Project Assignments

Andrea Passerini
andrea.passerini@unitn.it

Advanced Topics in Machine Learning and Optimization

DeepProbLog vs Deep Network

Assignment

- Consider the MNIST multi-digit addition example we saw in the lecture
- Design a purely neural architecture that predicts the result of the addition (e.g. CNN+LSTM, but feel free to invent)
- Compare results of the neural architecture with those of DeepProbLog in the same setting:
  - supervision only on the result of the addition
  - generalization to longer numbers
- Try training the neural network with a larger training set than the one used for DeepProbLog (check how many examples are need to match DeepProbLog performance)

Notes

- **Contact**: Andrea Passerini
- **Can be selected multiple times**

Probabilistic verification of Neural Networks

Assignment

- **Why** Standard evaluation methods for ML models give no guarantee on their out-of-sample behaviour.
- **What** Contribute to techniques that provide probabilistic guarantees on the correctness of NNs (pick one):
  - Explore different domain: fairness, safety
  - Push the scalability of probabilistic verification techniques

Notes

- **Contact**: Paolo Morettin
- **Extensible to thesis**
- **Can be selected multiple times**

Supernode GNN

Assignment

In Graph Neural Networks, the message passing operation is inherently local, gathering information solely from neighboring nodes. To enhance the expressiveness of these GNNs, a promising approach involves defining a vocabulary of concepts (e.g., cycles or paths within graphs) and aggregating information on these structures before applying the GNN.

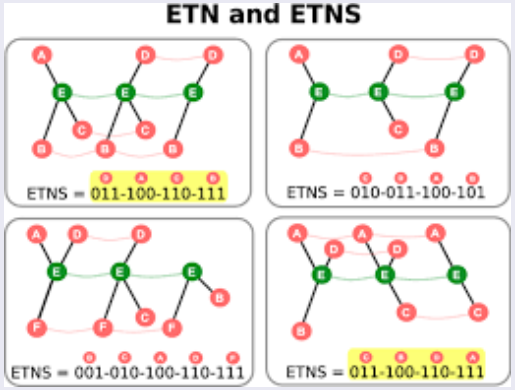
Notes

- **Contacts**: Azzolin Steve Ferrini Francesco

Approximate motifs counting in Temporal Networks

Assignment

Temporal motifs are statistically significant temporal substructures. Several methods already tackle the problem of counting temporal motifs in networks. On the other hand, approximate counting is less explored. The student should extend the notion of Egocentric Temporal Motifs Signature for approximate counting of temporal motifs.



Notes

- **Contact**: Antonio Longa
- **Extensible to thesis**
- recommended basics of Graphs/Networks
- link 1,link 2

Mitigation of Unfortunate Counterfactual Events (UCE)

Assignment

- **Counterfactual explanations** (CE) are an intuitive and model-agnostic class of explanations for ML models.
- They provide **actionable** feedback to users affected by automated decision systems. However, their validity is sensitive to model shifts (e.g., periodic re-training).
- Ferrario & Loi (2022) proposed counterfactual data augmentation (CDA) as a mitigation strategy, but considering only model shifts.
  - Building from the results of Ferrario & Loi (2022), evaluate CDA-improved models under **distribution shift**.

Notes

- **Contact**: Giovanni De Toni
- **Extensible to thesis**
- **Reference:** Ferrario & Loi (2022)'s paper

Addressing Reasoning Shortcuts with Information Theory

Assignment

- **Why** Reasoning Shortcuts exponentially increase when disentanglement cannot be guaranteed. We are looking at how to optimize for it in practice.
- **What** Information Theory can help disentangle concepts by decorrelating them. Much can be done in this respect:
  - Integrate a total correlation loss among concepts that
  - Investigate how existing entropy regularizations can help to reduce RSs.
- One case study is sufficient for the project while taking both would result in a higher mark.

Notes

- **Contact**: Emanuele Marconato, Samuele Bortolotti
- **Extensible to thesis**

Reasoning shortcuts with Unknown Knowledge

Assignment

- **Why** Reasoning Shortcuts do arise with prior-knowledge. There are new models, like ROAP and DSL that learn both concepts and knowledge.
- **What** Choose one model of interest and investigate:
  - Whether these models are also affected by Reasoning Shortcuts on knowledge
  - The effects of mitigation strategies, like concept supervision and entropy regularization
- Implementing both models would result in a higher mark.

Notes

- **Contact**: Emanuele Marconato
- **Extensible to thesis**

Causal Abstractions of Neural Networks solving logical operations

Assignment

- **Why** DAS is a powerful method for evaluating the degree to which a computational graph is a causal abstraction of a neural network. We want to investigate what happens in scenarios where we know that reasoning shortcuts can appear.
- **What** Replicate the learning of MNIST-Addition with two and three digits, but with a Deep Neural Network:
  - Estimate DAS for the graph of the sum operation
  - Can we spot eventual reasoning shortcuts?

Notes

- **Contact**: Emanuele Marconato, Samuele Bortolotti
- **Extensible to thesis**

Argumentative interaction with LLMs in Legal Sector

Assignment

- **Why** Unchecked use of LLMs in law is risky; involving humans in the loop of decision-making in legal tasks is more ethical and practical.
- **What** Develop an LLM-powered Intelligent Assistant (IA) to augment human decision-makers in the legal field. The IA will not make a final decision but rather help the legal expert make a final decision through an argumentative interaction.

Notes

- **Contact**: Burcu Sayin
- **Collaboration**: Marco Lippi, University of Modena and Reggio Emilia
- **Extensible to thesis**

xAI for Curriculum Learning on Graph/Euclidean data

Assignment

- Curriculum Learning is a learning paradigm that ranks the training samples according to a difficulty criterion, and a curriculum scheduler. An unexplored research trend regards the design of difficulty criterion depending on the explainability.
- Implement ***one*** *of the possible two options*:
  - **Option 1:** Curriculum Learning for GNNs [1] based on *Explainability on Graph data* [2]
  - **Option 2:** Curriculum Learning [3] based on *Explainability on Euclidean data (e.g. images)* [4]

Notes

- **Contact**: Vincenzo Marco De Luca
- Click on reference for external papers.

Highly-Explainable Graph Neural Networks

Assignment

- B-cos networks [1] proved that a straightforward transformation (B-Cos transformation) in the dot product enhances the explainability of CNNs.
- Tx he GNN community is focusing its efforts on GNNs' explainability [2], but *B-cos networks* in GNNs have not yet been tested.
- The student is asked to:
  - Apply *B-cos transformation* on GNNs
  - Test it on some datasets, GNN variants, and explainers.

Notes

- **Contact**: Vincenzo Marco De Luca, Steve Azzolin
- Click on reference for external papers.

Graph forecasting

Description

- Similar to time series forecasting: predict future development of graph evolution process
- Special case: predict development of information cascade in social network
- Train recurrent graph neural network models on information cascade data (e.g. $\mathbb{X}$)
- Apply trained network to predict future development of observed initial cascade

Notes

- **Contact**: Andrea Passerini
- **Collaboration**: Manfred Jaeger, Aalborg University, DK
- Possibility for internship abroad
- **Extensible to thesis**

Spuriosity Didn't Kill The Debugger

Description

- Interactive Debugging teaches models not to rely on spurious features, e.g., watermarks.
- Spurious features however can be useful as long as they are used in the proper context/domain!
- Combine recent techniques for *leveraging spurious features in the right domain* with explanatory interactive learning.

Notes

- **Contact**: Stefano Teso
- **Extensible to thesis**
- **Papers**: 1, 2

Smart Query Selection for Interactive Debugging

Description

- Interactive Debugging often builds on stock active learning strategies, e.g., uncertainty sampling.
- This is highly sub-optimal: uncertainty does not necessarily identify bugs!
- Make interactive debugging smarter (and less wasteful) by leveraging recent techniques for *automatic spurious feature detection* from the domain adaptation literature.

Notes

- **Contact**: Stefano Teso
- **Extensible to thesis**
- **Papers**: 1, 2

RLHF-style Debugging (Concept-based) Models

Description

- Interactive debugging requires users to supply feedback on individual images, e.g., “don’t look at the snow”.
- This is expensive! Rather, learn an RLHF-style surrogate that imitates and replaces user feedback.
- Evaluate whether RLHF+Debugging manages to simulate user corrections on a (possibly concept-based) model and data set of choice.

Notes

- **Contact**: Stefano Teso
- **Extensible to thesis**
- **Papers**: 1
- End-to-end training is okay, RL is not strictly required ;-)

Assignment

- Select one of the projects from the previous slides (or discuss with the teacher for custom projects)
- Complete it and prepare a report summarizing the methodology used and the results obtained.
- After completing the assignment send it via email to andrea.passerini@unitn.it
- Subject: ADVML2023
- Attachment: name_surname.zip containing:
  - the report (named report.pdf)
  - the code you wrote
  - the requirements needed to run the code

NOTE

- No group work
- Preliminary versions of the report can be sent for feedback
- The project is discussed asynchronously as soon as it is completed