Advancing Explainable Artificial Intelligence for Trustworthy Machine Learning Applications

Summary

Machine learning and Artificial Intelligence (AI) have found extensive applications across various sectors like energy and medicine. However, a major challenge lies in the lack of interpretability of these models, leading to distrust among users who prefer transparent decision-making processes. Explainable Artificial Intelligence (XAI) addresses this issue by making ML models understandable. The Department of Software Science seeks a PhD candidate to advance XAI techniques, focusing on developing novel frameworks that bridge mathematical complexity with human perception. This role aims to enhance trust in AI systems through improved interpretability, particularly in critical domains like energy, medicine, and cybersecurity.

Research field:	Information and communication technology
Supervisors:	Dr. Sven Nõmm Prof. Dr. Juri Belikov
Availability:	This position is available.
Offered by:	School of Information Technologies Department of Software Science
Application deadline:	Applications are accepted between June 01, 2024 00:00 and June 30, 2024 23:59 (Europe/Zurich)

Description

Machine learning and Artificial Intelligence are things we have all heard a lot about. They have wide range of applications, including energy and medicine sectors. However, one of the problems is the lack of interpretability of such models, meaning we do not understand how they make their decisions. This results in a lack of trust from the user's point of view, who would typically prefer to stick with proven tools that they understand. Explainable Artificial Intelligence (XAI) is a concept that tries to close this gap. These techniques are used to make an ML model explainable, allowing the user to understand how it reaches decisions.

In recent years, the Department of Software Science has developed competence in applying XAI techniques in areas such as energy, medicine, and cybersecurity. We are looking for a prospective PhD candidate who will be working on generalizing and developing these competences. The aim is to develop novel XAI frameworks with an emphasis on goodness metrics and relations between the underlying mathematical machinery and human perception.

The successful candidate will focus on proposing, developing, implementing, and validating explainable AI models with the following general objectives:

- Propose new validation techniques for complex ML models based on explainable AI. Existing taxonomies of explainer outputs are clearly biased towards human perception, allowing little or no relation to the mathematical or algorithmic backend.
- Propose new evaluation metrics for the quality of explanations.
- Develop new standardization approach and clear definitions. AI researchers, domain experts, policymakers, and people who use machine learning all use ML models and XAI in different ways. All these people use machine learning techniques for different reasons and at different levels of abstraction.
- Explore the problem of tradeoff between accuracy and interpretability of a model.

Main responsibilities of the prospective PhD candidate:

- Conduct and disseminate research in the area of XAI and publish achieved results in Q1 journals and top-level conferences.
- Support the teaching activities of the supervisors.
- Co-supervise bachelor and master-level theses.

Requirements:

- M.Sc. degree or equivalent in Computer Science, Mathematics, or a related field.
- Clear interest in the research topic, demonstrated through a motivation letter, preferably supported by the research plan.
- Proficiency in Python, MATLAB, and R programming.
- Excellent English communication skills, both written and verbal.
- Strong analytical and research skills.
- Capacity to work independently and collaboratively in an international team.
- Preferred: Experience in programming and deep learning, showcased through GitHub projects.

Supervisors: Prof. Sven Nõmm, Prof. Juri Belikov

References:

[1] R. Machlev, L. Heistrene, M. Perl, K. Y. Levy, J. Belikov, S. Mannor, and Y. Levron. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: review, challenges and opportunities. Energy and AI, 9, 100169−13 pp. 2022. DOI: 10.1016/j.egyai.2022.100169.

[2] R. Machlev, M. Perl, J. Belikov, K. Y. Levy, and Y. Levron, Measuring Explainability and Trustworthiness of Power Quality Disturbances Classifiers Using XAI—Explainable Artificial Intelligence, IEEE Transactions on Industrial Informatics, vol. 18, no. 8, pp. 5127-5137, Aug. 2022, doi: 10.1109/TII.2021.3126111

[3] M. Meas, R. Machlev, A. Kose, A. Tepljakov, L. Loo, Y. Levron, E. Petlenkov, and J. Belikov. Explainability and transparency of classifiers for air-handling unit faults using explainable artificial intelligence (XAI). Sensors, 22(17), 2022.

[4] A. Guerra-Manzanares, S. Nõmm, and H. Bahsi. Towards the integration of a post-hoc interpretation step into the machine learning workflow for IoT botnet detection. Proceedings 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019 : 16-19 December, Boca Raton, Florida, USA. Ed. Wani, M.Arif; Khoshgoftaar, Taghi M.; Wang, Dingding; Wang, Huanjing; Seliya, Naeem (Jim). Piscataway, NJ: IEEE, 1162−1169, 2019. DOI: 10.1109/ICMLA.2019.00193.

[5] S. Nõmm. Towards the Linear Algebra Based Taxonomy of XAI Explanations. arXiv preprint arXiv:2301.13138, 2023.

To get more information or to apply online, visit https://taltech.glowbase.com/positions/751 or scan the the code on the left with your smartphone.