CS-E4875 Research Project in ML, DS & AI – Topics 2023

Topics for the course **CS-E4875 Research Project in Machine Learning, Data Science and Artificial Intelligence** for the academic year 2023-2024.

Topic #1: Face Presentation Attack Detection Using Deep Learning

Background: The convenience and high accuracy of face recognition technology have led to its widespread adoption in everyday interactive tasks like mobile devices, border control, banking, surveillance, and monitoring. However, face spoofing poses a significant vulnerability to face recognition systems due to its capacity to deceive and manipulate the technology. Spoofing involves presenting manipulated or fake images to the system, which can lead to unauthorized access or impersonation. The goal of this project is to explore different deep learning architectures (Transformers, CNNs, RNNs, etc), as well as data augmentation techniques, to improve the performance against spoofing attacks.

This topic has plenty of exciting tasks for enthusiastic students. Some activities could include a literature search, coding of different parts (pre-processing, face detection, evaluation), and displaying real-time results using the webcam.

Related Work:

Sharma, Deepika, and Arvind Selwal. "A survey on face presentation attack detection mechanisms: hitherto and future perspectives."*Multimedia Systems* 29.3 (2023): 1527-1577.

Liu, Ajian, et al. "FM-ViT: Flexible Modal Vision Transformers for Face Anti-Spoofing."*IEEE Transactions on Information Forensics and Security* (2023).

Prerequisites: Basics of deep learning, Python or MATLAB

Language requirements: English

Supervisor name and email: Usman Muhammad (usman.muhammad@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Jorma Laaksonen

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_one_instance

Topic available also for a group: yes (max 2 students in the group)

Graded as something else than CS-E4875: no

Topic #2: Efficient Methods for Personalized Federated Learning

Background: Some application domains for federated learning, such as personalized healthcare or high-precision weather prediction, generate scattered collected of local datasets. It is often not obvious which of these local datasets should be used for the training of a personalized model. This project studies different techniques for finding out which local datasets should be included in the training for a personalized model.

Prerequisites: Must have: Linear algebra, probability theory (up to concentration inequalities for functions of sub-exponential random variables), convex optimization. Nice to have: online learning (contextual bandits), active learning.

Related Work:

- Werner, M., He, L., Praneeth Karimireddy, S., Jordan, M., and Jaggi, M., “Provably Personalized and Robust Federated Learning”, <i>arXiv e-prints</i>, 2023. doi:10.48550/arXiv.2306.08393.
- *Laurent Jacob, Jean-philippe Vert, Francis Bach*, Clustered Multi-Task Learning: A Convex Formulation Part of Advances in Neural Information Processing Systems 21 (NIPS) 2008
- SarcheshmehPour, Y., Tian, Y., Zhang, L., and Jung, A., “Clustered Federated Learning via Generalized Total Variation Minimization”, <i>arXiv e-prints</i>, 2021. doi:10.48550/arXiv.2105.12769.

Language requirements: English (oder Deutsch)

Supervisor name and email: Alexander Jung (alex.jung@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one:

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_many_instances (max 3 instances)

Topic available also for a group: yes (max 2 students in the group)

Graded as something else than CS-E4875: no

Topic #3: Trade-Off Between Interpretability/Explainability and Accuracy

Background: Some ML applications require more than high accuracy of trained models. In particular, some applications require also ML models to be interpretable or explainable. This project studies the trade-off between a precise notion of subjective interpretability and the achievable accuracy of ML methods in important settings.

Prerequisites: Must have: Linear algebra, probability theory (covariance structure of Gaussian processes), convex optimization.

Language requirements: English (oder Deutsch)

Related Work:

- A. Jung and P. H. J. Nardelli, "An Information-Theoretic Approach to Personalized Explainable Machine Learning," in *IEEE Signal Processing Letters*, vol. 27, pp. 825-829, 2020, doi: 10.1109/LSP.2020.2993176
- Zhang, L., Karakasidis, G., Odnoblyudova, A., Dogruel, L., and Jung, A., “Explainable Empirical Risk Minimization”, <i>arXiv e-prints</i>, 2020. doi:10.48550/arXiv.2009.01492.

Supervisor name and email: Alexander Jung (alex.jung@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one:

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_many_instances (max 3 instances)

Topic available also for a group: yes (max 2 students in the group)

Graded as something else than CS-E4875: no

Topic #4: Interactive AI for retrosynthesis planning

Background: Retrosynthesis planning is the core task in synthetic chemistry, in which chemists recursively deconstruct a target molecule to find a set of reactants that make up the target. It is challenging in that the search space is vast, and chemists are often lost in the process. Existing AI models can achieve automatic retrosynthesis planning fast, but they only work on relatively simple targets, which leaves complex molecules under chemists’ expertise. The goal of this project is to build a human-AI collaborative system through a participatory design process. This project would involve exploring and developing e.g. the following topics: (i) develop a scoring model based on retrosynthetic route datasets, (ii) improve the generative model through continuous integration of expert feedback, (iii) develop an interactive algorithm and interface for human-AI collaboration. Knowledge in (probabilistic) machine learning is expected.

Prerequisites:

Language requirements: English

Special conditions (if any):

Supervisor name and email: Yujia Guo (yujia.guo@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Samuel Kaski

Extent of the project: 5–10 credits

Preferred timing: autumn 2023

Topic available: yes_one_instance

Topic available also for a group: yes

Graded as something else than CS-E4875: no

Topic #5: Few-shot learning for simulation-based inference

Background: Simulation-based inference (SBI) methods are used to fit complex, simulator-based models with intractable likelihood function to data. Such models appear in a various field of science and engineering such as population genetics, radio propagation, and cosmology. Neural network-based SBI methods, such as Greenberg et al. 2019, use training data generated from the simulator to learn the posterior distribution. However, inference becomes challenging when the simulator is computationally expensive, thus limiting the number of training data points available. In this project, you will apply few shot learning methods from the deep learning literature to solve the problem of SBI under limited training budget.

References:
https://www.pnas.org/doi/10.1073/pnas.1912789117
http://proceedings.mlr.press/v97/greenberg19a/greenberg19a.pdf(Greenberg et al. 2019)

Prerequisites:

Language requirements: English

Special conditions (if any):

Supervisor name and email: Ayush Bharti (ayush.bharti@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Samuel Kaski

Extent of the project: 5–10 credits

Preferred timing: autumn 2023

Topic available: yes_one_instance

Topic available also for a group: yes

Graded as something else than CS-E4875: no

Topic #6: Bayesian Topological Deep Learning

Background: Topological deep learning (TDL) is a rapidly growing field that combines tools from (algebraic) topology and machine learning to build powerful predictive models. Recently, TDL has led to breakthroughs in many scientific domains, including computer vision, drug discovery, material science, recommender systems, and physics. Despite this success, current TDL methods fail to effectively model uncertainty in the underlying domains (e.g., graphs, simplicial complexes). In this project, we will study Bayesian approaches for TDL, aiming to reduce overfitting/oversmoothing/over-squashing, enable learning from small datasets, and improve uncertainty quantification of predictions. Our investigation will start with methods based on message passing, such as Message-Passing Simplicial Networks (MPSNs). We expect this project will support the design of better, more nuanced models that are both theoretically well-grounded and practically efficacious.

References:

1. Architectures of Topological Deep Learning: A Survey on Topological Neural Networks. ArXiv, 2023.
2. Bayesian Graph Neural Networks with Adaptive Connection Sampling. ICML, 2020.

Prerequisites:

Language requirements: English

Special conditions (if any):

Supervisor name and email: Amauri Souza (amauri.souza@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Samuel Kaski

Extent of the project: 5–10 credits

Preferred timing: autumn 2023

Topic available: yes_one_instance

Topic available also for a group: yes

Graded as something else than CS-E4875: no

Topic #7: Human-AI interaction in RL for molecular optimisation

Background: Reinforcement Learning (RL) has shown promise in advancing molecular design and therapeutic research. Domain expertise in medicinal chemistry is crucial to design the therapeutic goal (reward) and successfully guide the design process. Recently, human-in-the-loop RL was introduced to the molecular design field, enabling interactive reward learning from human feedback [1]. While a wide variety of human-AI interaction approaches for reward learning in RL were proposed in the existing literature, only very few were applied or adapted to the case of molecular optimisation. In this thesis project, you will study various human-AI interactions for reward learning that were successfully used in applications such as image recognition, text generation etc., and further investigate the potentially useful and feasible ones for the molecular optimisation case. You will work on modelling and implementing the process of human-AI interaction and human feedback integration for reward learning, and engage in discussions with the UI designer in charge of providing the interface for future experiments with chemists. The main aim of this project is to propose a model of human-AI interaction for reward learning in molecular optimisation and validate it with experiments through the UI.

References:

[1] Human-in-the-loop assisted de novo molecular design
[2] Where to add actions in human-in-the-loop RL
[3] A survey of human-in-the-loop for machine learning

Prerequisites: understanding of reinforcement learning and probabilistic inference methods. Some knowledge about computational molecular representations would be a plus!

Language requirements: English

Special conditions (if any):

Supervisor name and email: **Yasmine Nahal** (yasmine.nahal@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Samuel Kaski

Extent of the project: 5–10 credits

Preferred timing: autumn 2023

Topic available: yes_one_instance

Topic available also for a group: yes

Graded as something else than CS-E4875: no

Topic #8: SAM Enhancements for Audio-Visual tasks

Background: The Segment Anything Model (SAM) has demonstrated remarkable effectiveness in visual segmentation tasks. However, there has been relatively limited investigation into SAM's performance in video-related tasks, particularly those involving audio-visual components such as active speaker tracking, audio-visual event localization, and segmentation. In this research, we aim to enhance SAM's capabilities for video-based tasks by incorporating cross-modal features.

References:

1- Segment Anything https://arxiv.org/abs/2304.02643

2- Track Anything: Segment Anything Meets Videos https://arxiv.org/abs/2304.11968

Prerequisites: Good python programming. Proficiency in PyTorch is a plus.

Language requirements: English

Special conditions (if any):

Supervisor name and email: Jalil (abduljalil.saif@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Jorma Laaksonen

Extent of the project: 5–10 credits

Preferred timing: autumn 2023/spring 2024

Topic available: yes_one_instance

Topic available also for a group: yes (max 2 students in the group)

Graded as something else than CS-E4875: no

Topic #9: Large language models for acceleration of physical systems modelling

Background: Large language models (LLMs) changed the landscape of natural language processing (NLP) over the last few years. However, the capabilities of LLMs can be utilized in a number of applications beyond NLP tasks. In this project, we will study different approaches on how LLMs can be utilized in the context of physical system modelling, such as finding the analytic form and coefficients of the differential equation that models certain data points. As known, the performance of pre-trained LLMs can be significantly boosted by providing tuned prompts for solving the desired tasks. We will work on incorporation of knowledge about the phenomenon given in the natural language in the prompts. The project can potentially involve the creation of an ad-hoc dataset for similar tasks. For example one can use a combination of the Wikipedia API, ChatGPT and Python scripts to parse famous differential equations or probability distributions together with their description and then sample data points from them. These (data points, descriptions) pairs can be used as test data for the modeling methods described above.

References:

1. Mirchandani, S., Xia, F., Florence, P., Ichter, B., Driess, D., Arenas, M.G., Rao, K., Sadigh, D. and Zeng, A., 2023. Large language models as general pattern machines. arXiv preprint arXiv:2307.04721. https://arxiv.org/abs/2307.04721
2. Yang, C., Wang, X., Lu, Y., Liu, H., V. Le, Q., Zhou, D., Chen, X., 2023. Large Language Models as Optimizers. arXiv preprint arXiv:2309.03409. https://arxiv.org/abs/2309.03409

Prerequisites: Good Python programming, understanding of Deep Learning and LLMs is a plus.

Language requirements: English

Special conditions (if any):

Supervisor name and email: Katsiaryna Haitsiukevich (katsiaryna.haitsiukevich@aalto.fi), Nicola Dainese (nicola.dainese@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Pekka Marttinen

Extent of the project: 5–10 credits

Preferred timing: autumn 2023

Topic available: yes_one_instance

Topic available also for a group: yes (max 2 students in the group)

Graded as something else than CS-E4875: no

Topic #10: Detecting Moving Vehicles for Self-Driving: Leveraging Semi-Supervised Learning for Scalable Data Annotation

Background: This project focuses on employing Semi-Supervised Learning (SSL), specifically pseudo-labeling [1], to enhance the detection of moving vehicles in self-driving scenarios. It aims to use a small manually annotated dataset [2] to bootstrap the creation of a larger audio-visual dataset with pseudo-labels. This approach addresses the challenge of scaling up vehicle detection without the need for costly manual annotations.

References:

1- In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-Label Selection Framework For Semi-Supervised Learning https://github.com/nayeemrizve/ups

2- Self-Supervised Moving Vehicle Detection from Audio-Visual Cues http://av-vehicles.informatik.uni-freiburg.de/#

Prerequisites: Good python programming. Proficiency in PyTorch is a plus.

Language requirements: English

Special conditions (if any):

Supervisor name and email: Jalil (abduljalil.saif@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Jorma Laaksonen

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_one_instance

Topic available also for a group: yes (max 2 students in the group)

Graded as something else than CS-E4875: no

Topic #11: Uncertainty quantification of deep neural networks for high-dimensional data

Background:

In safety-critical application domains like personalized medicine and drug discovery, the reliability of a model prediction is at least as important as the prediction itself [1].

However, most of the research in machine learning focuses on methodologies producing only simple predictions.

This project will compare different methods for quantifying the uncertainty of predictions from deep neural networks applied to high-dimensional data.

In particular, this work will practically utilize uncertainty quantification methods like conformal prediction [2] and employ high-dimensional data like molecular and genomics data [3].

Finally, this study will consider applications like pairwise drug synergy prediction or classification [4].

Related Work:

[1] https://doi.org/10.1016/j.inffus.2021.05.008

[2] https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial_notebooks/DL2/Bayesian_Neural_Networks/dl2_bnn_tut2_student_with_answers.html#Conformal-prediction

[3] depmap.org

[4] https://doi.org/10.1093/bioinformatics/btx806

Prerequisites: PyTorch, understanding well related works 1 and 2.

Language requirements: English

Supervisor name and email: Gianmarco Midena (gianmarco.midena@aalto.fi)

Supervisor's research group leader's name unless the supervisor him/herself is one: Juho Rousu

Extent of the project: 5 - 10 credits

Preferred timing: any

Topic available: yes_one_instance

Topic available also for a group: no

Graded as something else than CS-E4875: no

Topic #12: Applied Machine Learning

Background: Pick some application domains of your interest (hobby, research project, master/phd thesis or work-related) and model it as a machine learning problem according to Chapter 2 of my textbook Ref1. In particular, you must define data points, their features and the quantity of interest ("label"). Based on the problem formulation you can apply different ML methods (ranging from plain old linear regression to the latest super-fancy media-hyped deep learning model) and compare their performance in numerical experiments.

Prerequisites: CS-EJ3211 Machine Learning with Python (or equivalent).

Language requirements: English (oder Deutsch)

Related Work:

Ref1 A. Jung, "Machine Learning: The Basics," Springer, Singapore, 2022 (available electronically at Aalto lib https://primo.aalto.fi/permalink/358AALTO_INST/ha1cg5/alma999673293406526).

Supervisor name and email: Alexander Jung (alex.jung@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one:

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_many_instances (max 99 instances)

Topic available also for a group: no (max 2 students in the group)

Graded as something else than CS-E4875: no

Topic #13: Deep learning in speech and language processing and large language and speech models

Background: Deep learning and large pre-trained self-supervised language and speech models have changed the ways how speech and language data can be processed and represented. Several specific topics are available either for experimenting with new deep learning architectures in real-word data or applications, such as automatic speech recognition, understanding and language learning and modeling. The topic can be selected together with the student.

Prerequisites: Aalto's basic course in either speech recognition or natural language processing or corresponding knowledge. Knowledge in deep learning. Experience in scientific programming, e.g. in Python**.**

Language requirements: English

Supervisors name and email: Mikko Kurimo mikko.kurimo@aalto.fi, Tamas Grosz tamas.grosz@aalto.fi and the automatic speech recognition (ASR) research group

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_multiple_instances (max 3)

Topic available also for a group: yes (max 2 in a group)

Graded as something else than CS-E4875: ELEC-E5541 Special Assignment (can substitute CS-E4875 for Macadamia)

Topic #14: Uncertainty Quantification for Deep Visual Geo-localization

Background: Visual Geo-localization (VG) is a computer vision task that coarsely finds the geographical position of a given picture. The task finds applications in autonomous driving and mobile robotics.
VG is commonly approached as an image retrieval problem. Given a query image to localize, the goal is to match it against a database of geo-tagged photos that cover a specific geographical area. Most recent works perform the retrieval phase as a k-nearest-neighbour (kNN) search in a learned embedding space. In practice, a deep neural network learns to extract global descriptors of each image with contrastive learning approaches, which are matched against the descriptors of the database of photos of the geographical area of interest.
The project aims to acquaint the student with the VG literature and focus on designing new methods to get uncertainty estimates for the retrieved results.

Prerequisites: Deep Learning - Computer Vision, experience with PyTorch

Language requirements: English (or Italian)

Related Work:
Visual Geo-localization:

- Deep Visual Geo-localization Benchmark, G. Berton, R. Mereu, G. Trivigno, C. Masone, G. Csurka, T. Sattler, and B. Caputo, CVPR 2022 - GitHub Page
- A survey on deep visual place recognition, C. Masone and B. Caputo, IEEE Access 2021

Uncertainty Quantification in image retrieval:

- Bayesian Triplet Loss: Uncertainty Quantification in Image Retrieval, F. Warburg, M. Jørgensen, J. Civera and S. Hauberg, ICCV 2021
- Bayesian Metric Learning for Uncertainty Quantification in Image Retrieval, F. Warburg, M. Miani, S. Brack, S. Hauberg, preprint

Supervisor name and email: Riccardo Mereu (riccardo.mereu@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Arno Solin (arno.solin@aalto.fi)

Extent of the project: 5–10 credits

Preferred timing: negotiable

Topic available: yes_one_instances

Topic available also for a group: no

Graded as something else than CS-E4875: no

Topic #15: Autoregressive models for T cell receptor generation

Background:
T cells are white blood cells and have a vital role in the adaptive immune system. They protect mammalian organisms against invading pathogens and cancer. For an immune response to be elicited, the receptor characterizing the T cell, the T cell receptor (TCR) must bind to an epitope (antigen fragment) presented by an MHC molecule of an infected cell. T cells are generated and selected in the human thymus by somatic V(D)J-recombination resulting in a vast diversity (~10^{20}) of TCRs. The TCR consists of two chains α and β that furthermore contains three complementary determining regions (CDRs). The CDR3 is the most variable of the three regions and has the largest importance in TCR-epitope binding.

The aim of this work is to devise and evaluate generative models for TCR repertoire generation conditioned on V and J genes. To this end, existing Protein Language Models (ProtBERT, ProtT5), can be further optimized for βCDR3 generation based on V and J gene information. Inspired by the success of recent Large Language Models, the models may be adapted to generate the sequences in an autoregressive fashion.

Prerequisites:
Deep Learning (CS-E4890); Pytorch

Language requirements: English

Supervisor name and email: Dani Korpela and Alexandru Dumitrescu (fristname.lastname@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Harri Lähdesmäki

Extent of the project: 5–10 credits

Topic available: yes_one_instance

Topic available also for a group: 2

Graded as something else than CS-E4875: no

Topic #16: Diffusion generative models with coarse-to-fine inductive biases

Background: Diffusion generative models are the current state-of-the-art models in image and audio synthesis. In their standard form, they generate data by iteratively denoising an initial white noise data point. Recently, some work has explored the use of *deblurring* (or inversion of the heat dissipation PDE over the image surface), in an attempt to incorporate the natural multi-scale structure of image data into the generative process, with one of the most recent work, the *Blurring Diffusion Model*, outperforming standard diffusion models on popular benchmark data sets.The first goal in this project would be to implement the Blurring Diffusion Model and reproduce some of the results. Aside from this, some further steps could include:

- Applying the model to audio data instead (the image generation part can be skipped entirely if you consider this more interesting). Can we obtain similar improvements?
- Trying to apply the model to larger image data sets with large resolution, e.g., 512x512 or lower-resolution ImageNet. Given that this may be not be trivial to do successfully in practice, careful evaluation and analysis of the model could be a valuable research contribution.

Related work:

Generative Modelling With Inverse Heat Dissipation, ICLR 2023, S. Rissanen, M. Heinonen, A. Solin (https://arxiv.org/pdf/2206.13397.pdf)

Blurring Diffusion Models, ICLR 2023, E. Hoogeboom, T. Salimans (https://openreview.net/pdf?id=OjDkC57x5sz)

Example of audio diffusion models:

DiffWave: A Versatile Diffusion Model for Audio Synthesis, ICLR 2021, Z. Kong, W. Ping, J, Huang, K. Zhao, B. Catanzaro (https://arxiv.org/pdf/2009.09761.pdf)

Prerequisites:
Deep Learning (CS-E4890); Pytorch

Language requirements: English

Supervisor name and email: Severi Rissanen (severi.rissanen@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Arno Solin

Extent of the project: 5–10 credits

Topic available: yes_one_instance

Topic available also for a group: 2

Graded as something else than CS-E4875: no

Topic #17: Accelerating reinforcement learning with GPU-based simulators

Background:

Deep reinforcement learning (DRL) has shown potential in training humanlike videogame agents [0][1] and physics based controllers for physically complex creatures.

The use of DRL is often limited by the time it takes to train new models. As a solution, GPU-based physics simulators [2][3] promise to increase the sample throughput by several orders of magnitude, allowing models to be trained in a matter of minutes. However, these GPU-based are very limited in features, and can therefore not be directly used to develop games.

The goal of this topic is to explore ways for training models in GPU-based simulators and transferring them to other game engines. Topic includes hands-on experimenting with reinforcement learning, GPU-based simulators and more traditional game engines, such as Unity.

References:

[0] Mnih, V. *et al.* (2015) 'Human-level control through deep reinforcement learning', *Nature*, 518(7540), pp. 529–533, https://doi.org/10.1038/nature14236

[1] Baker, B. *et al.* (2019) 'Emergent Tool Use From Multi-Agent Autocurricula', https://openai.com/research/emergent-tool-use

[2] Rudin, N. *et al.* (2021) 'Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning', http://arxiv.org/abs/2109.11978

[3] Shacklett, B. *et al.* (2023) 'An Extensible, Data-Oriented Architecture for High-Performance, Many-World Simulation', *ACM transactions on graphics*, https://doi.org/10.1145/3592427

Prerequisites:

Basics of deep learning, Python, Interest towards game engines (Unity, MuJoCo, …)

Language requirements: Finnish or English

Special conditions (if any):

Supervisor name and email: Anton Debner (anton.debner@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Vesa Hirvisalo

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_one_instance

Topic available also for a group: yes (max 2 students in the group)

Graded as something else than CS-E4875: no

Topic #18: Uncertainty quantification in additive Gaussian process models

Background: Gaussian processes are a flexible nonparametric model. Using them as building blocks in additive models allows us to decompose complex functions into understandable (interpretable) lower-dimensional components (see [1]). However, this relies on assumptions such as smoothness and feature independence, and in reality the model may still be overconfidently wrong. This project could go multiple directions:

- Study their calibration on out-of-distribution data or when features are correlated, and compare against similar methods (Generalised Additive Models using decision trees or Neural Additive Models).
- Another benefit of modelling a sum of low-dimensional components is that this may be more data-efficient. This could be used in Bayesian optimization to improve optimization of black-box functions (some familiarity with Bayesian optimization required).

If you are interested, please get in touch to arrange a meeting and discuss whether it could be a suitable project for you.

[1] Additive Gaussian Processes Revisited, Lu et al., ICML 2022 (https://arxiv.org/abs/2206.09861)

Prerequisites: Strong understanding of Gaussian processes and strong practical experience with Python and TensorFlow or JAX (or Julia ecosystem).

Language requirements: English

Supervisor name and email: Ti John <ti.john@aalto.fi>

Supervisor's research group leader's name unless supervisor him/herself is one: Pekka Marttinen

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_one_instance

Topic available also for a group: no

Graded as something else than CS-E4875: no

Topic #19: Explainable text similarity through word attribution weighting

Background:

One use case of (large) language models is to encode texts (e.g. sentences) into high-dimensional vectors, or *embeddings*. These embedding representations can further be used to efficiently compute and rank how similar texts are to each other – for example for the purpose of query-based sentence or document retrieval, or other semantic similarity assessment tasks. Using a contrastive training approach, as in the Sentence-BERT architecture [1], we achieve results that show high correlation to human similarity ratings (see also [2]). However, understanding *why* such a language model believes that two texts are similar or not is not a trivial task, yet potentially very helpful to the user.

In this project the aim is to explore the use of model explainability techniques in the form of word attribution weighting (see e.g. [3]) to help in explaining the *why*.

References:

[1] Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Reimers, N., & Gurevych, I. (2019). EMNLP 2019. https://aclanthology.org/D19-1410/

[2] Text embeddings by weakly-supervised contrastive pre-training. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., ... & Wei, F. (2022). *arXiv preprint arXiv:2212.03533*. https://arxiv.org/abs/2212.03533

[3] Axiomatic attribution for deep networks. Sundararajan, M., Taly, A., & Yan, Q. (2017). PMLR 2017. http://proceedings.mlr.press/v70/sundararajan17a.html

Prerequisites: Good Python programming skills, understanding of Deep Learning and LLMs is a plus.

Language requirements: English

Special conditions (if any):

Supervisor name and email: Hans Moen (hans.moen@aalto.fi)

Supervisor's research group leader's name unless supervisor him/herself is one: Pekka Marttinen, Samuel Kaski

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_one_instance

Topic available also for a group: no

Graded as something else than CS-E4875: no

Topic #20: Conformal Bayesian computation with Pareto smoothed importance sampling

Background: Conformal inference can provide better calibrated predictive intervals in case of model misspecification. In case of Bayesian models and Markov chain Monte Carlo methods, add-one-observation-in importance sampling can be used for efficient conformal inference. The project involves short review of the method, implementation of the method in Stan probabilistic programming ecosystem, and making experiments and a case study to illustrate the usefulness of the approach for model checking.

Prerequisites: Bayesian inference and Monte Carlo. R skills.

Language requirements: English

Supervisor name and email: Aki Vehtari (aki.vehtari@aalto.fi)

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_one_instance

Topic available also for a group: yes

Graded as something else than CS-E4875: no

Topic #21:

Background: Diagnostics and variance reduction for Bayesian leave-one-out cross-validation (LOO-CV). LOO-CV is a popular approach for assesing and selecting models based on predictive performance. As LOO-CV makes minimal assumptions about the future data distribution it tends o have relaivelу high variance. Adding some additional weak assumptions we may be able significanlty reduce that variance or gain additional informaion on the trustworthiness given the current data. The project involves short review of the method, implementation of the method in Stan probabilistic programming ecosystem, and making experiments and a case study to illustrate the usefulness of the approach for model assessment and comparison.

Prerequisites: Bayesian inference and Monte Carlo. R skills.

Language requirements: English

Supervisor name and email: Aki Vehtari (aki.vehtari@aalto.fi)

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_one_instance

Topic available also for a group: yes

Graded as something else than CS-E4875: no

Topic #22: Learnable filterbanks

Background: Filterbanks such as Mel or Bark are commonly used as fixed frontend transformations in many speech and audio processing tasks. The goal of this project is to implement a neural network layer that can be initialized as commonly used filterbanks, but whose weights can be updated through training and hence tailored to a specific audio processing task.

Prerequisites: Python, some background in speech and audio.

Language requirements: English

Supervisor name and email: Esteban Gómez (esteban.gomez@aalto.fi) or Tom Bäckström (tom.backstrom@aalto.fi)

Extent of the project: 5–10 credits

Preferred timing: any

Topic available: yes_one_instance

Topic available also for a group: yes

Graded as something else than CS-E4875: ELEC-E5541 Special Assignment (can substitute CS-E4875 for Macadamia)
