separation of hardware and software from many vendors with individual configurations
- Virtualized RAN applications share resources with others on multi-purpose hardware
- Analysis of huge metrics from multiple sources
- Difficult to know if the system behaves normally

System logs

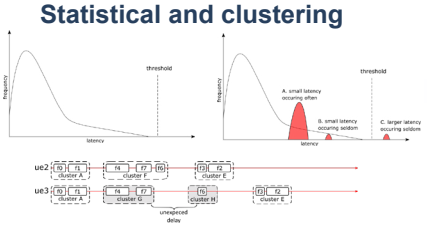
AdaBoost ensemble of LSTM



Detect changes in procedures

LSTM models learn the sequential order of procedures in the system log. Anomalies are detected as the order of events changes or abnormal events occur.

Statistical and clustering

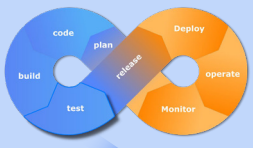


Detect abnormal delays

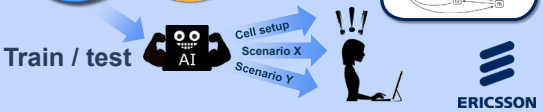
The methods provide a detailed view of where abnormal delays and large variations occur.

PROOF OF CONCEPT

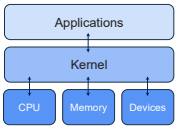
Continuous Integration



Methods are detecting anomalies in customers data. Speeds up fault localization process.

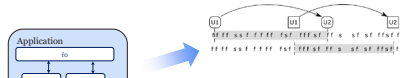


Root-cause detection



To fully understand the root cause of all software bugs in RAN we need to monitor both the function and kernel calls.

CallGraph



Learns what kind of function and kernel calls that occur between debug printouts in system log.

MultiSpace

Creates a call and mean time vector for each function.

- Find deviation in function call pattern.
- Find largest deviation in time for each function.

Scan me

And learn more about my research
http://www.qvistigt.com/

UMEÅ UNIVERSITET
Department of Computing Science, SE-901 87 Umeå, Sweden
Email: {sundqtob, monowar, elmroth}@cs.umu.se

Tabakovic, Selma
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Co-clustering of Tensor Data Using Sparse Tensor Factorisation

With the ever-increasing amounts of data generated from new sources and scientific methods, e.g. high throughput genome sequencing methods in bioinformatics, powerful tools for exploratory data analysis are required. One such tool is clustering, i.e. grouping together coherent observations in data, which is important for categorising vast amounts of observations into a more manageable format for further analysis. However, this task is subject to new challenges as tensor data, i.e. multidimensional data, has become a frequent occurrence in many applications. For tensor data, a clustering approach called co-clustering has recently attracted research attention. Co-clustering means that the clustering is performed on all the tensor dimensions simultaneously, which enables the detection of joint data expressions that only occur under special circumstances. Here a method for co-clustering of tensor data using a sparse CP decomposition is proposed.

Tabakovic, Selma
Chalmers

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Co-clustering of Tensor Data Using Sparse Tensor Factorisation

Selma Tabakovic, Chalmers University of Technology
Department of Applied Mathematics and Statistics
Supervisors: Rebecka Jörnsten (Chalmers and University of Gothenburg)

CHALMERS UNIVERSITY OF TECHNOLOGY

Motivation & Research Goals

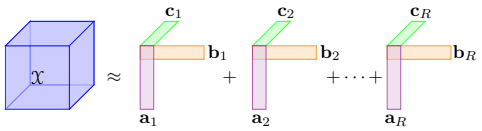
With the ever-increasing amounts of data generated from new sources and scientific methods, e.g. high throughput genome sequencing methods in bioinformatics, powerful tools for exploratory data analysis are required. One such tool is clustering, i.e. grouping together coherent observations in data, which is important for categorising vast amounts of observations into a more manageable format for further analysis. However, this task is subject to new challenges as tensor data, i.e. multidimensional data, has become a frequent occurrence in many applications. For tensor data, a clustering approach called co-clustering has recently attracted research attention. Co-clustering means that the clustering is performed on all the tensor dimensions simultaneously, which enables the detection of joint data expressions that only occur under special circumstances. Here a method for co-clustering of tensor data using a sparse CP decomposition is proposed.

Methods

The *CP decomposition* splits a tensor into a finite sum of rank one tensors, i.e. tensors that can be written as an outer product of N vectors Finding the CP decomposition of a three-way tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ may be formalised as

$$\min_{\hat{\mathcal{X}}} \|\mathcal{X} - \hat{\mathcal{X}}\| \quad \text{where} \quad \hat{\mathcal{X}} = \sum_{r=1}^{R} \lambda_r \, \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \,,$$

where R is a positive integer, λ_r is a scaling factor, and $\mathbf{a}_r \in \mathbb{R}^I$, $\mathbf{b}_r \in \mathbb{R}^J$ and $\mathbf{c}_r \in \mathbb{R}^K$ for $r = 1, 2 \ldots R$, and are all normalised to length one

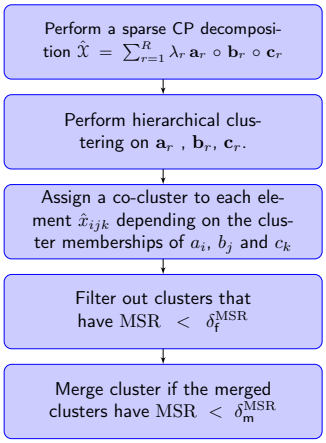


The sparse CP decomposition from [1] regularises the factors by a l_1 norm penalty and calculates them in a sequential manner.

The *mean square residue* (MSR) is used to measure the coherence of a co-cluster $\mathcal{D}$, and is defined as [2]

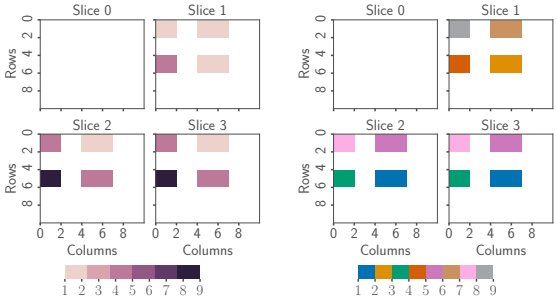
$$\mathrm{MSR}(\mathcal{D}) = \frac{1}{PQR} \sum_{p \in \mathcal{P}, q \in \mathcal{Q}, r \in \mathcal{R}} (d_{pqr} - d_{p\mathcal{Q}\mathcal{R}} - d_{\mathcal{P}q\mathcal{R}} - d_{\mathcal{P}\mathcal{Q}r} + 2d_{\mathcal{P}\mathcal{Q}\mathcal{R}})^2,$$

where $d_{p\mathcal{Q}\mathcal{R}}$, $d_{\mathcal{P}q\mathcal{R}}$ and $d_{\mathcal{P}\mathcal{Q}r}$ denotes the mean of the pth row, qth column and kth tube, respectively, and $d_{\mathcal{P}\mathcal{Q}\mathcal{R}}$ denotes the mean of $\mathcal{D}$.

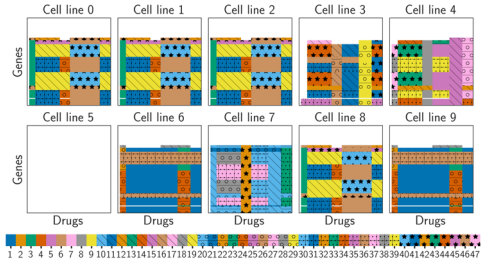


Selected Results

Different kinds of co-clusters were simulated. **Left**: Eight constant co-clusters were simulated. **Right**: The co-clusters found, corresponding to the true co-clusters.



The method was applied on a three-way genomic data set, containing dESeq2 normalised estimates of the fold changes, for different genes, drugs and cell lines.



The method has the potential to detect several types of additive coherent co-clusters. Applying it to real genomic data revealed several interesting co-clusters, which provides the biological researches with interesting gene-drug co-clusters to investigate further. Thus the method could be a useful tool for detecting coherent co-clusters in tensor data, and for exploratory data analysis.

References

[1] Sparse higher-order principal components analysis
Allen, Genevera
Artificial Intelligence and Statistics, 2012

[2] δ-TRIMAX: Extracting Triclusters and Analysing Coregulation in Time Series Gene Expression Data
Bhar, Anirban and Haubrock, Martin and Mukhopadhyay, Anirban and Maulik, Ujjwal and Bandyopadhyay, Sanghamitra and Wingender, Edgar
International Workshop on Algorithms in Bioinformatics, 2012

Taha, Mariam
Umeå University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Developing privacy-aware ML based on ML model spaces

The problem of model selection in machine learning can be seen as a search problem, where we have a space of models and we search for a model that best represents the data but also satisfies other criteria such as privacy.
The nature of dynamic databases that change frequently leads to an interaction between the space of the database and the space of models.
Previous studies show that models that appear frequently are better for integral privacy. In order to build privacy-aware machine and statistical learning, we plan to study the relationship between the space of data (possible databases, sets of databases) and the space of models. As a first step of this research direction, the construction of the two metric spaces and their relative distances is proposed.

Taha, Mariam
Umeå University

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Developing privacy-aware ML based on ML model spaces

Mariam Taha, Doctoral Student, Umeå University
Dept. of Computing Science
Supervisors: Prof. Vicenç Torra, Prof.Lili Jiang

UMEÅ UNIVERSITY

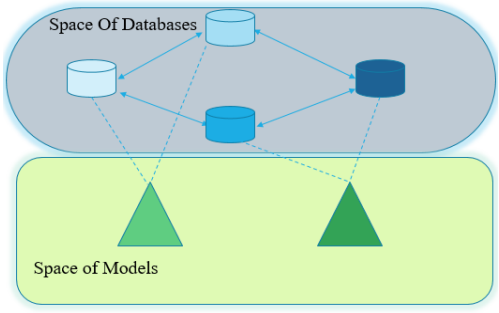
Abstract

The problem of model selection in machine learning can be seen as a search problem, where we have a space of models and we search for a model that best represents the data but also satisfies other criteria such as privacy. The nature of dynamic databases that change frequently leads to an interaction between the space of the database and the space of models. Previous studies show that models that appear frequently are better for integral privacy. In order to build privacy-aware machine and statistical learning, we plan to study the relationship between the space of data (possible databases, sets of databases) and the space of models. As a first step of this research direction, the construction of the two metric spaces and their relative distances is proposed.

Motivation

Several privacy models have been proposed in the literature, some of them are data driven models such as Differential privacy and Integral privacy [1]. Previous results have shown that when databases are sample, some machine learning models appear more frequently than others, these models are called recurrent, which are found to better from a privacy perspective specially for Integral privacy.
Since data are naturally dynamic, there is a need to derive machine learning models that are persistent in time but also satisfy other criteria such as privacy. In order to achieve this, it is necessary to study how the space of data interacts with the space of models.
More particularly, that any decision on the space of models has to take into account relationships between the space of data that generate these models.We consider that this perspective is of great interest in the following areas.

- In model selection for statistical and machine learning. In this area the goal is to select models that better generalize data and avoid overfitting. This also relates to the effect of outliers and influential points in learning. It is important to understand generalization and overfitting in terms of the relationship between the space of models and the space of data.



Interaction between Databases & Models

- In privacy preserving data mining and machine learning.

The need to study the relationship between the two spaces was first proposed in [2], in the context of integral privacy . In short, a model is integrally private if it can be generated by a large number of databases which are diverse enough. This is to avoid some type of privacy attacks on machine learning models.
Therefore, we want to investigate models with good accuracy, that does not have overfitting, that are not vulnerable to membership attacks and that are near to models with similar generators.

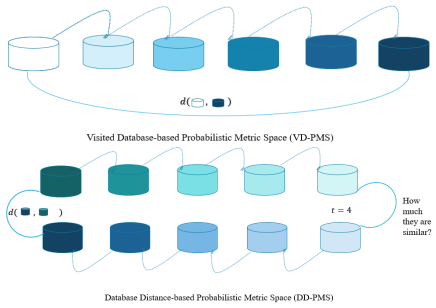
Aim & Objectives

The research focus on the study of spaces of data and models, formalizing these relationships and developing methodologies to select privacy-aware machine learning models.This research will require at least, the establishment of:

- Mathematical models including definition of metric spaces on both data and model spaces.
- Computational approaches to numerically compute distances between objects in these two spaces.
- Methodologies to select an appropriate model taking into account the relationships between spaces.

Selected Results

[3] In this paper we have proposed the use of Markov chains and transition matrices to model transitions between databases, and used them to define a probabilistic metric space for models. Our modelization uses two definitions to construct the probabilistic metric space. The first definition considers the distance between two databases in terms of the probability of being transformed into the second one (VD-PMS). The second one considers the distance between two databases in terms of their evolution(DD-PMS).



References

[1] Integral Privacy
Torra, V., Navarro-Arribas, G.
Proc. CANS 2016, LNCS 10052 661-669.

[2] Probabilistic Metric Spaces for Privacyby Design Machine Learning Algorithms: Modeling Database Changes
Torra, V., Navarro-Arribas, G.. DPM 2018/CBT 2018, LNCS 11025

[3] The space of models in machine learning: using Markov chains to model transitions
Vicenç Torra, Mariam Taha, Guillermo Navarro-Arribas
Progress in Artificial Intelligence volume 10, pages321–332 (2021)

Tarle, Magnus
KTH / Hitachi Energy

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Learning to Control Multiple Power Electronic Converters

The PhD project aims to identify a data-driven control architecture to meet the growing challenges of the electrical power system. In particular, the focus is on optimizing the control of multiple power electronic converters. The main objectives are to achieve greater utilization of the power system capacity, higher power system stability and improved power quality.

Tarle, Magnus
KTH / Hitachi Energy

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Learning to Control Multiple Power Electronic Converters

Magnus Tarle (KTH/Hitachi Energy)

Supervisors: Mårten Björkman (KTH), Mats Larsson (Hitachi Energy)
Co-supervisors: Lars Nordström (KTH), Gunnar Ingeström (Hitachi Energy)

Stiftelsen för Strategisk Forskning

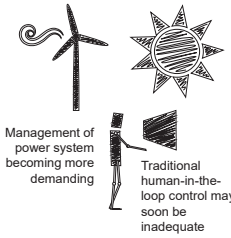
Hitachi Energy

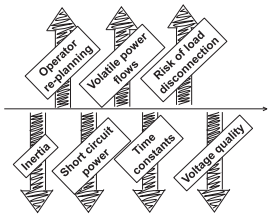
Abstract

The PhD project aims to identify a data-driven control architecture to meet the growing challenges of the electrical power system. In particular, the focus is on optimizing the control of multiple power electronic converters. The main objectives are to achieve greater utilization of the power system capacity, higher power system stability and improved power quality.

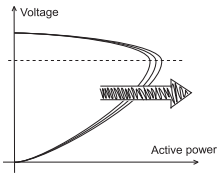
Motivation & Background

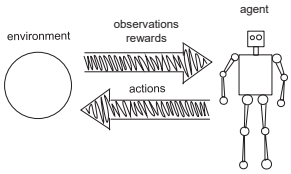
- The electrical power system is evolving and more coordinated, faster and accurate control alternatives are foreseen to be needed





- Using heuristics and engineering, a coordinated control of power electronic converters can increase transfer capability [2]
- Reinforcement learning can potentially self-learn such a control scheme

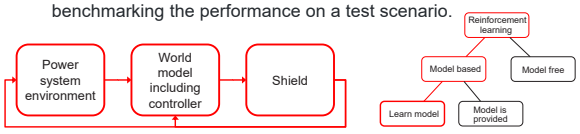


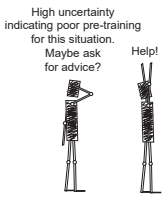


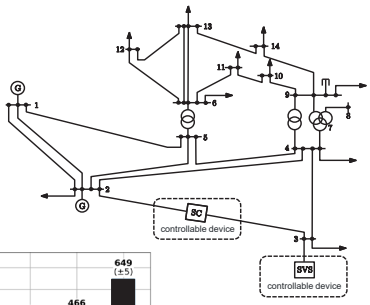
- Control challenges include:
  - Non-convex and sequential-decision-making problem
  - Model and measurement uncertainty
  - Partial observability
  - Variations in topology and dynamics
  - Safety and reliability crucial

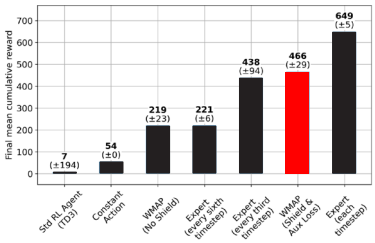
Methods & Preliminary Results

- We introduced a world model based reinforcement learning architecture (WMAP) in [1]. Contributions include:
  - Adapting the world model architecture from [3] to the power system domain
  - Increasing safety by adding a shield to make the architecture ask for guidance whenever uncertainty above threshold
  - Performing a case study on the IEEE 14-bus system and benchmarking the performance on a test scenario.









Results showed improved performance compared to model free reinforcement learning algorithm, fixed control setpoints and non-frequent expert setpoint updates.

More results and benchmark variants in the paper.

References

[1] Magnus Tarle, Mårten Björkman, Mats Larsson, Lars Nordström, Gunnar Ingeström, "A World Model Based Reinforcement Learning Architecture for Autonomous Power System Control", *SmartGridComm, 2021. In press.*

[2] M. Larsson, C. Rehtanz, and D. Westermann, "Improvement of Crossborder Trading Capabilities through Wide-area Control of FACTS", *In Bulk Power Syst. Dyn. Control VI, 2004.*

[3] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution", *Adv. Neural Inf. Process. Syst., vol. 2018-Decem, no. C, pp. 2450–2462, 2018.*

Future Work

Considerations for future work include:

- Continual learning
- Curriculum learning
- Scalability
- Improved safety
- Planning
- Benchmarking against state-of-the-art
- Hardware-in-the-loop implementation and assessment

Terra, Ahmad
KTH

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

BEERL: Both End Explanations for Reinforcement Learning

Some explainable reinforcement learning (XRL) methods are applied at the input side/end of the model to generate explanation(s) analyzing the contribution of the input features to the agent's decision. While some other methods like reward decomposition are applied to the output end of the model to explain the contribution of the output's components to the final reward value.
In this work we focus on a correlation between input and output explanations. This results in a finer granularity of explanation and also reveals reward prioritization effect which allows for a better adjustment for the RL agent. Additionally, we present a focus value measuring the fulfilment of the explanation to the desired properties.

Terra, Ahmad
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

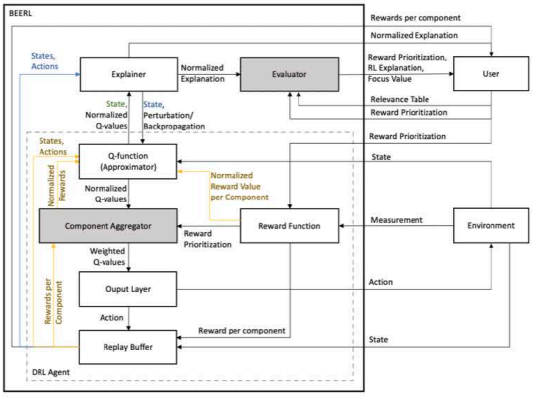
BEERL: Both End Explanations for Reinforcement Learning

Ahmad Terra, KTH - Ericsson
Machine Design, ITM - GFTL ER AI Trustworthy
Advisor: Elena Fersman, Martin Törngren – Rafia Inam

ERICSSON

Abstract

Some explainable reinforcement learning (XRL) methods are applied at the input side/end of the model to generate explanation(s) analyzing **the contribution of the input features to the agent's decision**. While some other methods like reward decomposition are applied to the output end of the model to explain **the contribution of the output's components to the final reward value**.
In this work we focus on a **correlation between input and output explanations**. This results in a finer granularity of explanation and also reveals reward prioritization effect which allows for a better adjustment for the RL agent. Additionally, we present a focus value measuring the fulfilment of the explanation to the desired properties.



Main elements of BEERL and interactions among them. The blue line is only executed on generating global explanation while the green text (state information) is only passed on generating local explanation. The yellow lines are only executed during training process i.e. updating Q-function parameters. The grey boxes show the proposed elements in modifying the existing solution

Methods

Generating input explanation (feature importance) per reward component (j) instead of the total reward:

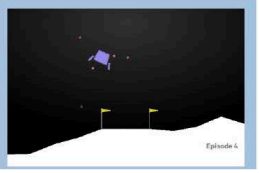
$$M_j = E\big(Q_j(s,a), s, a\big)$$

- Action
- State
- Q-function of reward component j
- XAI method

- E can be any explainer, in this experiment, we use Deeplift [1].
- Q function consists of several component constructing it.
- The final layer sums all the components as done by HRA [2].

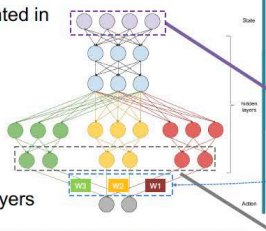
Use Case

Lunar Lander
State : Coordinate, angle, speed
Action : Nop, Fire [Left, Main, Right]
Reward : [-100 .. 100]
Algorithm: DQN with reward decomposition

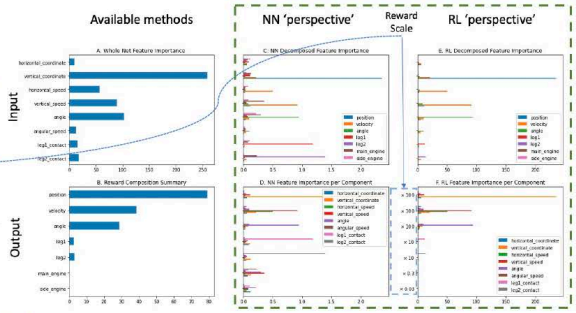


Reward decomposition can be implemented in two possible architectures:

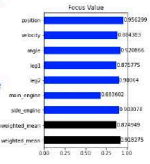
- **Individual network**
  - More parameters
  + Modular
- **Branched network**
  + Less parameters
  - Not modular
  + Mimic Feature extraction for initial layers



Results



The focus value (F_j) is the result of element-wise multiplication between the generated explanation (M_{ij}) and user-defined relevance table (N_{ij}) then averaged per component ($F_j = \frac{\sum_i M_{ij} \odot N_{ij}}{\sum_i M_{ij}}$). The unweighted mean value (U) is merely the mean of all components ($U = \frac{\sum_j F_j}{j}$) while the weighted mean (W) considers the reward prioritization ($W = \frac{\sum_j F_j \cdot R_j}{\sum_j R_j}$) where R_j is the reward weight or reward prioritization.



References

1. Avanti Shrikumar, et al, Learning Important Features through Propagating Activation Differences. In Proceedings of the 34th International Conference on Machine Learning (ICML'17), 2017.
2. Harm Van Seijen, et al, Hybrid Reward Architecture for Reinforce-ment Learning, In Advances in Neural Information Processing Systems, 2017.
3. Z. Juozapaitis, et al, Explainable Reinforcement Learning via Reward Decomposition, In proceedings at the International Joint Conference on Artificial Intelligence, 2019.

Wallin, Erik
Chalmers

DoubleMatch: improving Semi-Supervised Learning with a Self-Supervised Loss

Following the success of supervised learning, semi-supervised learning (SSL) is now becoming increasingly popular. SSL is the family of methods which in addition to a labeled training set, also use a large set of unlabeled data for training the model. Most of the recent successful SSL methods are based on pseudo-labeling approaches: letting confident model predictions act as training labels. While these methods have shown impressive results on many benchmark datasets, a drawback of this approach is that not all unlabeled data are used during training. We propose a method: DoubleMatch, which combines the pseudo-labeling approach with a self-supervised loss, enabling the model to utilize all unlabeled data through all stages of the training process. We show that this method achieves state of the art accuracies on multiple benchmark datasets while also reducing training times compared to previous SSL methods.

Wallin, Erik
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

DoubleMatch: improving semi-supervised learning with self-supervision

Erik Wallin, Ind. PhD, Saab AB and Chalmers
Supervisors: Lars Hammarstrand, Fredrik Kahl and Lennart Svensson

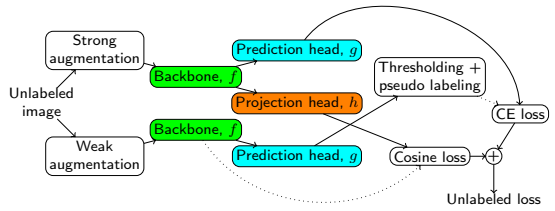
CHALMERS UNIVERSITY OF TECHNOLOGY

SAAB

Motivation & Research Goals

Supervised learning has gained a lot of attention in recent years because of remarkable achievements in fields such as image classification, object detection and natural language processing. The great results within supervised learning are typically fueled by huge amounts of labeled data. In practical applications, however, labeled data might be scarce, expensive, or require expert domain knowledge to attain. In contrast, *unlabeled data*, is often much easier to acquire, through e.g. web scraping or unsupervised sensor recordings. **Semi-supervised learning**, using both labeled and unlabeled data for fitting a model, has recently shown impressive results with methods such as FixMatch [1] and UDA [2]. These methods however suffer from not leveraging all unlabeled data during training. We propose a method, DoubleMatch, that takes inspiration from work in *self-supervised learning* to better utilize all unlabeled data. With this method we hope to 1) reduce converge rates of previous methods 2) increase test accuracy on benchmark datasets.

Methods



We build upon the FixMatch framework for semi-supervised learning. FixMatch uses the traditional cross-entropy loss for labeled data. For unlabeled data it utilizes both weak, α, and strong, β, data augmentations. A confident prediction on the weakly augmented sample is used as a *pseudo label* for a strong augmentation of that same sample:

$$l_p = \frac{1}{\mu B}\sum_{i=1}^{\mu B} \mathbb{1}\{\max(p(y|\alpha(x_i)) > \tau\} H(\text{argmax}\,(p(y|\alpha(x_i))), p(y|\beta(x_i)))$$

where τ is the confidence threshold, μB is the unlabeled batch size, and x_i are the unlabeled samples.
As we can see, only data with confident predictions on weakly augmented data are used to calculate this loss. We propose adding a feature loss on the output from the penultimate layer of the classification network using the cosine similarity:

$$l_s = -\frac{1}{\mu B}\sum_{i=1}^{\mu B} \frac{h(v_i) z_i}{|h(v_i)||z_i|} = -\frac{1}{\mu B}\sum_{i=1}^{\mu B} \cos(h(v_i), z_i).$$

Here, z_i is the feature vector for weakly augmented sample i and v_i is the feature vector for strongly augmented sample i. h is a trainable linear transformation to allow for different feature representations between weakly and strongly augmented data.

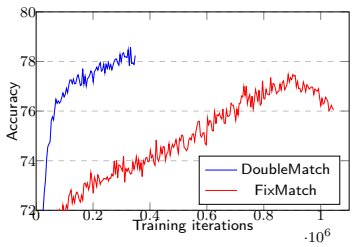
Selected Results

We evaluate our model on benchmark datasets for image classification using different sizes for the labeled training set. We achieve SOTA error rates despite running our method for fewer training iterations than the methods we use as comparisons.

Method	CIFAR10		
	40 labels	250 labels	4000 labels
Π-model	-	54.26±3.97	14.01±0.38
Pseudo-Labeling	-	49.78±0.43	16.09±0.28
Mean Teacher	-	32.32±2.30	9.19±0.19
MixMatch	47.54±11.50	11.05±0.86	6.42±0.10
UDA	29.05±5.93	8.82±1.08	4.88±0.18
ReMixMatch	19.10±9.64	5.44±0.05	4.72±0.13
FixMatch (CTA)	**11.39**±3.35	**5.07**±0.33	**4.31**±0.15
DoubleMatch (ours)	13.59±5.60	5.56±0.42	4.65±0.17

Method	CIFAR100			STL10
	400 labels	2500 labels	10000 labels	1000 labels
Π-model	-	57.25±0.48	37.88±0.11	26.23±0.82
Pseudo-Labeling	-	57.38±0.46	36.21±0.19	27.99±0.83
Mean Teacher	-	53.91±0.57	35.83±0.24	21.43±2.39
MixMatch	67.61±1.32	39.94±0.37	28.31±0.33	10.41±0.61
UDA	59.28±0.88	33.13±0.22	24.50±0.25	7.66±0.56
ReMixMatch	44.28±2.06	27.43±0.31	23.03±0.56	5.23±0.45
FixMatch (CTA)	49.95±3.01	28.64±0.24	23.18±0.11	5.17±0.63
DoubleMatch (ours)	**41.83**±1.22	**27.07**±0.26	**21.22**±0.17	**4.35**±0.20

To illustrate our increase in training speed the below figure shows test accuracy as a function of training iteration for DoubleMatch and FixMatch during a training run on Cifar100 with 10000 labeled data:



Our method performs well across many datasets. However seem to perform worse in the very low-label regime (e.g. Cifar10 with 40 labels). Our hypothesis is that high quality pseudo-labels is more important than self-supervision for the low-label datasets. Another weakness of our method is that the weight for the self-supervised loss needs to be tuned for each dataset.

References

[1] FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence
Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H. and Raffel, C.
NeurIPS 2020

[2] Unsupervised data augmentation for consistency training
Xie, Q., Dai, Z., Hovy, E., Luong, M.T. and Le, Q.V.
NeurIPS 2020

Willemsen, Bram
KTH

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Collecting visually-grounded dialogue data with a new vision+language task

We created a new vision+language task to collect visually-grounded dialogue data. The task is framed as a cooperative game in which two players have to come to an agreement on how to rank a set of images given some sorting criterion. The task is designed in such a way that it should lead to naturally-flowing conversations between participants discussing visual information and at the same time enable the study of the grounding and generation of referring expression in the face of distractors.

Willemsen, Bram
KTH

There are few visually-grounded dialogue datasets containing symmetric interactions of an unrestricted nature that capture commonly-observed dialogue phenomena and provide the opportunity to study the generation and grounding of referring expressions in the face of distractors, so we created a new vision+language task and decided to collect one

Bram Willemsen / bramw@kth.se
Dmytro Kalpakchi, Gabriel Skantze

"Which cat has the best beard?"

Task: a cooperative image ranking game in which two players have to reach an agreement on how images should be ranked given some sorting criterion

NB: the example dialogue is a highly-simplified artificial example for illustrative purposes only

Robot learning of symbol grounding in multiple contexts through dialog

Yang, Quantao
Örebro University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Learning Impedance Actions for Safe Reinforcement Learning in Contact-Rich Tasks

1. Extend the action space of RL policy by incorporating variable impedance
2. Our method can be safely deployed on the real robot directly

Yang, Quantao
Örebro University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Learning Impedance Actions for Safe Reinforcement Learning in Contact-Rich Tasks

Quantao Yang[1], Alexander Dürr[2], Elin Anna Topp[2], Johannes A. Stork[1], Todor Stoyanov[1]
[1]AMM Lab, Örebro University, Sweden
[2]Dept. of Computer Science, Lund University, Sweden

Contributions

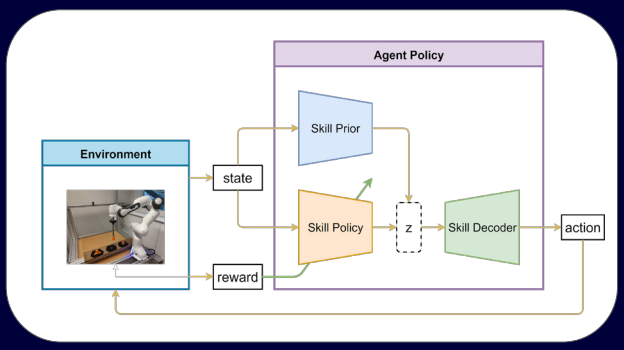
- Extend the action space of RL policy by incorporating ***variable impedance***
- Our method can be safely deployed on the real robot ***directly***

Reinforcement Learning with Skill Priors

- The evidence lower bound (ELBO) to learn a low-dimensional skill latent space[1]:

$$\log p(a) \geq \mathbb{E}_q[\log p(a \mid z) - \beta(\log q(z \mid a) - \log p(z))]$$

- Skill prior model $p_a(z \mid s_t)$ is used to guide Soft Actor-Critic (SAC) policy



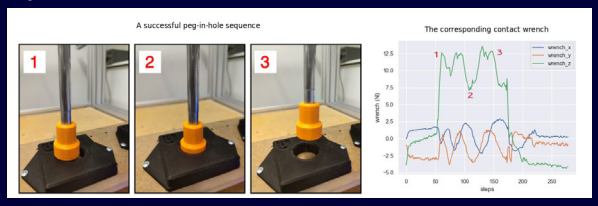
Variable Impedance Action Space

- The RL agent predicts the stiffness for contact-rich tasks
- We extend the policy action as the combination of end-effector pose $x \in SE(3)$ in Cartesian space and variable stiffness matrix $K \in \mathbb{R}^{6\times6}$

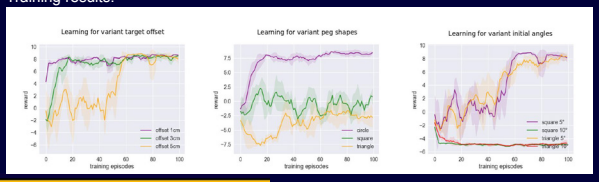
$$F_a = K(x - x_d) + D(\dot{x} - \dot{x_d}) + M(\ddot{x} - \ddot{x}_d)$$

Experiment Results

Peg-in-hole contact-rich task:



Training results:



This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

[1] Karl Pertsch, Youngwoon Lee, and Joseph J. Lim. Accelerating reinforcement learning with learned skill

Zangeneh Kamali, Fereidoon
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Camera Pose Posterior Inference for Visual Localisation

Visual localisation is the problem of estimating the camera pose in an environment from a camera image. Learning-based solutions, such as end-to-end camera pose regression, propose to solve this problem with the help of deep learning. One limitation of such approaches, whether wrapped in a probabilistic formulation or not, is that they assume a uni-modal solution to the pose estimation problem. While this assumption might hold in environments with unique visual features, it falls apart in presence of repetitive structures in the environment, where images from multiple camera poses appear visually similar. In this work we propose to learn inference of the complete pose posterior distribution that is desirable in such scenarios with multi-modal solutions, via variational inference of a simple posterior in a latent space, and learning a map from the latent space to SE(3).

Zangeneh Kamali, Fereidoon

KTH

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Camera Pose Posterior Inference for Visual Localisation

Fereidoon Zangeneh, Ind. PhD, Univrses & Kungliga Tekniska högskolan
Division of Robotics, Perception and Learning
Supervisors: Patric Jensfelt, Mårten Björkman (KTH),
Alessandro Pieropan, Amit Dekel (Univrses)

Motivation & Research Goals

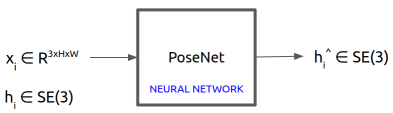
Visual localisation is the problem of estimating the camera pose in an environment from a camera image. Learning-based solutions, such as end-to-end camera pose regression, propose to solve this problem with the help of deep learning. One limitation of such approaches, whether wrapped in a probabilistic formulation or not, is that they assume a uni-modal solution to the pose estimation problem. While this assumption might hold in environments with unique visual features, it falls apart in presence of repetitive structures in the environment, where images from multiple camera poses appear visually similar. In this work we propose to learn inference of the complete pose posterior distribution that is desirable in such scenarios with multi-modal solutions, via variational inference of a simple posterior in a latent space, and learning a map from the latent space to SE(3).

Background

Visual localisation solutions [1] traditionally fall into 2 groups:

- Image-based (i.e. image retrieval in a database),
- Structure-based (i.e. feature-matching in a map).

Recent learning-based methods attempt to improve the visual localisation solutions to achieve invariance towards visual conditions such as seasonal and lighting variations. PoseNet [2] is one of the first methods that attempted to cast visual localisation as an end-to-end absolute pose regression problem.



Problem: PoseNet and its variants all assume a uni-modal (e.g. Gaussian) solution to the visual localisation problem.

References

[1] Long-Term Visual Localization Revisited
C. Toft, W. Maddern, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, T. Pajdla, F. Kahl, T. Sattler
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020

[2] PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization
Alex Kendall, Matthew Grimes, Roberto Cipolla
IEEE International Conference on Computer Vision, 2015

[3] NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis
B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, R. Ng
European Conference on Computer Vision, 2020

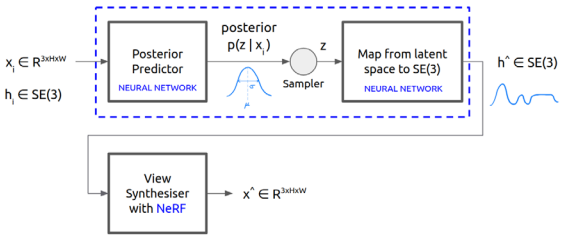
Method

Objective: Environments with repetitive structures call for a localisation solution that accommodates multi-modal hypotheses.

This requires a pipeline that for a query image produces the posterior distribution over possible camera poses. To avoid restricting the posterior to a fixed parametric form, we opt for a sampling-based solution that allows the learned distribution to take any form.

We formulate this in a pair of deep networks:

1. First network predicts a posterior distribution for the input query image in latent space with a simple parametric form such as Gaussian.
2. Second network maps the latent space to $SE(3)$.



The camera pose posterior distribution for a query image can be simulated by drawing samples from the predicted posterior distribution in the latent space and passing them through the learned map.

Training of the pair of networks is done via 2 supervision signals:

1. KL-divergence between the posterior distribution in the latent space and a prior (e.g. standard Gaussian)
2. photometric error between the query image and the image(s) generated from the camera pose(s) sampled from the posterior distribution

Generative model: A differentiable renderer such as a Neural Radiance Field (NeRF) model [3] pretrained on the scene is used with volume rendering techniques and an appropriate camera model to take $SE(3)$ samples back to image data space.

Challenge: Photometric error as a loss function has a small region of attraction, which requires the samples drawn from the posterior to be close enough the true mode(s) during training.

Zhang, Chi
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Learning the Pedestrian-Vehicle Interaction for Pedestrian Trajectory Prediction

The prediction of pedestrian behavior can help the drivers and the automated vehicles to make smarter and safer decisions and hence to protect the pedestrians from hazardous situations. The interaction between pedestrians and vehicles is an essential factor that influences pedestrian behavior. In this research, we propose a novel design called the Pedestrian-Vehicle Interaction (PVI) extractor for learning this interaction from data, and implement the proposed PVI extractor on sequential approaches (LSTMs) and non-sequential approaches (CNNs). We use the Waymo Open Dataset consisting of real-world traffic scenes with pedestrians and vehicles. The models using our proposed PVI extractor outperform the state-of-the-art models. The results show that the proposed PVI extractor can capture the interactions between pedestrians and vehicles.

Zhang, Chi
Chalmers

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Learning the Pedestrian-Vehicle Interaction for Pedestrian Trajectory Prediction

UNIVERSITY OF GOTHENBURG
CHALMERS

Chi Zhang, University of Gothenburg | Chalmers
Department of Computer Science and Engineering

* This paper is accepted in 2022 the 8th International Conference on Control Automation and Robotics (ICCAR). 2022.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement 860410

SHAPE-IT

Abstract

The prediction of pedestrian behavior can help the drivers and the automated vehicles to make smarter and safer decisions and hence to protect the pedestrians from hazardous situations. The interaction between pedestrians and vehicles is an essential factor that influences pedestrian behavior. In this research, we propose a novel design called the Pedestrian-Vehicle Interaction (PVI) extractor for learning this interaction from data, and implement the proposed PVI extractor on sequential approaches (LSTMs) and non-sequential approaches (CNNs). We use the Waymo Open Dataset consisting of real-world traffic scenes with pedestrians and vehicles. The models using our proposed PVI extractor outperform the state-of-the-art models. The results show that the proposed PVI extractor can capture the interactions between pedestrians and vehicles.

Introduction

Motivation:

Accurately predicting the trajectories of pedestrians in urban traffic scenarios is essential for automated vehicles to prevent hazardous situations. The interaction between pedestrians and vehicles is a key factor that influences the behavior of pedestrians. However, there is limited research on this topic that focuses on pedestrian-vehicle interaction.

Problem Definition:

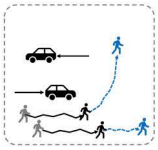
Use past trajectories of pedestrians and vehicles:

$$X_t^i = (x_t^i, y_t^i), 1 \le t \le T_{obs}, i \in \{1, \dots, N_p\}$$
$$V_t^j = (x_t^j, y_t^j), 1 \le t \le T_{obs}, i \in \{1, \dots, N_v\}$$

to predict future pedestrians' trajectories:

$$\hat{Y}_t^i = (x_t^i, y_t^i), \quad T_{obs} + 1 \le t \le T_{pred}, i \in \{1, \dots, N_p\}$$

where N_p and N_v are the number of pedestrians and vehicles in each frame.



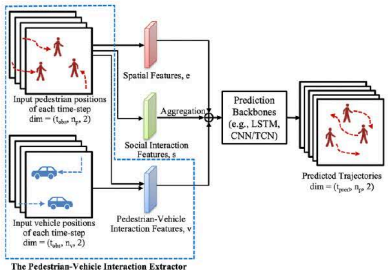
Challenges:

- Multiple factors (features) that influence pedestrian behavior;
- Multiple factors that influence pedestrian-vehicle interaction;
- Difficult to model the interactions between multiple vehicles and multiple pedestrians;
- Use in real-world urban traffic scenarios.

Overall Framework

The features we considered for trajectory prediction:

- Spatial Embedding;
- Social Interaction;
- Pedestrian-Vehicle Interaction.



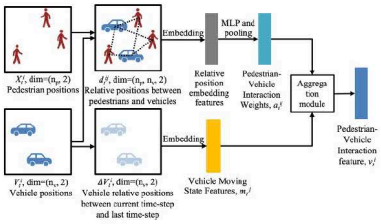
The Pedestrian-Vehicle Interaction Extractor

The PVI Extractor

The Pedestrian-Vehicle Interaction (PVI) extractor:

The factors that we consider in the pedestrian-vehicle interaction:

- Vehicle moving state;
- Relative positions between pedestrians and vehicles;
- Relative velocities between pedestrians and vehicles.



Experiments and Results

Dataset: the urban scenarios in Waymo Open Dataset, consists 450 real-world recording collected in urban traffic scenarios (374 training, 76 test), and each recording spans 20 seconds.

Settings:

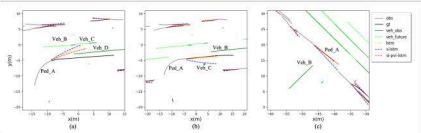
- Input: 3.2 sec in the past.
- Output: 4.8 sec in the future.

Evaluations:

- Quantitative evaluation shows the error (in meters) is smaller (better).
- Qualitative evaluation shows the algorithm can handle different scenarios.

TABLE I
THE ADE/FDE METRICS FOR BASELINE METHODS COMPARED TO PROPOSED METHODS ON THE WAYMO OPEN DATASET.

Prediction Method	Model Name	ADE	FDE	Interactions Used in the Model
Sequential	LSTM	0.392	0.844	No interactions.
	Social-LSTM (2016) [4]	0.402	0.840	Social interaction
	Social-GAN (2018) [5]	0.386	0.826	Social interaction
	SI-PVI-LSTM (ours)	**0.372**	**0.796**	Pedestrian-vehicle interaction
Non-sequential	Linear	0.412	0.892	No interactions.
	Social-STGCNN (2020) [6]	0.334	0.530	Social interaction
	Social-IWSTCNN (2021) [16]	0.329	**0.540**	Social interaction
	SI-PVI-Conv (ours)	0.327	0.543	Social interaction, Pedestrian-vehicle interaction



Future Work

- Joint prediction: predict both trajectory and crossing intention of the pedestrian in one framework.
- Hybrid Algorithms: for the pedestrian-vehicle interaction, combine the deep learning-based model with knowledge-based conventional model, and develop a model that is accurate and explainable.

Other Information

Contact details: chi.zhang@gu.se, zchi@chalmers.se
Supervisor: Associate Professor Christian Berger, GU
Co-Supervisor: Professor Marco Dozza, Chalmers
Our previous work: Social-IWSTCNN, published and presented on IEEE-IV 21 conference. The accessible version (or scan the QR code): https://arxiv.org/pdf/2105.12436.pdf

References

1. A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, F.-F. Li, andS. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp. 961–971.
2. A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2255–2264.
3. A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14 424–14 432
4. C. Zhang, C. Berger, and M. Dozza, "Social-iwstcnn: A social interaction-weighted spatio-temporal convolutional neural network for pedestrian trajectory prediction in urban traffic scenarios," in 2021 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2021, pp. 1515–1522.
5. P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui,J. Guo, Y. Zhou, Y. Chai, B. Caineet al., "Scalability in perception for autonomous driving: Waymo open dataset," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020, pp. 2446–2454.

Åkerblom, Niklas
Chalmers / Volvo Car Corporation

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Online Learning for Energy Efficient Navigation in Stochastic Transport Networks

Energy-efficient navigation constitutes an important challenge for the electrification of personal transportation, due to the limited battery capacity of electric vehicles. In this project, we employ a Bayesian approach to model the energy consumption at road segments for efficient navigation. In order to learn model parameters, we develop an online learning framework and investigate several exploration strategies, such as Thompson Sampling and Upper Confidence Bound. We also extend the framework to a multi-agent setting, where multiple vehicles adaptively navigate and learn the parameters of the energy model. To establish performance guarantees, we analyze combinatorial Thompson Sampling and derive upper bounds on the expected regret incurred in single-agent and multi-agent settings, through analysis of the algorithm under batched feedback. We demonstrate the performance of our methods via simulation experiments on several real-world city road networks.

Åkerblom, Niklas

Chalmers / Volvo Car Corporation

Online Learning for Energy Efficient Navigation in Stochastic Transport Networks

Niklas Åkerblom

Abstract

Energy-efficient navigation constitutes an important challenge for the electrification of personal transportation, due to the limited battery capacity of electric vehicles. In this project, we employ a Bayesian approach to model the energy consumption at road segments for efficient navigation. In order to learn model parameters, we develop an online learning framework and investigate several exploration strategies, such as Thompson Sampling and Upper Confidence Bound. We also extend the framework to a multi-agent setting, where multiple vehicles adaptively navigate and learn the parameters of the energy model. To establish performance guarantees, we analyze combinatorial Thompson Sampling and derive upper bounds on the expected regret incurred in single-agent and multi-agent settings, through analysis of the algorithm under batched feedback. We demonstrate the performance of our methods via simulation experiments on several real-world city road networks.

Background

Widespread adoption of electric vehicle technology is limited by range anxiety. Among other things, public perception is affected by factors like:

- Energy storage capacity limited due to expensive batteries.
- Fast-charging infrastructure still sparse.
- Still comparatively time consuming to charge batteries.

One way to mitigate worries is through better navigation methods. This requires knowledge about the environment!

The problem

- How to utilize a limited set of resources (e.g. autonomous vehicle fleet) to efficiently learn about energy consumption from environment.
- **Energy-efficient navigation problem:** Finding path which minimizes energy consumption through road network graph.
- We consider the energy consumption at different road segments to be stochastic and a priori unknown.
- We want to learn (explore) the parameters of the energy model adaptively while simultaneously solving the navigation problem.

Energy consumption model

- A road network graph may contain millions of edges, hence it is beneficial to utilize prior knowledge for more efficient exploration.
- **Bayesian approach:** Assume prior over expected energy consumption θ_e (e.g. Gaussian conjugate prior) for each edge e:

$$p(\theta_e \mid \mu_{0,e}, \sigma^2_{0,e}) = \mathcal{N}(\theta_e \mid \mu_{0,e}, \sigma^2_{0,e})$$

- Prior uses simplified deterministic energy consumption model based on longitudinal vehicle dynamics.
- Includes knowledge about the vehicle (e.g. mass m, front surface area A) and the road segment (e.g. length l_e, average speed v_e, slope angle α_e):

$$\mu_{0,e} = \frac{mgl_e \sin(\alpha_e) + mgC_r l_e \cos(\alpha_e) + 0.5 C_d A \rho l_e v_e^2}{3600\eta}$$

- Incrementally update posterior with new observations.

Shortest path problem as multi-armed bandit

We consider the stochastic **combinatorial semi-bandit** setting, a multi-armed bandit (MAB) problem where an agent sequentially selects sets of actions (subject to combinatorial constraints) instead of individual actions. The objective is to maximize the long-term expected reward.

- Road network graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \boldsymbol{\theta}^*)$, where each edge e has expected edge cost $\theta^*_e = -\mathbb{E}[r(e)]$ (we view energy consumption / cost as negative reward $-r(e)$).
- We want to find an optimal action set (edge sequence or path) $\boldsymbol{a}^*$ in the set of all paths $\mathcal{P}$ between two fixed nodes in $\mathcal{V}$ (characterizing the problem instance), such that:

$$\boldsymbol{a}^* = \arg\min_{\boldsymbol{a} \in \mathcal{P}} \sum_{e \in \boldsymbol{a}} \theta^*_e$$

- Bayesian regret (time horizon T, prior π_0 over mean vector $\boldsymbol{\theta}^*$, and expected reward function $f_{\boldsymbol{\theta}}(\boldsymbol{a}) := \sum_{i \in \boldsymbol{a}} \theta_i$):

$$\text{BayesRegret}(T) = \mathbb{E}_{\boldsymbol{\theta}^* \sim \pi_0}\left[\mathbb{E}\left[\sum_{t=1}^{T} \left(f_{\boldsymbol{\theta}^*}(\boldsymbol{a}^*) - f_{\boldsymbol{\theta}^*}(\boldsymbol{a}_t)\right) \,\middle|\, \boldsymbol{\theta}^*\right]\right]$$

Online learning framework

- At time t, use exploration method like Thompson sampling (TS) to compute weights $\boldsymbol{w}_t$.
- Find shortest path $\boldsymbol{a}_t$ w.r.t. weights, e.g. using Dijkstra's algorithm.
- Traverse path $\boldsymbol{a}_t$, observe energy consumption $-r_t$ and update posterior parameters.

Algorithm 1 Online learning for energy-efficient navigation

Require: $\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0$

for $t \leftarrow 0, 1, \ldots, T$ **do**
- $\boldsymbol{w}_t \leftarrow$ GetEdgeWeights$(t, \boldsymbol{\mu}_t, \boldsymbol{\sigma}_t)$
- $\boldsymbol{a}_t \leftarrow$ FindShortestPath$(\boldsymbol{w}_t)$
- $\boldsymbol{r}_t \leftarrow$ ApplyActionAndObserveReward$(\boldsymbol{a}_t)$
- $\boldsymbol{\mu}_{t+1}, \boldsymbol{\sigma}_{t+1} \leftarrow$ UpdateParameters$(\boldsymbol{a}_t, \boldsymbol{r}_t, \boldsymbol{\mu}_t, \boldsymbol{\sigma}_t)$

Assign edge weights using bandit algorithms (e.g., TS, BayesUCB):

Algorithm 3 Combinatorial Thompson Sampling

procedure GetEdgeWeights$(t, \boldsymbol{\mu}_t, \boldsymbol{\sigma}_t)$
- **for** each edge $e \in \mathcal{E}$ **do**
  - $\tilde{\theta}_e \leftarrow$ Sample from posterior $\mathcal{N}(\mu_{t,e}, \sigma^2_{t,e})$
  - $w_{t,e} \leftarrow \mathbb{E}[y_e]$ where $y_e \sim \mathcal{N}^R(\tilde{\theta}_e, \gamma_e^2)$
- **return** $\boldsymbol{w}_t$

Algorithm 4 Combinatorial BayesUCB

procedure GetEdgeWeights$(t, \boldsymbol{\mu}_t, \boldsymbol{\sigma}_t)$
- **for** each edge $e \in \mathcal{E}$ **do**
  - $w_{t,e} \leftarrow \max\left(0, Q\left(\frac{1}{T}, \mathcal{N}(\mu_{t,e}, \sigma^2_{t,e})\right)\right)$
- **return** $\boldsymbol{w}_t$

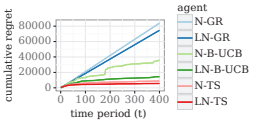
Regret analysis

We have derive upper bounds on the Bayesian regret of combinatorial TS for the case where reward feedback is delayed and received in batches of size K:

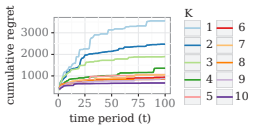
- Combinatorial TS for batched feedback: $\text{BayesRegret}(T) \leq \tilde{\mathcal{O}}(|\mathcal{E}|K + |\mathcal{E}|\sqrt{T})$.
- Extends to non-delayed case ($K = 1$): $\text{BayesRegret}(T) \leq \tilde{\mathcal{O}}(|\mathcal{E}|\sqrt{T})$.
- Also extends to multi-agent (vehicle fleet of size K) scenario with synchronous sharing of observations, with upper bound on fleet (Bayesian) regret: $\tilde{\mathcal{O}}(|\mathcal{E}|K + |\mathcal{E}|\sqrt{TK})$.

Experiments

- We utilize SUMO traffic simulation data from real-world traffic networks to provide realistic traffic patterns and vehicle speed distributions for simulations.



Single-agent algorithm comparison of cumulative regret



Multi-agent scenarios (K agents) with average cumulative regret per agent in each scenario

- We evaluate exploration through measured cumulative regret and visual inspection.

Exploration with combinatorial TS (explored paths in red)

Exploration with combinatorial BayesUCB (explored paths in red

Publications

Åkerblom, N., Chen, Y., and Chehreghani, M. H. (2020).
An online learning framework for energy-efficient navigation of electric vehicles.
In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2051–2057.

Åkerblom, N., Chen, Y., and Chehreghani, M. H. (2021).
Online learning of energy consumption for navigation of electric vehicles.
arXiv preprint arXiv:2111.02314.

niklas.akerblom@chalmers.se