Universität Stuttgart
Institut für Industrielle Fertigung
und Fabrikbetrieb IFF

Master Thesis

Machine Learning: Concept Extraction Validation Benchmark

Context Machine Learning (ML) models are reaching a maturity level that allows their operational use in businesses. However, in some areas, this use is limited by their "black box" nature: the decision-making logic and potential errors of a model are not transparent, making it unsuitable for safety-critical applications or those requiring trust in the model. The field of Explainable Artificial Intelligence (XAI) addresses this by providing methods to make model behavior more interpretable. Among these, concept-based and prototype-based methods show promise in offering intuitive insights into model decisions. To truly build trust and ensure safe deployment of models, however, it is not enough for XAI methods to be intuitive — they must must also meet some key requirements. For example, the methods need to be reliable and their explanations need to be faithful to the model, while having a complexity level appropriate for human users. To ensure that these properties are met, XAI methods must be rigorously validated. Furthermore, such an evaluation should be systematic, allowing to compare most methods on the same ground. A framework for this is still largely missing in current XAI pipelines.

Topic This thesis investigates the systematic benchmarking of concept-based explanation methods for machine learning models. It adapts an existing benchmarking framework, originally developed for prototype methods, to support the evaluation of concept-based explanations. The project also includes the empirical testing of concept extraction methods, evaluating their effectiveness and reliability using diverse metrics and datasets. The work contributes toward standardizing the evaluation of XAI techniques to ensure that generated explanations are meaningful and faithful to the underlying model.

The candidate will first conduct a literature review to identify desirable properties of trustworthy explanations and corresponding evaluation criteria. This includes analyzing existing benchmarks, theoretical foundations, and practical requirements of concept-based XAI methods. Based on this, suitable evaluation metrics will be selected or developed and integrated into the benchmarking pipeline. The newly implemented metrics will then be used to evaluate a concept extraction method in various scenarios. This requires proficiency in Python and familiarity with modern ML libraries.

Scope

- Identifying and formalizing evaluation properties for concept-based XAI methods.
- Adapting an existing benchmark suite for prototype methods to accommodate concept-based explanations.
- Implementing and testing relevant evaluation metrics.
- Empirical benchmarking of a selected concept extraction method across multiple datasets and models.

Qualifications

- Solid understanding of machine learning
- Strong programming skills in Python
- Ideally, prior experience with explainability or XAI methods
- Independent, reliable, and result-oriented working style
- Good english communication skills
- Optional: Good german communication skills

Benefits

- Interesting tasks in applied research
- Intensive support during the project
- Collaboration project with University of Stuttgart IFF and RWTH Aachen University DSME

Institute of Industrial Manufacturing and Management (IFF)

The Institute for Industrial Manufacturing and Management (IFF) at the University of Stuttgart is one of the largest, most research-intensive, and traditional institutes at the university. Led by Prof. M. Huber, the "Cognitive Production Systems" group specializes in developing and applying advanced control and prediction algorithms to improve robustness and reliability in production systems. IFF offers a diverse array of practical research projects for students and industry partners and maintains strong collaborations with the Fraunhofer Institute for Manufacturing Engineering and Automation IPA to ensure the practical application of innovative research. The project is open to applicants from all institutions.

Institute for Data Science in Mechanical Engineering (DSME)

DSME is a highly dynamic, and internationally oriented institute at RWTH Aachen University affiliated with the Faculty of Mechanical Engineering and the Department of Computer Science. DSME is headed by Prof. S. Trimpe and has close ties to several national and international partners. We aim for the publication of top-quality research results - including student projects - at international conferences. The project is open to applicants from all institutions.

Contact: Do not hesitate to contact us if you are interested in this project. When applying, please include your CV, short motivation, grade transcript, and optionally other documents helpful to evaluate your background.

Helena Monke, helena.monke@ipa.fraunhofer.de
Antonia Holzapfel, holzapfel@dsme.rwth-aachen.de