Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization
TAILOR
Grant Agreement Number 952215

Connectivity Fund

Document type (nature)	Report
Deliverable No	D10.6
Work package number(s)	WP10
Date	Due M36
Responsible Beneficiary	TUE, ID 12
Author(s)	Joaquin Vanschoren
Publicity level	Public
Short description	This deliverable repeats every year and presents regular updates about the outcomes of the open call of the connectivity fund.

Document History			
Revision	**Date**	**Modification**	**Author**
1.0	15/09/2023	first version	Joaquin Vanschoren

Document Review		
Reviewer	**Institution**	**Date of report approval**
Marc Schoenauer	Inria Saclay	29/9/2023
Peter Flach	Bristol University	29/9/2023

This document is a public report. However, the information herein is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

Table of Content

Useful links

Hosted website for the connectivity fund:
https://tailor-network.eu/connectivity-fund/
Gallery of funded research visits and workshops:
https://tailor-network.eu/connectivity-fund/funded-projects/

Introduction

The TAILOR network includes many of Europe's top AI labs. However, we also want to reach out to the many other excellent labs and organizations across Europe to work together and create new breakthroughs in AI. The Connectivity Fund is a key instrument in this mission.

To establish a truly vibrant network, the Connectivity Fund provides funding to AI researchers from across Europe for research visits or workshops that bring together researchers from TAILOR labs and non-TAILOR labs. It especially aims to support young researchers to gain valuable experience and nurture the next generation of AI researchers. The goals, scope, organization, proposal evaluation, and legal framework have all been described in earlier deliverables D10.1-10.5.

This series of deliverables provides updates on the status of the connectivity fund. Since September 2021, these updates are done yearly. It will detail the number of submissions, the evaluation process, and the outcomes, funded visits, and workshops in a transparent way. Since the connectivity fund operates using a continuous open call, with cut-offs every 4 months, this report covers the results of the following cut-off rounds:

- 15th of November, 2022
- 15th of March, 2023
- 15th of July, 2023

Hence, this deliverable covers the period between 15th of July 2022 and 15th of July 2023. In the remainder of this document, we will detail the received proposals, their evaluation outcome, dissemination activities and organizational changes in this reporting period.

Overview

The connectivity fund has funded 37 research visits and 10 workshops so far. A geographical overview is shown below.

In the period from 15th of July 2022 to 15th of July 2023 we received and evaluated 38 proposals, of which 30 were funded. 5 were rejected and 3 were cancelled after acceptance by the recipient because of unforeseen circumstances.

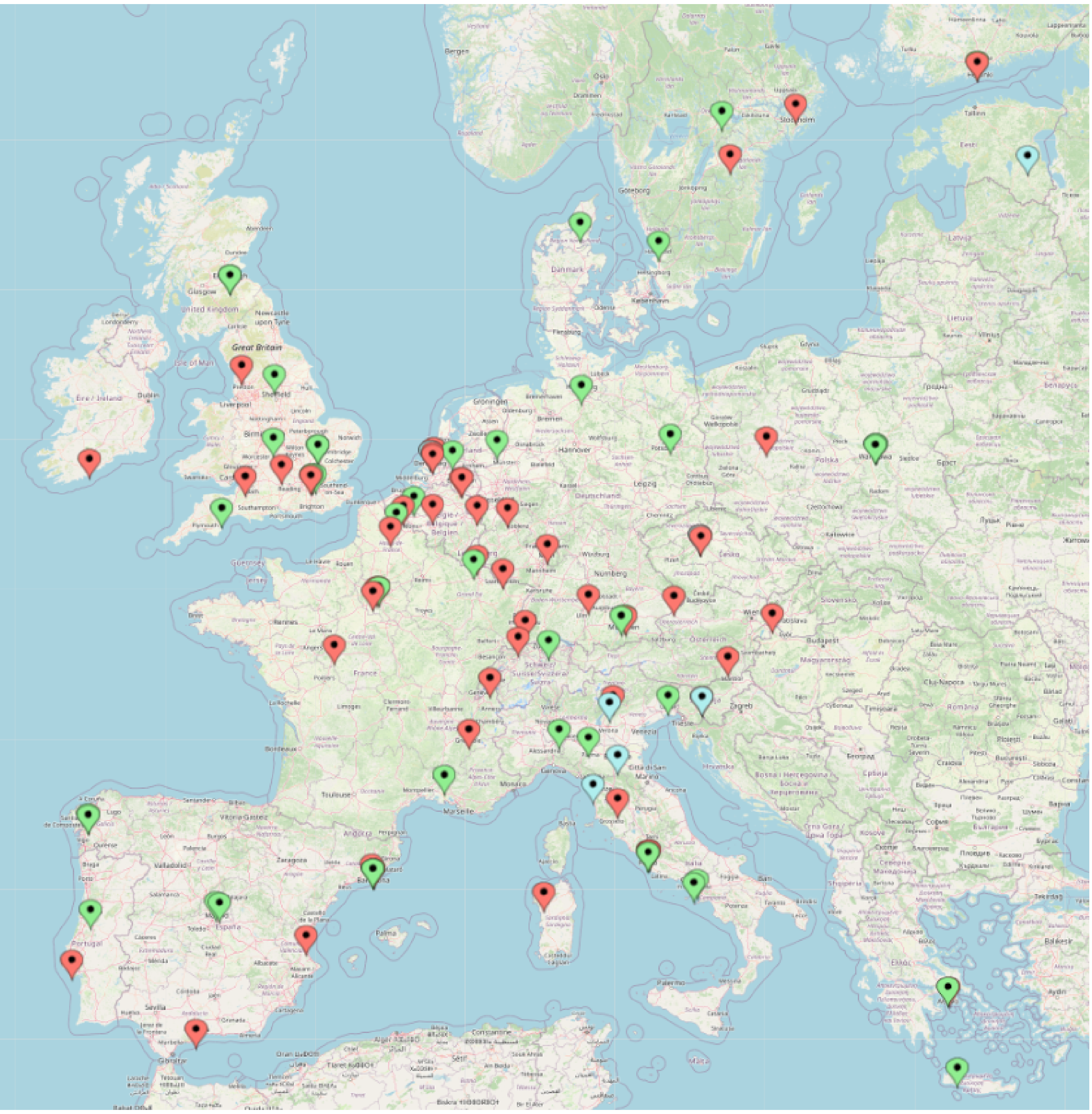


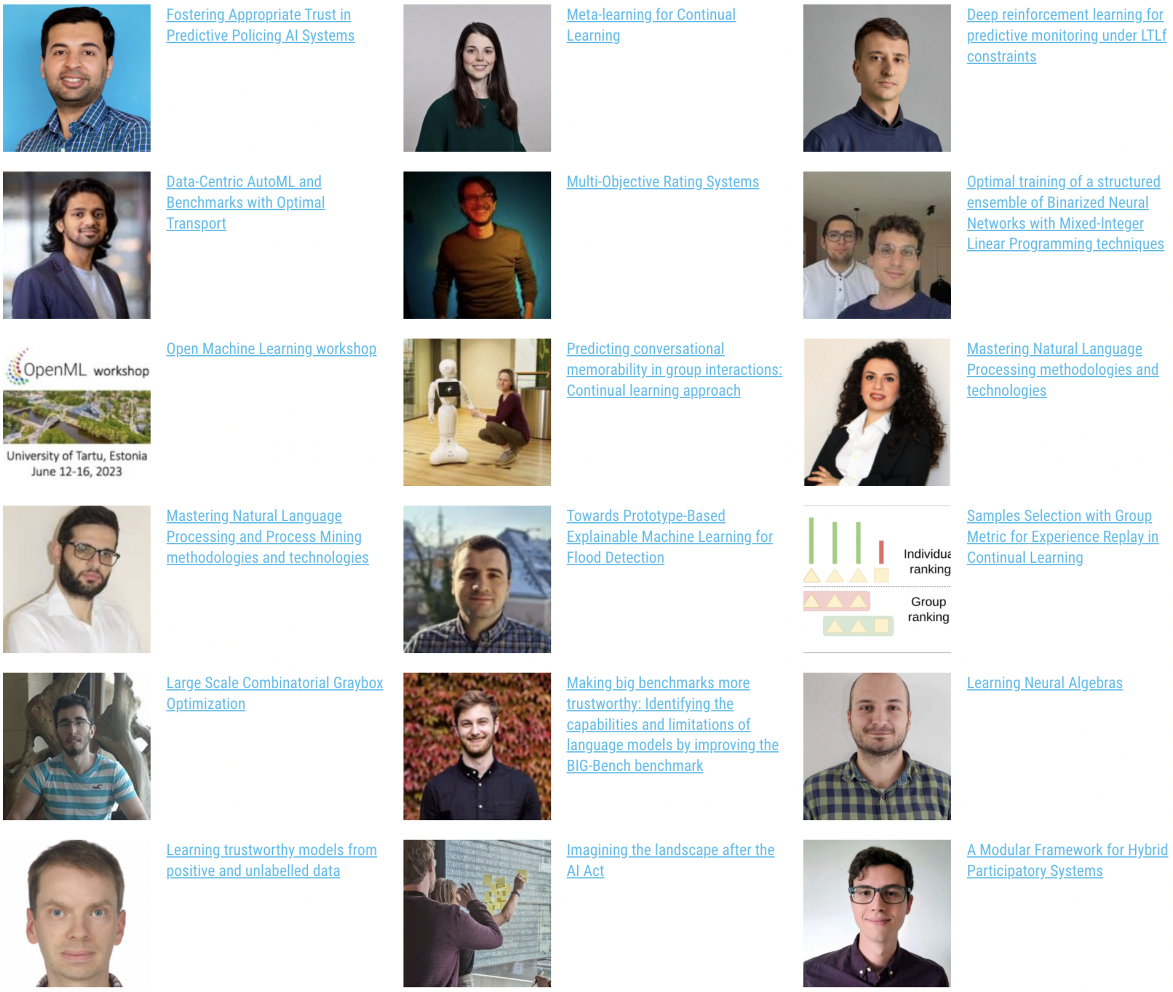
Figure 1. Geographical overview of Connectivity Fund visits, with all TAILOR labs in red, the Connectivity Fund beneficiaries (non-TAILOR labs) in green, and workshops in blue. Note that visits, workshops and labs within the same city can overlap.

Dissemination activities

In this reporting period, we have continued to disseminate the Connectivity Fund activities to make sure that it is widely known in the European AI community.

Connectivity Fund Website

The Connectivity Fund website, seamlessly integrated in the TAILOR project website, has all the latest information about the aims of the fund, how to apply, the evaluation procedure, and other useful information for applicants. It was repeatedly updated to streamline the application process. It also contains a gallery of all funded projects to inspire would-be applicants. For each project, we show the abstract of the research proposal and a short bio of the researcher who won the award. A screenshot of this gallery, containing the latest research visits amd workshops, is shown below.



Fostering Appropriate Trust in Predictive Policing AI Systems

Meta-learning for Continual Learning

Deep reinforcement learning for predictive monitoring under LTLf constraints

Data-Centric AutoML and Benchmarks with Optimal Transport

Multi-Objective Rating Systems

Optimal training of a structured ensemble of Binarized Neural Networks with Mixed-Integer Linear Programming techniques

Open Machine Learning workshop

Predicting conversational memorability in group interactions: Continual learning approach

Mastering Natural Language Processing methodologies and technologies

Mastering Natural Language Processing and Process Mining methodologies and technologies

Towards Prototype-Based Explainable Machine Learning for Flood Detection

Samples Selection with Group Metric for Experience Replay in Continual Learning

Large Scale Combinatorial Graybox Optimization

Making big benchmarks more trustworthy: Identifying the capabilities and limitations of language models by improving the BIG-Bench benchmark

Learning Neural Algebras

Learning trustworthy models from positive and unlabelled data

Imagining the landscape after the AI Act

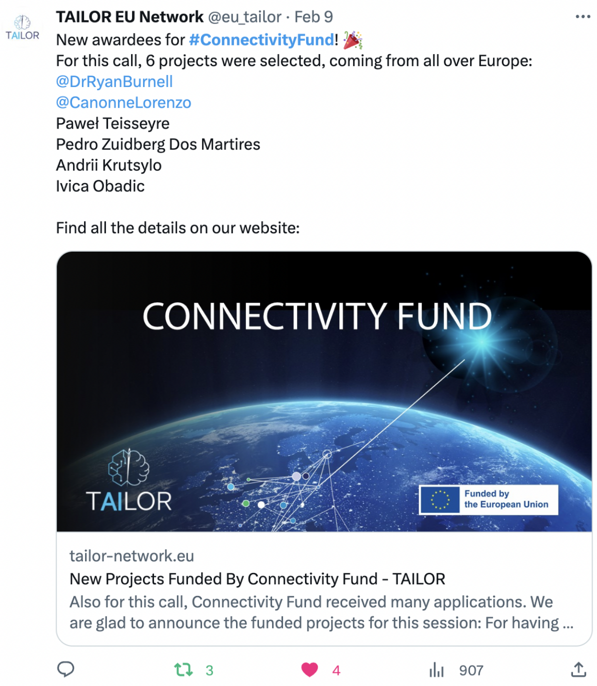
A Modular Framework for Hybrid Participatory Systems

Figure 2. Gallery of successful Connectivity Fund projects on the TAILOR website. This figure only shows a subset of them.

General dissemination

We have spread the word about the connectivity fund through various communication channels:

- Regular announcements on the upcoming Connectivity Fund cut-off dates in the TAILOR newsletters
- Social media (e.g. Twitter). Examples are shown below.
- The TAILOR Conference in Siena where the audience could learn more about the fund.
- An open call submitted to the European Commission Funding and tenders website.



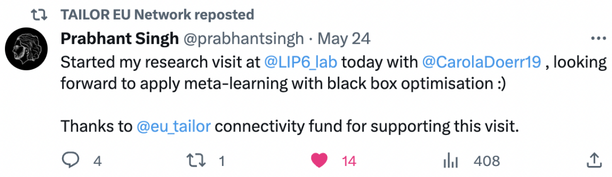
TAILOR EU Network @eu_tailor · Feb 9

New awardees for **#ConnectivityFund**! 🎉
For this call, 6 projects were selected, coming from all over Europe:
@DrRyanBurnell
@CanonneLorenzo
Paweł Teisseyre
Pedro Zuidberg Dos Martires
Andrii Krutsylo
Ivica Obadic

Find all the details on our website:

tailor-network.eu
New Projects Funded By Connectivity Fund - TAILOR
Also for this call, Connectivity Fund received many applications. We are glad to announce the funded projects for this session: For having ...

3 4 907



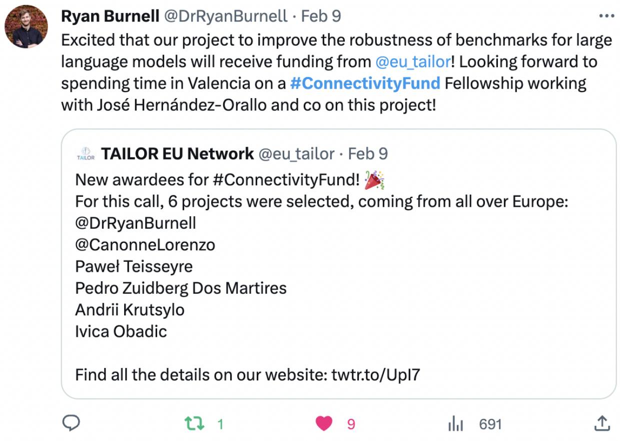
TAILOR EU Network reposted

Prabhant Singh @prabhantsingh · May 24

Started my research visit at @LIP6_lab today with @CarolaDoerr19 , looking forward to apply meta-learning with black box optimisation :)

Thanks to @eu_tailor connectivity fund for supporting this visit.

4 1 14 408



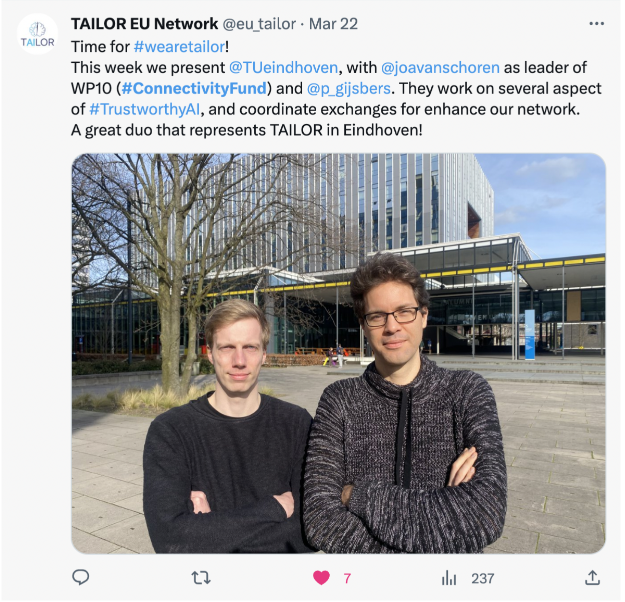
Ryan Burnell @DrRyanBurnell · Feb 9

Excited that our project to improve the robustness of benchmarks for large language models will receive funding from @eu_tailor! Looking forward to spending time in Valencia on a **#ConnectivityFund** Fellowship working with José Hernández-Orallo and co on this project!

TAILOR EU Network @eu_tailor · Feb 9
New awardees for #ConnectivityFund! 🎉
For this call, 6 projects were selected, coming from all over Europe:
@DrRyanBurnell
@CanonneLorenzo
Paweł Teisseyre
Pedro Zuidberg Dos Martires
Andrii Krutsylo
Ivica Obadic

Find all the details on our website: twtr.to/UpI7

1 9 691



TAILOR EU Network @eu_tailor · Mar 22

Time for #wearetailor!
This week we present @TUeindhoven, with @joavanschoren as leader of WP10 (**#ConnectivityFund**) and @p_gijsbers. They work on several aspect of #TrustworthyAI, and coordinate exchanges for enhance our network.
A great duo that represents TAILOR in Eindhoven!

7 237

Figure 3. Twitter activity

New applications received

Table 1 lists the eligible applications in this reporting period. Proposals 1-17 were discussed in previous deliverables. The institute shown in **bold** is the *hosting lab*, hosting the researcher(s) during their visit or organising a workshop.

The projects cover a wide range of topics all centred around trustworthy AI. These topics are visualised in Figure 4. Some are visits and some are workshops (highlighted in blue).

The maximum recommended funding is 15.000 EUR, in order to realise about 100 research visits and workshops on different topics and in different countries. Some visits have a slightly higher budget because of added overhead costs. One workshop (the ESSAI summer school) has a much larger budget. This was approved by the TAILOR coordinator since it is a key TAILOR activity and some participants from the TAILOR network received a free registration.

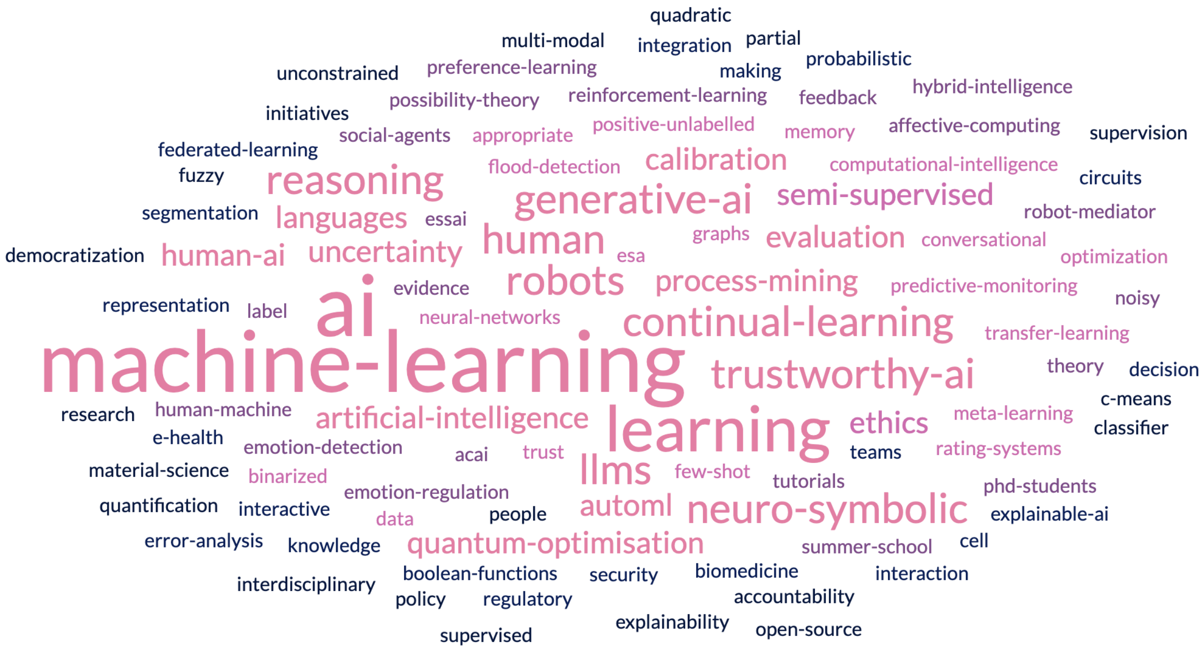


Figure 4 . Word Cloud of the most frequent keywords of all proposals

Table 1. Overview of all proposals received during this reporting period. The hosting lab is shown in bold. The non-TAILOR lab is always the recipient of the funding. Workshops are highlighted in blue.

ID	Title and applicant	TAILOR lab	Non-TAILOR lab	Funding
18	**Learning trustworthy models from positive and unlabelled data** *Paweł Teisseyre (Warsaw University of Technology)*	**KU Leuven**	Warsaw University of Technology	€ 4900
19	**Learning Neural Algebras** *Pedro Zuidberg Dos Martires (Orebro University)*	**University of Trento**	Orebro University	€ 6824
20	**Making big benchmarks more trustworthy: Identifying the capabilities and limitations of language models by improving the BIG-Bench benchmark.** *Ryan Burnell (University of Cambridge)*	**UP Valencia**	University of Cambridge	€ 14850
21	**Mastering experience in Natural Language Processing, Federated Learning, Time-Series Analysis and Process Mining** *Luigi Colucci Cante (University della Campania)*	**IIIA-CSIC Barcelona**	Universita della Campania	€ 15000
22	**Large Scale Combinatorial Graybox Optimization** *Lorenzo Canonne (Université de Lille)*	**University of Malaga**	Universite de Lille	€ 16900
23	**Samples Selection with Group Metric for Experience Replay in Continual Learning** *Andrii Krutsylo (Polish Academy of Sciences)*	**University of Pisa**	Polish Academy of Sciences	€ 9350
24	**Efficient Meta-Learning in Neural Networks** *Mike Huisman (Leiden University)*	U Leiden	**University of Edinburgh**	€ 5800
25	**Towards Prototype-Based Explainable Machine Learning for Flood Detection** *Ivica Obadic (LMU Munich)*	**University of Lancaster**	U Munich	€ 4150
26	**Automating Reliability Check for Machine Learning Models** *Xin Du (University of Edinburgh)*	**TU Eindhoven**	U Edinburgh	€ 15000
27	**Deep reinforcement learning for predictive monitoring under LTLf constraints** *Efrén Rama Maneiro (U Santiago de Compostela)*	**University Sapienza Rome**	U Santiago de Compostela	€ 8120

28	**Mastering Natural Language Processing and Process Mining methodologies and technologies** *Luigi Colucci Cante (University della Campania)*	**IIIA-CSIC Barcelona**	Universita della Campania	€ 23440
29	**Multi-objective Rating Systems** *Paolo Turrini (University of Warwick)*	Vrije Universiteit Brussels	**University of Warwick**	€ 8100
30	**TAILOR workshop on Open Machine Learning** *Meelis Kull (University of Tartu, Estonia)*	TU Eindhoven	**University of Tartu**	€ 15000
31	**Graph learning and applications considering high-order interactions** *Andrei Buciulea (Universidad Rey Juan Carlos, Spain)*	TU Delft	**Universidad Rey Juan Carlos**	€ 8150
32	**Fostering Appropriate Trust in Predictive Policing AI Systems** *Siddharth Mehrotra (TU Delft, Netherlands)*	TU Delft	**Univeristy of Hamburg**	€ 11700
33	**Optimal training of a structured ensemble of Binarized Neural Networks with Mixed-Integer Linear Programming techniques** *Simone Milanesi (University of Pavia, Italy)*	TU Delft	**University of Pavia**	€ 9500
34	**Predicting conversational memorability in group interactions: Continual learning approach** *Maria Tsfasman (TU Delft, Netherlands)*	TU Delft	**Univeristy of Cambridge**	€ 14948
35	**Meta-learning for Continual Learning** *Anna Vettoruzzo (Högskolan i Halmstad)*	**TU Eindhoven**	Högskolan i Halmstad	€ 16250
36	**AI-Assisted Quantum Optimisation Algorithms: Application to Smart City Problems** *Zakaria Abdelmoiz Dahi (Universidad de Malaga)*	Universidad de Malaga	**University of Exeter**	€ 14364
37	**Data Centric AutoML and Benchmarks with optimal transport** *Prabhant Singh (TU Eindhoven)*	TU Eindhoven	**Sorbonne University**	€ 13350
38	**Towards Efficient Energy Management in Microgrids: AI-Based Modeling of Household Power Consumption** *Petra Vrablecova (KINIT)*	Universita de Bologna	**KINIT, Bratislava**	€ 7000
39	**Connecting TAILOR and CLAIRE Rising Research Network: Strengthening AI Research through Collaborative Summer School Event** *Nicolo Brandizzi (La Sapienza, Rome)*	La Sapienza, Rome	**CLAIRE, The Hague**	€ 8650

40	**Continual Learning Unconference** *Vincenzo Lomonaco (University of Pisa)*	**University of Pisa**	ContinualAI (non-profit)	€ 5600
41	**ESSAI & ACAI 2023 summer school** *Vida Groznik (University of Ljubljana)*	All TAILOR project partners	**University of Ljubljana**	€ 46300
42	**Leveraging Social Agents as Mediators to Foster Comprehension and Control of Affective Engagement with Digital Content** *Sergio Muñoz (Universidad Politécnica de Madrid)*	**Istituto Superior Técnico, Lisboa**	Universidad Politécnica de Madrid	€ 5800
43	**Robust and safe reinforcement learning against uncertainties in human feedback** *Taku Yamagata (University of Bristol)*	University of Bristol	**LMU Munich**	€ 2960
44	**The First Workshop on Hybrid Human-Machine Learning and Decision Making (HLDM' 2023)** *Andrea Passerini (University of Trento)*	University of Trento	**Scuola Normale Superiore, Pisa**	€ 14500
45	**Trustworthy Probabilistic Machine Learning Models** *Stefano Teso (University of Trento)*	University of Trento	**University of Edinburgh**	€ 7500
46	**Boolean Seminar Liblice 2023** *Ondrej Cepek (Charles University, Prague)*	Charles University, Prague	**Czech Academy of Sciences**	€ 5600
47	**Holistic Evaluation of AI-assisted Biomedicine: A Case Study on Interactive Cell Segmentation** *Wout Schellaert (UP Valencia)*	UP Valencia	**Ghent University**	€ 15000
48	**Ethics, Norms and AI: Navigating the Regulatory Domain of Artificial Intelligence in Healthcare with a Focus on Federated Learning** *Gennaro Junior Pezzullo (University of Campania)*	**IIIA-CSIC Barcelona**	University of Campania	€ 15000
49	**Explainable Semi-Supervised Fuzzy C-Means** *Kamil Kmita (Polish Academy of Sciences)*	**La Sapienza, Rome**	Polish Academy of Sciences	€ 10850
50	**5th Young Researchers' Workshop on Machine Learning for Materials Science** *Saso Dzeroski (JSI)*	Jozef Stefan Institute	**SISSA (IT)**	€ 15000

51	**Uncertainty-aware calibration for trustworthy AI: Combining approaches from machine learning and speaker verification** *Paul-Gauthier Noé (Avignon University)*	**University of Bristol**	Avignon University	€ 10000
52	**International Conference on AI for People: Democratizing AI** *Marta Ziosi (University of Oxford)*	University of Oxford	**University of Bologna**	€ 3874
53	**Supervised Learning for Enhancing the Quantum Approximate Optimisation Algorithm** *Zakaria Abdelmoiz Dahi (Universidad de Malaga)*	Universidad de Malaga	**University of Exeter**	€ 11756
54	**Temporal & Sequential Neurosymbolic AI** *Nikolaos Manginas (NCSR "Demokritos")*	**KU Leuven**	NCSR "Demokritos"	€ 9900

Review process and outcome

First, all applications submitted within the deadline were evaluated for formal eligibility by the call management (Joaquin Vanschoren). The eligibility criteria are specified in deliverable D10.1. All proposals except one passed the eligibility test.

Next, each proposal was reviewed by two members of the scientific board. All members of the board have been actively involved. We checked for any conflicts of interest in the assignment. The final results of the evaluation are summarised in Table 2. All proposals were evaluated according to 5 criteria (AI excellence, scientific track record of the candidate, scientific step-up, the suitability of the host, and appropriateness of the activity duration). The final score is a weighted average of all scores, using the weighting described in deliverable D10.1.

Scores per criteria are on a scale from 0-10:

- **0-1** Application fails to address the criterion or cannot be assessed due to missing or incomplete information
- **2-3 Poor** – criterion is inadequately addressed or there are serious inherent weaknesses
- **4-5 Fair** – application broadly addresses the criterion, but there are significant weaknesses
- **6-7 Good** – application addresses the criterion well, but a number of shortcomings are present
- **8-9 Very good** – application addresses the criterion very well, but a small number of shortcomings are present
- **10 Excellent** – application successfully addresses all relevant aspects of the criterion. Any shortcomings are minor.

As per the Connectivity Fund rules, proposals must achieve a minimum of 70% of the maximal score to receive funding. All eligible applications passed this threshold. The total requested funding is also well within the budget earmarked for the second year of the connectivity fund (500k EUR divided over 4 cut-offs). Based on these outcomes, all these applications were accepted for funding.

ID	AI Excellence	Science track record	Science step-up	Host lab	Visit length	Final Score	Decision
18	9,6	9,8	9,6	10,8	10,5	7.9	accept
19	10,10	10,9	10,10	10,10	8,10	9.8	accept
20	6,10	7,9	7,10	9,8	9,10	8.4	accept
21	6,7	8,8	5,7	5,10	4,7	6.8	reject
22	7,8	8,9	8,9	9,9	7,10	8.3	accept
23	8,9	7,8	9,8	9,10	7,8	8.3	accept
24	8,7	7,8	8,9	9,9	8,8	8	canceled
25	8,10	8,8	8,9	8,9	7,7	8.3	accept

ID	AI Excellence	Science track record	Science step-up	Host lab	Visit length	Final Score	Decision
26	7,7	7,7	5,5	7,7	5,7	6.4	reject
27	9,9	8,9	9,9	10,9	9,9	8.9	accept
28	3,7	7,9	9,8	10,6	10,9	7.6	accept
29	9,10	7,10	9,10	10,10	6,10	9.1	accept
30	9,10	10,10	9,10	10,10	9,10	9.7	accept
31	8,8	8,8	8,8	8,8	7,7	7.9	accept
32	10,9	9,7	1,8	10,7	8,8	7.6	accept
33	7,9	6,6	8,9	8,8	8,8	7.6	accept
34	10,9	10,9	10,9	10,10	9,9	9.5	canceled
35	8,9	7,7	8,9	8,10	7,9	8.1	accept
36	6,6	6,7	6,6	7,8	7,7	6.4	reject
37	8,7	7,8	8,9	9,9	8,7	7.9	accept
38	10,4	8,7	8,5	9,8	8,7	7.3	canceled
39	7,7	7,7	8,8	9,8	8,8	7.6	accept
40	7,7	7,8	8,8	8,7	8,8	7.6	accept
42	10,3	10,4	10,3	10,3	8,2	6.4	reject
43	8,8	8,7	8,8	9,9	7,4	7.7	accept
44	8,8	9,9	9,9	8,8	10,10	8.8	accept
45	6,6	8,8	7,7	8,8	8,8	7.3	accept
46	6,8	8,9	7,7	8,10	8,10	7.9	accept
47	9,10	10,10	9,10	8,8	9,8	9.3	accept
48	10,10	8,8	9,9	10,10	7,7	8.9	accept
49	5,7	7,8	8,9	8,8	8,7	7.4	accept
50	10,8	10,9	10,9	10,9	10,7	9.3	accept
51	10,9	8,7	10,10	10,9	8,9	9	accept
52	9,6	10,4	9,4	10,5	9,7	7.2	accept
53	2,10	8,9	7,9	6,8	9,9	7.6	accept
54	9,10	7,9	9,10	10,10	8,10	9.1	accept

Table 2. Overview of TAILOR Connectivity fund applications and their evaluation. Each cell shows the evaluation results by two independent reviewers. Hence, '7,9' means that the first reviewer rated the proposal 7/10 and the second reviewer rated it 9/10.

Organisational streamlining

To streamline the management of the connectivity fund, a number of changes were made during this reporting period.

Overhead cost
Since the 15th of March 2023, beneficiaries are allowed to charge overhead costs to cover the administrative overhead and unforeseen costs. This was a requirement for many institutes to be able to administer the Connectivity Fund funding. This is done according to a flat rate of 25%, as is in compliance with the TAILOR general agreement.

Submission procedure
Since the 1st of July 2023, we moved the submission platform from the Easychair platform to a set of simple Google forms. This allowed us to automate several steps of the proposal handling and follow-up via scripts and collaborative spreadsheets.

Continuous review
Since the 15th of July 2023, we have moved from submission round cut-offs every four months to a continuous evaluation, with all new proposals being evaluated within a month. This was done since some research visits and workshops cannot wait until the next cut-off date. This continuous evaluation enables more research visits and workshops to be supported by the Connectivity fund, which is especially useful in the last year of the program.

Impact evaluation

The Connectivity fund is a key mechanism to foster collaboration and allows researchers all over Europe to work on the core research problems addressed by the TAILOR network. It also allows TAILOR to open up to a wider section of the AI community. To measure its impact, for all proposals since 2022, we ask all participants to send a structured scientific report containing:

- A summary of the research objectives
- Technical approach, findings, and future work
- A self-assessment of the impact of the research visit on AI excellence and their own careers, as well as the suitability of the host and the visit length.
- A list of publications and other outcomes of the visit.

These reports (for the completed visits) are attached with this deliverable.

Basic Information

Project title: Rotation-invariant vision transformers for trustworthy and sample efficient computer vision
Period of project: 10.2022-12.2022
Period of reporting: full project
Author(s): Mohammadreza Amirian
Organisation: Zurich University of Applied Sciences (ZHAW)
Host organisation: Swiss Federal Institute of Technology Lausanne (EPFL)

Public summary

After the breakthrough of transformers in natural language processing, these models are now adapted for computer vision and image classification tasks. Transformer-based models showed at least equal descriptive properties compared to convolutional models; however, initial specimens required more data for generalisation than convolutional models. Furthermore, these models didn't produce translation, rotation, and scale equivariant features. In a recent study, researchers introduced rotation equivariant transformers for a discrete rotation group, although these models' generalisation properties and sample efficiency were not well investigated yet. During this project, we aimed to extend the concept of rotation invariance for continuous rotation to improve the trustworthiness of the decisions and robustness of the vision transformer models. However, computer vision models demonstrated vulnerability towards variations in the angle and scale of the input images. This weakness leads to reduced trust in models' decisions in some circumstances that we addressed through furthering reliability and robustness using rotation invariance. Furthermore, gains in sample efficiency and improvement in generalisation are expected as the features show consistency over different variations in the rotation of the original image.

Research objectives

Objectives

Inductive biases such as translation-invariance undeniably accelerated the rapid advances of modern vision models through parameter sharing and improving sample efficiency. However, state-of-the-art models can only partially incorporate rotation-invariance. Recent attempts to develop rotation-invariant techniques mainly face the challenge of high memory requirements or limiting the original model capacity. This research proposes an embedding layer for vision transformers to leverage the invariance of self-attention layers to the order of tokens and train robust models against local and global rotation. The proposed image embedding technique requires negligible memory overhead to train rotation invariance models on large datasets such as ImageNet. The paper presents the proposed method's merit in improving the sample efficiency and robustness of vision transformers on small and larger datasets on classification and segmentation tasks.

Impact

Reproducible research work with source code for rotation-invariant vision transformers has been developed. The developed method provides more robust vision models against rotation variations of the input images, contributing to training more robust and trustworthy vision models. Furthermore, the models aim at rotation invariance, and the proposed transformers can cope with a range of transformed input images; thus, we expect a higher sample efficiency in future experiments. The developed method will be published in a Ph.D. dissertation as a proof of concept with the possibility of application in future research.

Technical approach

Detailed description

Deep vision models demonstrated vulnerability against rotation and scaling, starting from early research works. Many researchers attempted to tackle this problem using data-driven techniques or by adding rotation invariance inductive bias to models. Although data-driven approaches turned out to be more straightforward to implement, rotation-invariant models show their merits regarding sample efficiency and small data applications. The memory required for current state-of-the-art vision transformer techniques aiming at rotation invariance shows a linear increase with the size of the rotation group. The memory requirement of these models based on lie-groups increases linearly with the size of the group and only applies to discrete rotations with a limited number of angles. This project resulted in a novel patch embedding method for transformers that is robust to the continuous global rotation of the input, shows minimal memory overhead, and can achieve acceptable performance on several benchmark datasets for image recognition with a minor drop in accuracy.

Scientific outcomes

The overview of the present literature and techniques for rotation invariance, together with the insights found in rotation invariant and equivariant transformers, was presented both at the machine learning and optimization (MLO) lab group meeting at EPFL and computer vision, perception, and cognition (CVPC) lab at ZHAW.

Future plans

This research work might be published in future publications, or there are chances for collaboration between participants at the postdoc level.

Self-assessment

Please provide your own final assessment of the effective progress against the goals stated in the proposal, according to the following points:

- **AI Excellence**: The visit helped me to have a better understanding of vision transformers and rotation invariant and equivariant models. We found a tokenization technique for vision transformers to gain rotation invariance in object recognition. The proposed method achieved robustness against rotation in object recognition tasks with competitive performance with a minor compute overhead and drop in accuracy.
- **Scientific step-up**: I worked in a fully functional and productive team with a positive dynamic and vibe. I tried my best to learn leadership and research skills from Martin. The visit contributed to my personal development by learning new scientific content, working on a new problem, and getting in touch with scientists at the summit of their niche.
- **Suitability of the host**: I felt welcomed in the MLO lab and experienced an excellent work environment. In addition, I had good access to Martin and worked closely with his Ph.D. student. Hence, I enjoyed high-level and low-level support for concept development and hands-on work. I also had access to MLO's compute resources.
- **Suitability of the visit length**: The visit was too short for the project considering the time for paperwork, settling down in Lausanne, and new years' holidays. However, the extension was not also an option due to the other projects' commitments and limitations of migration rules and regulations in Switzerland and universities of applied sciences.

List of publications, meetings, presentations, patents,...

Presentations:

- Equivariant Neural Networks: machine learning and optimization lab (MLO), 4th of January 2022, Lausanne, Switzerland
- Equivariant Neural Networks: computer vision, perception and cognition group (CVPC), 23th of February 2022, Winterthur, Switzerland

Additional comments

This is the report of the first project supported by TAILOR's connectivity fund. From initial communications, it wasn't clear that the funding doesn't support working hours. Hence, the budget from other projects, personal overtime, and holidays of the applicant filled the working hours for this project. Unfortunately, this was not healthy for the applicant and closed the opportunity to continue the project after the end of the stay. Therefore, it is highly recommended to check the existence of base funding in the future to guarantee fair working conditions for applicants and ensure the project's success without early stopping.

1st International Joint Conference on Learning & Reasoning

25 - 27 October 2021
Virtual

ILP 2020-21

30th International Conference on Inductive Logic Programming

NeSy 2020-21

15th International Workshop on Neural-Symbolic Learning and Reasoning

StarAI 2020-21

10th International Workshop on Statistical Relational AI

AAIP 2020-21

10th International Workshop on Approaches and Applications of Inductive Programming

1ST INTERNATIONAL JOINT CONFERENCE ON LEARNING & REASONING

7 March 2022

Nikos Katzouris

National Center for Scientific Research “Demokritos”

The rapid progress in machine learning has been the primary reason for a fresh look in the transformative potential of AI as a whole during the past decade. A crucial milestone for taking full advantage of this potential is the endowment of algorithms that learn from experience with the ability to consult existing knowledge and reason with what has already been learned. Integrating learning and reasoning constitutes one of the key open questions in AI, and holds the potential of addressing many of the shortcomings of contemporary AI approaches, including the black-box nature and the brittleness of deep learning, and the difficulty to adapt knowledge representation models in the light of new data. Integrating learning and reasoning calls for approaches that combine knowledge representation and machine reasoning techniques with learning algorithms from the fields of neural, statistical and relational learning.

Aiming to address such challenges, the 1st International Joint Conference on Learning & Reasoning (IJCLR 2021), which was sponsored by TAILOR, took place as a virtual conference from October 25-27 2021. IJCLR 2021 brought together, for the first time, four international conferences and workshops, addressing various aspects of integrating machine learning and machine reasoning:

- The 30th International Conference on Inductive Logic Programming (ILP).
- The 15th International Workshop on Neural-Symbolic Learning & Reasoning (NeSy).
- The 10th International Workshop on Statistical Relational Artificial Intelligence (StarAI).
- The 10th International Workshop on Approaches and Applications of Inductive Programming (AAIP).

The conference featured presentation of cutting-edge research in a number of parallel sessions for each participating event, in addition to a number of joint invited talks from leading researchers in the field, tutorials, poster sessions and a panel discussion on "Future Challenges in Learning & Reasoning".

The virtual conference was organized by the Institute of Informatics of the National Center for Scientific Research (NCSR) "Demokritos", in Athens, Greece. The conference had more than 550 registrants and a very high participation overall.

The video recordings from IJCLR 2021 are available online for everyone to watch.

← Previous Post Next Post →

Funded by EU Horizon 2020.

Contact: Coordinator Fredrik Heintz,
fredrik.heintz@liu.se.

Final Report

for the research stay of dr. Réka Markovich in CNR, Pisa, 2021 titled
"Abstraction and Implementation: Toward a Context-Dependent Conflict Resolution
Algorithm for the Ethics of AI"

The goal of dr. Markovich's research stay at the CNR in Pisa was to proceed her investigations about a context-dependent conflict resolution algorithm which can be used in Machine Ethics.

The problem of moral disagreement is a serious barrier to the advancement of ethical AI as such [Bostrom 2014, Brundage 2014, Etzioni and Etzioni 2017, Formosa and Ryan 2020, Gabriel 2020]. People agree that the soon-to-be-developed autonomous intelligent systems (AIS) should obey our moral rules, but do not agree about what these rules are. The most well-known results about this problem are those of the Moral Machine experiment [Awad et al., 2018] exemplifying the differences of what behavior of an autonomous system people from all over the world find morally acceptable/desirable. But the situation is not better with the professionals when deciding what norms these systems should follow: as Jobin et al. [2019] wrote after analyzing 84 ethical guidelines for AI, while there were some emerging values, "no single ethical principle appeared to be common to the entire corpus of documents". The shared opinion of ethicist seems to be that AIS's operation should be restricted to those occasional cases where stakeholders actually agree on the moral rules [Anderson, 2011]. The problem of moral disagreement, therefore, threatens the promising aspects of AIS. What is more, if we think of that the technological development cannot be stopped, the conclusion is even (much) more threatening: with moral disagreement we lose moral control over these systems, so the demand for a solution is even more urgent.

Dr. Markovich has been working on a conflict-resolution algorithm which is based on a branch of law's model, the so-called Conflict of Laws. It is part of Private International Law, and its task is exactly to provide rules in international situations where more than one nation's laws could be applied: the Conflict of Laws rules tell which one's should be. Dr. Markovich has already published a paper, "On the Formal Structure of Rules in Conflict of Laws", in which she proposes a formalization of the rules of Conflict of Laws using an approach, language, and semantics taken from the so-called Input/Output Logics [Markovich, 2019]. Input/Output Logic is a rule-based system, a theoretical framework for reasoning about conditional norms and investigating normative systems. [Makinson and van der Torre, 2000]. It is a formalism that, modelling the essential structure of rule-based reasoning, allows the definition and analysis of different possible kinds of normative reasoning. However, it does not really take into consideration the expressivity needed for particular applications, and, consequently, it does not consider computational costs and ease of implementation.

The direct goal of dr. Markovich's research stay was to investigate a family of formal languages that are more expressive and more implementation-oriented: the OWL family of languages (https://www.w3.org/OWL/), and to see how it could be used for the modelling of Conflict of Laws and then later to the conflict-resolution algorithm to be applied in Machine Ethics. OWL (Web Ontology Language) is the most popular formalism in formal ontologies. Ontological modelling tools are very popular in many areas, even in legal informatics [Casanovas et al., 2016]. The OWL languages have a solid logic background, represented by the Description Logics (DLs), a family of logical systems that has allowed the development of efficient reasoners (http://owl.cs.manchester.ac.uk/tools/list-of-reasoners/). As in the case of normative reasoning, a big part of the DLs formalism (the TBox part) is of conditional nature in the language and in the reasoning processes. However, classical DLs rely on classical logical semantics. These are usually considered as not appropriate for modelling normative reasoning which exhibits properties that are not shared by classical logics: defeasible conclusions, different priorities among the norms, context-dependent interpretations… [Rotolo et al., 2017].

Dr. Markovich has chosen to visit ISTI-CNR because of the experts in the areas of Knowledge Representation and Reasoning and in formalisms for the Semantic Web. In particular, she was collaborating with
- Dr. Umberto Straccia (http://www.umbertostraccia.it/cs/)
- Dr. Giovanni Casini (https://www.isti.cnr.it/en/about/people-detail/85/Giovanni_Casini)

These researchers have provided a main contribution to the area of extending DLs with non-classical reasoning [Casini and Straccia 2013, Casini et al., 2019]. Actually, the Defeasible DL system developed by the researcher at ISTI-CNR has already been used to define an architecture to solve potential conflicts between the different l egal codes (national, regional, and so on) valid in Brazil [Rodrigues et al. 2019].

During the research stay, the researchers investigated the applicability of the methods used in [Rodrigues et al. 2019] to the given problem, identifying some limitations in such a proposal, and rethinking a possible approach in the framework of defeasible DLs toward a formalization of the CoL conflict resolution methodology. The desired properties of the mechanism are correctness, ease of implementation and relatively low computational costs.

The main goal of the research stay was to investigate whether the use of Defeasible DLs is a feasible approach toward the formalization of the CoL. Dr. Markovich and dr. Casini worked on an alternative model of Conflict of Laws, with the support of dr. Straccia, to proceed toward the conflict-resolution algorithm with the desiderata indicated above. The work focused on some portions of the Hungarian CoL, that dr. Markovich identified as particularly problematic from the point of view of formalization and reasoning. They identified an appropriate vocabulary (concepts, roles) and a first draft of an appropriate T-box (the part of the ontology containing all the general rules and constraints expressed, in this case, by a CoL code. We identified those norms in Conflict of Laws that do not seem expressible in description logic. We are still looking for a correct formalization in DLs, but if we do not find any, we are considering combining the DL ontology with a Datalog rule-based system, that would allow the formalization of such norms. Some complex constructors in the language turned out to be necessary in order to develop a proper ontology formalizing CoL rule: qualified number restrictions, inverse and transitive roles and nominals, among others. An expressive DL like SHOIQ [Horrocks and Sattler 2005] appears sufficiently expressive for the task.

On the one hand, as foreseen in the proposal’s review, one month was enough only to start the process and identify some cornerstones and directions. Therefore, the researchers plan to submit a TAILOR proposal in 2022 too to continue the collaboration. On the other hand, however, the visit did already provide dr. Markovich with knowledge on the practical implementation of normative reasoning systems for conflict resolution and the identified aspects crucially contributed to the elaboration of her research plans about the algorithm, which she implemented in an ERC Starting Grant proposal submitted in January of 2022. Also, the researchers are preparing a submission to JURIX or its accompanying workshops on AI and Law, and will later a journal submission.

References:

[Anderson, 2011] Anderson, S. (2011). Philosophical Concerns with Machine Ethics. In M. Anderson & S. Anderson (Eds.) *Machine Ethics*. Cambridge: Cambridge University Press. pp. 162–167.
[Awad et al., 2018] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, F. Bonnefon, and I. Rahwan. The Moral Machine experiment. *Nature*, 536:59., 2018
[Bostrom, 2014] Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
[Brundage, 214] Brundage, M. 2014. Limitations and Risks of Machine Ethics. *Journal of Experimental & Theoretical Artificial Intelligence* 26(3): 355–372.
[Casanovas et al., 2016] P. Casanovas et al., Special Issue on the Semantic Web for the Legal Domain. Editorial: The Next Step. *Semantic Web*, 7(3), 2016
[Casini and Straccia, 2013] G. Casini and U. Straccia. Defeasible Inheritance-Based Description Logics. *Journal of Artificial Intelligence Research*, 48, pp. 415-473, 2013.
[Casini et al., 2019] G. Casini, U. Straccia, T. Meyer. A polynomial Time Subsumption Algorithm for Nominal Safe ELO⊥ under Rational Closure. *Information Sciences*, 501, pp. 588-620, 2019.
[Casini et al., 2015] G. Casini, T. Meyer, K. Moodley, U. Sattler, I. Varzinczak. Introducing Defeasibility into OWL Ontologies. In *Proceedings of the International Semantic Web Conference* (ISWC - 2015), Vol. 2, pp. 409-426, Springer, 2015.
[Etzioni and Etzioni, 2017] Etzioni A. and Etzioni O. 2017. Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics* 21: 403–418.
[Formosa and Ryan, 2021] Formosa, P., Ryan, M. Making moral machines: why we need artificial moral agents. *AI & Society* 36, 839–851 (2021)
[Gabriel 2020] Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30: 411–437
[Horrocks and Sattler 2005] Ian Horrocks and Ulrike Sattler. A tableaux decision procedure for SHOIQ, *Proceedings of IJCAI 2005*, pp. 448-453, Professional Book Center, 2005.
[Jobin et al., 2019] Jobin, A.; Ienca, M.; and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1:389–399.
[Makinson and van der Torre, 2000] D. Makinson and L. van der Torre. Input/output logics. *Journal of Philosophical Logic*, 29(4), pp.383–408, August 2000.
[Markovich, 2019] R. Markovich. On the Formal Structure of Rules in Conflict of Laws. *Legal Knowledge and Information Systems. Frontiers in Artificial Intelligence and Applications, Vol 322*, pp. 199 - 204. IOS Press Books, 2019
[Rodrigues et al. 2019] C. M. O. Rodrigues, E. P. da Silva, F. Freitas, I J. da Silva Oliveira, I. Varzinczak. LEGIS: A Proposal to Handle Legal Normative Exceptions and Leverage Inference Proofs Readability. *FLAP*, 6(5), pp. 755-780, 2019.
[Rotolo et al., 2017] A. Rotolo et al., Deliverable D1.1 MIREL Project, Report, 2017, http://www.mirelproject.eu/publications/D1.1.pdf

Basic Information

Project title: Design of Matheuristic Techniques for Timetabling Problems
Period of project: 13/09/2021 - 30/06/2022
Period of reporting: 13/09/2021 - 30/06/2022
Author(s): Roberto Maria Rosati (Uniud)
Other key persons: Christian Blum (IIIA-CSIC), Andrea Schaerf (Uniud)
Organization: Dipartimento Politecnico d’Ingegneria e Architettura, Università degli Studi di Udine, via delle Scienze 206, Udine (UD), ITALY.
Host organization: Institut d’Investigació en Intel·ligència Artificial, Consejo Superior de Investigaciones Científicas, Campus de la UAB, E-08193, Bellaterra (Barcelona), SPAIN.

Public summary

Matheuristics are a class of metaheuristic algorithms that have recently emerged as a promising research branch in combinatorial optimization. They consist in the hybridization of heuristic procedures with exact methods. Thanks to this collaboration supported by TAILOR connectivity fund, we designed and applied “construct, merge, solve and adapt” (CMSA), a novel method based on instance reduction that is under research at IIIA in Barcelona [1], to a variety of combinatorial optimization problems previously studied at University of Udine. Our CMSA obtained very good results on the dominating sets problem (MDDSP), a NP-hard problem on graphs with applications in management of wireless sensors networks, and on a real-world bus driver scheduling (BDS) problem with complex break constraints. Though the results are preliminary, experimental data clearly shows that CMSA is a flexible general approach able to produce advancements on a wide set of problems.

Research objectives

Objectives

Our objective for this research stay was to contribute to new advancements in the design and implementation of CMSA, a novel general-purpose algorithm, and to improve the current state-of-the-art algorithms for certain scheduling and timetabling problems.

Impact

Applications of CMSA concern potentially any combinatorial optimization problem. It is, indeed, a general procedure that envelopes some problem-dependent components (e.g. a probabilistic constructive heuristic for the “construct” phase and an ILP model for the “solve” phase) inside a general, problem-agnostic metaheuristic procedure. Though our main focus was in scheduling and timetabling, we showed that CMSA works well and even improves the state-of-the-art on a variety of problems.

Technical approach

Detailed description

Matheuristics are a particular class of metaheuristics which consist in the hybridization of

exact methods, such as mixed integer programming (MIP) solvers, with heuristics algorithm. Often, the combination of these two different approaches produces good results for the solution of combinatorial optimizations problems.
CMSA works roughly as follows: at each iteration, a number of solutions are generated probabilistically, usually through a randomized greedy procedure, and solutions components are merged into a sub-instance, which is solved by an exact solver. After each iteration an aging procedure is applied to progressively discard solution components that are not chosen by the exact solver. Those four steps are labeled as: construct, merge, solve, and adapt.

We developed our version of CMSA in C++, as framework that already contains the main CMSA loop, while the problem-specific components, namely the probabilistic solution generator and the integer linear programming (ILP) model for the exact solver, are declared as virtual methods and are left to the user for implementation. We took advantage of the templates and inheritance functionalities of C++, to enhance modularity and reusability of the code for different problems.

So far, we applied CMSA on three different problems: sport timetabling, maximum disjoint dominating sets and bus driver scheduling [3]. For the solution phase, we developed an ILP model which was solved by Cplex 20.1 for those problems. For the generation of the solutions in the construct phase, the simulated annealing that we designed for the ITC2021 competition was used in the case of the sport timetabling [2], while for the MDDSP and for the BDS we used a randomized greedy procedure.

Finally, the CMSA takes as input various parameters, so that a statistically-principled tuning is necessary. We adopted two different tools: json2run and irace.

Scientific outcomes

The main scientific results obtained during and as an immediate consequence of the research stay are listed hereafter. Related publications are listed in Section “List of publications, meetings, presentations”.

- For the maximum disjoint dominating sets problem:
  - We showed that CMSA improves previous state-of-the-art results.
  - It is, furthermore, the first metaheuristic method for this problem.
  - We published a new set of instances, as well as an instance and solution validator.
- For the Bus Driver Scheduling Problem:
  - We showed that CMSA performs better than all other metaheuristics methods previously developed for this problem and improve the overall state-of-the art (including Branch and Bound) on larger instances.

Additionally, we are still studying applications of CMSA to other problems such as complex sport timetabling, as well as adaptive versions of CMSA.

Future plans

Future work goes in two directions. On the one hand, we plan to extend the works that we have realized, on the other hand we plan to test CMSA on new problems.

Regarding extensions for MDDSP and BDS, we plan to test the algorithm on larger problem instances. Additionally, we want to incorporate other available greedy heuristics in the construction phase, with a probability assigned to each greedy. The probability will be either tuned or adapted during the search. Finally, we plan to investigate which CMSA components are having the most impact through a component-based analysis and to perform a feature-based tuning to face sensitivity to instance size.

Concerning new problems, we plan to work on the application of CMSA to other complex scheduling problem. This includes the sport timetabling problem from ITC2021, or other complex problems that have been studied at University of Udine, such as examination timetabling or home health care routing and scheduling.

Progress against planned goals

Planned milestone	Planned date	Actual date
Start of project	September 2021	13th September 2021
Repository created	October 2021	3rd October 2021
Conference paper	February 2022	- MDDSP: submitted on 5th April 2022, presented at Metaheuristics International Conference (MIC) on July 2022 - BDS: submitted on 10th August 2022 for AIxIA
Main solver release	May 2022	April 2022
Journal paper	June 2022	Plan to submit journal paper for MDDSP work in September/October 2022.
End of Visit	June 2022	June 2022

List of publications, meetings, presentations

- Rosati, R.M., Bouamama, S., Blum, C., “Construct, Merge, Solve and Adapt applied to the maximum disjoint dominating sets problem”:
  - presented at Metaheuristics International Conference (MIC2022, https://www.ants-lab.it/mic2022/), in Siracusa, Italy, 11-14 July 2022.
  - It will be published in *post-proceedings* in Lecture Notes in Computer Science series by Springer.
- Rosati, R.M., Kletzander, L., Blum, C., Musliu, N., Schaerf, A., “Construct, Merge, Solve and Adapt Applied to a Bus Driver Scheduling Problem with Complex Break Constraints”:
  - the work has been submitted for conference AIxIA (https://aixia2022.uniud.it, Udine, Italy, 28th November-2nd December 2022)

Additional comments

The author is grateful to the TAILOR connectivity fund, that sponsored this collaboration of high scientific profile and let him strengthen its scientific network at a European level.

References

[1] Blum, C., Pinacho, P., López-Ibáñez, M., Lozano, J.A., 2016. Construct, merge, solve & adapt a new general algorithm for combinatorial optimization. Computers & Operations Research 68, 75–88.

[2] Rosati, R. M., Petris, M., Di Gaspero, L., & Schaerf, A. (2021). Multi-Neighborhood Simulated Annealing for the Sport Timetabling Competition ITC2021. In Proceedings of the 13th International Conference on the Practice and Theory of Automated Timetabling-PATAT (Vol. 2).

[3] Kletzander, L., & Musliu, N. (2020, June). Solving large real-life bus driver scheduling problems with complex break constraints. In Proceedings of the International Conference on Automated Planning and Scheduling (Vol. 30, pp. 421-429).

Report on the TAILOR connectivity fund supported research visit "Neuro-symbolic integration for graph data"

Manfred Jaeger
Aalborg University

June 13, 2022

1 Technical data

Visitor: Manfred Jaeger, Aalborg University
Host: Andrea Passerini, Trento University
Duration: March 14 - April 15, 2022

2 Activities during the visit

The objective of the visit was to serve as a seed event that initiates a longer term research program on neuro-symbolic methods for learning and reasoning with graph data. To this end, the activities during the visit consisted of the following main components:

Development of research agenda: Andrea and Manfred had nearly daily meetings for knowledge exchange and the development and initiation of concrete research plans. The discussions were mostly based on prior joint work [3, 2], and a recent tutorial article that exhibits the already existing close linkage between graph neural networks and the statistical relational learning framework of relational Bayesian networks (RBNs) [1]. Two specific lines of investigation were initiated:

- Neural network learning of multi-relational slotchain dependencies: in our earlier work [3] we have developed an approach to learn how a target label of an entity depends on attributes of other entities connected by a certain chain of relational connections (e.g., learn that the *thesis topic* of a student depends on the *research area* attribute of professors the student is connected to by the chain of *takes(student, course)*, *teaches(course,professor)* relations). During the visit we developed an adaptation of our method, which was originally developed in the context of statistical relational learning, to graph neural networks (GNNs).
- Integration of RBNs with GNNs: based on preliminary findings detailed in [1] we developed ideas and a proof-of-concept implementation for how GNNs trained in a neural network framework such as Pytorch geometric can be exported as an RBN, and thereby become amenable to a richer class of reasoning tasks than the specialized classification or link prediction tasks that GNNs usually are limited to.

Recruitment of Master thesis students: to support the two lines of research described above, two master students at Trento university were recruited who will write their master theses on these respective topics. Both students have procured an ERASMUS+ grant to conduct part of their thesis research as visiting students at Aalborg university. Both also are possible candidates for continuing this line of research as PhD students.

Further collaboration initiatives: apart from neuro-symbolic integration as the core area of interest, two other forms of collaborations took shape:

- During Manfred's visit it turned out that there was a strong common interest with Luciano Serafini and Sagar Malhotra from Fondazione Bruno Kessler on the topic of projectivity of statistical relational models. This common interest was explored in an extensive exchange of ideas with a view towards a continuing research collaboration.
- We explored the possibility of participating in a EU grant proposal on the topic of robustness and verification of neural networks in aviation.

3 Continuing activities

Since the main objective of the visit was to initiate a longer term research collaboration, the main outcomes of the visit are the continuing activities:

- The two master students working on the topics described above will visit Manfred at Aalborg in the periods June 15 - September 15, and September 15 - December 15, respectively. They will be co-supervised by Andrea and Manfred, and are expected to complete their Master theses in Trento within a short period after their return. It is envisioned that at least one of these students will continue under our co-supervision as a PhD student.
- Luciano, Sagar and Manfred continue to have regular online meetings on the subject of projective models. It is envisioned that Sagar may come to Aalborg as a visiting PhD student, and that this collaboration extends into a post-doc phase of Sagar.

References

[1] Manfred Jaeger. Learning and reasoning with graph data: Neural and statistical-relational approaches. In *International Research School in Artificial Intelligence in Bergen (AIB 2022).*

[2] Manfred Jaeger, Marco Lippi, Andrea Passerini, and Paolo Frasconi. Type extension trees for feature construction and learning in relational domains. *Artificial Intelligence*, 204:30–55, 2013.

[3] Marco Lippi, Manfred Jaeger, P. Frasconi, and A. Passerini. Relational information gain. *Machine Learning*, 83:219–239, 2011.

New uncertainty quantification algorithms in metric spaces with strong theoretical guarantees and computational feasibility

Short report,

Marcos Matabuena, CiTIUS, Universidad de Santiago de Compostela

Abstract

Statistical and machine learning approaches transform industry, healthcare, and the digital marketplace. A key point in applying these data analysis strategies in decision-making is quantifying the limits of trustworthy statistical and machine learning models with an uncertainty analysis. Conformal inference maybe is the most dominant general framework for prediction intervals and obtains a predictive model outputs reliability measure. However, one of the significant open challenges of conformal inference is to build new general methods that remain valid with multivariate responses or even complex ones such as curves or graphs that may arise in modern medical applications of precision medicine. In addition, from a computational point of view, there are critical problems in building the prediction intervals in large and high-dimensional datasets. There is also a lack of solid theoretical results in many settings. The purpose of this article is to try to break this gap in the literature by proposing a new uncertainty quantification method when the response of the regression algorithm lives in a metrics space, and it can handle straightforwardly and efficiently large datasets with new theoretical strong results that it cannot reach with traditional conformal inference approaches. Two algorithm versions are proposed; the first is global and designed to handle the homoscedastic case, while the second is local and works in the heteroscedastic signal noise regime. The potential advantages of the new method are illustrated in the context of the global Fréchet regression model when to the best of our knowledge, any method exists to propose a level set of uncertainty. In this case, we analyze relevant medical examples with complex statistical objects as responses that appear in modern precision and digital medicine problems.

1 Introduction

A critical issue when we model many of these real world applications, is the considerable uncertainty in the outputs of a predictive model. In this context, to create trustworthy machine learning models, quantifying this uncertainty is crucial to determine the models' limits and and to specify when we can obtain reliable results [18].

In medicine, it is common that the patient's responses to the same clinical treatment can present a high variability at the individual level [19]. The same large variability arises in many machine learning models' output when we try to predict future patient status in the long term to improve the early diagnosis of diseases and screening campaigns. For example, our recent work on diabetes [11] tries to predict the A1C (glycosilated hemoglobin; the primary diabetes biomarkers to control and diagnose the disease) in five years. However, the considerable uncertainty in the problem does not allow our algorithm to obtain reliable results along the subset of patients, which is crucial to determine the development of diabetes disease accurately. Then, with our uncertainty analysis, we will provide a new stratification of subjects based on their glucose uncertainty, promoting new personalized patient follow-ups with more complex medical tests and new interventions that avoid the development of diabetes. In our work, to develop this modeling goal, we have extended conformal inference techniques in the setting of missing responses, allowing us to obtain prediction intervals of the A1C predictions.

Conformal inference techniques [20] constitute a general and unique framework for measuring the uncertainty of statistical and machine learning models with the estimation of prediction intervals in multiple situations. This research topic has become of great interest in the statistical and machine learning community in recent years. Multiple extensions of the usual techniques have appeared to model causality, missing data, and survival analysis problems [9, 11], and as even dependent-data situations [2]. Despite noteworthy advances in this area, as far we are concerned, there is only one work handling multivariate data [8]. In the case of classification problems, the first contributions to handle the multivariate responses have also been recently proposed (see for example [1]).

A critical aspect when increasing the realism of conformal inference techniques is to provide a local interval of predictions. Achieving this goal requires estimating local variability [7] or running traditional conformal inference algorithms locally [6, 10].

In personalized medicine applications, it is increasingly common to register patients' health with complex statistical objects such as curves to record patients' physiological functions and graphs, to measure the dynamic evolution of the patients' brain connectivity. For example, in our recent work, we have coined the concept of "glucodensity" [13], a new compositional distributional representation of a patient's glucose profiles that improve the existing methodology data analysis in the area. This type of representation was also helpful to obtain better results with accelerometer data [5, 12]. In this context, Fréchet's linear model has recently been proposed to analyze these predictors [15]. However, so far, there is no methodology to provide an interval of predictions in this setting. In addition, the disease definitions often depend on several biomarkers, and it is crucial to estimate the uncertainty about the evolution of patients' conditions, introducing the correlation structure of multiple biomarkers from a multivariate perspective.

There is hardly literature on conformal inference with stochastic processes, and this is restricted to the case of geometry of supremum norm, e.g., in the case of simulation models of epidemics [14] or with euclidean functional data objects [3].

Given the need for predicting several variables simultaneously and even complex objects such as curves and graphs, this research project aims to provide new uncertainity quantification stratifications which are valid for multivariate and even complex data in metrics spaces. A key point in the new uncertainity quantification strategies that we will propose is that they allow to obtain local prediction intervals.

In order to illustrate the need to quantify the uncertainty and provide predictions beyond the conditional mean, Figure 1 shows the levels set of the uncertainty of the linear regression model with the response to the distributional representations of glucose profiles [11]. We show that the conditional mean is constant with an increase in body mass index; however, the uncertainty increases dramatically with a variation of the body-mass index. Unfortunately, to the best of our knowledge, we do not know any prior methodology to address the uncertainty quantification for this problem.

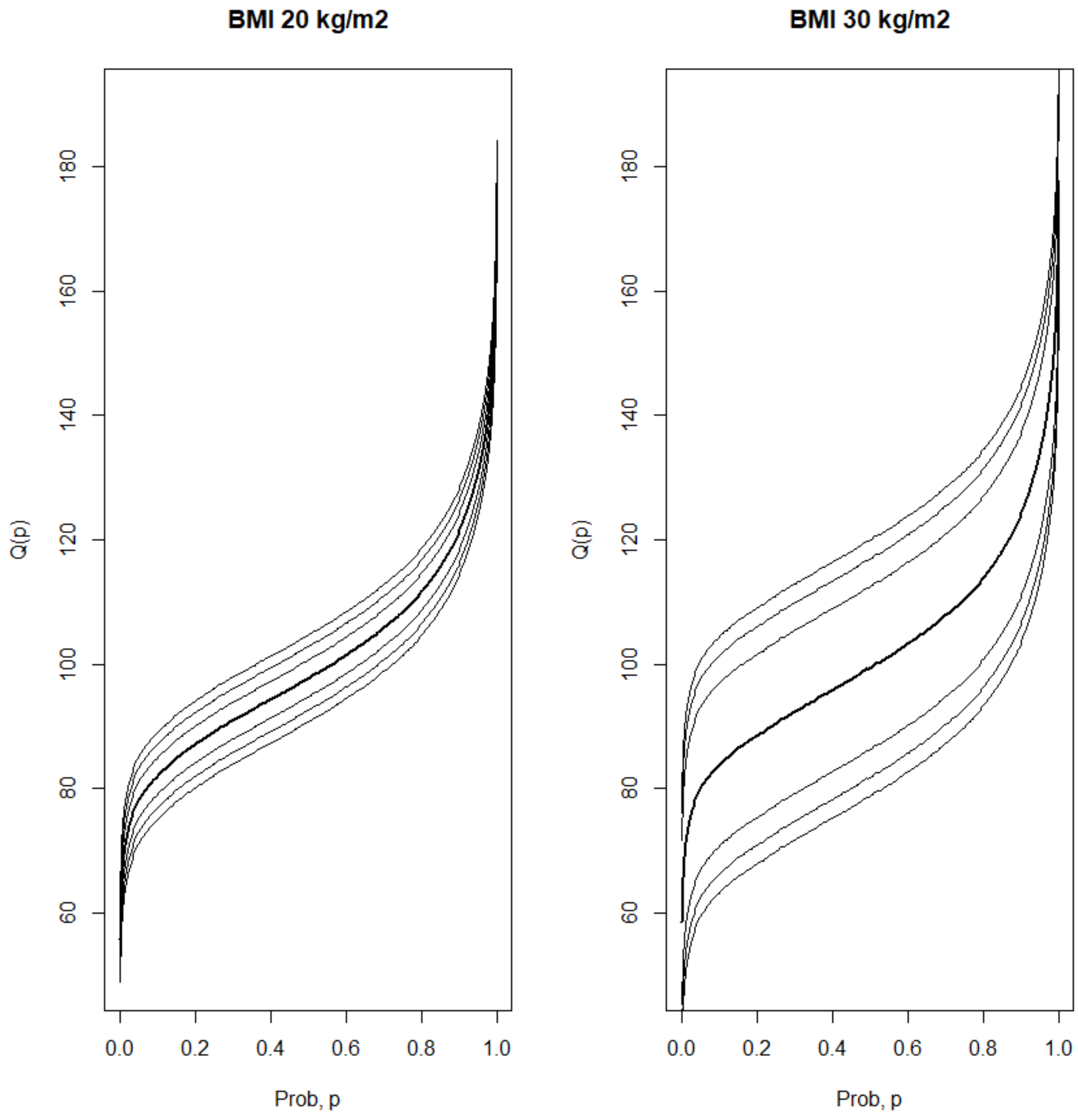


Fig. 1: Levels set of predicting distributional representations of glucose profiles according to our algorithm in two individuals: i) Individual with a body mass index of 20 mg/dL ii) Individual with a body mass index of 30 mg/dL..

1.1 Notation and problem definition

Suppose that we observe a random sample i.i.d $\mathcal{D}_n = \{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} : 1 \leq i \leq n = n_1 + n_2\}$ from the random variable $(X, Y) \sim P = P_X \otimes P_{Y|X}$, where in our setting $\mathcal{X} = \mathbb{R}^p$, and $\mathcal{Y}$ denote a arbitrary separable metric space equipped with a specific metric $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$.

In this paper, to connect the spaces $\mathcal{X}$ and $\mathcal{Y}$, we consider a function $m:\mathcal{X}\to\mathcal{Y}$, that in practice, is the regression function we consider to be estimated. Mathematically, we formally this problem as solve the following M-estimator problem:

$$m(\cdot)=\arg\min_{y\in\mathcal{Y}} M(\cdot,y), \tag{1}$$

where in practice, $m(\cdot)=\arg\min_{y\in\mathcal{Y}} M(\cdot,y)=E\left(d^2(Y,y)\,|X=\cdot\right)$ is the conditional Fréchet mean or the Fréchet median $m(\cdot)=\arg\min_{y\in\mathcal{Y}} M(\cdot,y)=E\left(d(Y,y)\,|X=\cdot\right)$, that constitute the natural mathematical notions of center in métric spaces.

Given a new random point X_{n+1}, our modeling goal is to estimate a prediction region of the random response variable Y, such that $C^{\alpha}(X_{n+1})\subset\mathcal{Y}$, mimic the following property of the population bands,

$$C^{\alpha}(X_{n+1})=\mathcal{B}\left(m(X_{n+1}),r(X_{n+1})\right)=\arg\min_{\mathcal{B}(m(X_{n+1}),r):\mathbb{P}(Y\in\mathcal{B}(m(X_{n+1}),r)|X=X_{n+1})\geq 1-\alpha} r, \tag{2}$$

where $\mathcal{B}(m(x),r(x))$ denote a ball of mean $m(x)$, and radius $r(x)$. $\mathbb{P}$ stands over the randomness of both the random pair (X_{n+1},Y) and the full procedure.

Our final estimator takes the following structure

$$\tilde{C}^{\alpha}(x)=\mathcal{B}\left(\tilde{m}(x),\tilde{r}(x)\right)\ \ \forall x\in\mathcal{X}, \tag{3}$$

where $\tilde{m}(x)$, and $\tilde{r}(x)$ denote two estimators of the center and radius of the ball from the random sample $\mathcal{D}_n$, that hold the following property

$$\int_{\mathcal{X}}\mathbb{P}\left(Y\in C^{\alpha}(x)\,\triangle\tilde{C}^{\alpha}(x)\,|X=x,\mathcal{D}_n\right)P_X(dx)=o_p(1), \tag{4}$$

that is, that we recover in some uniform sense the right optimal region sets.

In the rest of this paper, we will split $\mathcal{D}_n$ in three disjoint subset, $\mathcal{D}_n=\mathcal{D}_{train}\cup\mathcal{D}_{test}$, where $|\mathcal{D}_{train}|=n_1$, $|\mathcal{D}_{test}|=n_2$. We denote the set of indexes $\mathcal{J}_{\mathcal{D}_{train}}:=\{i\in\{1,\dots,n\}:(X_i,Y_i)\in\mathcal{D}_{train}\}$, and $\mathcal{J}_{\mathcal{D}_{test}}:=\{i\in\{1,\dots,n\}:(X_i,Y_i)\in\mathcal{D}_{test}\}$.

1.2 Paper motivation: Quantify the uncertainty with global Fréchet regression model in medical applications

Complex statistical objects such as graphs, strings, probability distributions, compositional data, or other functional data objects appear naturally in recording information in medical and related fields. For example, practitioners can register with new and more sophisticated profiles of patients' conditions using these more complex mathematical constructions, enabling improved and refined clinical decision-making using new, more advanced clinical models with these objects.

A general framework for predicting these complex objects is provided by the recent global Fréchet model, which would be the equivalent in metric spaces to the notion of the regression model in Euclidean spaces. Bellow, we introduce the mathematical details briefly.

Let $(X,Y)\sim P$ be a multivariate random variable, where $X\in\mathbb{R}^p$, and $Y\in\mathcal{Y}$ a separable bounded metric spaces.

For a fixed $X = x$, the population value of the Global-Fréchet model is given by solving the following optimization problem:

$$m(x) = \arg\min_{y \in \mathcal{Y}} E\left[\left[1 + (X - x)\Sigma^{-1}(x - \mu)\right]\left(d^2(Y, y)\right)\right], \tag{5}$$

where $\Sigma = Cov(X, X)$, and $\mu = E(X)$.

Suppose that a random sample i.i.d $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ from a distribution P is available, we can obtain a estimation of conditional mean as follows:

$$\tilde{m}(x) = \arg\min_{y \in \Omega} \frac{1}{n}\sum_{i=1}^{n}\left[1 + (x - X_i)\tilde{\Sigma}^{-1}\left(x - \overline{X}\right)d^2(y, Y_i)\right], \tag{6}$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, and $\tilde{\Sigma} = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^T\left(X_i - \overline{X}\right)$.

The following results characterize the ratios of convergence of a global Fréchet regression model.

Proposition 1. *Suppose that, for a fixed $x \in \mathbb{R}^p$ and the conditions s $2 - 4$ (see Supplemental Material for more details) are hold. Then*

$$d(\tilde{m}(x), m(x)) = O\left(n^{-\frac{1}{2(\beta-1)}}\right). \tag{7}$$

Furthermore, for a given $B > 0$, if $5 - 7$ hold,

$$\sup_{\|x\| \leq B} d(\tilde{m}(x), m(x)) = O\left(n^{-\frac{1}{2(\alpha-1)}}\right), \tag{8}$$

for any $\alpha' > \alpha$.

Recently a new random forest was proposed in this setting, extending this M-estimator theory to the context of infinite-order U-statistics [16].

In both cases, to the best of our knowledge, no statistical methodology exists to provide regions of predictions. Bellow, we introduce the three examples that we use to illustrate the potential of our proposal with the Global Fréchet-regression model.

1.2.1 Laplacian space graphs

Neuroimaging and related fields are another critical scientific branch to obtain relevant examples that analysis of metric spaces and, in particular, graphs connectivity brain structures can drive promising scientific progress about how the brain works and consequently draw new insights on how we can optimize brain behavior.

In this paper, we consider the space $\mathcal{Y} = (\Omega, d)$ where Ω is the set of networks with a fixed number, say r, of nodes. One can view networks as adjacency matrices, graph Laplacians equipped with the Frobenius metric $d_{\text{FRO}}(A, B) = \left(\sqrt{\sum_{i=1}^r \sum_{j=1}^r (A_{ij} - B_{ij})^2}\right)$ for all $A, B \in \mathcal{Y}$.

Recently, in the setting of the global Fréchet-regression model with this metric, an efficient projection strategy was proposed in [22] to estimate the conditional mean regression function.

Predicting brain graph structures is a fundamental problem in a medical image. Uncertain quantification is a significant step in mathematical modeling, for example, to determine whether cerebral connections vary over different test conditions or stimuli with a certain confidence level and as a consequence of reliability.

1.2.2 Probability distributions of physical activity data

In this work, we summarized the information of dynamic wearable time series in their specific representation in the space of univariate probability distributions. In particular, we use data from a continuous glucose monitor and consider the density function as a representation. Our recent work [11] showed the potential advantages of considering this representation in different predictive tasks concerning existing diabetes summary metrics to predict some diabetes biomarkers.

Mathematically, we consider the space $\mathcal{Y} = (\Omega, d)$ where Ω is the set of univariate probability distributions on a compact support in $T \subset \mathbb{R}$. We choices as metric popular 2-Wasserstein distance $d_{\mathcal{W}_2}$ that is defined as $d^2_{\mathcal{W}_2}(F, G) = \int_0^1 \left(F^{-1}(t) - G^{-1}(t)\right)^2 dt$, for all $F, G \in \mathcal{Y}$.

The scientific challenge of predicting these representations allows us to elucidate which factors modify glucose profiles in the whole range of both high and low concentrations, i.e., hypo- and hyperglycemia. In this mathematical modeling task, determining the predictive limits of the algorithms is crucial to know how variable the glucose values are with the modification of some variables related to the patients' status.

1.2.3 Multivariate Euclidean data

Multivariate Euclidean data is another critical example in real applications, where sufficiently satisfactory methods that scale computationally efficiently and adapt adequately to the local geometry of the data have not yet been provided.

In this example, consider the space $\mathcal{Y} = (\Omega, d)$ where $\Omega = \mathbb{R}^p$, and we use as a metric the Mahalanobis distance that introduce the local geometry of the response variable on the randon variables Y as follows, $d(x, y) = \sqrt{(x - \mu)\, \Sigma^{-1} (x - \mu)}$, where $\mu = E(Y)$ and $\Sigma = Cov(Y, Y)$.

The example introduced here is also related to Diabetes Mellitus Disease and $p = 2$. Using a large dataset, we try to predict the biomarkers used to control and diagnose the disease FPG together with the physical activity levels of individuals under multivariate response linear models. The new methods have the potential to determine the model's predictive limits in different clinical phenotypes of patients. In case of large uncertainty, we must use more complex models or, if necessary, resort use more complex diabetes biomarkers or medical tests to predict the mentioned variables.

1.3 Summary of methodological contributions

We will propose two general uncertainty quantification algorithms in the context of regression models that work in the setting of response Y take values in separable metrics spaces $\mathcal{Y}$. More specifically, we propose two specific algorithms in basis on the signal-noise regime of the basic regression model specified by the function m.

1. Homocedastic set-up: We assume that $\mathbb{P}(Y \in \mathcal{B}(m(x), r) | X = x) = \phi(r)$, that is the probability mass of ball is invariant to the point $X = x$ selected.

2. Heterocedastic set-up: We assume that $\mathbb{P}(Y \in \mathcal{B}(m(x), r) | X = x) = \phi(r, x)$, that is the probability mass of ball depend locally point $X = x$ selected.

A primary characteristic of new algorithms proposed is that it works independently for any predictive learning algorithm that, in practice, is specified by the problem of estimating the function m.

Our theoretical results are summarized bellow:

1. We introduce a consistency results so in the homoscedastic and heteroscedastic case that we can recover the optimal regions set in some uniform sense:

$$\int_{\mathcal{X}} \mathbb{P}\left(Y \in C^{\alpha}\left(x\right) \triangle \tilde{C}^{\alpha}\left(x\right) | X = x, \mathcal{D}_n\right) P_X\left(dx\right) = o_p\left(1\right), \tag{9}$$

   The previous result are stronger than those presented in the literature for the univariate response set-up that not conditioned to the random sample $\mathcal{D}_n$ (see for example [4]).

2. In heterocedastic and homocedastic case, fixed $X = x$, according the convergence rate of estimator $\tilde{m}$ selected of function m, and the the radius $r\left(x\right)$

   we establish the following rate of convergence for the problem of estimated predictive region set

$$E\left(\left|\mathbb{P}\left(Y \in \tilde{C}^{\alpha}\left(x\right) | X = x, \mathcal{D}_n\right) - \left(1-\alpha\right)\right|\right) = \cdots \tag{10}$$

2 Our novel uncertainity quantification methods in metrics spaces

2.1 Homocedastic case

In this setting, to obtain $\tilde{C}_{\alpha}(x)$, we propose to use the following two-step algorithm:

Algorithm 1 Uncertainty quantification algorithm homocedastic set-up

1. Estimate the function $m(\cdot)$, $\tilde{m}(\cdot)$ using the random sample $\{(X_i, Y_i) : i \in \mathcal{J}_{\mathcal{D}_{train}}\}$.
2. For all $i \in \mathcal{J}_{\mathcal{D}_{test}}$, evaluate $\tilde{m}(X_i)$ and define $\tilde{r}_i = d(Y_i, \tilde{m}(X_i))$.
3. Estimate the empirical distribution $\hat{F}_{n_3}(t) = \frac{1}{n_3}\sum_{i\in\mathcal{J}_{\mathcal{D}_{test}}} 1\{\tilde{r}_i \leq t\}$ and denote by $\tilde{q}_{1-\alpha}$ the empirical quantile of level $1-\alpha$.
4. Return as estimation of region band $\tilde{C}_{\alpha}(x) = \mathcal{B}(\tilde{m}(x), \tilde{q}_{1-\alpha})$

Bellow introduces some technical assumptions that guarantee that the method is asymptotically consistent.

Assumption 1. *Suppose that the following hold:*

1. $\{(X_i, Y_i)\}_{i\in\mathcal{J}_{\mathcal{D}_{test}}}$ *is* iid *and* $|\mathcal{J}_{\mathcal{D}_{test}}| \to \infty$.
2. $\mathbb{P}(Y \in \mathcal{B}(m(x), r) | X = x) = \phi(r)$.
3. $\tilde{m}$ *is a consistent estimator in the sense that* $E(d(\tilde{m}(X), m(X)) | \mathcal{D}_{train}) \to 0$, *in probability as* $|\mathcal{J}_{\mathcal{D}_{train}}| \to \infty$, *that is, the next random variable hold in probability*

$$\int_{\mathcal{X}} d(m(x), \hat{m}(x)) P_X(dx) = o_p(1). \tag{11}$$

Lemma 2. *Assume that Assumptions 1 are hold. Then*

$$\frac{1}{|\mathcal{J}_{\mathcal{D}_{test}}|} \sum_{i\in\mathcal{J}_{\mathcal{D}_{test}}} |d(Y_i, m(X_i)) - d(Y_i, \tilde{m}(X_i))| = o_p(1) \tag{12}$$

and

$$\sup_{v\in\mathbb{R}^+} \left|\tilde{G}^*(v) - G^*(v)\right| = o_p(1) \tag{13}$$

where $\tilde{G}^*(v) = \frac{1}{|\mathcal{J}_{\mathcal{D}_{test}}|}\sum_{i\in\mathcal{J}_{\mathcal{D}_{test}}} 1\{d(Y_i, \tilde{m}(X_i)) \leq v\}$, $G^*(v) = \frac{1}{|\mathcal{J}_{\mathcal{D}_{test}}|}\sum_{i\in\mathcal{J}_{\mathcal{D}_{test}}} 1\{d(Y_i, m(X_i)) \leq v\}$.

Proof. Observe that

$$\frac{1}{|\mathcal{J}_{\mathcal{D}_{test}}|} \sum_{i\in\mathcal{J}_{\mathcal{D}_{test}}} |d(Y_i, m(X_i)) - d(Y_i, \tilde{m}(X_i))| \leq \frac{1}{|\mathcal{J}_{\mathcal{D}_{test}}|} \sum_{i\in\mathcal{J}_{\mathcal{D}_{test}}} |d(m(X_i), \tilde{m}(X_i))|, \tag{14}$$

where we infer trivially that

$$\frac{1}{|\mathcal{J}_{\mathcal{D}_{test}}|}\sum_{i\in\mathcal{J}_{\mathcal{D}_{test}}} \left|d\left(Y_i, m\left(X_i\right)\right) - d\left(Y_i,\tilde{m}\left(X_i\right)\right)\right| = o_p\left(1\right)$$

.

For the second part, define quantities $R_T = \sup_{v\in\mathbb{R}}\left|G^*\left(v\right)-G\left(v\right)\right|$ and $W = \sup_{x_1\neq x_2}\frac{|G(x_1)-G(x_2)|}{|x_1-x_2|}$.
Let $A = \{i \in \mathcal{J}_{\mathcal{D}_{test}} : \left|d\left(Y_i,\tilde{m}\left(X_i\right) - d\left(Y_i, m\left(X_i\right)\right)\right)\right| \geq \delta\}$.
Fix $x\in\mathbb{R}$. Then

$$\begin{aligned}\left(|\mathcal{J}_{\mathcal{D}_{test}}|\right)\left[\tilde{G}^*\left(x\right) - G^*\left(x\right)\right] \leq \left|\sum_{i\in A}\left(1\{d\left(Y_i,\tilde{m}\left(X_i\right)\leq x\right)\} - 1\{d\left(Y_i,m\left(X_i\right)\right)\leq x)\}\right)\right| + \quad \text{(15)}\\ \left|\sum_{i\in A^c}\left(1\left(\{d\left(Y_i,\tilde{m}\left(X_i\right)\leq x\right))\} - 1\{d\left(Y_i,m\left(X_i\right)\right)\leq x)\}\right)\right| \leq \quad \text{(16)}\\ |A| + \left|\sum_{i\in A^c}\left(1\left(\{d\left(Y_i,\tilde{m}\left(X_i\right)\leq x\right))\} - 1\{d\left(Y_i,m\left(X_i\right)\right)\leq x)\}\right)\right|, \quad \text{(17)}\end{aligned}$$

where the last inequality follows by the fact the difference of two indicators take values $\{-1,1,0\}$. We must note that for $i\in A^c$, $d\left(Y_i,m\left(X_i\right)\leq x\right)-\delta \leq d\left(Y_i,\tilde{m}\left(X_i\right)\leq x\right) \leq d\left(Y_i,\tilde{m}\left(X_i\right)\leq x\right)+\delta$. Therefore,

$$\sum_{i\in A^c} 1\{d\left(Y_i,m\left(X_i\right)\right)\leq x-\delta\} \leq \sum_{t\in A^c} 1\{d\left(Y_i,\tilde{m}\left(X_i\right)\right)\leq x\} \leq \sum_{i\in A^c} 1\{d\left(Y_i,m\left(X_i\right)\right)\leq x+\delta\} \tag{18}$$

Since, $\sum_{i\in A^c} 1\{d\left(Y_i,m\left(X_i\right)\right)\leq x-\delta\} \leq \sum_{t\in A^c}1\{d\left(Y_i,m\left(X_i\right)\right)\leq x\} \leq \sum_{t\in A^c}1\{d\left(Y_i,m\left(X_i\right)\right)\leq x+\delta\}$, it is follow that

$$\begin{aligned}&\left|\left(\sum_{i\in A^c}1\{d\left(Y_i,\tilde{m}\left(X_i\right)\right)\leq x\}\right) - \left(\sum_{i\in A^c}1\{d\left(Y_i,m\left(X_i\right)\right)\leq x\}\right)\right| && \text{(19)}\\ &\leq \left|\left(|\mathcal{J}_{\mathcal{D}_{test}}|\right)\left[\tilde{G}^*\left(x+\delta\right) - G^*\left(x-\delta\right)\right]\right|| - && \text{(20)}\\ &\left(\left(\sum_{i\in A^c}1\{d\left(Y_i,m\left(X_i\right)\right)\leq x+\delta\}\right) - \left(\sum_{i\in A^c}1\{d\left(Y_i,m\left(X_i\right)\right)\leq x-\delta\}\right)\right) && \text{(21)}\\ &\leq \left(|\mathcal{J}_{\mathcal{D}_{test}}|\left(G\left(x+\delta\right)\right) - G\left(x-\delta\right)+2R_T)\right) + |A| && \text{(22)}\\ &\leq |\mathcal{J}_{\mathcal{D}_{test}}|\left(2\delta W + 2R_T\right) + |A|. && \text{(23)}\end{aligned}$$

Using the previous inequality, we can show

$$|\mathcal{J}_{\mathcal{D}_{test}}|\left|\tilde{G}^* - \tilde{G}\left(x\right)\right| \leq |\mathcal{J}_{\mathcal{D}_{test}}|\left(2\delta W+2R_T\right)+|A| \tag{24}$$

Since the right-hand does not depend on x, we have that

$$\sup_{x\in\mathbb{R}}\left|\tilde{G}^{*}\left(x\right)-\tilde{G}\left(x\right)\right|\leq\frac{2}{|A|}\mathcal{J}_{\mathcal{D}_{test}}+\left(2\delta W+2R_{T}\right) \tag{25}$$

and we can bound the behavior $|A|$ using the fact that $\frac{1}{|\mathcal{J}_{\mathcal{D}_{test}}|}\sum_{i\in\mathcal{J}_{\mathcal{D}_{test}}}\left|d\left(Y_i,m\left(X_i\right)\right)-d\left(Y_i,\tilde{m}\left(X_i\right)\right)\right|\leq \frac{1}{\left|\mathcal{J}_{\mathcal{D}_{test}}\right|}\sum_{i\in\mathcal{J}_{\mathcal{D}_{test}}}\left|d\left(m\left(X_i\right),\tilde{m}\left(X_i\right)\right)\right|$.

Using again the triangle inequality, we have:

$$\sup_{x\in\mathbb{R}}\left|\tilde{G}^{*}\left(x\right)-G\left(x\right)\right|\leq\sup_{x\in\mathbb{R}}\left|\tilde{G}^{*}\left(x\right)-\tilde{G}\left(x\right)\right|+R_{T} \tag{26}$$

and we conclude the results using the assumptions introduced.

□

Lemma 3. *For any $\alpha\in(0,1)$, as $|\mathcal{J}_{\mathcal{D}_{test}}|\to\infty.$, $\tilde{q}_{1-\alpha}\to q_{1-\alpha}$.*

Proof. As $\tilde{G}^{*}\left(q_{1-\alpha}\right)=G\left(q_{1-\alpha}\right)+o_p\left(1\right)$, it is consequence of strong law of large numbers. □

Theorem 4. *Assume that Assumptions 1 are hold. Then, $\tilde{C}_{\alpha}\left(\cdot\right)$ estimated with the Algorithm 1 hold*

$$\int_{\mathcal{X}}\mathbb{P}\left(Y\in C^{\alpha}\left(x\right)\triangle\tilde{C}^{\alpha}\left(x\right)|X=x,\mathcal{D}_n\right)P_X\left(dx\right)=o_p\left(1\right).$$

Proof. For each $x\in\mathcal{X}$, we define $\tilde{C}^{\alpha}_{m}\left(x\right)=\mathcal{B}\left(m\left(x\right),\tilde{q}_{1-\alpha}\right)$ and $\tilde{C}^{\alpha}_{q}\left(x\right)=\mathcal{B}\left(\tilde{m}\left(x\right),q_{1-\alpha}\right)$.

For a fixed $x\in\mathcal{X}$ by the properties of the metric induced by the symmetric difference of two sets, it is hold that

$$\mathbb{P}\left(Y\in C^{\alpha}\left(x\right)\triangle\tilde{C}^{\alpha}\left(x\right)|X=x,\mathcal{D}_n\right)\leq \tag{27}$$

$$\mathbb{P}\left(Y\in C^{\alpha}\left(x\right)\triangle\tilde{C}^{\alpha}_{m}\left(x\right)|X=x,\mathcal{D}_n\right)+\mathbb{P}\left(Y\in \tilde{C}^{\alpha}_{m}\left(x\right)\triangle\tilde{C}^{\alpha}\left(x\right)|X=x,\mathcal{D}_n\right)+ \tag{28}$$

$$\mathbb{P}\left(Y\in C^{\alpha}\left(x\right)\triangle\tilde{C}^{\alpha}_{q}\left(x\right)|X=x,\mathcal{D}_n\right)+\mathbb{P}\left(Y\in \tilde{C}^{\alpha}_{q}\left(x\right)\triangle\tilde{C}^{\alpha}\left(x\right)|X=x,\mathcal{D}_n\right). \tag{29}$$

We will show:

$$\int_{\mathcal{X}}\mathbb{P}\left(Y\in C^{\alpha}\left(x\right)\triangle\tilde{C}^{\alpha}_{m}\left(x\right)|X=x,\mathcal{D}_n\right)P_X\left(dx\right)=o_p\left(1\right), \tag{30}$$

$$\int_{\mathcal{X}}\mathbb{P}\left(Y\in \tilde{C}^{\alpha}_{m}\left(x\right)\triangle\tilde{C}^{\alpha}\left(x\right)|X=x,\mathcal{D}_n\right)P_X\left(dx\right)=o_p\left(1\right), \tag{31}$$

$$\int_{\mathcal{X}}\mathbb{P}\left(Y\in C^{\alpha}\left(x\right)\triangle\tilde{C}^{\alpha}_{q}\left(x\right)|X=x,\mathcal{D}_n\right)P_X\left(dx\right)=o_p\left(1\right), \tag{32}$$

$$\int_{\mathcal{X}}\mathbb{P}\left(Y\in \tilde{C}^{\alpha}_{q}\left(x\right)\triangle\tilde{C}^{\alpha}\left(x\right)|X=x,\mathcal{D}_n\right)P_X\left(dx\right)=o_p\left(1\right), \tag{33}$$

To do this, we analyze the four terms separately.

Case 1:

We define $q^{m}_{1-\alpha}$ as the empirical quantile of the empirical distribution $G^{*}(v) = \frac{1}{\left|\mathcal{J}_{\mathcal{D}_{test}}\right|}\sum_{i\in\mathcal{J}_{\mathcal{D}_{test}}} 1\{d(Y_i, m(X_i)) \leq v\}$ and related with the ball $\tilde{C}^{\alpha}_{q_m}(x)$. Then

$$\mathbb{P}\left(Y \in C^{\alpha}(x) \triangle \tilde{C}_{m^{\alpha}}(x) \,|X = x, \mathcal{D}_n\right) \tag{34}$$

$$= \mathbb{P}\left(\{d(Y, m(x)) > q_{1-\alpha}, d(Y, m(x)) \leq \hat{q}_{1-\alpha}\}\,|X = x, \mathcal{D}_n\right) + \tag{35}$$

$$\mathbb{P}\left(\{d(Y, m(x)) \leq q_{1-\alpha}, d(y, m(x)) > \hat{q}_{1-\alpha}\}\,|X = x, \mathcal{D}_n\right) \tag{36}$$

$$\tag{37}$$

$$\leq \mathbb{P}\left(Y \in C^{\alpha}(x) \triangle \tilde{C}^{\alpha}_{q_m}(x) \,|X = x, \mathcal{D}_n\right) + \mathbb{P}\left(Y \in \tilde{C}^{\alpha}_{q_m}(x) \triangle \tilde{C}^{\alpha}_{m}(x) \,|X = x, \mathcal{D}_n\right) = \tag{38}$$

$$\mathbb{P}\left(\{d(Y, m(x)) > q_{1-\alpha}, d(Y, m(x)) \leq q^{m}_{1-\alpha}\}\middle| X = x, \mathcal{D}_n\right) \tag{39}$$

$$+\mathbb{P}\left(\{d(Y, m(x)) \leq q_{1-\alpha}, d(y, m(x)) > q^{m}_{1-\alpha}\}\middle| X = x, \mathcal{D}_n\right) \tag{40}$$

$$\mathbb{P}\left(\left\{d(Y, m(x)) > \hat{q}_{1-\alpha}, d(Y, m(x)) \leq q^{m}_{1-\alpha}\right\}\,|X = x, \mathcal{D}_n\right) \tag{41}$$

$$+\mathbb{P}\left(\{d(Y, m(x)) \leq \hat{q}_{1-\alpha}, d(Y, m(x)) > q^{m}_{1-\alpha}\}X = x, \mathcal{D}_n\right) \leq \tag{42}$$

$$2\left|G\left(q^{m}_{1-\alpha}\right) - G(q_{1-\alpha})\right| + 2\left|G(\hat{q}_{1-\alpha}) - G\left(q^{m}_{1-\alpha}\right)\right|. \tag{43}$$

Therefore

$$\int_{\mathcal{X}} \left|G\left(q^{m}_{1-\alpha}\right) - G(q_{1-\alpha})\right| + 2\left|G(\hat{q}_{1-\alpha}) - G\left(q^{m}_{1-\alpha}\right)\right| P_X(dx) = o_p(1). \tag{44}$$

Case 2:

$$\mathbb{P}\left(Y \in \tilde{C}^{\alpha}_{m}(x) \triangle \tilde{C}^{\alpha}(x) \,|X = x, \mathcal{D}_n\right) = \mathbb{P}\left(\{d(Y, m(x)) > \hat{q}_{1-\alpha}, d(y, \hat{m}(x)) \leq \hat{q}_{1-\alpha}\}|X = x, \mathcal{D}_n\right) \tag{45}$$

$$+\mathbb{P}\left(\{d(Y, m(x)) \leq \hat{q}_{1-\alpha}, d(y, \hat{m}(x)) > \hat{q}_{1-\alpha}\}|\, X = x, \mathcal{D}_n\right) \leq \tag{46}$$

$$\mathbb{P}\left(\{d(Y, \hat{m}(x)) \leq \hat{q}_{1-\alpha} < d(\hat{m}(x), m(x)) + d(Y, \hat{m}(x))\}|X = x, \mathcal{D}_n\right) \tag{47}$$

$$\mathbb{P}\left(\{d(Y, m(x)) \leq \hat{q}_{1-\alpha} < d(\hat{m}(x), m(x)) + d(y, m(x))\}|\, X = x, \mathcal{D}_n\right) \tag{48}$$

$$\leq G^{*}(\hat{q}_{1-\alpha} + d(\hat{m}(x), m(x))) - G^{*}(\hat{q}_{1-\alpha}) + G(\hat{q}_{1-\alpha}) + d(\hat{m}(x), m(x)) - G(\hat{q}_{1-\alpha}), \tag{49}$$

Case 3:

$$\mathbb{P}\left(Y \in C^{\alpha}(x) \triangle \tilde{C}^{\alpha}_{q}(x) \,|X = x, \mathcal{D}_n\right) = \mathbb{P}\left(\{d(Y, m(x)) > q_{1-\alpha}, d(Y, \hat{m}(x)) \leq q_{1-\alpha}\}|\, X = x, \mathcal{D}_n\right) \tag{50}$$

$$+\mathbb{P}\left(\{d(Y, m(x)) \leq q_{1-\alpha}, d(y, \hat{m}(x)) > q_{1-\alpha}\}|\, X = x, \mathcal{D}_n\right) \tag{51}$$

that it is similar to the prior case except that we exchange the role $\hat{q}_{1-\alpha}$ by $q_{1-\alpha}$. Therefore, we have

$$\mathbb{P}\left(Y \in C^{\alpha}(x) \triangle \tilde{C}^{\alpha}_{q}(x) \,|X = x, \mathcal{D}_n\right) \leq \tag{52}$$

$$\leq G^{*}(q_{1-\alpha} + d(\hat{m}(x), m(x))) - G^{*}(q_{1-\alpha}) + G(q_{1-\alpha} + d(\hat{m}(x), m(x))) - G(q_{1-\alpha}) \tag{53}$$

and as a consequence

$$\int_{\mathcal{X}}\left[G^{*}\left(q_{1-\alpha}+d\left(\hat{m}\left(x\right),m\left(x\right)\right)\right)-G^{*}\left(q_{1-\alpha}\right)+G\left(q_{1-\alpha}+d\left(\hat{m}\left(x\right),m\left(x\right)\right)\right)-G\left(q_{1-\alpha}\right)P_{X}\left(dx\right)\right]=o_{p}\left(1\right).\text{(54}$$

Case 4:

$$\begin{aligned}\mathbb{P}\left(Y\in\tilde{C}_{q}^{\alpha}\left(x\right)\triangle\tilde{C}^{\alpha}\left(x\right)|X=x,\mathcal{D}_{n}\right)&=\mathbb{P}\left(\left\{y:d\left(y,\hat{m}\left(x\right)\right)>q_{1-\alpha},d\left(y,\hat{m}\left(x\right)\right)\leq\hat{q}_{1-\alpha}\right\}\mid X=x,\mathcal{D}_{n}\right)\text{(55)}\\&+\mathbb{P}\left(\left\{y:d\left(y,\hat{m}\left(x\right)\right)\leq q_{1-\alpha},d\left(y,\hat{m}\left(x\right)\right)>\hat{q}_{1-\alpha}\right\}\mid X=x,\mathcal{D}_{n}\right),\text{(56)}\end{aligned}$$

that is similar to Step 1, the main diference is consider the random variable $d\left(Y,\tilde{m}\left(X\right)\right)|X=x$ instead of random variable $d\left(Y,m\left(X\right)\right)|X=x$. Therefore we have

$$\mathbb{P}\left(Y\in\tilde{C}_{q}^{\alpha}\left(x\right)\triangle\tilde{C}^{\alpha}\left(x\right)|X=x,\mathcal{D}_{n}\right)\leq 2|G^{*}(q_{1-\alpha}^{m})-G^{*}(q_{1-\alpha})|+2|G^{*}(\hat{q}_{1-\alpha})-G^{*}(q_{1-\alpha}^{m})|=4o_{p}\left(1\right).\text{(57)}$$

Combining the prior results, we have

$$\int_{\mathcal{X}}\mathbb{P}\left(Y\in C^{\alpha}\left(x\right)\triangle\tilde{C}^{\alpha}\left(x\right)|X=x,\mathcal{D}_{n}\right)P_{X}\left(dx\right)=o_{p}\left(1\right).$$

□

	Age	BMXWAIST	Diastolic Blood Pressure	Systolic blood pressure	Glucose
1	66	97.70	68	130	87
2	73	101.90	82	118	95
3	53	101.00	64	116	83
4	67	101.20	68	114	160
5	31	134.40	76	142	95
6	79	88.70	76	134	147

2.2 Heteroscedasticity case

In this setting, in order to obtain a estimator $\tilde{C}_{\alpha}(x)$, we propose to use the following two-step algorithm that, that unlike to Algorithm 1, we make a local approximation of the radius of the ball with k-nearest neighbors algorithm.

Algorithm 2 Uncertainty quantification algorithm heterocedastic set-up

```
1. Estimate the function $m(\cdot)$, $\tilde{m}(\cdot)$ using the random sample $\{(X_i, Y_i) : i \in \mathcal{J}_{\mathcal{D}_{train}}\}$.

2. For a fixed $X = x$, denote by $X_{(1,n_2)}(x), \cdots, X_{(n_2,n_2)}(x)$ the orders elements in decreasing
   order by a particular distance to the point $x$. Evaluate $\tilde{m}(x)$, and for all $i \in \mathcal{J}_{\mathcal{D}_{train2}}$, define
   $\tilde{r}_i = d(Y_i, \tilde{m}(x))$. We denote by $\tilde{r}_{(i,n_2)}(x)$ the pseudo-residuals related with the $i$-order-
   observation. Define by $\hat{F}_{n_2,k}(x,t) = \frac{1}{k}\sum_{i=1}^{k} 1\{\tilde{r}_{(i,n_2)}(x) \leq t\}$ the empirical conditional
   distributional and denote by $\tilde{q}_{1-\alpha}(x)$ the empirical quantile.

3. For a fixed $k$, return as estimation of prediction region $\tilde{C}_{\alpha}^{k}(x) = \mathcal{B}(\tilde{m}(x), \tilde{q}_{1-\alpha}(x))$.
```

3 Our model in global Fréchet linear regresion model

3.1 Multivariate value-responses

In this example, consider the space $\mathcal{Y} = (\Omega, d)$ where $\Omega = \mathbb{R}^p$, and we use as a metric the Mahalanobis distance that introduce the local geometry of the response variable on the randon variables Y as follows, $d(x,y) = \sqrt{(x-\mu)\Sigma^{-1}(x-\mu)}$, where $\mu = E(Y)$ and $\Sigma = Cov(Y,Y)$.

The example introduced here is also related to Diabetes Mellitus Disease and $p = 2$. Using a large dataset of more than 5000 patients, we try to predict the biomarkers used to control and diagnose the disease A1C (glycosylated hemoglobin) together with the physical activity levels of individuals under multivariate response linear models. For this purpose, we consider a bivariate regression model whose response is the mentioned A1C and TAC, and the predictors are age, waist, diastolic Blood Pressure, systolic blood pressure, and glucose. Figure 2 shows the levels set for six patients with basis protectors are found in Table 3.1- Generally, we observe that if the glucose is altered, the uncertainty increases.

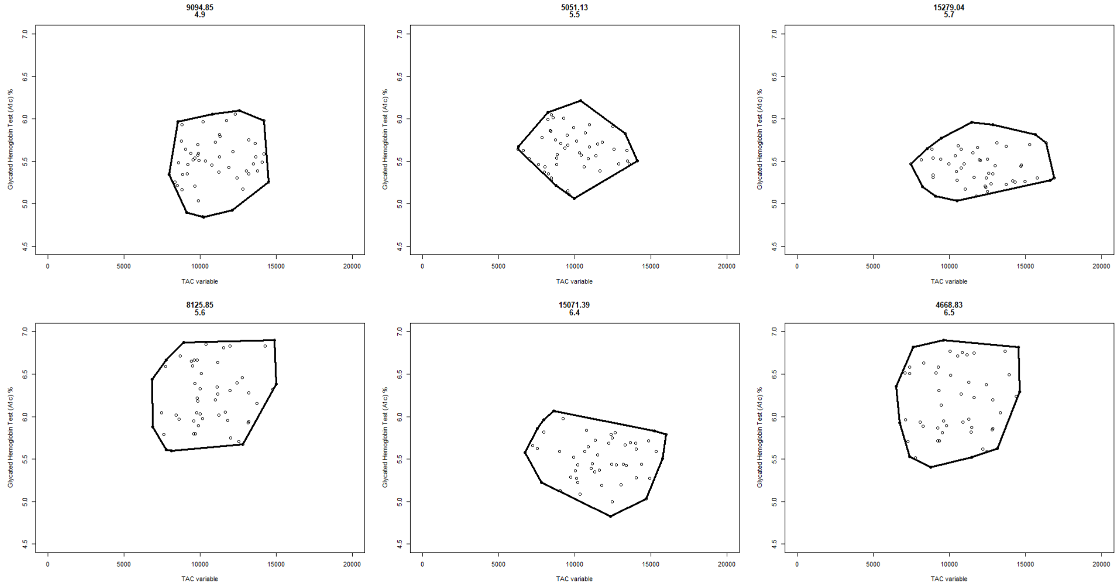


Fig. 2: Prediction region interval from six patients in the total activity time variable (TAC) and glycosilated haemoglobin (A1C).

3.2 Probability distribution with 2-Wassertien metrics in diabetes example

Using non-diabetes individuals from [21], we estimate for the first time using the new distributional representations the expected glucose confidence at a level of 80 percent of confidence by age groups. Figure 3 shows that the mean glucose values do not modify. However, the width of the bands yes.

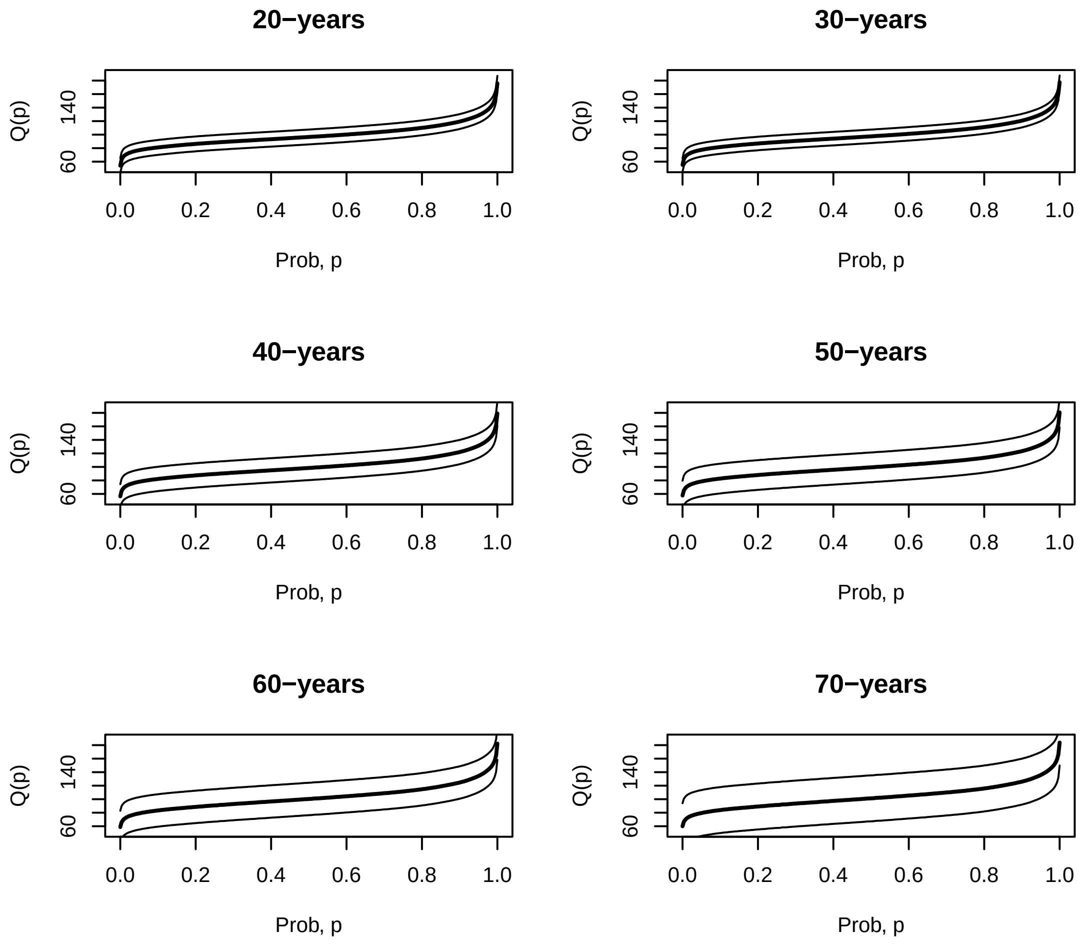


Fig. 3: Left: Expected glucose profiles by age

.

3.3 Graphs with Laplacian metrics

Using the first schizophrenia dataset from [17], we fitted a conditional fréchet model with a Laplacian metric by mental status and applied the uncertainty quantification algorithm. In this case, we use the homoscedastic algorithm. Figure 4 shows the Laplacian matrix's different conditional mean and uncertainty estimations of the mentioned matrix.

4 Discussion and final comments

In this work, to define the optimal uncertainty region set, we define the mathematical problem using balls covering specific probabilities and minimizing the diameter. However, many other geometrical sets for this purpose can be considered. The new methods provide more robust results than classical conformal inference algorithms with the cost that we do lose the theoretical guarantee of exact control with finite samples. In addition, the new methods are computational efficiently and can handle large datasets.

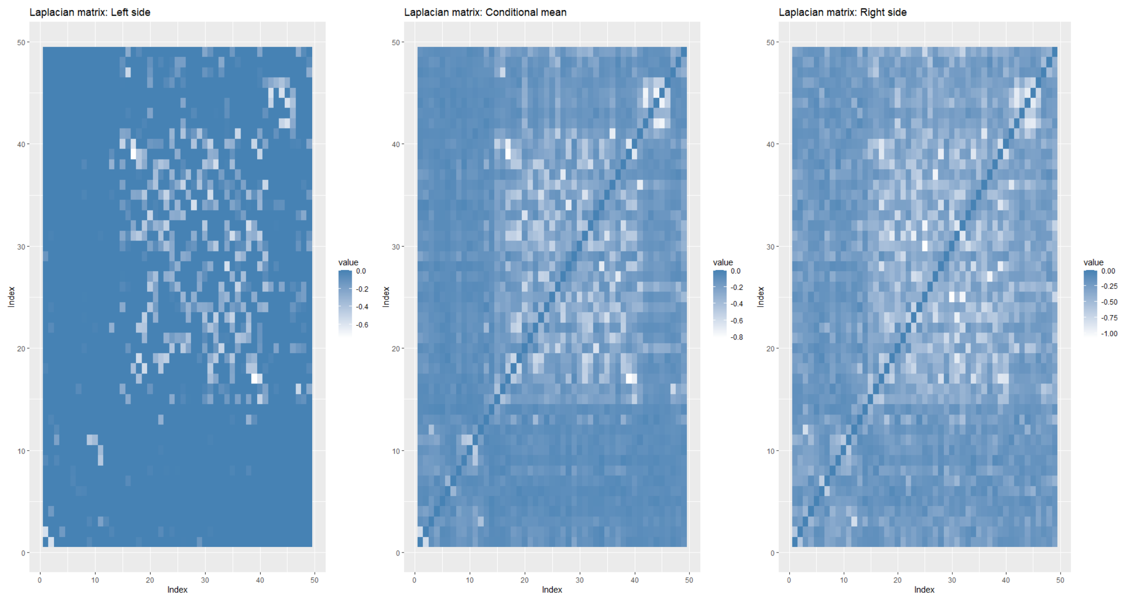


Fig. 4: Left: The CGM recording from a normoglycemic patient. Right: The corresponding glucodensity.

Appendix A Technical conditions global Fréchet model

Define the following quantity:

$$M(\omega, x) := E\left(\left[1 + (X - x)\Sigma\widetilde{\Sigma}^{-1}(x - \mu)\right]\left(d^2(Y, \omega)\right)\right). \tag{58}$$

For a fixed $x \in \mathbb{R}^p$, in order to guarantee the existence of population conditional mean, the convergence of empirical estimators, and ratios of convergence, we require to introduce the following assumptions.

Assumption 2. *The objects $m(x)$ and $\tilde{m}(x)$ exists and are unique, the latter almost surely, that is, for any $\epsilon > 0$, $\inf_{d(\omega, m(x)) > \epsilon} M(\omega, x)$.*

Assumption 3.

Let $B_\delta(m(x)) \subset \Omega$ be the ball of radius δ centered at $m(x)$ and $N(\epsilon, B_\delta(m(x)), d)$ be its covering number using balls of size ϵ. Then

$$\int_0^1 \sqrt{1 + log\left[N(\epsilon, B_\delta(m(x)), d)\right]} d\epsilon = O(1) \text{ as } \delta \to 0. \tag{59}$$

Assumption 4. *There exist $\eta > 0$, $C > 0$ and $\beta > 1$, possibly depending on x, such that, whenever $d(m(x), \omega) < \eta$, we have $M(\omega, x) - M(m(x), x) \geq Cd(\omega, m(x))^\beta$.*

In order to establish results in a uniform sense, we must introduce more strong assumptions. Let $\|\cdot\|$ be the Euclidean norm on $\mathbb{R}^p$ and $B > 0$.

Assumption 5. *Almost surely, for all $\|x\| \leq B$, the objects $m(x)$ and $\tilde{m}(x)$ exists and are unique. Additionally, for any $\epsilon > 0$,*

$$\inf_{\|x\| \leq B} \inf_{d(\omega, m(x)) > \epsilon} M(\omega, x) - M(m(x), x) > 0, \tag{60}$$

and there exist $\gamma = \gamma(\epsilon) > 0$ such that

$$P\left(\inf_{\|x\| \leq B} \inf_{d(\omega, m(x)) > \epsilon} M(\omega, x) - M(m(x), x) \geq \gamma \right) = 1 \tag{61}$$

Assumption 6.

With $B_\delta(m(x))$ and $N(\epsilon, B_\delta(m(x)), d)$,

$$\int_0^1 \sup_{\|x\| \leq B} \sqrt{1 + log\left[N(\epsilon, B_\delta(m(x)), d)\right]} d\epsilon = O(1) \text{ as } \delta \to 0. \tag{62}$$

Assumption 7. *There exist $\theta > 0$, $D > 0$ and $\alpha > 11$, possibly depending on B, such that,*

$$\inf_{\|x\| \leq B} \inf_{d(m(x), \omega) < \theta} \{M(\omega, x) - M(m(x), x) - Dd(\omega, m(x))^\alpha\} \geq 0 \tag{63}$$

References

[1] Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81):1–42, 2021.

[2] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 0(0):1–16, 2021.

[3] Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for multivariate functional data. *arXiv preprint arXiv:2106.01792*, 2021.

[4] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

[5] Rahul Ghosal, Vijay R Varma, Dmitri Volfson, Jacek Urbanek, Jeffrey M Hausdorff, Amber Watts, and Vadim Zipunnikov. Scalar on time-by-distribution regression and its application for modelling associations between daily-living physical activity and cognitive functions in alzheimer's disease. *arXiv preprint arXiv:2106.03979*, 2021.

[6] L GYÖFI and Harro Walk. Nearest neighbor based conformal prediction, 2020.

[7] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[8] Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

[9] Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, n/a(n/a).

[10] Benjamin LeRoy and David Zhao. Md-split+: Practical local conformal inference in high dimensions. *arXiv preprint arXiv:2107.03280*, 2021.

[11] Marcos Matabuena, Paulo Félix, Carlos Meijide-Garcia, and Francisco Gude. Kernel methods to handle missing responses and their application in modeling five-year glucose changes using distributional representations. `https://github.com/mmatabuena/Marcos-Matabuena-website/blob/main/Matabuena21web.pdf`, 2021.

[12] Marcos Matabuena and Alex Petersen. Distributional data analysis with accelerometer data in a nhanes database with nonparametric survey regression models. *arXiv preprint arXiv:2104.01165*, 2021.

[13] Marcos Matabuena, Alexander Petersen, Juan C Vidal, and Francisco Gude. Glucodensities: A new representation of glucose profiles using distributional data analysis. *Statistical Methods in Medical Research*, 30(6):1445–1464, 2021. PMID: 33760665.

[14] Marcos Matabuena, Pablo Rodríguez-Mier, Carlos García-Meixide, and Victor Leborán. Covid-19: Estimation of the transmission dynamics in spain using a stochastic simulator and black-box optimization techniques. *Computer methods and programs in biomedicine*, 211:106399, 2021.

[15] Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019.

[16] Rui Qiu, Zhou Yu, and Ruoqing Zhu. Random forests weighted local fr\'echet regression with theoretical guarantee. *arXiv preprint arXiv:2202.04912*, 2022.

[17] Jesús D Arroyo Relión, Daniel Kessler, Elizaveta Levina, and Stephan F Taylor. Network classification with applications to brain connectomics. *The annals of applied statistics*, 13(3):1648, 2019.

[18] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32:3543–3553, 2019.

[19] André J Scheen. Precision medicine: the future in diabetes care? *Diabetes research and clinical practice*, 117:12–21, 2016.

[20] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[21] Viral N Shah, Stephanie N DuBose, Zoey Li, Roy W Beck, Anne L Peters, Ruth S Weinstock, Davida Kruger, Michael Tansey, David Sparling, Stephanie Woerner, et al. Continuous glucose monitoring profiles in healthy nondiabetic participants: a multicenter prospective study. *The Journal of Clinical Endocrinology & Metabolism*, 104(10):4356–4364, 2019.

[22] Yidong Zhou and Hans-Georg Müller. Dynamic network regression. *arXiv preprint arXiv:2109.02981*, 2021.

Basic Information

Project title: “Private Continual Learning from a Stream of Pretrained Models”
Period of project: 2022-06-01 / 2022-07-31
Author(s): Antonio Carta
Organization: University of Pisa
Host organization: Computer Vision Center, Barcelona

Public summary

Today, machine learning models are ubiquitous. These models are trained on large datasets and can solve a wide range of tasks, such as automatic translation, speech recognition, or activity recognition. At the same time, personal devices such as smartphones collect lots of data that can be used to improve the quality of such models. Personal devices are becoming powerful enough to be used to train personalized machine learning models adapted and customized to the user.

There are several constraints that we need to consider in this scenario. First, we don’t want to share personal data, so the user data should never leave the device. Second, we want to limit the computational cost such that this process could run in the background without disrupting the normal usage. Finally, we would like to share the knowledge between different devices (in a private way). We also want the user to keep control over this process and have the ability to limit or disable it.

None of the currently available solutions work in this setting since they either require sharing the data or to surrender the control to a central server and synchronize frequently with it. Instead, we propose a novel method that allows fast local adaptations and private sharing of common knowledge. The heavy computations can be offloaded to the cloud (if necessary) but the data never leaves the device. Unlike previous solutions, each device has full control over its own training process and the model is able to adapt over time to the user’s behavior.

Research objectives

Objectives

The project aims to extend continual learning (CL) techniques to a multi-agent scenario. Continual learning is a machine learning problem where an agent learns from a non-stationary stream of data. In a multi-agent scenario, we have a set of independent machine learning models (agents), each one learning from an independent non-stationary stream of data. Each agent keeps its own data private but the agents can communicate by sharing the model’s weights.
This scenario combines several challenges. First, we have the problem of distribution shifts because each stream is changing over time and each agent learns from a different stream. Then, we have the problem of knowledge sharing in a data-free setting. Finally, we have sparse communication, because the weights are shared only after completing the training on a task, unlike federated learning, where the weights are shared every few iterations. The

combination of distribution shifts and sparse communication means that most knowledge distillation and federated learning methods fail in this scenario.

The project aims to formally define the multi-agent continual learning problem, propose a suitable set of benchmarks, and explore different solutions for the problem.

Impact

The problem under study encapsulates several properties of real world applications, such as low computational resources, limited connectivity, and hard privacy constraints. Our solution provides a novel alternative to learning in such constrained systems. Unlike previous frameworks, such as federated learning, this approach allows the user to retain full control over the training process.

Technical approach

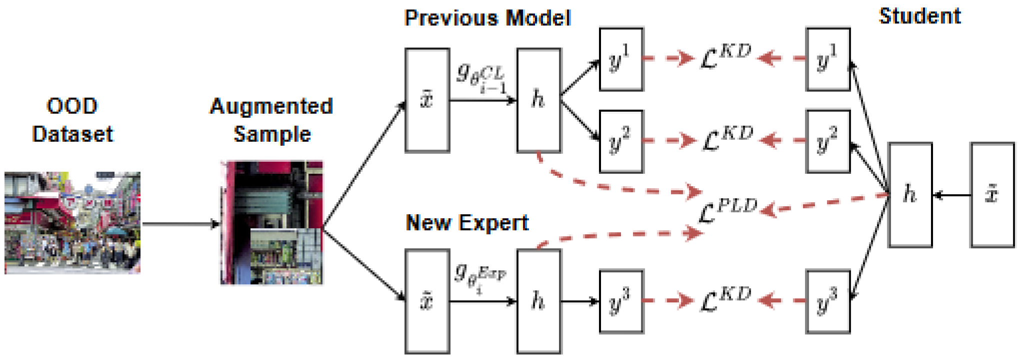
Detailed description

Our approach is based on the idea that we can separate the continual learning process into two steps: *adaptation*, where the agent learns a new task, and *consolidation*, where the new knowledge is integrated into a joint model with all the other tasks. In our case, the adaptation phase uses the original data, while the consolidation uses *data-free knowledge distillation*. This choice allows to share knowledge between agents without sharing the data and even to offload the consolidation to the cloud.
We studied the multi-agent continual learning problem in two simplified settings. In the *sequential* setting, at time *t* the new model is initialized with the weights learned at time *t-1*. In the independent setting, each task is learned starting from a common initialization.

Our approach, called *Data-Free Consolidation (DFC)* is based on the idea that it is possible to do knowledge distillation with external data, as long as we train for a long time [1] and use heavy augmentations [2]. We also propose a novel regularization term, called *Projected Latent Distillation (PLD)*, that allows to perform a double knowledge distillation over the latent space even when the two teachers are trained on different tasks. The idea is that the new latent space of the student encodes a linear transformation of the two latent spaces. With two teachers, encoding the original latent spaces would not be possible because the two teachers will give completely different targets.

We are evaluating DFC in multi-task continual learning benchmarks. The preliminary results show state-of-the-art results for SplitCIFAR100 in the 5, 10, and 20 tasks configurations. Somewhat surprisingly, the method is competitive even against non-data-free baselines. We hypothesize that the separation between the adaptation and consolidation steps makes it easier to balance adaptation and forgetting, while other methods combine both steps and often have to trade off between the two.

The figure below shows a schematic view of DFC.



[1] Beyer, Lucas, et al. "Knowledge distillation: A good teacher is patient and consistent." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
[2] Asano, Yuki M., and Aaqib Saeed. "Extrapolating from a Single Image to a Thousand Classes using Distillation." *arXiv preprint arXiv:2112.00725* (2021).

Scientific outcomes

The results of this project provide important insights:

- We showed that data-free knowledge distillation is a viable method for continual learning. Even simple out-of-distributions sources, such as a single image, can be sufficient provided there is enough diversity.
- We proposed a novel strategy that uses data-free double knowledge distillation in the output and latent space.
- The proposed method is state-of-the-art in popular multi-task continual learning benchmarks.
- The separation into explicit adaptation and consolidation steps seems helpful for continual learning, and we believe that it should be studied more in depth and applied to new methods.

Future plans

Currently, we are running more experiments to provide a more thorough evaluation of the approach. In the future, we plan to continue extending the work to more difficult settings.

Self-assessment

- **AI Excellence**: The results of the project provide a promising approach for a user-centered personalization, which allows to share knowledge privately while keeping the user in full control of the process. With our method, it is also possible to offload a part of the computation to a server without sharing the data.
- **Scientific step-up**: The visit helped me to connect with the LAMP group at CVC, which is one of the top European labs in continual learning and computer vision.
- **Suitability of the host**: The host lab helped me with the design of the methodology and experiments. In particular, the lab has a strong expertise in computer vision, while my main focus is on continual learning.

- **Suitability of the visit length**: The visit length (2 months) was sufficient to start a collaboration with the lab. Apart from the main project discussed above, the visit allowed me to start a collaboration with another CVC member.

List of publications, meetings, presentations, patents,...

We are currently working on the submission of two publications. The first one is about Data-Free Consolidation, as described above. The second one is a collaboration with Albin Soutif, a CVC Ph.D. student, and Joost van de Weijer. The paper proposes a method to solve the problem of the continual evaluation gap.

Basic Information

Project title: Modelling others for cooperation under imperfect information
Period of project: 01/06/2022 - 30/06/2022
Author(s): Nieves Montes, Nardine Osman, Carles Sierra
Organization: Artificial Intelligence Research Institute (IIIA-CSIC)
Host organization: King's College London

Public summary

During this visit, we have developed an agent model capable of using Theory of Mind to understand the reasons behind the behaviour of other agents. This means that agents can, at any given point in time, change the perspective of the world by the perspective that they estimate other agents have ("I believe that this is the way in which agent *x* views the world"). Furthermore, they can switch their perspective *recursively*, as in "I believe that this is the way in which agent *x* thinks that agent *y* views the world". This ability is called *higher-order* Theory of Mind, and it is seamless integrated into our model. Thanks to this, our agents are able to display *empathy*, since they are able to operate not on their view of the world, but on the view of the world that someone else has.

However, an agent switching its perspective does not, on its own, entail any benefit. For this reason, we have further integrated techniques for *abductive reasoning* into our agent model. Essentially, after an agent has changed its perspective of the world by that of someone else, it infers the *explanations* for the actions performed by the agent whose perspective it has adopted. This type of reasoning from observations (in this case, observed actions) to explanations is called *abduction*. These explanations put the observing agent in a better-informed position when it comes to its own decision-making later on.

Research objectives

Objectives

The scientific objective that this research visit has pursued is that outlined by Work Package 6 (*Social AI*), Task 6.1 (*Modelling social cognition, collaboration and teamwork*), which is led by IIIA-CSIC: "To study the modelling of agent's cognitive capabilities that integrate individual knowledge (...) with knowledge available to and from other agents (possibly obtained at different times and from different perspectives) (...). To achieve that, agents should have the capability of understanding others, reason about them (for example have Theory of Mind) and be able to act in a team (...)".

The goal of this visit has been to develop an agent model, on top of the well-known Belief-Desire-Intention (BDI) architecture, that displays the abilities outlined by the task description above. To do this, we proposed to combine Theory of Mind (i.e. the ability to switch its perspective on the state of the system by the perspective that they estimate other agents have) with abductive reasoning (i.e. the inference from observation to explanations).

Inspiration for such an agent model originated in the Hanabi game. This is a cooperative card game where players collaborate to build stacks of coloured cards as high as possible. Players do not see their own cards, however they can give hints to one another that partly reveal their identity. The game presents researchers with an interesting challenge and a benchmark to develop techniques for modelling others in cooperative domains. Nonetheless, the objective of this visit is to develop an agent model that is completely domain-independent. Another feature that we set out to include in our model is the ability to engage in Theory of Mind of higher order. This means that agents should be able to switch their perspective recursively an arbitrary number of times.

Impact

As modern society evolves into a massive sociotechnical system that includes humans, autonomous agents and humans that augment their capacities with software assistants, it is important to design autonomous agents in such a way that they are able to engage in social interactions (with other humans as well as software agents) in a meaningful way. It is a well-established fact that Theory of Mind plays an essential role in human social interactions. Therefore, in order to design agents that engage in interactions in the most human-like way possible, it is absolutely necessary to include Theory of Mind capabilities. The work carried out during the course of this research visit constitutes a step in that direction.

Technical approach

Detailed description

The agent model we propose presents the following functionalities. There, an observer agent i, upon observing action a_l by actor agent l, engages in a *Theory of Mind + abduction task*, manages by function TomAbductionTask. This function includes calls to replace the agent’s current belief base by an estimation of the belief base of the observer, which is generated thanks to pre-coded domain knowledge. Then, this function calls to an abductive meta-interpreter that generates the explanations for the observer actions. Then, these explanations are revised by the *explanation revision function* (ERF). Finally, they are transformed into a format suitable to be appended to the agent’s original belief base.

Additionally, the agent model also includes functionalities to update previous explanations based on new observations of the state of the world through the *explanation update function* (EUF). Last of all, a deliberation function (SelectAction) that may use Theory of Mind during action selection has also been included. We would like to note that the model that we present is highly flexible as many of the mentioned functions are customizable, and we only provide default implementations.

Related work symbolic-based Theory of Mind had been developed previously developed, for purposes of deception [2,3,4] and personnel training [1]. A notable different between our work and theirs is the *focus* that the Theory of Mind capabilities have. In our work, Theory of Mind is mostly oriented towards *sensing*, i.e. extracting information about the state of the system. Meanwhile, related work uses Theory of Mind to *deliberate* about which action to take next. Nevertheless, Theory of Mind can also be included in the deliberation step of our model through customization of the SelectAction function. Another outstanding feature of our model in comparison with previous work is the ability of our agents to engage in higher-order Theory of Mind with minimal memory burden.

Scientific outcomes

The agent model we have developed has been implemented in Jason (a Java-based BDI agent language). The code is public[1] and thoroughly documented, so other researcher or practitioners can use it.

Future plans

Future work will focus on the use of the developed model in tasks and domains where higher-order Theory of Mind may entail a performance benefit. It will attempt to theoretically derive the features of domains for higher-order Theory of Mind to have an impact, and empirically validate these findings with simulations using our developed model. Furthermore, interest has been expressed by partners in TAILOR's Work Package 6 to collaborate on an integration of our agent model with embodiment.

Self-assessment

The research visit contributed to trustworthy AI by advancing the state-of-the-art on socially-oriented agent architectures. Trust in one another goes hand in hand with a *mutual understanding* of each other. Therefore, software agents that are deserving of the trust of the humans they act for should display the ability to understand their point of view, which is the central capability of the agent model we have developed.

The extended (4 months long) research visit of Nieves Montes, a third-year PhD student at IIIA-CSIC, proved very positive for her professional development. It provided her with ample opportunities to network at her host institution and present her work at the seminar series hosted at the Center for Doctoral Training run by the Department of Informatics at King's College London. For Nardine Osman and Carles Sierra, who are senior researchers, the visit was an opportunity to catch up with colleagues working in the areas of multi-agent systems and safe and trusted AI.

The choice of host proved to be a good fit for the research that this visit carried out. Hosts Michael Luck and Odinaldo Rodrigues are experienced researchers in the area of multi-agent systems, and their input during the development of the model was very valuable. Additionally, the interests of research staff at the wider Department of Informatics at the host institution are also aligned with those of the visitors.

The lengths of the visits (4 months for Nieves Montes, 1 week for Carles Sierra and Nardine Osman) were adequate. It was enough time for Nieves Montes to develop and test the agent model to its current form. During the time all visitors were in London, there was enough time to have multiple meetings to discuss possible directions for future research.

[1] https://github.com/nmontesg/tomabd

List of publications, meetings, presentations, patents,...

A paper outlining a preliminary account of the work carried out during the research visit was presented at the 19th European Conference on Multi-Agent Systems:

Montes, N., Osman, N., & Sierra, C. (2022). Combining Theory of Mind and Abduction for Cooperation Under Imperfect Information. In *Multi-Agent Systems* (pp. 294–311). Springer International Publishing. https://doi.org/10.1007/978-3-031-20614-6_17

An extended paper detailing all the work carried out during the visit is currently under review at the Journal for Autonomous Agents and Multi-Agent Systems.

References

[1] Harbers, M., Bosch, K.v.d., Meyer, J.-J.: Modeling agents with a theory of mind. In: 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, vol. 2, pp. 217–224 (2009). https://doi.org/10.1109/WI-IAT.2009.153

[2] Panisson, A., Mcburney, P., Parsons, S., Bordini, R., Sarkadi, S.: Lies, bullshit, and deception in agent-oriented programming languages. In: Proceedings of the 20th International Trust Workshop Co-located with AAMAS/IJCAI/ECAI/ICML (AAMAS/IJCAI/ECAI/ICML 2018) (2018)

[3] Panisson, A.R., Sarkadi, S., McBurney, P., Parsons, S., Bordini, R.H.: On the formal semantics of theory of mind in agent communication. In: Agreement Technologies, pp. 18–32. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-17294-7_2

[4] Sarkadi, S., Panisson, A.R., Bordini, R.H., McBurney, P., Parsons, S., Chapman, M.: Modelling deception using theory of mind in multi-agent systems. AI Communications 32, 287–302 (2019). https://doi.org/10.3233/AIC-190615

Basic Information

Project title: Workshop proposal: Imagining the AI landscape after the AI Act
Period of project: 13 june 2022
Period of reporting: 13 June 2022
Author(s): Francesca Naretto
Organization: Scuola Normale Superiore
Host organization: Vrije Universiteit Amsterdam (HHAI 2022 Conference – Hybrid Human Artificial Intelligence conference)

Public summary

The TAILOR Connectivity fund in analysis was used for organizing the workshop: "Imagining the AI landscape after the AI Act" at the first conference on Hybrid Human Artificial Intelligence.

The workshop started with a short welcome moment, in which the organizers presented the schedule of the day. Then, there were the two invited speakers: first Prof. Virginia Dignum and then Prof. Mireille Hildebrandt. Regarding Prof. Dignum, she presented different perspectives regarding the AI Act, focusing on several limitations that she still sees in this Act. Prof. Dignum started her presentation by asking an interesting question: what does the AI Act represent for people who work with AI? This is a particularly difficult question to answer since for now we do not even have an agreement on the definition of AI. Hence, Prof. Dignum started to reason about this question considering the fact that the AI Act focuses mostly on data and the users: in fact, the problems considered all originated from the data, which may be dirt and incorrect. From them, bias and discrimination may arise, creating several problems. Following this line of thinking, she categorized the prohibitions in the AI Act into three categories: (i) the use of social scoring, (ii) distortion of user behavior, and (iii) biometric identification. She also pointed out a wide number of limitations in the AI Act, such as the lack of sustainability and power consumption, as well as difficulty in allowing innovation and development.

The second talk was from prof. Hildebrandt. In this case, the keynote was a series of questions, all regarding the AI Act, collected from the participants. Involving them made the discussion even more interesting, raising several questions but also possible solutions. Overall, Prof. Hildebrandt has an optimistic view: she reckons that implementing all the requirements in the AI Act is a difficult task, but finding the best level of abstraction and, with time, it is feasible.

After the invited speakers, we moved to the presentation of the accepted papers. We split the presentations of the works into two sessions: Session 1, in the morning, was titled "Technical Aspects of the AI Act" in which the papers focused on the technicalities required by the AI Act, as well as the requirements for the individuals; while the Session 2, in the afternoon, was titled "Ethical and Legal Aspects about the AI Act" and hence included contributions focused more on the ethical and legal problems of the AI Act. At the end of every session we left some time for open mike, in which the participants were able to talk more about the contributions presented.

Between the two sessions of papers, we also organized a group activity held by Dr. Tommaso Turchi. The group activity is called "MiniCoDe", which stands for "MINImize algorithmic bias in COllaborative decision making with DEsign fiction". It is a project lead by Prof. Alessio Malizia, in which the aim is to tackle social injustice in future algorithmic-based decision-making applications. This project was already used in several different context, both in academic settings and in companies. In our case, the participants were split into small groups and were asked to think about a possible scenario which involved Artificial Intelligence as well as ethical problems. At the beginning, the participants were asked to reason first alone and then as a group, with the goal of proposing a solution. In particular, the group activity was a structured brainstorming around how to implement a process/methodology to be compliant with Art.14 on Human Oversight. More specifically, participants were presented with a fictional narrative describing how postcode bias might lead to discrimination against the poor. This type of bias is more subtle compared to other types of biases such as gender or race bias, so enabling human oversight is more difficult. The discussions allowed participants to have a deeper understanding of the implications of the AI Act and EU digital policies.

The proposed workshop found great interest from researchers in a variety of fields, from computer science to law, psychology and economics. We received 17 submissions and we accepted, through a peer reviewed process, 11 contributions, with regular papers, short papers, and extended abstracts (more details about the accepted contributions can be found at http://iail2022.isti.cnr.it/#program). Among them, 3 were regular papers (i.e., 12+ pages), 6 short papers, and 2 abstracts. The abstract can contain preliminary or already published work, while papers must contain original work.

Overall, we are confident that our workshop was successful. The participants at the workshop were 30-35 we are particularly happy to have gathered 11 contributions, all of which were extremely interesting and multidisciplinary, but most of all to have initiated very interesting discussions on the topic, which we are confident will bring new contributions in this area. Due to the success of this first workshop, also demonstrated by the several grateful emails, we have received, we hope to have a workshop again in the coming year.

Research objectives

Maximum 1 page.

Objectives

Scientific high-level description of the scientific goals of this research visit or workshop. Which scientific challenge are you addressing and why is it important.

The workshop's main goal was to help the community understand and reason over the implications of an AI regulation: what problems does it solve, what problems does it not solve, what problems does it cause, and propose new approaches that solve the new challenges.We decided to conduct this workshop with the objective of collecting ideas regarding how new AI regulation will shape the AI technologies of the future. We aimed to collect contributions regarding how to operationalize the AIA requirements and how to guarantee privacy, fairness and explainability. In particular, our most significant concerns regarded how to guarantee individual rights while achieving the requirements stated in the AIA, and how to assess the AI risk. This is due to the fact that the AI Act considers different classes of AI systems, depending on how risky the Ai under analysis is. But the AIA does not clearly state how to assess an AI system's risk. In addition, in the AIA, several ethical values are cited and considered, but also in this case it is not clear how to guarantee them while preserving the performance of the AI model. In practice, we were afraid that the technologies available right now were not enough to fulfill the general and high-level requirements proposed in the AIA.

Impact

Please describe the expected impact of this work on society.

Firstly, our goal was to bring together legal experts, tech experts and other interested stakeholders for constructive discussions. Hence, we aimed at having a multidisciplinary setting in which persons with different backgrounds and levels of expertise were brought together and encouraged to share different points of view on the subject of the AI Act. For this reason, in our workshop schedule, we planned several moments of discussion. Before the workshop, we collected the participants' questions to ask Prof. Hildebrandt, intending to conduct a discussion with her on the themes most important for the audience. Then, we encouraged people to ask questions at the end of each keynote speaker, as well as after every paper presentation. Lastly, we organized a moment of open mike and a group activity to encourage exchanging ideas and constructive discussion. In addition, we aimed at stakeholder and geographical balance. These goals dictated the choice of hosting the workshop in the Hybrid Human Artificial Intelligence conference. In fact, this conference was multidisciplinary, with special attention to the human in interaction with the AI systems and the ethical aspects related to this subject, such as guaranteeing fairness, protecting privacy, and ensuring accountability while providing secure and sustainable AI systems.

Technical approach

Maximum 1 page.

Detailed description

Please detail the technical approach that you followed during your research visit. Include a comparison with the state of the art. For workshops, describe the performed work as well as possible.

We received 17 submissions, and we accepted, through a peer-reviewed process, 11 contributions, with regular papers, short papers, and extended abstracts (more details about the accepted contributions can be found at http://iail2022.isti.cnr.it/#program). Among them, 3 were regular papers (i.e., 12+ pages), 6 short papers, and 2 abstracts. The abstract can contain preliminary or already published work, while papers must contain original work. These contributions will be published in the proceedings of Ceur (http://ceur-ws.org/). All accepted papers presented their work. Our workshop was in-person (as the main conference), but allowed a hybrid mode for those who had difficulty being present. Most of the papers were presented in person, creating an excellent dialogue among all participants. We split the presentation of the contributions into 2 sessions: the first one, titled "Technical aspects of AI Act" was more technical, while the second one, titled "Ethical and Legal Aspects about the AI Act" focused more on the several legal and ethical implications of the AIA. In the first session, we had 5 contributions.

The first one, titled "Using Sentence Embeddings and Semantic Similarity for Seeking Consensus when Assessing Trustworthy AI" was presented by Dennis Vetter and dealt with the difficult task of assessing the trustworthiness of the AI systems. This work addressed this problem by considering sentence embeddings and semantic similarity during the consensus phase of the assessment. The contribution was interesting and well presented; the audience asked several questions, which allowed for an in-depth discussion during the coffee break. The second contribution was titled "FutureNewsCorp, or how the AI Act changed the future of news" and was presented by the single author Natali Helberger. In this case, the paper focused more on the technical aspects of news, newspapers and journalists in the new era of AI. The author started her presentation by presenting a possible future scenario, which was quite intriguing and scary, thus providing several constructive discussions. Then, there was

"Federated Learning as an Analytical Framework for Personal Data Management", a paper presented by Maciej Zuziak. This author has a legal background, but he is now collaborating with computer scientists to merge his knowledge with the more technical aspects of Federated Learning. The paper "The forgotten human autonomy in Machine Learning" was presented by Prof. Oriol Pujol. It considers several limitations of AI systems when dealing with human autonomy. This theme was also a key topic of the conference; thus, it allowed for several comments and discussions. Lastly, Prof. Francesca Carroccia presented the paper "AI Act and Individual Rights: A Juridical and Technical Perspective", in which she tackled the problem of the gap between the requirements of the AIA and the technical capacities now available.
Moving to the second session of the paper presentation, we first had an extremely interesting presentation by Prof. Marc Anderson for his paper "Some Ethical Reflections on the EU AI Act". This contribution considered several ethical aspects in relation to the requirements of the AIA, highlighting the limitations now present in this context from a legal point of view. Then, we had Jonne Mas presenting "A Neo-republican Critique of AI ethics". This presentation considered the ethical problems from a broader perspective with respect to the works considered so far. The themes considered by Jonne were also linked to the next presentation from Prof. Jerome De Cooman of his work titled "Without Any Prejudice? The Antitrust Implication of the AI Act". Following, Pietro Dunn and Giovanni De Gregorio presented "The Ambiguous Risk-Based Approach of the Artificial Intelligence Act: Links and Discrepancies with Other Union Strategies": in this case, the AIA was compared against other European regulations. Then, the paper "The Artificial Intelligence Act. A Jurisprudential View" was presented. This paper focuses mostly on the legal aspects of the AIA, dealing with the actual actuation of all the requirements when considering the different European countries and their different needs. Linked to this talk, there was also the last one from Farhana Ferdousi Liza. She presented "Challenges of Enforcing Regulations in Artificial Intelligence Act" in which the problem of enforcing the regulation is tackled from the point of view of the problems that may arise.

Scientific outcomes

Please detail the outcomes of this work. This includes any novel insight, discoveries, transfer of technology, or other types of knowledge sharing. Publications can be listed below.
The papers accepted to this workshop will be published as proceedings in Ceur (http://ceur-ws.org/).. In the following there is a list of the papers accepted:

1. Without Any Prejudice? The Antitrust Implication of the AI Act, Jerome De Cooman
2. Challenges of Enforcing Regulations in Artificial Intelligence Act — Analyzing Quantity Requirement in Data and Data Governance, Farhana Ferdousi Liza
3. Using Sentence Embeddings and Semantic Similarity for Seeking Consensus when Assessing Trustworthy AI, Dennis Vetter, Jesmin Jahan Tithi, Magnus Westerlund, Roberto V. Zicari and Gemma Roig
4. Federated Learning as an Analytical Framework for Personal Data Management – a proposition paper, Maciej Zuziak, Salvatore Rinzivillo
5. The forgotten human autonomy in Machine Learning, Paula Subías-Beltrán, Oriol Pujol and Itziar de Lecuona
6. AI Act and Individual Rights: A Juridical and Technical Perspective, Costanza Alfieri, Francesca Caroccia and Paola Inverardi
7. The Ambiguous Risk-Based Approach of the Artificial Intelligence Act: Links and Discrepancies with Other Union Strategies, Pietro Dunn and Giovanni De Gregorio
8. Some Ethical Reflections on the EU AI Act, Marc M. Anderson

9. The Artificial Intelligence Act: A Jurisprudential Perspective, Michał Araszkiewicz, Grzegorz J. Nalepa, and Radosław Pałosz

Future plans

Please detail future plans or new collaborations based on this work.
The proposed workshop greatly interested researchers in various fields, from computer science to law, psychology and economics. Overall, we are confident that our workshop was a success: we brought together people from different backgrounds, creating a constructive dialogue, which we are sure will lead to interesting works in the future. In addition, we were able to publish the contributions of the workshop in proceedings, hence more people, also outside the conference audience, will be able to see them and follow-up on these ideas and projects. We also plan to propose the workshop at the next HHAI 2022 conference to provide a follow-up on the several interesting ideas discussed in this workshop.

Progress against planned goals

Maximum 1 page. Only for intermediate reporting (projects running longer than 6 months)

Please provide an indication of your progress towards the stated research goals. What have you already achieved? Are you on track to attain all goals? Did issues occur that made you adapt previous plans?

All the organizers of the workshop are happy to say that the workshop achieved the expected results: our main goal was to create a constructive discussion about the AI Act and the problems and limitations we are going to face when it will be actualized. In particular, we were interested into understanding what is missing in terms of technology for achieving a satisfying assessment of the AI risk and how to protect ethical requirements in this context. The workshop we organized tackled exactly these themes as well as all the publications produced. In addition, another important requirement for us was to have multidisciplinary both in terms of audience and of the publications gathered. This requirement for us was extremely important since the Ai Act will tackle everyday life, involving both legal and ethical experts, as well as different kinds of technicians to achieve the requirements stated in the regulation.

Self-assessment

Maximum 1 page. Only for final reporting (the project has finished)

Please provide your own final assessment of the effective progress against the goals stated in the proposal, according to the following points:
The proposed workshop was a success: several people complimented us, both in person and via e-mails. The hybrid version of the event allowed several people to follow and to present even if they were not able to join us in presence. We found it a really good way to include more people, however, for the group activity we were not able to provide it online due to the impossibility of provide the necessary material to the people online. For this reason, for the next edition we plan to propose a full in-presence workshop. In addition, we would like to propose a two day workshop, so that we are going to have more time for the group activates.

In fact, at the beginning, we had several ideas about different group activities, but then due to the time constraints we were able to propose just one group activity.
To conclude, this workshop consider several topics related to Trustworthy AI.

List of publications, meetings, presentations, patents,...

Please detail the outcomes of this work. This includes any produced deliverables, e.g. papers or technical reports, transfer of technology, or other types of knowledge sharing.
Please note that all dissemination, including publications, with financing from TAILOR needs to acknowledge this. Also, TAILOR is required to publish with open access. Also mention any intellectual property rights (IPR), such as patents, based on the performed work.

The papers accepted to this workshop are going to be published as proceedings. The list of all the papers is in section "Technical approach – scientific outcomes".

Additional comments

Any additional comments that you'd like to share.

Basic Information

Project title: A Modular Framework for Hybrid Participatory Systems
Period of project: 01/10/2022 - 01/12/2022
Author(s): Enrico Liscio
Organization: Delft University of Technology, the Netherlands
Host organization: University of Barcelona, Spain

Public summary

Human values are the abstract motivations that drive our opinions and actions, including concepts such as *safety* or *freedom.* Thus, human values are at the core of human societies. As AI becomes increasingly embedded in our society, we must ensure that its behavior aligns with our values. However, value alignment must be preceded by *value inference*: in practice, we must first identify which values are important to different society members, and how different people prioritize values. Only then can we design value-aligned AI.

During this research visit, we delineate the challenge of value inference and propose our joint view on how it can be performed in a society where humans and artificial agents coexist. We introduce a holistic framework that connects the technical components necessary for value inference introduced in different subfields of AI. Subsequently, we discuss how hybrid intelligence -- the synergy of human and artificial intelligence -- is instrumental to the success of value inference. Finally, we illustrate how value inference both poses significant research challenges and provides novel research opportunities that span multiple research fields, ranging from AI to social sciences.

Research objectives

Objectives

Values are the abstract motivations that drive our opinions and actions. Different values may compete when we ought to take a decision, and the relative importance we ascribe to values (our *value preferences*) guides our actions. However, how different individuals prioritize values is significantly influenced by the socio-cultural environment and the decision context. For instance, consider how the conflict between the values of freedom and safety has shaped the conversation around COVID-19.
Ethical artificial agents must behave in accordance with stakeholders' values. Thus, values are the centerpiece of ethical sociotechnical systems (STSs). An important step toward realizing a value-aligned STS is *value inference*, the process of identifying values and reasoning about stakeholders' value preferences. Value inference is a prerequisite for creating systems that align with stakeholders' value preferences. There is an increasing body of AI literature on value inference, focusing on the identification of values, the classification of values in text, the estimation of individual value preferences, and the societal aggregation of value preferences. However, real-world applications often require a combination of these functionalities. The current literature does not offer a holistic view on how the pieces of value inference fit together.

The main goal of this research visit is to define and detail the concept of value inference. We do so by writing a position paper that targets the agent and multiagent community, for two reasons: (1) the importance of values is well established in this community, and (2) authors in both hosting and visiting research groups have recently published similar content in this community. First, we outline the challenges of value inference, and unify them in a modular framework. Then, we investigate how the interactions among humans and artificial agents are instrumental in improving the effectiveness of value inference by fostering self-reflection and deliberation. Finally, we show that value inference is a major research challenge not only for AI and multiagent systems, but it is an interdisciplinary research endeavor that concerns other areas such as sociology and ethics.

Impact

Value inference is an essential component towards the creation of value-aligned STSs, as it allows to gauge how stakeholders of an STS understand and prioritize values. For instance, value inference allows to understand how stakeholders in a hospital STS prioritize values, with the goal of designing value-aligned norms and protocols that both the humans and the artificial agents that operate in the hospital ought to respect. Similarly, value inference can allow policy makers to re-iteratively investigate how citizens prioritize values around COVID-19 policies, and hence decide whether to update the related regulations.
In practice, value inference is followed by the operationalization of values, both at agent and system level, which has been extensively explored in the multiagent community. Values have been used for modeling an individual agent's behavior, eliciting appropriate trust, plan selection, negotiation, social simulation, and engineering normative systems. We envision value inference and operationalization as actively influencing each other throughout the lifecycle of an STS. An example of such connection is the evaluation of norm compliance, i.e., assessing whether the implemented norms align with the inferred values.

Technical approach

Detailed description

We outline the challenge of value inference as a modular framework (Figure 1) consisting of the steps necessary to go from the behavioral data to the individual and aggregated value preferences. The dark blocks represent the processes, and the light blocks represent the data these processes consume or produce. This modularization has two advantages. First, the separation of concerns into processes delineates research challenges. Second, the interdependencies between processes expose research challenges that can otherwise fall through the gaps. For example, although value identification influences value preferences estimation and aggregation, these connections are largely unexplored.

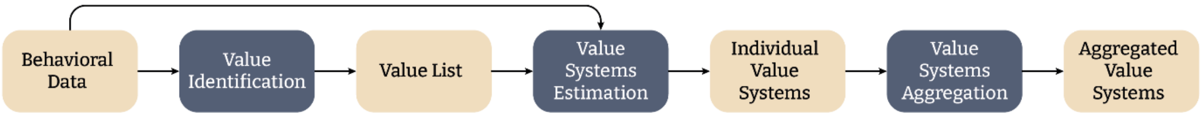


Figure 1: Value inference processes (dark-colored blocks) and data (light-colored blocks) as a modular framework

In our framework, we consider stakeholders' *actions* (e.g., how they choose over competing alternatives) and the *justifications* they provide for those decisions, as the behavioral data that constitutes the input to the value inference framework. We identify three fundamental processes of value inference as (1) value *identification* (determining which values are relevant to a context), (2) value preferences *estimation* (assessing how each stakeholder prioritizes values), and (3) value preferences *aggregation* (deriving a societal consensus from individual preferences). We describe the processes and the data they generate. Then, we discuss how value inference, as a purely AI task, is not likely to yield good estimates of individual and societal value systems. This is because value preferences are often implicit to humans and thus not easily observable in behavioral data. Hence, we must actively engage humans for successful value inference. This makes value inference a *hybrid intelligence* endeavor, requiring human and artificial intelligence to augment each other. Finally, we relate these challenges to emerging research topics in AI to demonstrate that value inference is a topic that can contribute to and benefit from interdisciplinary research.

Scientific outcomes

The outcome of our collaboration is a position paper where we explore the challenge of value inference. We outline the components of value inference (identification, estimation, and aggregation), and motivate how a hybrid intelligence approach is instrumental in performing value inference. Finally, we present the research challenges and opportunities spurred by value inference that span multiagent systems, other AI fields, but also other disciplines including ethics and social sciences.

Future plans

An extension of the work has been discussed, to broaden the scope from the multiagent community to the field of AI in general. Further, the authors were in parallel collaborating on another paper that connects value estimation to value aggregation, which has been submitted to Knowledge-Based Systems in December 2022. This work is an example of the connections among the components of value inference that our proposed framework spurs.

Self-assessment

- **AI Excellence**: The work we conducted during this research visit contributes to Trustworthy AI in three ways. First, the ultimate goal of value inference is the creation of value-aligned AI, a crucial step towards the design of beneficial Social AI (which is TAILOR’s WP6). With our work, we intend to facilitate the creation of AI that aligns with a plurality of value preferences. Second, our framework facilitates the combination of different AI paradigms, such as learning, reasoning, and optimization, in line with TAILOR’s WP4. Finally, we connect the challenge of value inference to several emerging topics that span across TAILOR’s work packages, such as explainability, robustness, fairness, and responsible autonomy.
- **Scientific step-up**: This research visit helped me grow professionally for two reasons. First, it allowed me to expand my network within the Barcelona AI landscape, including the host lab, the IIIA-CSIC institute, the VALAWAI project, the Decidim ecosystem, the AIRA initiative, and the Citibeats company (the latter of which led to an additional collaboration). Second, it pushed me to closely interact with researchers that think and work differently, due to different backgrounds and cultures, which is essential for becoming an experienced international researcher.

- **Suitability of the host**: The host lab was chosen for the complementarity of our expertise. Considering the three processes of value inference described above in the Technical Approach (identification, estimation, and aggregation), our group at TU Delft has expertise in the first two, and the host lab in the third. Further, the collaboration was extended to researchers at the IIIA-CSIC institute in Barcelona, who often collaborate with the host lab on value aggregation. Their combined expertise was crucial for providing a complete overview of value inference.
- **Suitability of the visit length**: In hindsight, strictly considering the work on the project, the visit length could have been 10 days shorter (i.e., ~15% shorter). However, the (slightly) more relaxed time frame has allowed me to spend time on expanding my network, by attending meetings and seminars that I would have instead avoided with a stricter timeline. I believe that, considering the combination of project work and networking, the duration was ideal.

List of publications, meetings, presentations, patents, …

The main outcome of the collaboration is a paper titled “Value Inference in Sociotechnical Systems”, submitted at the Blue Sky Ideas track of the AAMAS 2023 conference. The paper is currently under review, with author notification on February 19th, 2023. If accepted, I will present the work at the conference in London, between May 29th and June 2nd, 2023. Further, I presented the work at the IIIA-CSIC institute in Barcelona in November 2022, and I will present it at the Hybrid Intelligence Center (of which I am part) reading group in February 2023.

Additional comments

When submitting the application, it was not clear that the budget would be transferred from TAILOR to the host university, and from the host university to me (instead of being directly transferred from TAILOR to me). Furthermore, the response to the scholarship application was delayed by several weeks. The combination of these two issues has caused quite some struggle with the logistical aspects of the research visit. However, the contact persons within TAILOR (Joaquin Vanschoren and José Jong) have been kind and helpful.

Basic Information

Project title: Learning trustworthy models from positive and unlabelled data
Period of project: 17 April 2023- 17 May 2023
Period of reporting: -
Author(s): Paweł Teisseyre
Organization: Institute of Computer Science, Polish Academy of Sciences
Host organization Machine Learning Research Group, KU Leuven

Public summary.

Positive unlabelled (PU) learning is a subdomain of machine learning aiming in building binary classification model based on partially labelled training data. In PU learning, it is assumed that only some observations in training data are assigned label, which is positive, whereas the remaining observations are unlabelled and can be either positive or negative. PU datasets appear naturally in many domains, for example, in medical databases, some patients have been diagnosed with a disease (positive observations), whereas the remaining patients have not been diagnosed. However the absence of a diagnosis does not mean that the patient does not have the disease in question. Further examples include detecting illegal or detrimental content in social networks, under-reporting, image and text classification among many others. The goal of the research stay was to explore biased PU learning in which the probability of being labelled for the positive example may depend on certain characteristics of the given observation. This is particularly important in medical applications, where the probability of diagnosing the disease in a person who really has the disease may strongly depend on several factors such as age, availability of health care facilities, social status, level of education and others. To solve the problem, most existing approaches aim to estimate the propensity score function, which is a probability of being labelled for the positive example. Accurate estimation of the propensity score allows effective estimation of the posterior probability of the true class variable. We explored a method which is based on slightly different idea. It turns out that, under additional assumptions on the distribution of the feature vector and the form of aposteriori probability, PU learning is still possible, without direct estimation of the propensity score function. This leads to the three-step algorithm. In the first step, we rank the observations using biased learning model. In the second step, we find the threshold that separates the positive and negative observations. Finally, we label examples in training data according to the threshold and train new model using those labels. Experiments indicate that the proposed method performs similarly or better than competing methods based on propensity score estimation.

Research objectives

Objectives

Using standard supervised classification models requires fully labelled data, i.e. data for which each example in training data is assigned label (positive or negative). However, nowadays in a large number of practical situations, training data is labelled only partially. Fitting models based on such data is often called weak label learning. Learning trustworthy models using incompletely labelled data is challenging. The aim of the research stay was to explore positive-unlabelled (PU) scenario [1], which is one of the most important and frequently occurring variants of weak label learning. In PU scenario, it is assumed that some positive observations in training data are assigned labels whereas the remaining observations are unlabelled and they can be either positive or negative. In PU learning, usually some labelling mechanism is assumed, which decides which positive instances are labelled. The labelling mechanism can be described by propensity score function defined as probability of assigning label for positive instance. Prior works in PU learning [2, 3, 17, 18] assumed that all positive instances are equally likely to be labelled or, in other words, that propensity score is constant (Selected Completely At Random assumption: SCAR). Unfortunately, the SCAR is clearly violated in many real-life situations, where the propensity score is instance-specific, i.e. it depends on values that particular features take on. For example, in medical applications [4], certain features such as age, availability of health care facilities or presence of specific symptoms may be important factors in observing the diagnosis of a disease (e.g. prostate cancer). Assuming SCAR, when it is not true, leads to inaccurate prediction of the model, which in turn may lead to misleading conclusions for scientists who use the model in practice. Therefore, in the project we focused on learning PU models under selection bias, i.e., without assuming SCAR. The main objective was to design novel methods which overcome the existing ones when SCAR is not met.

Impact

In machine learning community, PU learning under selection bias has been identified as an important problem [5,6,7,8]. Accurately estimating the propensity score allows learning trustworthy models in situations when labelling bias is present. This is particularly important in medical applications, where the probability of diagnosing the disease in a person who really has the disease may strongly depend on several factors such as age, availability of health care facilities, social status, level of education and others. The labelling bias naturally appears in multi-label case, for example the probability of diagnosing the disease in a person who really has the disease may additionally depend on the presence of other diseases. Several application domains may potentially benefit from our research advances. The proposed PU learning method can be used to address important problems occurring in medical applications, such as handling databases containing under-reported diseases [14]. More specifically, the proposed method allows to predict the probability of the disease in a given patient in a situation when the model is trained on incomplete dataset containing under-reported diseases. The PU methods may be valuable to detect 'false negative' patients, i.e. those who have the disease but the disease is undiagnosed [9]. Further examples include application of the proposed method in image and text classification based on incompletely labelled datasets [12,13], dealing with under-reporting in surveys [14], detecting illegal content in social network analysis [15], among others.

Technical approach

Detailed description

Among the scientists working on PU learning, there is full agreement that accurate estimation of the propensity score function plays a crucial role and opens the door to effective estimation of the posterior probability of the true class variable which in turn allows finding the Bayes optimal classifier, i.e., the classifier which minimizes the expected 0-1 loss function [19].

In the proposed method we show that under additional assumptions made on (1) the form of the aposteriori probability and (2) the distribution of feature vector, recovering the Bayes optimal classifier is possible without direct estimation of the propensity score. We assume that aposteriori probabilities for both true and observed class variables are unknown functions of the linear combination of features. Such assumption is commonly used in robust statistics [11]. Importantly, we make no assumptions about the form of the aposteriori probability for the observed class variable as well as on the propensity score function. One can consider any function, including decreasing or non-monotonic functions. Thus, the common assumptions used in PU learning, such as SCAR [2], PG (Probabilistic Gap) [7] and IOO (invariance of order) [16] are special cases. On the other hand, we make an additional assumption on the distribution of the feature vector, namely that it is elliptical [10].The elliptical distributions are a broad generalization of Gaussian distributions, including distributions with heavy tails.

It turns out that under such assumptions, the direction of the hyperplane corresponding to the Bayes rule can be recovered using the naïve method, in which unlabelled examples are treated as negative. Unfortunately, the naïve method does not allow to determine the shift coefficient in a consistent way. Estimation of the shift coefficient is crucial as it is related to optimal threshold in Bayes rule. Hence, we propose to optimize the threshold by maximizing the mutual information between the linear combination of the features estimated by the naive model and the predicted class variable depending on the threshold. The final model is learnt using the predicted label vector as class variable. The assumption of ellipticity of the distribution seems strong, but it can be circumvented by applying feature vector transformation as a pre-processing step.

The proposed method is compared with state of the art methods: standard naïve method, PG [7], PUSB [16], EM [5], LBE [6] and oracle method that assumes the knowledge of true labels. The oracle method is used as a reference method. In experiments we use different labelling strategies corresponding to different assumptions considered in the literature (SCAR, PG, IOO). Experiments were performed on 20 datasets from UCI repository as well as on image datasets (MNIST, CIFAR, etc.). We also performed experiments on artificial datasets to analyse how the distribution of the feature vector affects the results.

Scientific outcomes

Experiments show that the proposed method allows to obtain comparable or even better results than methods based on direct estimation of the propensity function, even when the assumptions are not necessarily met. At the same time, the proposed method is much faster than the competitive methods (EM and LBE) which require learning two models (one for true class variable and one for the propensity score) in many iterations. The most important conclusion is that, under assumptions which are more general than existing ones (SCAR, PG, IOO), the naïve method can be successfully used to recover the correct ordering of observations. This, combined with the novel method of choosing the threshold, allows to recover the Bayes optimal classifier.

Future plans

The one-month research stay allowed me to gain additional experience in PU learning and establish cooperation with the research group at KU Leuven. The future plans are as follows.

- Completion of the publication describing the method and the experiments. Currently the manuscript is 80% ready, we plan to submit it at AAAI conference (deadline is 15 August 2023).
- Developing further methods based on the proposed one. This includes modifications of the proposed methods as well as application of the proposed method in multi-label setting.

- Developing novel methods of propensity score estimation, which have been discussed during my research stay, e.g. the method which estimates the propensity score function using sigmoid function depending on the posterior probability.

Self-assessment

Please provide your own final assessment of the effective progress against the goals stated in the proposal, according to the following points:

- **AI Excellence**: Did the visit contribute to Trustworthy AI? In what way? Nowadays, PU learning is one of the most intensively studied topic in machine learning. Unfortunately, choosing the incorrect method (e.g. the naive model or the model based on SCAR assumption, when SCAR is not true) can lead to poor performance of the model and misleading conclusions. To learn trustworthy PU learning models, it is necessary to gain a deeper understanding of labelling mechanisms. The method developed during the research visit is a significant step in this direction as we show precisely under what conditions the naive method can be used and how to modify the threshold to make it work.
- **Scientific step-up**. For research visits: did the visit help you grow professionally and reinforce your scientific reputation?
  The research stay allowed me to gain additional experience in PU learning and establish cooperation with the research group at KU Leuven. We intend to continue our cooperation. The joint publication will certainly help to reinforce my scientific reputation. The research visits are recommended and appreciated by my universities in Poland (Warsaw University of Technology and Polish Academy of Sciences). Finally, I had the opportunity to present my research work during the seminar at KU Leuven and discuss it with different researchers.
- **Suitability of the host**: How did the host lab (or workshop venue) help you achieve the proposed research?
  The Machine Learning research group at KU Leuven (host lab) is world leading group in the field of learning from PU data. Prof. Jesse Davis and Dr Jessa Bekker are authors of highly cited survey paper [1] on PU learning and other related papers, e.g. on class prior estimation from PU data [3] and propensity score estimation using EM algorithm [5]. The host lab contributed greatly to the new method, design of experiments and analysis of the results. The joint discussions helped to significantly improve the description of the results.
- **Suitability of the visit length**: In hindsight, was the visit length (or workshop length) adequate to reach your goals?

  During one-month research visit, I managed to achieve the main assumed goal, namely to develop a new method of PU learning, working under sample selection bias. This included: design of the algorithm, implementation, performing experiments on real and artificial datasets, writing manuscript. Due to the short duration of the visit, some issues are implemented after its completion (completion of the manuscript, additional experiments, application of the method to the multi-label case).

List of publications, meetings, presentations, patents,...

- Pawel Teisseyre, Timo Martens, Jessa Bekker, Jesse Davis, Bayes optimal positive-unlabeled learning via threshold calibration, manuscript, planned submission to the AAAI'24 conference.
- Pawel Teisseyre, On selected challenges in positive-unlabelled learning, presentation at the DTAI Seminar, KU Leuven.

References

1. Bekker, J. and Davis, J., Learning from positive and unlabeled data: a survey, Machine Learning, 2020.
2. Elkan, C. and Noto, K., Learning Classifiers from Only Positive and Unlabeled Data, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.
3. Bekker, J. and Davis, J., Estimating the Class Prior in Positive and Unlabeled Data through Decision Tree Induction, Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
4. Li, F., et. al., Positive-unlabeled learning in bioinformatics and computational biology: a brief review, Briefings in Bioinformatics, 2021.
5. Bekker, J. and Robberechts, P. and Davis, J., Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data, Proceedings of the 2019 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2019.
6. Gong, C. and Wang, Q. and Liu, T. and Han, B. and You, J. and Yang, J. and Tao, D., Instance-Dependent Positive and Unlabeled Learning with Labeling Bias Estimation, IEEE Trans Pattern Anal Mach Intell, 2021.
7. Gerych, W. and Hartvigsen, T. and Buquicchio, L. and Agu, E. and Rundensteiner, E., Recovering The Propensity Score From Biased Positive Unlabeled Data, Proceedings of the AAAI Conference on Artificial Intelligence, 2022.
8. Gong, C. and Zulfiqar, M. I. and Zhang, C. and Mahmood, S. and Yang, J., A recent survey on instance-dependent positive and unlabeled learning, Fundamental Research, 2022.
9. Walley, N. M., et.al., Characteristics of undiagnosed diseases network applicants: implications for referring providers, BMC Health Services Research, 2018.
10. Cambanis, S. and Huang, S. and Simons, G., On the theory of elliptically contoured distributions, Journal of Multivariate Analysis, 1981.
11. Li, K. and Duan, N., Regression Analysis Under Link Violation, Annals of Statistics, 1989.
12. Chiaroni, F. and Rahal, M-C. and Hueber, N. and Dufaux, F., Learning with A Generative Adversarial Network From a Positive Unlabeled Dataset for Image Classification, Proceedings of the IEEE International Conference on Image Processing, 2018.
13. Fung, G. P. C. and Yu, J. X. and Lu, H. and Yu, P. S., Text Classification without Negative Examples Revisit, IEEE Transactions on Knowledge and Data Engineering, 2006.

14. Sechidis, K. and Sperrin, M. and Petherick, E. S. and Luján, M. and Brown, G., Dealing with under-reported variables: An information theoretic solution, International Journal of Approximate Reasoning, 2017.
15. Chang, S., et. al., Positive-Unlabeled Learning in Streaming Networks, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
16. Kato, M. and Teshima, T. and Honda, J., Learning from positive and unlabeled data with a selection bias, Proceedings of International Conference on Learning Representations, 2019.
17. Lazecka, M. and Mielniczuk, J. and Teisseyre, P., Estimating the class prior for positive and unlabelled data via logistic regression, Advances in Data Analysis and Classification, 2021.
18. Teisseyre, P. and Mielniczuk, J. and Lazecka, M., Different strategies of fitting logistic regression for positive and unlabelled data, Proceedings of the International Conference on Computational Science, 2020.
19. Shalev-Shwartz, S. and Ben-David, S., Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2013.

Multi-Objective Statistically Robust Algorithm Ranking

Final report TAILOR connectivity fund grant

Authors: Jeroen Rook, Holger Hoos, Heike Trautmann
Date: 30-03-2023

This document reports on the research visit of Jeroen Rook, a PhD student at the Data Management and Biometrics Group at the University of Twente, to the Alexander von Humboldt Chair for AI Methodology (AIM) at RWTH Aachen University led by Prof. Holger Hoos. The research visit lasted three months and took place from 26 September 2022 until 31 December 2022. All costs to accommodate this visit were pledged to be covered by the TAILOR connectivity fund after approval of our proposal. In this final report, we reflect on the objectives set in that proposal. In summary, the visit was in many ways successful, provided promising additional aspects for future research, and the expenses for the visit remained within budget.

Scientific report

The primary purpose of the research visit was to create a methodology to evaluate and rank algorithms based on multiple performance objectives. Based on bootstrap sampling, this ranking method aims to obtain a more informed and statically robust understanding of how algorithms empirically relate to each other, which is helpful in various areas within AI.

In the proposal, we stated several results we expected to obtain. We reflect on each of them here:

- *An open-source, statistically robust algorithm ranking method that supports multiple performance objectives.* We created a Python package that does this, and we will release it, once the method and the manuscript have passed peer-reviewing.
- *An Implementation of the SO and MO algorithm ranking methods in Sparkle.* To make the method accessible to non-expert users, we planned to implement the single-objective and multi-objective ranking methods into the Sparkle platform. Both methods will be made available in the next release of Sparkle, before the end of 2023.
- *Case studies of the methods on at least two different problem domains.* Not many AI competitions assess algorithms based on multiple performance objectives, which makes finding suitable use cases challenging. We have one use case from the SAT competition and are still working on obtaining an ML competition use case.
- *A submission to a high-impact journal or conference.* We are currently preparing a manuscript to submit to ECAI 2023 (CORE ranking: A).

Besides the promising impact of the robust ranking method, Jeroen's visit also achieved some other benefits. One was establishing a pan-European collaboration between groups at the University of Twente and RWTH Aachen University. Another direction was to broaden Jeroen's research network and foster new collaboration partners for future research ideas. During the visit, at least two directions for research building on Jeroen'' work in Aachen were identified, on which we will work together with researchers from the AIM group at RWTH Aachen University in the future. In the context of his research, Jeroen was also integrated into the internal COSEAL network, which brings together researchers working on topics related to automated algorithm selection and configuration and is hence highly relevant for WP7 within TAILOR.

Financial report

The projected costs for the visit were estimated to be €4617.00, as can be seen in Table 1. These costs covered housing, travel and sustenance. The actual costs of €4400.44 remained well within the budget. The underlying budget categories, however, show some differences between projected and realised costs, due to the following reasons:

- The housing budget exceeded the projected costs, because of the limited availability of short-time rental properties in and near Aachen. Most rental properties demanded a minimal stay of 6 months or more. Due to availability of the rental apartment, Jeroen had to stay at a hotel in Aachen for the first week of his visit.
- The travel budget exceeded the projected budget, because only the travel costs for the intermediate return weekends were taken into account and not the trip to and from Aachen at the beginning and the end of the visit, respectively.
- There were no expenses for sustenance, since there was no need to use public transport to get from and to the university, and there was the possibility of preparing meals in the rented apartment.

Table 2 gives a detailed description of the expenses that have been incurred.

Table 1: Projected and realised budget

	Projected	**Realised**
Housing	€ 2550.00	€ 4025.00
Travel	€ 267.00	€ 375.44
Sustenance	€ 1800.00	€ -
Total	*€ 4617.00*	*€ 4400.44*

Table 2: Detailed expense sheet

	Declared expense	**Description**
Hotel	*€ 425.00*	*26-09-2022 tm 01-10-2022*
Apartement	*€ 1200.00*	*01-10-2022 tm 31-10-2022*
Apartement	*€ 1200.00*	*01-11-2022 tm 31-11-2022*
Apartement	*€ 1200.00*	*01-12-2022 tm 31-12-2022*
Travel	€ 140.79	*Outward trip + 1 return weekend*
Travel	*€ 234.65*	*Return trip + 2 return weekends*

Samples Selection with Group Metric for Experience Replay

Andrii Krutsylo

18.07.23

Basic Information

Project title: Samples Selection with Group Metric for Experience Replay
Period of project: 15.02.23 – 30.04.23
Period of reporting: 15.02.23 – 30.04.23
Author(s): Andrii Krutsylo
Organization: Polish Academy of Sciences
Host organization: University of Pisa

Public summary

In the field of artificial intelligence (AI), a crucial challenge is achieving continual learning, the ability of an AI model to incrementally learn from new data while retaining previously acquired knowledge. An obstacle in this task is the phenomenon of *catastrophic forgetting*, where new learning tends to overwrite older ones.

To mitigate this, we use a method called replay-based continual learning. Think of it as an AI that keeps a small notebook or a memory buffer of key learning points from previous tasks. As new tasks are learned, the AI revisits these stored notes, effectively *replaying* the learning process, hence mitigating the forgetting of prior knowledge.

Previous strategies mainly focused on selecting the most impactful individual samples (or notes) from the memory buffer. However, we found that the collective contribution of a group of samples in a batch is equally important, if not more. This led us to focus on selecting the most effective replay batch.

To do this, we introduced a new metric that evaluates a group of samples. Specifically, we measure the cosine distance between the hidden representations of group of samples before and after the model update, i.e., after new data is learned. This process quantifies how much the learning of a new task affected these samples. The batch that shows the most changes, indicated by the highest cosine distance, is deemed to be the most impactful for replay.

This research has significant implications for the development of trustworthy AI. By enabling models to retain older knowledge as they learn new tasks, we are creating systems that are more reliable and efficient. This contributes to AI behaving more predictably and consistently over time, which is essential for building trust in these systems. Our approach offers an innovative path for creating AI systems that are more proficient at continual learning.

Research objectives

Objectives

The primary goal was to modify the naive Eexperience Replay in a way that the new sample selection strategy would make the replay of past experiences more effective. This leads to two objectives: 1) creating a new metric capable of estimating the efficiency of a batch, and 2) finding the optimal batch using the new metric.

The challenge of the first objective is that, according to [5], the batch has three measures that influence its potential performance: informativeness, diversity and representativeness. It is hard to predict the importance of each measure at a specific training step of the model and approximate all into one metric. Therefore, the problem was reduced to combining or simultaneously improving only informativeness and diversity, while representativeness should be ensured by the method selecting samples to the memory buffer, which in our case is reservoir sampling.

The proposed metric is based on selecting stored in memory samples that have their loss increased after a model update on the new data. This approach was proposed in MIR [2] and since it focuses only on the informativeness of the samples, it requires the random selection of a relatively small subset (whose size is an adjustable hyperparameter) and then selecting the samples with the highest MIR value from this subset. This enhances the diversity of the batch, as similar samples would have similar MIR values. The same problem was encountered in Consistency Aware Sampling [6], where a certain number of samples with the highest CAS value were selected and then the batch was randomly sampled from this subset.

The proposed metric is similar to MIR, but instead of tracking changes in the samples' losses, I focus on hidden representations. Despite the lack of improvement as a measure of individual sample informativeness, it allows one to identify parts of the representations that changed the most and ensure that the final batch covers different parts, thus increasing diversity. During extensive research, I discovered that the optimal metric is the cosine distance applied to normalized representations of the last hidden layer for both the ResNet-18 and MLP models.

In the setting used [8], the number of possible combinations sampled from the memory buffer with a size of 1000 is greater than 263 trillion for a batch size of 10. The second objective was to improve performance while finding the optimal batch in a reasonable number of trials. For this purpose, I implemented Hill Climbing and Genetic algorithms designed to accelerate the search, but surprisingly, only ten random trials were sufficient to find a batch that would improve performance over the MIR baseline. Increasing the number of trials does not necessarily improve the result. This makes combinatorial searches with complicated methods unnecessary. Overall, it is a great success that the proposed approach remains fast, so this objective is also considered to be completed.

Impact

The whole branch of methods that uses past experiences of the model for replay or regularization has significantly improved by being able to select or generate samples in diverse batches in addition to their metrics of individual sample impact.

Technical approach

Detailed description

The continual learning approach developed in this research uses an online learning setting [8] applied to MNIST and CIFAR-10 datasets. Each dataset is divided into tasks, each consisting of two classes. For MNIST, only 1,000 training samples per class are considered [8]. Both datasets are processed with a batch size of 10, with a corresponding replay batch size of 10 chosen from the memory buffer. Each online batch is exposed to the model only once. I utilize the ResNet-18 model architecture for CIFAR-10 and a Multilayer Perceptron with a 400-neuron hidden layer for MNIST.

Current selection strategies in replay-based learning, such as Gradient-Based Sample Selection [1], Representation-Based Sample Selection [11], MIR, and CAS emphasize individual sample properties, often neglecting the overall impact of the batch on model performance. In response, I propose a batch-level optimization method that assesses the combined effect of all samples within a batch.

My method diverges from conventional selection techniques by focusing on the cumulative effect of samples within a batch. This method uses the cosine distance between the averaged hidden representations of a set of samples before and after the update of the model as a batch-wise impact measure. The cosine distance offers a measure of how the learning of new tasks has affected these samples collectively.

My sampling method follows a random and uniform distribution, avoiding samples that belong to the current task. I then extracted hidden representations of these samples both before and after the model update. I measured the similarity between averaged hidden representations before and after the model update using the cosine similarity metric. This metric is computed by normalizing the two sets of embeddings and calculating the cosine similarity between them. This process is repeated for several batches, and the batch with the highest cosine similarity is selected to be replayed. I am also evaluating a batch produced by MIR as a strong baseline, but it outperforms the random one only in average 17.3% cases and in 35.82% they share at least one common sample.

Scientific outcomes

Table 1: The average accuracy across all five tasks of the split MNIST and CIFAR-10 settings, evaluated after learning the whole sequence. Each value is the average of 20 runs with standard deviation.

CIFAR-10

Memory	ER	Class-balanced	MIR	**Batch Sample**
200	22.54 ± 2.00	23.38 ± 1.64	24.39 ± 1.85	**27.74 ± 3.45**
500	26.51 ± 2.76	26.13 ± 2.74	32.06 ± 3.51	**36.79 ± 2.73**
1000	28.53 ± 4.17	28.87 ± 2.86	40.09 ± 3.52	**42.15 ± 2.08**

MNIST

Memory	ER	Class-balanced	MIR	**Batch Sample**
200	79.50 ± 5.00	80.37 ± 2.72	80.86 ± 4.09	**81.81 ± 2.70**
500	84.91 ± 3.61	85.06 ± 1.88	85.09 ± 5.61	**87.01 ± 3.25**
800	86.22 ± 2.29	85.64 ± 2.88	**89.47 ± 1.94**	89.33 ± 2.49

The Table 1 shows that the proposed batch selection method can improve the performance of Experience Replay in Class Incremental [10] setting for different training data and memory budgets.

Insights

1. Random selection of samples leads to fair balance between informativeness and diversity, but picking the most informative samples in most cases will harm the diversity of the batch.

2. It is possible to improve batch diversity without relying completely on randomness.

3. The improved diversity of the batch does not necessarily mean the diversity across a few batches, which was left for future work.

4. The data complexity metrics [9] that were successfully used instead of the classification error to assess the goodness of subsets of features [3] are too slow to be applied after each update of the model, but could be used in future work for meta-learning.

Future plans

I have established perspective collaboration with my host laboratory PAILab in the University of Pisa that already leads to us working on a new research and opens the possibility of my future postdoc studies in this institution.

The incoming publication ”Batch Sampling for Experience Replay” could be further improved to address the issue of diversity not only inside one batch but across a sequence of selected batches.

Self-assessment

Please provide your own final assessment of the effective progress against the goals stated in the proposal, according to the following points:

- AI Excellence: Did the visit contribute to Trustworthy AI? In what way?

  I have successfully implemented an easy to use, highly modifiable, and compatible tool to improve arguably the most powerful technique of Continual Learning, the field that makes an AI more trustworthy in many ways [4].

- Scientific step-up: For research visits: did the visit help you grow professionally and reinforce your scientific reputation?

  It was the most valuable experience I have, in terms of both professional growth and development of scientific reputation. During my visit, I have learned a lot of new tools and developed strong baselines along with an expandable codebase that I am planning to use in the next research. Participation in laboratory meetings allowed me to familiarize myself with the most promising publication in Continual Learning and the recent developments in PAILab and the Italian laboratories partnering in this field.

  From scientific career perspective my research visit was highly evaluated at my research institute and according to my thesis advisor, along with published results, it would be the locomotive of my CV and the first topic for internships and postdoctoral interviews.

- Suitability of the host: How did the host lab (or workshop venue) help you achieve the proposed research?

  In my host lab, I learned from the developers of the most popular CL framework [7] how to optimize my code, which would greatly speed up my experiments and would have a long-term impact on all of my work. It was also very beneficial to work on this project with people who had already faced the problem of selecting the diverse samples in their previous research and who already have some unpublished insights and recommendations.

- Suitability of the visit length: In hindsight, was the visit length (or workshop length) adequate to reach your goals?

The visit length fits perfectly. The research appears to be more complicated than expected, so I was unable to submit it to the scientific conference during my stay in the host lab. But the main part that requires intensive collaboration was completed successfully and the article was finished remotely shortly after the visit.

List of publications

1. (in review) Andrii Krutsylo. "Batch Sampling for Experience Replay"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952215.

References

[1] Rahaf Aljundi et al. "Gradient based sample selection for online continual learning". In: *Neural Information Processing Systems*. 2019.

[2] Rahaf Aljundi et al. "Online Continual Learning with Maximally Interfered Retrieval". In: *ArXiv* abs/1908.04742 (2019).

[3] Verónica Bolón-Canedo, Guillermo Castillo García, and Laura Morán-Fernández. "Feature selection for transfer learning using particle swarm optimization and complexity measures". In: *ESANN 2022 proceedings* (2022).

[4] Andrea Cossu, Marta Ziosi, and Vincenzo Lomonaco. "Sustainable Artificial Intelligence through Continual Learning". In: *CoRR* abs/2111.09437 (2021). arXiv: `2111.09437`. URL: `https://arxiv.org/abs/2111.09437`.

[5] Adrian Englhardt et al. "Finding the Sweet Spot: Batch Selection for One-Class Active Learning". In: *SDM*. 2020.

[6] Julio Hurtado et al. "Populating Memory in Continual Learning with Consistency Aware Sampling". In: 2022.

[7] Vincenzo Lomonaco et al. "Avalanche: an End-to-End Library for Continual Learning". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2021), pp. 3595–3605.

[8] David Lopez-Paz and Marc'Aurelio Ranzato. "Gradient Episodic Memory for Continual Learning". In: *NIPS*. 2017.

[9] Ana Carolina Lorena et al. "How Complex Is Your Classification Problem?" In: *ACM Computing Surveys (CSUR)* 52 (2019), pp. 1–34.

[10] Marc Masana et al. *Class-incremental learning: survey and performance evaluation on image classification*. 2020. URL: `https://arxiv.org/abs/2010.15277`.

[11] Vinay Venkatesh Ramasesh, Ethan Dyer, and Maithra Raghu. “Anatomy of Catastrophic Forgetting: Hidden Representations and Task Semantics”. In: *ArXiv* abs/2007.07400 (2021).

Basic Information

Project title: Towards Prototype-Based Explainable Machine Learning for Flood Detection
Period of project: 16.01.2023 - 24.02.2023
Author(s): Ivica Obadic
Organization: Technical University of Munich (TUM), Chair of Data Science in Earth Observation
Host organization: Lancaster University, Lancaster Intelligent, Robotic and Autonomous Systems Centre (LIRA)

Public summary

Extreme flooding events appear with increasing frequency in recent years. Automated flood detection systems can be of utmost importance in the early detection of floods and efficient disaster management. While deep learning models trained on remote sensing data have shown to be a promising approach for accurate flood detection, the lack of interpretability for their decisions limits the real-time usage of these models in such critical applications. During this research visit, we conceptualized our approach based on graph neural networks for automated and interpretable flood detection from remote sensing data. The initial results demonstrate its potential for accurate flood prediction. Furthermore, the introduced graph representation learning helps present the prediction in an interpretable way that captures the importance of individual fragments of the image. For example, our explainability analysis reveals that the proposed model captures that objects near the flood water typically bear high importance for flood detection. In an ongoing work after the end of this research visit, we further improve the transparency of the proposed approach by identifying the relevant relationships and prototypical patterns, with an ambition to design a prototype-based graph neural network for flood detection. In summary, this research visit has enabled a valuable exchange of knowledge and expertise between the two labs and established the basis for a long-term collaboration that will foster the development of novel and trustworthy deep learning approaches for the nowadays critical environmental problems like flood detection.

Research Objectives

Objectives

Deep learning approaches based on remote sensing data produce state-of-the-art flood detection results in recent years. These approaches typically detect the flood extent by training a convolutional neural network (CNN) that extracts relevant flood information from satellite images [1]. However, due to their complex nature, deep learning models are known to be black-box models that do not provide human-understandable explanations for their predictions [2]. Moreover, CNNs are known for their bias towards texture rather than geometric shapes for object recognition [3] and they also struggle to capture the contextual relations between the different objects in the scene [4]. These limitations represent a serious challenge when relying on CNN for flood detection where the recognition of certain shapes

and modeling the complex interactions between the natural and built-up environments (e.g. houses or cars floating in water) is crucial to capture the flood semantics [1].
With this research visit, we aim to address the above shortcomings by proposing an interpretable deep learning approach based on graph neural network (GNN) for accurate flood detection from remote sensing data. GNNs are capable of modeling the complex relationships inherent in the data, and as such, they excel in a wide range of applications like traffic forecasting, recommender systems, or drug discovery [5]. Therefore, our proposed approach that combines GNN with explainable artificial intelligence (xAI) has the following research objectives:

1. Investigate the capability of GNNs to efficiently model and uncover the complex patterns and relationships that describe floods.
2. Introduce a self-interpretable GNN for flood detection based on prototypical flood patterns.

Impact

Our research objectives contribute to improving the accuracy and transparency of flood detection systems. We believe that this will lead to higher stakeholder acceptance and increased real-time usage of these systems, and ultimately, more efficient disaster management. Moreover, to the best of our knowledge, our work represents the first effort that combines GNNs and xAI for flood detection, and as such, it unlocks new research perspectives for trustworthy flood detection.

Technical approach

Detailed description

Our proposed approach for interpretable flood detection based on GNNs is illustrated in Figure 1. In the first step, the remote sensing images are encoded into graphs. Next, a GNN is trained for flood prediction. Finally, the learned flood patterns are revealed with interpretability analysis.

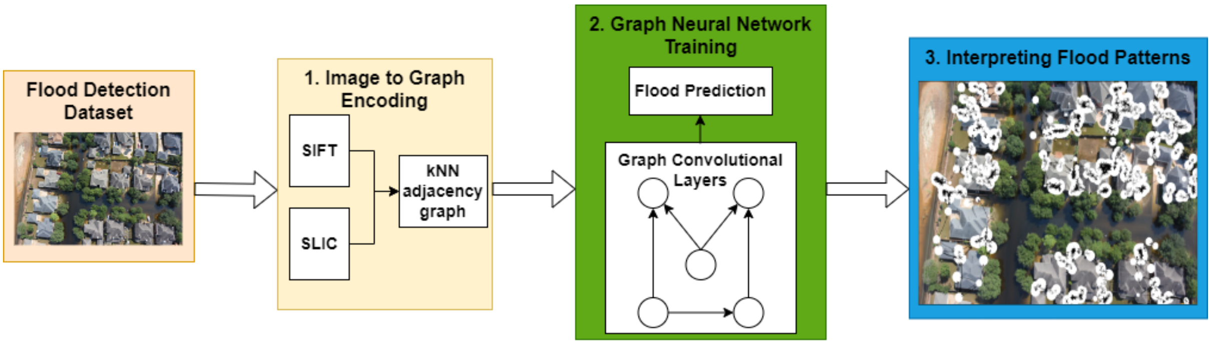

Flood Detection Dataset
1. Image to Graph Encoding
SIFT
SLIC
kNN adjacency graph
2. Graph Neural Network Training
Flood Prediction
Graph Convolutional Layers
3. Interpreting Flood Patterns

Figure 1: The proposed approach for interpretable flood detection based on GNNs. First, we encode the remote sensing images into graphs by using keypoint extraction methods followed by a construction of kNN adjacency matrices. Next, we train a GNN for flood prediction on these graphs. Finally, we reveal the learned patterns for flood detection by visualizing the graph node and edge importance onto the remote sensing image.

1. Image to Graph Encoding

We encode the remote sensing images into graphs by first extracting distinctive image key points that represent the graph nodes. We evaluated two baseline methods for extracting key points, namely SIFT [6] and SLIC [7]. The SIFT method extracts local key points from the image and computes a 128-dimensional feature descriptor that represents a histogram of the gradient distributions in the neighborhood of a key point. On the other hand, the SLIC algorithm segments the image into superpixels using K-means clustering. The key point positions correspond to the center of the superpixel with the feature representation being the average of the pixel values within the superpixel. Having extracted the graph nodes and their features, we construct two graphs for the used key point extraction methods with adjacency matrices based on the proximity of the node positions in the image and the similarity of their features.

2. Graph Neural Network Training

Our proposed GNN model consists of three graph convolution layers followed by an average pooling layer and a linear classification layer that performs the flood prediction. The graph convolutional layers create high-level feature representations for the nodes in the graph by aggregating the information from the node and its neighbors in the adjacency matrix [8]. Further, the pooling layer creates a representation for the entire graph by averaging the high-level feature representation for each node. Finally, the linear layer uses the average graph representation to predict the existence of a flood in the remote sensing image.

3. Interpreting Flood Patterns

We implemented the GNNExplainer algorithm to reveal the relevance of the nodes and edges in the graph towards flood prediction [9]. Next, we interpret the learned flood detection patterns from our proposed GNN by projecting the graph and the relevance of its nodes and edges onto the remote sensing image.

Scientific outcomes

We evaluated our approach on the FloodNet dataset which contains high-resolution Unmanned Aerial Vehicle images captured after the Harvey hurricane [10]. The dataset includes 2343 images with only 16% of the images having labels, presenting 51 examples of flood and 320 examples of non-flood. We used 60% of the examples for training our proposed GNN, and the rest 40% to assess its prediction performance. In our experiments, we evaluate the effect of the keypoint extraction methods, the number of nodes in the graph as well as the choice of adjacency matrix. The top-3 prediction scores are listed in Table 1 and they demonstrates that a GNN with SIFT-based graph encoding can be a powerful approach for modeling flood as our results slightly outperform the state-of-the-art accuracy of 0.94 reported in [10] achieved with the standard ResNet50 CNN network. Moreover, the reported F1 score for the flooded class shows that eventhough flood examples are underrepresented in the dataset, our approach can still identify floods with high accuracy.

Graph Encoding	Num. Nodes	Adjacency matrix	Accuracy	F1 Score
SIFT	1000	Feature similarity	0.95	0.82
SIFT	500	Feature similarity	0.95	0.77

Graph Encoding	Num. Nodes	Adjacency matrix	Accuracy	F1 Score
SIFT	500	Node proximity	0.94	0.76

Table 1: Prediction performance of our proposed GNN approach.

Having shown that our approach can lead to accurate flood detection, we uncover the transparency of the proposed GNN model by estimating the node and edge importance with the GNNExplainer method. In the below figure we visualize two examples of flooded images and their corresponding graphs. The white circles represent the graph nodes and the lines are the graph edges. Further, the size of the nodes and the opacity of the edges indicate their relevance to flood prediction. The both images show that typically, nodes representing objects like trees and houses that lie close to the flood water have high importance while the nodes inside the flooded water not related to any object have lower relevance. Furthermore, the edge importance is uniformly distributed in the both images which doesn't reveal specific relationships that contribute to flood decisions.

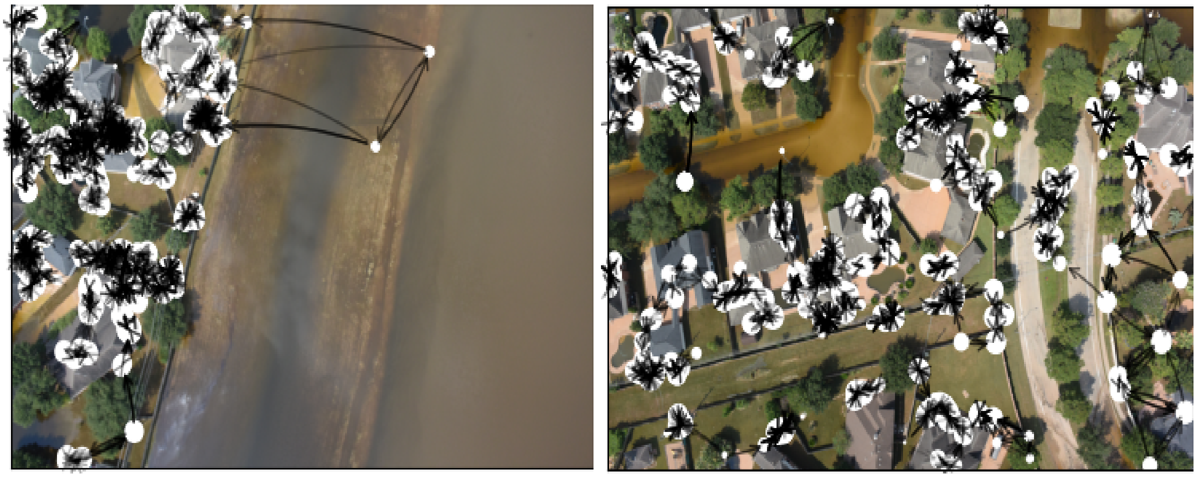


Figure 2: Examples of flooded images and the relevance of the graphs nodes and edges for flood prediction. The size of the circle represents the node importance and the line opacity describes the edge importance.

Future plans

Our findings indicate that the proposed approach attributes high relevance to objects close to the water for flood detection which leads to intuitive interpretation of the model workings. However, from the analysis of the edge importance, we have also seen that the proposed GNN does not reveal specific and interpretable relationships that describe flood. Therefore, to further improve improve the transparency of the proposed model, currently we evaluate novel deep learning key point extraction methods and introduce layers in our architecture capable to disentagle the relevant relationships. Uncovering these relationships is an essential step towards understanding the flood patterns and achieving our objective of designing a self-interpretable GNN based on flood prototypes. Prototype-based neural networks are getting increasingly popular as they enable inherent interpretability by directly relating the model decisions with intuitive prototypical patterns [11]. Hence, we believe that introducing a transparent prototype-based GNN approach is a novel research contribution that has high potential to enable automated trustworthy systems for flood detection. After the completion of these steps, we aim to summarize our contribution and results in a manuscript that will be submitted to the International Conference on Machine Learning.

Self-assessment

- **AI Excellence**: Did the visit contribute to Trustworthy AI? In what way?
    - This research visit helped in establishing a long-term collaboration focused on the development of an interpretable approach for flood detection. We believe that our joint efforts and the initial results towards designing a novel approach based on prototype-based GNN that reveals the common flood patterns contribute highly to trustworthy AI as they can provide accurate and self-interpretable systems for the critical problem of flood detection.
- **Scientific step-up**: For research visits: did the visit help you grow professionally and reinforce your scientific reputation?
    - This research visit enabled me to expand my professional network with experts in xAI. As a result, besides improving my methodological and technical knowledge, I also gained valuable insights for identifying relevant research questions, structurally approaching open problems, and improving my presentation skills. I believe that these skills are highly relevant for my further professional growth and that the planned joint publications with the LIRA center will have a tremendous impact on my scientific reputation as a researcher in xAI.
- **Suitability of the host**: How did the host lab (or workshop venue) help you achieve the proposed research?
    - The host lab encouraged frequent meetings and open exchanges where they shared their expert knowledge in xAI systems and we jointly brainstormed our proposed approach which helped in achieving the proposed research objectives. Further, I am very grateful for the invitation to present my work and insights on xAI in earth observation on the LIRA seminar which was a valuable experience in giving a scientific talk and also helped to further expland my professional network.
- **Suitability of the visit length**: In hindsight, was the visit length (or workshop length) adequate to reach your goals?
    - The visit length was sufficient to conceptualize our approach and establish the basis for a long-term collaboration between the two labs.

List of publications, meetings, presentations, patents,...

- Invited presentation on 18.01.2023 to the LIRA seminar on the topic: Overview of Common Explainable Artificial Intelligence (xAI) Methods and their Applications, Challenges and Outcomes in Earth Observation
- Ongoing recurring weekly progress update meetings with Dr. Dmitry Kangin from LIRA centre
- Ongoing recurring monthly progress update meetings with Prof. Dr. Plamen Angelov (LIRA center) and Dr. Dmitry Kangin (LIRA center) and Prof. Dr. Xiaoxiang Zhu (head of the Chair of Data Science in Earth Observation at the Technical University of Munich).

References

[1] Bentivoglio, Roberto, et al. "Deep learning methods for flood mapping: a review of existing applications and future research directions." Hydrology and Earth System Sciences 26.16 (2022): 4345-4378.
[2] Ras, Gabrielle, et al. "Explainable deep learning: A field guide for the uninitiated." Journal of Artificial Intelligence Research 73 (2022): 329-397.
[3] Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." arXiv preprint arXiv:1811.12231 (2018).
[4] Li, Yin, and Abhinav Gupta. "Beyond grids: Learning graph representations for visual recognition." Advances in neural information processing systems 31 (2018).
[5] Veličković, Petar. "Everything is connected: Graph neural networks." Current Opinion in Structural Biology 79 (2023): 102538.
[6] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60 (2004): 91-110.
[7] Achanta, Radhakrishna, et al. "SLIC superpixels compared to state-of-the-art superpixel methods." IEEE transactions on pattern analysis and machine intelligence 34.11 (2012): 2274-2282.
[8] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2016).
[9] Ying, Zhitao, et al. "Gnnexplainer: Generating explanations for graph neural networks." Advances in neural information processing systems 32 (2019).
[10] Rahnemoonfar, Maryam, et al. "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding." IEEE Access 9 (2021): 89644-89654.
[11] Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." Advances in neural information processing systems 32 (2019).

Basic Information

Project title: TAILOR workshop - Open Machine Learning
Period of project: June 12-16, 2023
Author(s): Meelis Kull, Joaquin Vanschoren, Peter Mattson, Pieter Gijsbers
Organization: University of Tartu

Public summary

The field of Machine Learning continues to grow tremendously and has a significant impact on society. As such, it is important to *democratize* machine learning, i.e. to make sure that software, datasets, models, and analyses are freely available for easy discovery, verifiability, reproducibility, reuse and meta-analysis. An ecosystem of platforms and tools has emerged to this end, including open platforms such as OpenML and HuggingFace that allow for easy sharing of datasets, models, and experiments, and automated machine learning tools that help people build better models.

This inclusive workshop has brought together researchers and practitioners from academia, industry, and anyone interested in contributing to this goal. It was organized as a full-week hackathon that includes the developers of these platforms and tools, as well as anyone who wishes to contribute or learn more about them, in order to further democratize machine learning, make them more interoperable and more useful for machine learning researchers.

Research objectives

We organized this workshop with the following objectives:

- Improved access to machine learning datasets, especially through interoperability with other ML data platforms. This could include progress towards shared best practices and increased standardization, but also actual implemented bridges between platforms, strengthening existing ML communities.
- Outreach to new subcommunities and researchers who previously had limited opportunities to interact with the people behind ML platforms. The level of participants varies from young to senior researchers, allowing a free transfer of knowledge between them.
- Early stage researchers will be able to make contributions with considerable impact that will strengthen their track record, as well as get to know new collaborators to help their future careers.
- Improvements to the ML platform itself, such as new features requested by many ML researchers and features related to trustworthiness, such as data and model quality checks. The ML platform becomes easier to use by students, researchers from other sciences, and the general public.
- Improvements in automated machine learning tools that make it easier for domain scientists to find out which ML techniques to try.
- Publications based on new ideas or specific work started during this workshop.
- New long-term collaborations are formed, resulting in papers co-authored by workshop participants that did not collaborate before.

Progress against planned goals

During the hackathon, we achieved the following results:

OpenML Website improvements

We worked on a new OpenML frontend (website) to make all the AI resources in OpenML (especially datasets, models, and benchmarks) easy to find and reuse. Significant progress was being made both on the website itself and 'under the hood'. We did a major refactoring that included the use of modern web development frameworks. This lead to a reduction of the codebase of over 4,000 lines of code. This will drastically facilitate the maintenance of the platform and to make it easier for new collaborators to understand the code and make their own contributions, which is a major aspect of open source development. In the coming weeks and months, we hope to finish this work and update the documentation so that we can deploy the new website and work with new contributors.

Croissant format

We kicked off community development of the Croissant ML dataset format at the OpenML hackathon. Croissant is a high-level format that combines metadata, data resources, data structure, and default ML semantics into a single file to make ML datasets easier to find, use, perform research on, and define tools for. Early contributors include representatives from ETH Zurich, Google, Hugging Face, Kaggle, King's College, Meta, MLCommons, and OpenML. The OpenML hackathon provided a great physical anchor for our kick off meeting which drew over thirty people from the broader ML community to learn more about the format and provide technical input to further its development.

Python Interface to OpenML

The 'openml-python' package provides programmatic access to all data on OpenML from Python environments. The package has been around for many years, and needed modernization. For the users, our next release (which we worked on during the hackathon) will prefer a modern data science software stack (i.e., use pandas dataframes) and use lazy loading by default (only downloading files if you really need to), which makes the experience faster and more efficient for users while reducing the strain on the server. For ease of maintenance and contributors, we have worked on modernising our code base by starting to provide stricter type annotations and migrating unit tests from Python's built-in unit test framework to the industry-standard 'pytest' framework. This combined effort ensures that users will have a modern experience and contributors can extend openml-python so it stays relevant for future use cases.

The AutoML benchmark

The AutoML benchmark is a software tool for benchmarking automated machine learning software. During the hackathon we added the ability to upload benchmark experiment results directly to OpenML. This makes results of the benchmark more easily obtainable and comparable while increasing reproducibility.

NLP Benchmarking

Currently OpenML focuses on numerical datasets and benchmarking, however, in recent years, there has been an explosion of research in the NLP space. There was an effort during the hackathon to enable the uploading of NLP datasets, eg: question answering datasets. During the hackathon, a new type of dataset was added to the testing server to enable this functionality.

Dataset availability

An important use case of OpenML is its feature as a dataset repository that provides its data in a machine-readable output format. However, a dataset repository is only as good as the datasets it provides. During the hackathon, an effort was made to extend the number of datasets on OpenML with new, relevant datasets that have not been on OpenML yet. As a result, we now have access to about a dozen datasets that have been used in recent studies, and also identified shortcomings of the dataset upload feature that prevent the upload of overly large datasets.

Dataset Versioning

Now OpenML contains several versions of the same dataset, which makes finding a dataset difficult. We had several discussions about how to make searching for datasets on the OpenML page more efficient.

Enriching Dataset Visibility and Accessibility on OpenML through Ontologies

Ontologies can serve as a crucial tool for navigating and matching different datasets. These ontologies, semantic structures that encapsulate a rich set of concepts and their interrelationships, provide a common terminology framework that ensures interoperability between disparate data sources. They also enable the accurate matching of different but related data entries across multiple datasets. For instance, in healthcare an ontology can bridge the variations between datasets by recognizing that "MI" in one dataset refers to the same concept as "myocardial infarction" in another. This consistent semantic environment allows researchers and engineers to effectively query, analyze, and interpret complex and heterogeneous data sources. Based on a pull request by Jan van Rijn, Katarzyna Woźnica and Luis Oala plan to test ontology support on ooenml.org for 10 exemplary tabular datasets from the healthcare domain.

Self-assessment

AI Excellence: Did the visit contribute to Trustworthy AI? In what way?

We managed to make real progress towards improving OpenML, which is a platform that in itself contributes to Trustworthy AI by making AI research research result reproducible by design, by democratizing access to AI resources and results, and by establishing we standards for AI workflows, such as the Croissant dataset format we are building together with other platforms such as TensorFlow, HuggingFace, and Kaggle.

We would have liked to have worked more with the AI Fairness community as well, but sadly our workshop coincided with the FAccT conference, which means that most people in this community couldn’t attend. However, we do have a fairness-related layer in the Croissant dataset format that will contribute to this, and we have this continued collaboration on our roadmap.

Scientific step-up: For workshops: how did the workshop help the participants?

Yes, these workshops are a great opportunity for people to free up some time to work on open-source projects such as OpenML, which is sometimes hard given all other academic duties. Everybody was very happy with that. We also had several new people attending the hackathon, partly supported by the diversity travel grants that we could offer thanks to the connectivity fund, and these people were able to build a much deeper understanding of OpenML and how to contribute to it. Making contributions to open science projects is a very positive aspect of many AI profiles and careers.

We would have liked to have welcomed more new people. Unfortunately, several people could not attend due to Visa issues (even if they were awarded a travel grant) or because their companies ultimately didn’t allow the travel. In the future we plan to leave more time for people to arrange travel and visas. This year this was difficult due to the timing between receiving the grant and the planning constraints at the host site.

Suitability of the host: How did the host lab (or workshop venue) help you achieve the proposed research?

The venue was fantastic. The facilities were very modern and spacious, which allowed both quick efficient breakout sessions and online discussions (e.g. the Croissant announcement which was streamed from the workshop). There were regular coffee breaks, well-organized social events, and the opportunity for people to really focus on their work (e.g. by bringing in pizzas to work through the evenings.) It was also great to meet the data science community in Estonia.

Suitability of the visit length: In hindsight, was the visit length (or workshop length) adequate to reach your goals?

Yes, a week is a good time for a hackathon such as this. Longer would have been great to get more work done, but that would make it harder for many people to attend. Shorter would probably not work well, since most people need a few days to learn the underlying concepts to meaningfully contribute.

TAILOR Technical Report

Optimizing AI: MILPs for Discrete Neural Networks

PhD students: Ambrogio Maria Bernardelli, Simone Milanesi
Host institution: TU Delft, in the research group of Dr Neil Yorke-Smith
Visiting duration: One month, from mid-April to mid-May

Abstract

The research topic is the training and simultaneous optimization of neural networks by means of exact discrete optimization solving technology. This project is divided into two main sections. For the first section, we leverage the expertise of Dr N. Yorke-Smith on Integer Neural Networks [7] to study the interactions of the methodology proposed in our previous paper [1]. Also, different paradigms for both robustness and simplicity are investigated and outcomes are described. To quickly determine the most promising direction, we conduct several experiments on small datasets. We analyze the results and plan to use them as the basis for larger-scale experiments, with the ultimate goal of improving our paper [1] for a journal version. The second section of our work focuses on a theoretical analysis of Graph Neural Networks (GNNs). While this latter section does not introduce any novel ideas, it serves as a foundation for a project that requires further exploration, and that will take place in a future collaboration.

Contents

1 Preliminaries

In this section, we summarize the key concepts of our previous work that will be used in this report, together with the main notation.

1.1 Binarized Neural Networks

Binarized Neural Networks (BNNs) are getting increasing attention thanks to their compactness and versatility. In this kind of neural network4, every neuron $j \in N_l$ is connected to every neuron $i \in N_{l-1}$ by a weight $w_{ilj} \in \{-1, 0, 1\}$. Given a value x for input neurons, the preactivation $a_{lj}(x)$ of neuron $j \in N_l$ and the activation $p_j(x)$ can be written, respectively, as

$$a_{lj}(x) = \sum_{i \in N_{l-1}} w_{ilj} \cdot p_{(l-1)i}(x), \tag{1}$$

$$p_{lj}(x) = \begin{cases} x_j & \text{if } l = 0, \\ +1 & \text{if } l > 0, a_{lj}(x) \geq 0, \\ -1 & \text{otherwise.} \end{cases} \tag{2}$$

Recent works [8] show that this kind of networks are hard to train with GD-based algorithms in a context of few-shot learning. Instead, MILP approaches are being researched.

1.2 Lexicographic Multi-Objective Function

A few MIP models are proposed in the literature to train BNNs efficiently. In this work, to train a single BNN, we use a lexicographic multi-objective function that results in the sequential solution of three different MIP models: the Sat-Margin (`SM`) described in [7], the Max-Margin (`MM`), and the Min-Weight (`MW`), both described in [8]. The first model `SM` maximizes the number of confidently correctly predicted data. The other two models, `MM` and `MW`, aim to train a BNN following two principles: robustness and simplicity. While robustness is achieved by maximizing the margin of each neuron, thus making it less sensitive to input perturbations, simplicity is achieved by minimizing the number of connections between neurons, that is, the number of non-zero weights. Our model is based on a lexicographic multi-objective function: first, we train a BNN with the model `SM`, which is fast to solve and is always feasible. Second, we use this solution as a warm start for the `MM` model, training the BNN only with the images that `SM` correctly classified. Third, we fix the margins found with `MM`, and minimize the number of active weights with `MW`, finding the lightest BNN with the robustness found by `MM`. At each step, the worst-case scenario is obtaining the solution found in the previous step. Given that the first step is always feasible, also the overall model is always feasible.

1.3 Structured Ensemble

Given a multiclassification problem, naming $\mathcal{I}$ the set of classes, we train one network for each pair of elements of $\mathcal{I}$. When given unseen data, we feed it to our list of networks obtaining a list of outputs and we then apply a majority voting scheme. We are basically using a one-versus-one scheme.

The idea behind this structured ensemble is that, given an input $\boldsymbol{x}^k$ labelled l $(= y^k)$, the input is fed into $\binom{n}{2}$ networks where $n - 1$ of them are trained to recognize an input with label l. If all of the

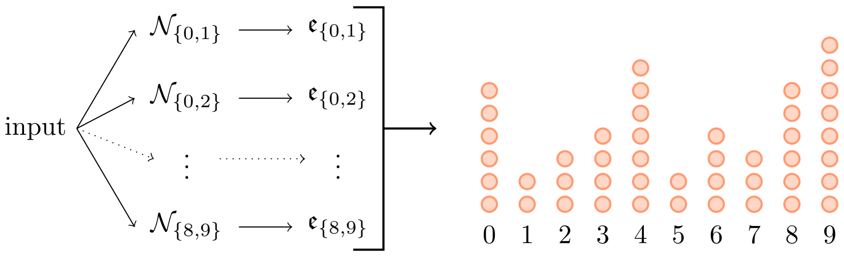


Figure 1: An example of the voting scheme for an input of the MNIST dataset. $\mathcal{N}_{i,j}$ is the network trained to distinguish between i-th and j-th digits, $\mathfrak{c}_{i,j}$ is the output of the network $\mathcal{N}_{i,j}$. On the right, each dot represents the vote of a single network.

networks correctly classify the input $\boldsymbol{x}^k$, then at most $n-2$ other networks can classify the input with a different label. With this approach, if we plan to use $r \in \mathbb{N}$ inputs for each label, we are feeding our BNNs a total of $2 \times r$ inputs instead of feeding $n \times r$ inputs to a single large BNN. It is much easier to train our structured ensemble of BNNs rather than training one large BNN. The downside of this approach is the large number of networks that have to be trained, even if the training can run in parallel.

After the training, we feed one input $\boldsymbol{x}^k$ to our list of BNNs, and we need to elaborate on the set of outputs.

If there is a clear winner label, for example as in the case depicted in Figure 1, we set it as our output. If two labels tie with the highest number of votes, our output is the same as the output of the network that was trained to distinguish between those two labels. In the other cases, we choose not to classify, for example labelling the input as -1 (if $-1 \notin \mathcal{I}$).

Note that the proposed structured ensemble alongside its voting scheme can also be exploited for regular neural networks.

2 Integer Neural Networks

Our goal is to train neural networks using discrete optimization solvers. The advantages and limitations of this idea have been studied in the literature in only the last few years, including in our own works.

In this section, we aim to investigate the potential of the lexicographic MILP approach introduced in [1], focusing on three distinct directions. Firstly, with the valuable expertise of Dr Neil Yorke-Smith, we study and compare Binarized Neural Networks (BNNs) with more general Integer Neural Networks (INNs). Special attention is paid to the analysis of their performances as the number of bits that each network parameter can utilize varies. Specifically, experiments are conducted on classification problems using the MNIST [4], heart disease [3], and breast cancer[9] datasets.

Secondly, in our preceding research [1], we emphasized the significance of simplicity and robustness in training and concurrently optimizing neural networks. In this report, we reformulate these fundamental principles by introducing novel objectives and activation functions.

The computational experiments and the resolution of MIP problems described in this report have been performed by Gurobi [2] version 10.0.0. Gurobi parameters have been left to default values unless otherwise specified. When the time limit parameter is specified for different models, note that whenever the optimum is reached within the time limit, the remaining time is added to the time limit of the subsequent model. Experiments are run on a computer with an 11th Gen Intel Core i7-1185G7 processor running at 3.00 GHz using 16,0 GB of RAM.

P	SM time	MM Gap	links	accuracy
1	5.61	11.53	**29.19**	75.60
3	3.59	11.48	29.69	75.68
7	2.59	**11.46**	29.61	76.16
15	3.18	**11.46**	29.64	**76.36**
31	**2.44**	**11.46**	29.64	75.88

Table 1: Comparison between the solutions found with different values of P. The 2-nd column reports the runtime to solve the first model SM; the 3-rd one refers to the percentage of the MIPGap at the second MIP model MM; the 4-th column is the percentage of non-zero weights after the solution of models MM and MW; the last column reports the percentage of accuracy obtained on the test images.

2.1 Comparison with BNNs and insights on the solutions

In this subsection, we compare MIP-training of BNNs with non-binarized, integer-valued NNs. We also give some insights into the solutions, in particular, regarding the distribution of the weights of the optimal networks.

2.1.1 A first comparison

The first experiment we perform aims at comparing the accuracy of BNNs with the one of INNs, where each INN have weights that vary in the set $\{-P, \dots, P\}$, for a certain integer P. We compare training BNNs, i.e. $P = 1$, to training non-trivial INNs, i.e. $P = 3, 7, 15, 31$. Each increase in range represents one extra bit needed to store a network's parameter in memory. In this experiment, we use the even digits of the MNIST dataset [4], so that only 10 INNs have to be trained instead of 45. Only 10 images per digit are used as training, with each network consisting of the architecture $[784, 4, 4, 1]$. We set a time limit of $75s$ for the Sat-Margin (SM) model, $75s$ for the Max-Margin (MM) model, $10s$ for the Min-Weight (MW) model. The test are performed over 500 test images per class. In Table 1, the column *SM time* shows the average optimization time for the first model SM, the column *MM Gap* shows the percentage of the average Gurobi MIPGap value after the time limit of the second model has been reached, the column *links* shows the percentage of non-zero weights after the MW model. Note that, although the increase in memory usage results in higher accuracy in some cases, this improvement is only a few tenths of a percent, so we do not consider it significant.

In summary, the comparison between BNNs and INNs limited to even-numbered instances of MNIST suggests that increasing the number of bits per parameter does not necessarily lead to an increase in accuracy. Importantly, this fact may be related to the quasi-binary nature of the input: the digits represent the 256 shades between black and white, but they tend to cluster around extreme values.

2.1.2 Weights distribution – A first look

Similar accuracy values can be obtained with different values of P. The goal of this subsection is to study the distribution of the weights over the values $\{-P, \dots, P\}$ to determine whether different distributions lead to comparable accuracy or if the weights are similar even for different values of P. We also aim to study the impact of INNs on the lexicographic multi-objective function used in our previous work [1]. For this purpose, we select two class of the MNIST dataset [4], the digit 4 and the digit 9, that are commonly misclassified. We use the network architecture of $[784, 4, 4, 1]$, trained with 10 images per class, with a time limit of $110s + 110s + 20s$ for SM+MM+MW. Notice that in the binarized case, namely $P = 1$, the MW model minimizes the sum of the absolute values of the weights, maximizing the number of zero weights. When $P \neq 1$, we have two possible generalizations: minimizing the sum of the absolute values of the weights (possibly reducing the number of used bits) or minimizing the number of non-zero weights. We choose the second option, thinking it better follows the idea of a

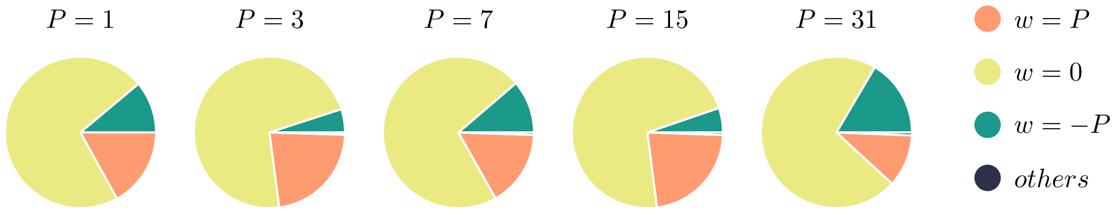


Figure 2: Weights distribution for the model `SM` + `MM` + `MW`. $w = P$ indicates the percentage of weights of the network equal to P, and so on. “Others” indicates the percentage of weights that are neither equal to P, 0, nor $-P$.

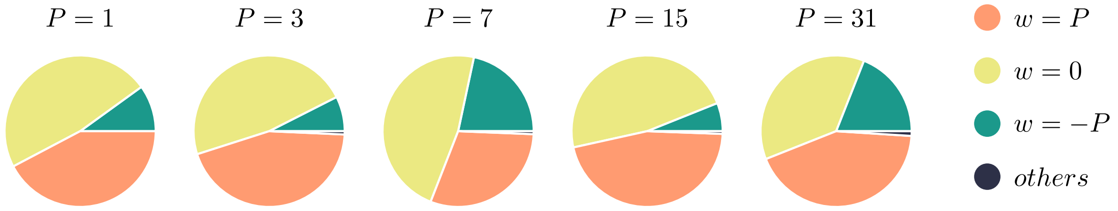


Figure 3: Weights distribution for the model `SM` + `MM`. Comparing these results with Figure 4, one can see how the `MW` model affect the percentage of weights set to zero.

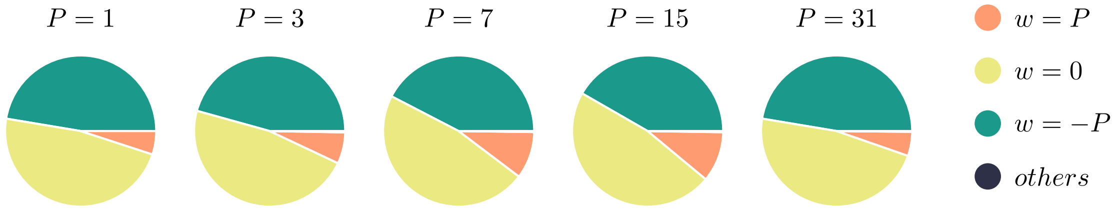


Figure 4: Weights distribution for the model `SM`.

lightweight NN. When using the three models subsequently, i.e. `SM`, `MM`, and `MW`, we find the weights distributions depicted in Figure 2.

Note that $w = P$ indicates the percentage of weights of the network equal to P, and so on. “Others” indicates the percentage of weights that are neither equal to P, 0, nor $-P$. The majority of the weights are set to zero, possibly thanks to the `MW` and to the following preprocessing technique: if, for all the 2×10 images given as input, the $i-$th pixel is zero, then all the weights connecting that pixel to all of the neurons of the first hidden layer are removed from the decision variables of the model. The most interesting fact is that the percentage of weights that are neither $-P$, P, nor 0 is extremely low, less than 0.70%. So the networks reach comparable accuracy having very different weights but with similar weights distribution, i.e. condensed in 0, P and $-P$. To investigate the impact of the models over the weights distribution, we run different experiments. When running the same experiment using only the first two models subsequently, i.e. `SM` and `MM`, the distributions in Figure 3 are obtained. Comparing these results with the previous ones, we can see the impact of the `MW` model. The weights distribution of the `SM` model alone are presented in Figure 4. Note that in this case the model is solved to optimality. The detailed percentages of these results are summarized in Table 2.

In summary, different models find extremal weight distribution while optimizing a network. In addition, results show that a higher value of P does not correspond to higher accuracy. BNNs generalize better since they are less prone to overfitting. Note that the use of the lexicographical model improves the accuracy.

Models	P	$w = P$	$w = 0$	$w = -P$	*others*	SM time	MM Gap	accuracy
SM	1	5.01	47.62	47.37	-	2.35	-	**73.70**
	3	6.72	47.31	45.69	0.28	2.37	-	68.40
	7	10.08	47.31	42.40	0.21	1.93	-	55.40
	15	10.77	47.28	41.70	0.25	2.29	-	59.00
	31	5.16	47.31	47.37	0.16	2.56	-	70.90
SM + MM	1	42.30	47.75	9.95	-	2.40	17.03	**75.20**
	3	44.39	47.47	7.45	0.69	2.09	16.95	74.50
	7	30.30	47.43	21.64	0.63	1.96	16.94	74.90
	15	45.98	47.40	6.05	0.57	2.05	16.93	74.20
	31	33.43	47.43	18.54	0.60	2.35	16.93	74.20
SM + MM + MW	1	16.95	71.96	11.09	-	2.48	17.03	***75.80***
	3	22.43	72.12	4.97	0.48	6.91	16.96	*75.60*
	7	16.29	71.93	11.25	0.53	7.17	16.94	*75.50*
	15	22.43	71.80	5.23	0.54	2.54	16.94	*75.10*
	31	11.19	71.67	16.51	0.63	6.50	16.93	*74.70*

Table 2: Comparison of different values of P together with an ablation study. Even with different models, the weights distribution looks condensed in extremal points. The *accuracy* column reports the percentage of correctly classified images and it is calculated on 500 test images per digit. The highest value of accuracy for a fixed P is written in italic, while the highest value for a fixed model is written in bold. Notice that for a fixed model the highest accuracy is obtained with $P = 1$ and that for a fixed P the highest accuracy is obtained with the SM + MM + MW model.

P	$w = P$	$w = 0$	$w = -P$	*others*	SM time	MM time	MW time	accuracy
3	12.83	82.92	4.06	0.19	0.70	31.18	11.36	99.90
7	12.83	82.73	4.31	0.13	0.99	23.10	29.16	99.90
15	13.91	82.67	3.20	0.22	1.04	114.73	6.63	99.90

Table 3: Weights distributions and technical insights for an almost-optimal experiment: distinguishing between 0 and 1 digits, with 4 input images per digit for the training.

2.1.3 Weights distribution – Optimality

Since the MIPGap of the MM model is significantly big, one can investigate if the weights distribution at optimality is the same as the one found within the time limit. We choose a simpler problem to solve: we select two classes of the MNIST dataset [4] that are easier to classify, the digit 0 and the digit 1, and we train our model with only 4 input images per digit. We use the network architecture of $[784, 4, 4, 1]$. P varies within the set $\{3, 7, 15\}$ and no time limit parameter was given to Gurobi. We observe that, in the case $P = 3$, after 14 seconds the model finds a solution that does not improve after 1800 seconds, lowering the MIPGap to 0.04%. We then choose to set the Gurobi parameter MIPGap to the value 0.001, making the hypothesis that the solver actually finds an optimal solution but cannot prove its optimality. Table 3 summarise the results obtained in our experiments.

Even at almost-optimality, the weights distribution does not change. Thus, we want to test a different model, to be optimized after SM, that looks for a as-uniform-as-possible distribution of the weights and see its accuracy on the MNIST dataset, to compare it with previous models, that find extremal distributions.

In summary, the observed distribution of extremal types does not appear to be influenced by a time constraint issue, as the solution remains consistent even at the optimal level. The forthcoming investigation will focus on assessing the performance when the parameter distribution is constrained to be uniform.

digits	P	`SM` time	`FD` time	`FD` obj	accuracy
4 VS 9	3	3.02	3.40	0.00	60.40
	7	1.98	56.31	0.00	70.30
	15	1.67	166.71	0.00	55.90
0 VS 1	3	1.59	3.73	0.00	88.30
	7	1.73	58.31	0.00	92.40
	15	2.43	197.64	0.00	94.00

Table 4: Results of the Fair-Distribution model optimized after the Sat-Margin model. Having a uniform distribution of the absolute values of the weights leads to a far worse accuracy with respect to a network with an extremal distribution of the weights.

2.1.4 Optimizing weight homogeneity

We describe the fair distribution (`FD`) model. We use the same structure constraints but optimize a new objective function. New constraints are added just to evaluate the terms in the objective function. Given the network weights w_{ilj}, we introduce new variables v_{ilj} and set them as follows:

$$v_{ilj} = |w_{ilj}|. \tag{3}$$

Note that is done in the `MW` model too. Then, for every $h = 1, \dots, P$, we introduce new binary variables $b^h_{ilj} \in \{0, 1\}$ so that the constraint

$$b^h_{ilj} = 1 \iff v_{ilj} = h, \tag{4}$$

is satisfied. This can be linearized with the following constraints

$$v_{ilj} = \sum_h h b^h_{ilj} \tag{5a}$$

$$\sum_h b^h_{ilj} \leq 1 \tag{5b}$$

and notice that this also implies that a specific weight is set to zero if and only if all of the b variables are set to zero, i.e.

$$\sum_h b^h_{ilj} = 0 \iff v_{ilj} = 0. \tag{6}$$

What we are doing is counting the number of weights whose absolute value equals h. We define A as set of the indexes of the weights that are not automatically set to zero with the preprocessing technique described above, that can be applied to any dataset. A represents then the set of weights that are actually decision variables inside the model. Our new objective function can be written in the following way

$$\min \sum_h |(\frac{1}{\sum_{(i,l,j) \in A, h} b^h_{ilj}} \sum_{(i,l,j) \in A} b^h_{ilj}) - \frac{1}{P}| \tag{7}$$

and, to avoid numerical errors, we can work with integer values by rewriting the objective function in the following way

$$\min \sum_h |(P \sum_{(i,l,j) \in A} b^h_{ilj}) - \sum_{(i,l,j) \in A, h} b^h_{ilj}|. \tag{8}$$

Note that we only optimize over the absolute values of the non-zero weights, to have a bigger research space, making the problem easier. The solution will still consist of a network with a non-extremal weights distribution. We test the model with the digits 4 and 9, given 10 images as input per digit, and testing the network on 500 images per digit. We use the network architecture of $[784, 4, 4, 1]$. The results of this experiment are summarized in Table 4. We are actually able to find a network whose absolute values of the weights have a uniform distribution over the set $\{1, \dots, P\}$. Note that

Models	P	$w = P$	$w = 0$	$w = -P$	*others*	SM time	MM Gap	accuracy
SM	1	43.33	20.00	36.67	-	0.06	-	65.50
	3	33.33	6.67	30.00	30.00	0.03	-	**74.25**
	7	6.67	6.67	20.00	66.66	0.06	-	71.75
	15	16.67	0.00	10.00	73.33	0.03	-	54.25
	31	10.00	0.00	20.00	70.00	0.03	-	67.25
SM + MM	1	56.67	20.00	23.33	-	0.07	0.00	*72.50*
	3	53.33	0.00	30.00	16.67	0.09	4.70	**79.25**
	7	50.00	0.00	26.67	23.33	0.08	9.07	*75.75*
	15	56.67	0.00	20.00	23.33	0.04	10.24	*74.50*
	31	56.67	0.00	23.33	20.00	0.03	9.52	*74.25*
SM + MM + MW	1	56.67	20.00	23.33	-	0.07	0.00	*72.50*
	3	53.33	6.67	26.67	13.33	0.13	4.70	***83.00***
	7	50.00	0.00	26.67	23.33	0.08	9.07	*75.75*
	15	56.67	0.00	20.00	23.33	0.04	10.24	*74.50*
	31	56.67	0.00	23.33	20.00	0.05	9.52	*74.25*

Table 5: Comparison of different values of P together with an ablation study for the Breast Cancer dataset. The highest value of accuracy for a fixed P is written in italic, while the highest value for a fixed model is written in bold. Notice that the model SM is outperformed for every P, and that $P = 3$ reaches the highest accuracy for a fixed model.

the accuracy decreases drastically, hinting that a network with an extremal weights distribution is able to better generalize on unseen data.

In summary, in order to investigate whether a more homogeneous distribution had a similar performance to the extreme distribution observed at the optimum, we set up a new MILP that favored the uniform distribution of parameters. As can be seen from the values shown in the tables above, performance worsens.

2.1.5 Different datasets

Regarding the weights distribution (Section 2.1.2), we perform other experiments on different datasets for two main reasons:

- the MNIST dataset is made of images that are in greyscale but the majority of the pixels are either black or high-valued greys (almost white), so data could be the cause of the extremal weights distribution;
- with a smaller dataset we believe we can actually achieve optimality for all three models within a reasonable time.

Breast cancer. This medical dataset [9] is obtained from a digitized image of a fine needle aspirate (FNA) of a breast mass. The data describe characteristics of the cell nuclei present in the image. The dataset is composed by 699 vectors of $\mathbb{R}^9$, subdivided in 2 classes: benign (458 data), and malignant (241 data), and after removing the incomplete data, we are left with 444 data and 239 data, respectively. Results with $110s + 110s + 20s$ of time limit, 10 data per class and a network architecture of $[9, 3, 1]$ are shown in Table 5.

With a low-dimension dataset, without too many extremal values in input data, the percentage of non-extremal weights increases. Altough, the solution found does not have a uniform weights distribution. Also, note that MM+MW considerably reduce the number of non-extremal weights. We performed some tests over 200 data per class. Note that SM alone has a lower accuracy then the other models.

	0	**1**	**2**	**3**	**4**	**Total**
Cleveland	164	55	36	35	13	303
Hungarian	188	37	26	28	15	294
Switzerland	8	48	32	30	5	123
Long Beach VA	51	56	41	42	10	200

Table 6: The Heart Disease dataset [3] is from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach VA. The first column represents values of healthy patients, while the remaining columns represent values of patients with cardiac abnormalities.

Models	P	$w = P$	$w = 0$	$w = -P$	*others*	SM time	MM time	accuracy
SM	1	33.33	38.10	28.57	-	0.39	-	***68.75***
	3	33.33	16.67	35.71	14.29	0.33	-	56.66
	7	28.57	7.14	33.33	11.90	0.24	-	68.75
	15	38.10	7.14	21.43	33.33	0.09	-	65.42
	31	21.43	0.00	54.76	23.81	0.12	-	63.33
SM + MM	1	54.76	28.57	16.67	-	0.16	3.92	60.42
	3	50.00	7.14	26.19	16.67	0.35	17.98	*62.08*
	7	66.67	4.76	16.67	11.90	0.24	3.06	**72.50**
	15	64.29	4.76	19.05	11.90	0.27	7.38	*70.00*
	31	66.67	4.76	11.90	16.67	0.33	21.51	*70.42*
SM + MM + MW	1	40.48	47.62	11.90	-	0.17	4.00	62.92
	3	57.14	16.67	9.52	16.67	0.45	32.98	60.83
	7	66.67	16.67	4.76	11.90	0.70	7.01	***72.92***
	15	61.90	14.29	11.90	11.91	0.36	7.68	*70.00*
	31	61.90	14.29	7.14	16.67	0.40	18.73	*70.42*

Table 7: Comparison of different values of P together with an ablation study for the Breast Cancer dataset. The highest value of accuracy for a fixed P is written in italic, while the highest value for a fixed model is written in bold.

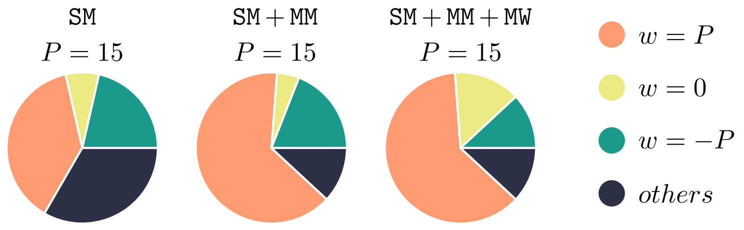


Figure 5: Different weights distributions for the Heart Disease dataset. The MM model reduces the percentage of non extremal weights. The MW model is still able to increase the percentage of zero-valued weights.

Heart disease. The dataset [3] consists of 920 vectors of $\mathbb{R}^{13}$, collected by four medical institutions and subdivided in 5 classes as shown below.

In our experiments, we use the Cleveland dataset. The data labelled with $1, 2, 3$, and 4 are collapsed in a unique class 5, such that we have a binary classification problem "with heart disease" (164 instances), "without heart disease" (139 instances), and after removing the incomplete data, we are left with 160 data and 137 data, respectively. We use 10 images for the training and 120 for the tests, with a network architecture of $[13, 3, 1]$. Detailed results are shown in Table 7.

Visually, some results are presented in Figure 5

Note that, as before, even if the percentage of non-extremal is significant, the solution found does not have a uniform weights distribution. MM+MW reduce the number of non-extremal weights only for certain values of P, while MW is able to set to zero a high percentage of weights. The following table show some technical insights and the accuracy over the tests. Note that every model is solved to optimality at each step. Except for the case $P = 1$, SM is outperformed by the other models.

2.1.6 Symmetry breaking constraints

To cut off a certain region of the research space and get rid of a subset of solutions that are a permutation of other solutions, consequently speeding up the code, one could try to add symmetry-breaking constraints. We try various experiments adding the following ones:

$$w_{1,1,j} \geq w_{1,1,j+1} \quad \forall j = 1, \ldots, N[1] - 1 \tag{9}$$

as done in [6] but the code does not seem to get any faster at reaching optimality nor to obtain a smaller MIPGap in the same time limit.

2.2 A different paradigm for simplicity: pruning neurons

This subsection looks at how a given (trained) NN can be optimized – specifically, simplified – again by using a MIP solver.

2.2.1 The idea

The MW model[8] optimizes the network simplicity by minimizing the number of connections, i.e., the number of non-zero weights. A second approach consists in the direct pruning of the neurons while training. In particular, one wants to remove those neurons whose out-going weights are zero. For every jth neuron of level l, we introduce a binary variable a_{lj} such that the following constraint holds:

$$a_{l-1,i} = 0 \Longrightarrow \sum_{j_1} v_{ilj_1} = 0 \tag{10}$$

and then it suffices to solve

$$\min \sum_{l,i} a_{l-1,i} \tag{11}$$

We notice that it is not restrictive to require the following constraint:

$$a_{l-1,i} = 0 \Longrightarrow \sum_{j_2} v_{j_2,l-1,i} = 0 \tag{12}$$

meaning that if a neuron is not active, then the in-going weights of that neuron can be set to zero. The two constraints can be then written as

$$a_{l-1,i} = 0 \Longrightarrow \sum_{j_1} v_{ilj_1} + \sum_{j_2} v_{j_2,l-1,i} = 0 \tag{13}$$

that can be linearized as

$$\sum_{j_1 \in N[l]} v_{ilj_1} + \sum_{j_2 \in N[l-2]} v_{j_2,l-1,i} \leq M a_{l-1,i} \tag{14}$$

with $M = P \times (n_l + n_{l-1})$. This gives us the Pruning Neurons (PN) model. We study it empirically in the sequel.

2.2.2 Computational results

This approach lead to a preliminary experiment with the following features: we solve a classification problem between 0 and 1 digits of the MNIST dataset [4]. We used 20 images per digit and trained a $[784, 10, 10, 1]$ architecture. The time limits have been set to $110s(\texttt{SM}) + 130s(\texttt{PN})$. The results are summarized in Table 8. In summary, an alternative formulation for simplicity is proposed. Instead

P	SM time	PN time	PN obj	PN accuracy	SM + MM accuracy
1	32.45	204.36	2.00	76.10	96.40
3	19.56	163.59	2.00	90.80	99.50
7	21.13	128.32	2.00	68.00	87.20

Table 8: Results for the Pruning-Neurons model. We notice a worsening in the accuracy with respect to the MW model, that is the previous model taking into account the simplicity principle.

of aiming to simplify the network by minimizing the number of non-zero weights, we pursued an optimization strategy that focuses on maximizing the presence of inactive neurons, which are neurons that solely output null weights. While the former approach promotes sparse linear operations, the latter enables us to reduce the dimensions of both the input and output of the linear applications. Limited to the conducted experiments, it is evident that the performance in the second case is inferior compared to the first case.

2.2.3 A theoretical insight

Although the results are not improving the previous accuracy, it is useful to notice a hidden advantage of this approach. The algorithm favours a low number of neurons per layer. In the case of binary classification, and when a layer has only one active neuron, the network can be "truncated" without loss of any information. Indeed, the only information that passes through the rest of the network is the one which is encapsulated within that single neuron. The following clarifies and formalizes in a mathematical way this intuition.

Lemma 1 (Network Reduction). *Suppose we have a Neural Network written in the form*

$$f := \rho_L \circ T_L \circ \rho_{L-1} \circ \cdots \circ T_2 \circ \rho_1 \circ T_1 \tag{15}$$

with $T_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}$, $N_0, \dots, N_L \in \mathbb{N}$, affine functions and $\rho_l(x) = \rho(x) = 2 \times \mathbb{1}(x \geq 0) - 1$ applied component-wise. Suppose the network correctly classifies a set of labelled data $\mathcal{T} = (x_k, y_k)_{1,\dots,t}$, $t \geq 2$, meaning

$$f(x_k) = y_k \in \{-1, 1\}^{N_L} \quad k = 1, \dots, t \tag{16}$$

and suppose there exist k_1, k_2 such that $y_{k_1} \neq y_{k_2}$, i.e. the classification task is non-trivial. If there exists l such that $N_l = 1$, then the network f can be truncated at layer l, meaning that if two data have different labels, the output of level l for those two data would be different, and if two data have the same label, the output of level l for those two data would be the same.

Proof. We can write the network in the form

$$f = g \circ \rho \circ T_l \circ h \tag{17}$$

then the function g is of the form

$$g : \{-1, +1\} \to \{-1, +1\}^{N_L} \tag{18}$$

Let us define $z_{-1} = g(-1)$ and $z_1 = g(1)$ and, since $y_k = g \circ \rho \circ T_l \circ h(x_k)$, we have that, for every k, either $y_k = z_{-1}$ or $y_k = z_1$. Since there exist k_1, k_2 such that $y_{k_1} \neq y_{k_2}$, $z_{-1} \neq z_1$ so g must take at least two different values and so it is a bijective function from its domain to its image. Since the outputs of the network are in one-to-one correspondence with the output of the only neuron in level l, by changing the encoding of the labels, the network can be truncated. □

Definition 1. Given a Neural Network

$$f := \rho_L \circ T_L \circ \rho_{L-1} \circ \cdots \circ T_2 \circ \rho_1 \circ T_1$$

where $T_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}$, $N_0, \dots, N_L \in \mathbb{N}$, affine functions, in particular $T_l(x) = W_l x + b_l \; \forall l \in \{1, ..., L\}$, and a Neural Network

$$\hat{f} := \rho_R \circ \hat{T}_R \circ \rho_{R-1} \circ \cdots \circ \hat{T}_2 \circ \rho_1 \circ \hat{T}_1$$

where $\hat{T}_r(x) = \hat{W}_r x + \hat{b}_r \; \forall r \in \{1, ..., R\}$, the network $\hat{f}$ is said to be a subnetwork of f if all of the following are satisfied:

- $R \leq L$;
- there exists a subset $\mathcal{I}_1 \subsetneq \{1, \dots, N_1\}$ of indexes such that $\hat{W}_1$ is obtained from W_1 removing the i-th row and $\hat{b}_1$ is obtained from b_1 removing the i-th entry, for every $i \in \mathcal{I}_1$;
- for every $k = 2, \dots, R$, there exists a subset $\mathcal{I}_k \subsetneq \{1, \dots, N_k\}$ such that $\hat{W}_k$ is obtained from W_k removing the i-th row and the j-th column and $\hat{b}_k$ is obtained from b_k removing the i-th entry, for every $i \in \mathcal{I}_k$, $j \in \mathcal{I}_{k-1}$.

The subnetwork is said to be proper if $R < L$ or $\bigcup_{k=1,\dots,R} \mathcal{I}_k \neq \emptyset$.

Remark. *Note that, intuitively, when removing a neuron of a layer l, one must remove all input links, removing the correspondent row of the affine function T_l, and all the output links, removing the correspondent column of the affine function T_{l+1}.*

Definition 2. Given a Neural Network f and a set of labelled data $\mathcal{T} = (x_k, y_k)_{1,\dots,t}$, f is said to be reducible for $\mathcal{T}$ if there exists a proper subnetwork $\hat{f}$ and a injective function g such that $g \circ \hat{f}(x_k) = f(x_k) = y_k$. The network f is said to be irreducible for $\mathcal{T}$ if it is not reducible.

Remark. *Lemma 1 can be generalized at the case where there exists l such that $N_l = p$ and there exists 2^p different labels, for example if $N_L = p$ and every vector encode a different class. Notice that in this case the network is irreducible only if $N_l > N_L$ for all $l < L$.*

Proof. Now g is a bijective function between the set $\{-1, +1\}^p$ and the set of labels. □

Theorem 1. *With the notation of Lemma 1, but allowing each level l to have its own activation function $(\rho_l)_{l=1,\dots,L}$, if there exists $\hat{l}$ such that*

$$|\rho_{\hat{l}}(\{\hat{x}_1, \dots, \hat{x}_t\})|^{N_{\hat{l}}} = |f(\{x_1, \dots, x_t\})| \tag{19}$$

where $\hat{x}_k = T_{\hat{l}} \circ h(x_k)$, then f is reducible.

Proof. Same as the Remark above. □

Remark. *This simply says that if the number of classes is equal to all the possible output values of a certain layer, the network can be truncated at that layer.*

In summary, reducing the number of neurons can also lead a decreasing in the number of layers, simplifying the network structure.

2.3 A different paradigm for robustness: a killer activation function

In this final subsection, we discuss the third topic of our visit, namely the reformulation of the robustness paradigm. The `MM` model[8] aims to make neurons less sensitive to small perturbations in the input by imposing that neuron inputs are as positive or as negative as possible. We introduce a new formulation of robustness: by choosing a different activation function, we can require that the neuron "kills" inputs that are not positive or negative enough. This second approach can be combined with the `MM` model or used independently.

2.3.1 The idea

Three-valued sign function. We consider a different activation function, i.e., $\rho : \mathbb{R} \to \mathbb{R}$ such that

$$\rho(x) = -\mathbb{1}(x < -\alpha) + \mathbb{1}(x \geq \alpha) =: -u + r$$

This activation function shares the same principle of the sign function, but with a threshold: if the input is *enough* positive, the output is 1, if the input is *enough* negative, the output is -1, otherwise it is 0. The concept of *enough* is explicited by the constant α. Note that, when $\alpha = 0$, this new activation function is the same as the one described before. The logical conditions that define the new activation function and that we want to model are

$$\begin{aligned} (u, r) = (0, 0) &\iff x \in [-\alpha, \alpha) \\ (u, r) = (1, 0) &\iff x < -\alpha \\ (u, r) = (0, 1) &\iff x \geq \alpha \\ (u, r) = (1, 1) &\iff \text{not feasible} \end{aligned}$$

This can be made by the linear constraints and decision variables reported here:

$$x \geq -Mu - \alpha + 2\alpha r \tag{20a}$$
$$x \leq Mr - 2\alpha u + \alpha - \epsilon \tag{20b}$$
$$u + r \leq 1 \tag{20c}$$
$$u, r \in \{0, 1\} \tag{20d}$$

Product Linearization. Defining the bilinear product ρw in terms of linear constraints is a critical aspect of neural network modeling. This can be achieved precisely by exploting McCormick cuts.

$$c = w(-u + r) \tag{21}$$

The strategy is to add two binary variables γ, δ writing the following constraints

$$c = \gamma + \delta, \tag{22a}$$
$$\gamma = -wu \tag{22b}$$
$$\delta = wr \tag{22c}$$

and then writing McCormick cuts for γ and δ, i.e.,

$$\gamma \leq Pu, \tag{23a}$$
$$\gamma \geq Pu - P - w, \tag{23b}$$
$$\gamma \geq -Pu, \tag{23c}$$
$$\gamma \leq -Pu + P - w, \tag{23d}$$
$$\delta \geq -Pr, \tag{23e}$$
$$\delta \leq -Pr + P + w, \tag{23f}$$
$$\delta \leq Pr, \tag{23g}$$
$$\delta \geq Pr - P + w. \tag{23h}$$

Combining, respectively, (23c) and (23e), (23a) and (23g), (23a) and (23f), (23d) and (23g), (23b) and (23e), (23c) and (23h), we obtain the following constraints:

$$c \geq -P(u + r), \tag{24a}$$
$$c \leq P(u + r), \tag{24b}$$

β	SM-α time	MM-α time	accuracy
0	8.55	1.72	33.16
0.1	6.68	1.45	32.72
0.3	6.14	1.32	34.80

Table 9: Results on the even digits MNIST with different activation functions. We notice a worsening in the accuracy with respect to the MM model, that is the previous model taking into account the robustness principle.

$$c \leq P(u-r) + P + w, \tag{24c}$$
$$c \leq P(r-u) + P - w, \tag{24d}$$
$$c \geq P(u-r) - P - w, \tag{24e}$$
$$c \geq P(r-u) - P + w. \tag{24f}$$

and we can get rid of the binary variables γ and δ. Notice that combining (23b) and (23h), (23d) and (23f) would give us redundant constraints.

2.3.2 Computational results

We perform an experiment on the even digits MNIST, using 20 images per digit in training, the architecture $[784, 4, 4, 1]$, $P = 1$, and different value for the constant $\alpha = \beta \times P \times \min([784, 4, 4]) = 4\beta$. Note that, for simplicity, we choose the same activation function for every layer and we set the parameter α proportional to the minimum number of neurons of all the layers but the last one. However, one could choose to use different activation functions setting different parameters α_l proportional to n_l. The time limit parameter is set to $110s + 130s$. The tests are performed over 500 images per digit. Results are shown in Table 9. Note that, in this case, the margin is optimized only on the last level of the network. We notice a significant decreasing in the accuracy, compared to the one obtained with the previous objective function.

In summary, we made modifications to the activation function, opting for a more versatile approach that takes into consideration the possibility to kill an output if it is not positive or negative enough. Although the concept was promising, experimental results indicate that a broader range of support for the activation function yields improved accuracy. It is important to note that tuning the hyperparameter for this modification is a delicate process, necessitating further considerations. Notably, customization of the hyperparameter becomes crucial for each specific level, and additional analysis is required to determine the optimal configuration. It is worth exploring various scenarios and conducting in-depth evaluations before drawing definitive conclusions.

3 Graph Neural Networks: preliminary theoretical analysis

In this second section, we present a preliminary approach to training Graph Neural Networks (GNNs) using discrete optimization solvers. This is based on a Master's thesis [5] co-supervised by Dr Neil Yorke-Smith. While this section does not introduce any novel ideas, it provides a foundational review of the GNN structure. Our intention is to expand on these ideas in future work, but for now, we highlight some of the main concepts concerning the GNN structure.

3.1 Structure of a Graph Neural Network (GNN)

Let us suppose we have a dataset composed of graphs $\mathcal{D} = \{\mathcal{G}_k\}_{k=1,...,|Tr|}$. Each graph is undirected and endowed with features. The graph $\mathcal{G}_k$ has M_k nodes and each node has F features. The set of the edges of $\mathcal{G}_k$ is denoted with $\mathcal{E}_k$. The dependency of k for M_k is usually mandatory, whereas we can suppose without loss of generality that all the nodes from all the graphs have the same number of features F. To index such quantities we use $m_k \in \{1, ..., M_k\}$ for the nodes of the graph $\mathcal{G}_k$, and

$f \in \{1, ..., F\}$ for the features.

We associate the adjacency matrix A_k to the graph $\mathcal{G}_k \forall k \in \{1, ..., |Tr|\}$. This matrix is $M_k \times M_k$ and each entry is either 0 or 1. In particular, $A_k(i,j) = 1 \iff (i,j) \in \mathcal{E}_k$ or $(j,i) \in \mathcal{E}_k \iff (i,j) \in \mathcal{E}_k$ and $(j,i) \in \mathcal{E}_k$. For example, if the nodes are $\{1, 2, 3\}$ and the edges are $\{(1, 2), (2, 3)\}$, the corresponding adjacency matrix is

$$A_k = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{25}$$

The nodes features of $\mathcal{G}_k$ are collected in a real matrix X_k, whose dimension is $F \times M_k$.

The main idea about a GNN is the following: each node owns a family of A neural networks, with A defined as the number of the epochs of the algorithm. The structure of the α-th network $R^{\alpha}_{m_k}$ is

$$[N_0^{m_k,\alpha}, N_1^{\alpha}, N_2^{\alpha}, ..., N_{L(\alpha)}^{\alpha}]$$

$\forall \alpha = 1, \ldots, A$. We notice that the only layer depending on m_k is the first one.

3.2 Modelling the nodes networks

The structural property of a GNN is that each network must take into account not only the features of the corresponding node, but also those of adjacent nodes. We describe two ways to accomplish this task.

(i) The most trivial way to do this is to concatenate the features. With this requirement, the only way to define $N_0^{m_k,\alpha}$ is

$$N_0^{m_k,\alpha} := (1 + \sum_{j=1}^{M_k} A_k(m_k, j))F^{\alpha-1}$$

where $F^0 := F$ and $F^{\alpha} := N^{\alpha}_{L(\alpha)}$. The main con of this approach is that there is a structural problem when considering graph with a different number of nodes or different edges, both during training and during test. For example, if we consider the graph (25), the second network takes as input vectors of dimension $(1 + 2) \times F$. If a second graph has the same nodes but edges equal to $\{(1, 2), (1, 3)\}$, the input of the second network should be $(1 + 1) \times F$, that is different from $3F$. Also,it would be more suitable that all the network of a fixed epoch α were equal. This allows to have test data with a different number of nodes with respect to the ones of training data. This approach leads to the next point.

(ii) If we exploit an 'aggregation' operator that allows us to have rigid control on N_0, the previous problems are solved. For example, one could define

$$N_0^{m_k,\alpha} := F^{\alpha-1}$$

and use a mean operator on adjacent features before passing them to the network. In this setting, the structure of the network $R^{\alpha}_{m_k}$ is

$$[F^{\alpha-1}, N_1^{\alpha}, N_2^{\alpha}, ..., N_{L(\alpha)}^{\alpha}]$$

and we lost the dependency of m_k.

This means that, when we fix an epoch, we can have the same network for each neuron. Also, different training (and test) data with different structural properties are now compatible with the fixed network structure.

After the A-th epoch, a 'classification' network, is applied. The name refers to the fact the input is the result of an aggregation function applied to the totality of the last features of the nodes and the output is mono-dimensional. Different aggregation function could be exploited, e.g., a sum, a max or a mean.

The linear formulation. In order to avoid the problems discussed, a first formulation[5] concerns the case where the training and test graphs all have the same number of nodes M. The exploited aggregation function is the sum.

$$x_i^\alpha = \rho(x_i^{\alpha-1} W_1^\alpha + \sum_{l \in \mathfrak{N}(i)} x_l^{\alpha-1} W_2^\alpha) \tag{26}$$

where x_i^α is the feature vector of the node i at the epoch α, $\mathfrak{N}(i)$ is the set of nodes that share an edge with i, and W_1, W_2 are the weights matrices that perform the linear operation. The network (endowed with a final aggregation layer) can be represented with a MILP formulation, as in the following:

$$x_i^{\alpha-1} W_{1f}^\alpha + \sum_{l=1}^{M} b_{il}^{\alpha-1} W_{2f}^\alpha = x_{if}^\alpha - S_{if}^\alpha \tag{27a}$$

$$x_{if}^\alpha \leq U_{if}^\alpha Z_{if}^\alpha \tag{27b}$$

$$S_{if}^\alpha \leq -L_{if}^\alpha (1 - Z_{if}^\alpha) \tag{27c}$$

$$x_l^{\alpha-1} - M(1 - A_{il}) \leq b_{il}^{\alpha-1} \tag{27d}$$

$$b_{il}^{\alpha-1} \leq x_l^{\alpha-1} + M(1 - A_{il}) \tag{27e}$$

$$- M(A_{il}) \leq b_{il}^{\alpha-1} \leq M(A_{il}) \tag{27f}$$

$$x_i^0 = x_i \tag{27g}$$

$$y_f = \sum_{i=1}^{M} x_{if} \tag{27h}$$

$$x_{if}^\alpha, S_{if}^\alpha \geq 0 \tag{27i}$$

$$Z_{if}^\alpha \in \{0, 1\} \tag{27j}$$

$\forall \alpha \in \{1, \ldots, A\}$, $\forall i \in \{1, \ldots, M\}$, $\forall l \in \{1, \ldots, M\}$, $\forall f \in \{1, \ldots, F^\alpha\}$. The variables b are necessary to model the conditional sum (26). In particular,

$$b_{il}^\alpha = \begin{cases} 0 & \text{if } A_{il} = 0, \\ x_l^\alpha & \text{if } A_{il} = 1. \end{cases}$$

Also, Eqs. (27a-27c) define the affine maps and linearize the ReLU function. (27d-27f) model the conditional sum in 26. Eq. (27g) specifies the input of the network. Eq. (27h) performs the last sum-aggregation function and gives the input for the classification network. With the aim of making the parameters of this formulation binary, we require that $W_{r,f}^\alpha \in \{-1, 0, 1\}$ $\forall \alpha \in \{1, \ldots, A\}$, $\forall r \in \{1, 2\}$, $\forall f \in \{1, \ldots, F^\alpha\}$, $\rho(x) = 2 \times \mathbb{1}(x \geq 0) - 1$. This choice leads to an exact linear reformulation concerning the products between decision variables. Testing will be carried out in the near future and results will be compared.

4 Conclusions

In conclusion, our project involved two main sections. In the first section, we collaborated with Dr N. Yorke-Smith to further our understanding of Integer Neural Networks and investigate different paradigms for achieving robustness and simplicity. Through experiments on small datasets, we identified promising directions for further research, which will be the basis for larger-scale experiments in the future. Our goal is to build on the methodology proposed in our previous paper [1] and improve it for a journal version.

Specifically, preliminary experiments on weights distribution suggest that more pervasive use of memory for parameters does not necessarily lead to improved accuracy on tests. In addition, the extremal distribution observed in the totality of cases needs to be investigated more. These points will be addressed more extensively through the conduct of new experiments, and the results obtained will be the main integrative contribution for the new version of the paper for *INFORMS Journal on Computing*. On the other hand, concerning the different approaches to simplicity and robustness paradigms, the results obtained showed no improvement over those obtained in our paper [1]. Despite this, the experiments conducted provided theoretical insight of independent interest. In particular, BNNs that possess a layer with only one active neuron can be truncated at that level.

The second section provides an overview of the theoretical framework underlying the structure of graph neural networks, and a possible way of proceeding is highlighted.

Taken together, the topics of this report are a valuable starting point for a collaboration with Dr Yorke-Smith and the STAR Lab at the TU Delft. In particular, our visit allowed us to identify a deep commonality of interests that could lead to new periods of visiting, both by masters and Ph.D. students. The relationship between STAR Lab and CompOpt Lab shows promising prospects.

References

[1] Bernardelli, A.M., Gualandi, S., Lau, H.C., Milanesi, S.: The BeMi Stardust: a Structured Ensemble of Binarized Neural Networks. arXiv preprint arXiv:2212.03659 (to appear in proceedings of LION17) (2022)

[2] Gurobi Optimization, LLC: Gurobi Optimizer Reference Manual (2022), https://www.gurobi.com

[3] Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R.: Heart Disease Data Set. UCI Machine Learning Repository (1988), http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] LeCun, Y., Cortes, C., Burges, C.J.: The MNIST database of handwritten digits (1998), http://yann.lecun.com/exdb/mnist

[5] Mc Donald, T.: Mixed Integer (Non-) Linear Programming Formulations of Graph Neural Networks. Master's thesis, TU Delft (2022), http://resolver.tudelft.nl/uuid:b1d7ce2f-f773-4593-ac72-6e086c4d2d11

[6] Schürholt, K., Schweidtmann, A.: Global Deterministic Training of Artificial Neural Networks. Master's thesis, RWTH Aachen University (2018)

[7] Thorbjarnarson, T., Yorke-Smith, N.: Optimal training of integer-valued neural networks with mixed integer programming. PLOS ONE **18**(2), 1–17 (02 2023). https://doi.org/10.1371/journal.pone.0261029, https://doi.org/10.1371/journal.pone.0261029

[8] Toro Icarte, Rodrigo, e.a.: Training binarized neural networks using MIP and CP. Principles and Practice of Constraint Programming: 25th International Conference, CP 2019, Stamford, CT, USA, September 30–October 4, 2019, Proceedings 25, Springer (2019)

[9] Wolberg, W.H., Street, W.N., Mangasarian, O.L.: Breast Cancer Wisconsin (Diagnostic) Data Set. UCI Machine Learning Repository (1995), https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Basic Information

Project title: Graph Gaussian Processes for Interactive Robot Task Learning
Period of project: 1/9/2022 to 28/02/2023
Period of reporting: 1/9/2022 to 28/02/2023
Author(s): Giovanni Franzese
Organization: Delft University of Technology
Host organization: University College London

Public summary

Maximum half a page.
Uncertainty-aware machine learning is a rapidly evolving field at the intersection of artificial intelligence and statistics that addresses a fundamental challenge in real-world decision-making: the presence of uncertainty. In many practical applications, such as autonomous driving, healthcare, finance, and natural language processing, it is essential to not only make predictions but also quantify the uncertainty associated with those predictions.

Traditionally, machine learning models have been designed to provide point estimates, which means they generate a single best guess for a given input. Uncertainty-aware machine learning seeks to remedy this limitation by incorporating probabilistic reasoning into the modeling process. The output distribution can capture various forms of uncertainty, including aleatoric uncertainty (inherent randomness) and epistemic uncertainty (uncertainty due to lack of data or model knowledge).

This project aims to enhance the quantification of uncertainty for node classification in graph structures by combining Gaussian Process (GP) techniques with Graph Neural Networks (GNN). The objective is to harness GNN's feature extraction capabilities along with GP's statistical principles to enhance the model's uncertainty estimation without sacrificing performance. Enhancing the calibration of probability predictions enables a more accurate assessment of when to have confidence in the model's predictions and when to exercise caution.

Research objectives

Maximum 1 pag

Objectives

When dealing with data represented as graphs or networks, a graph Neural Network (GNN) is a powerful machine learning model. Unlike traditional neural networks, GNNs are specifically tailored to handle structured data where entities are connected by edges, making them well-suited for tasks like social network analysis, recommendation systems, and molecular chemistry modeling but also robot task planning, multi-robot coordination, and robot manipulation. GNNs learn to propagate information across nodes in the graph enabling them to capture complex relationships and dependencies within the data. By iteratively aggregating and updating node information based on their neighbors, GNNs have proven to be remarkably effective in uncovering hidden patterns and making predictions in various domains, contributing significantly to the field of graph-based machine learning.

However, whether the results are trustworthy is still unexplored. Previous studies suggest that many modern neural networks are over-confident in their predictions, however, surprisingly, GNNs are primarily in the opposite direction, i.e., GNNs are under-confident. Therefore, researching how to perform calibration for GNNs is highly desired.

In machine learning, a classification model is considered well-calibrated when the predicted probabilities it assigns to its predicted class labels are representative of the true likelihood or probability that those instances belong to their respective classes. In other words, a well-calibrated classification model produces predicted probabilities that can be interpreted as confidence scores or estimates of the true probability of an instance belonging to a particular class. This project has the objective of better understanding how to calibrate the uncertainty prediction of modern graph neural networks.

Impact

Predicting outcomes within a graph structure, such as discerning the legitimacy of a bank transaction, gauging potential risks within a social network, classifying molecule properties, or forecasting actions in a multi-robot system, holds considerable societal significance. The presence of a machine capable of quantifying prediction uncertainty and aligning it effectively with the actual likelihood of correctness can enhance its trustworthiness and deployability in safety-critical applications like healthcare or human-centric robotics.

Technical approach

Maximum 1 page.

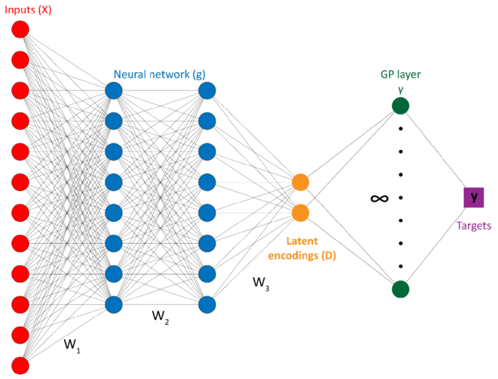
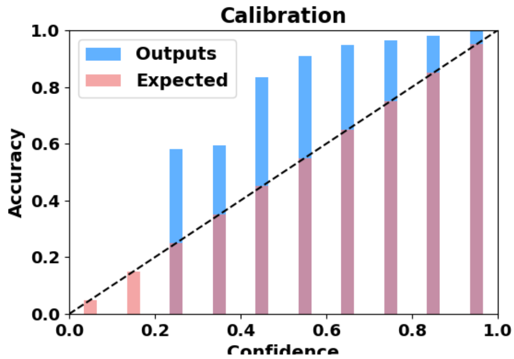


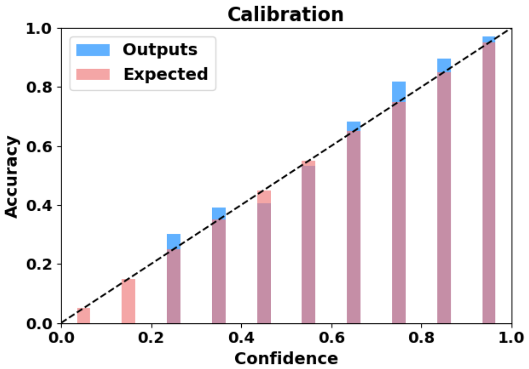
Fig 1. Deep Kernel Learning of Gaussian Process

Deep kernel learning, see Fig 1, merges the flexibility of kernel methods with modern deep learning architectures. We introduce a deep graph kernel learning model that is trained with a stochastic variational procedure. We simultaneously optimize the parameters of these base kernels and the deep network using an expected lower bound of the marginal likelihood objective. This approach allows us to make inferences with the Gaussian Process while still exploiting the message-passing paradigm of graph neural networks. This results in a higher accuracy compared to graph Gaussian processes and better calibration of the uncertainty prediction with respect to GNN alternatives. The results have been validated with popular graph benchmarks.

Scientific outcomes

Our results demonstrate enhanced performance compared to standalone graph neural networks or Gaussian Gaussian Processes across various classification benchmarks.
The use of Gaussian Process layer shows a natural calibration of the prediction probability of the class of each node. Additionally, the predicted entropy of the classifier was used as a decision theory tool, i.e. when it is larger than a value, the classifier prefers to not make predictions on the queried node showing an increase in the trustworthiness of the model. Additionally, a visual interpreter of the classification prediction leveraging the learned latent space is proposed as a tool to increase visual interpretability of why the machine learning tool predicted what it did. The following figures show the calibration curve of a graph attention network, on the left, and the result of the calibration using the proposed method, on the right. We can see that the use of the Gaussian Process remarkably improves the performance of the classifier, i.e. when the classifier predicts that a label is x with a probability p, it is going to be correct on average p times of the total attempts. This is indeed highly desired to make sense of the output of our machine learning model.





Future plans

Preliminary results of the proposed method in multi-robot coordination have been obtained but in the future steps, more experiments need to be done.

Self-assessment

Maximum 1 page. Only for final reporting (the project has finished)

- **AI Excellence**: Developing a statistical machine learning tool for the calibration of uncertainties in the classification of graph neural networks can increase the perceived trustworthiness of this in different engineering applications, from robotics to healthcare or drug design.
- **Scientific step-up**: Having the possibility of working with a lead scientist in Statistical Machine Learning and being in the AI center of UCL gave me multiple insights into the important problems in machine learning made me a better researcher and increased my knowledge of the topic in particular on Gaussian Process and Graph Neural Networks.
- **Suitability of the host**: Marc Deisenroth was the ideal host for this project. He is eager to try new ideas and always available for meetings and feedback.
- **Suitability of the visit length**: The length was good. However, being at the end of my Ph.D. made the finalization of the project overlap with the writing of my Ph.D. thesis.

List of publications, meetings, presentations, patents,...

The paper that summarizes the final results of the project is in the writing process.