Studying the soundness of SHAP explanations of machine learning predictions

Introduction. Machine learning (ML) adoption in various application domains has propelled the development of a wide range of techniques aimed at interpreting and explaining complex models. This emerging field, known as Explainable AI (XAI) aims at demystifying AI systems' decision-making processes [1]. It encompasses diverse methods designed to enhance transparency, facilitate human understanding, and foster trust in AI applications. In an era where AI-driven decisions influence critical aspects, XAI plays a crucial role in ensuring that AI models are not just powerful but also accountable and comprehensible.

In this context, SHAP (SHapley Additive exPlanations) is a prominent method for explaining machine learning predictions [2]. SHAP offers an innovative approach based on cooperative game theory, providing insights into the individual contributions of input features to model predictions. By decomposing complex predictions into understandable components, SHAP explanations empower users to grasp AI-driven decisions' rationale. Despite its widespread use and trust in practice, there is a growing need to rigorously assess the soundness and reliability of SHAP explanations. This is typically in terms of their alignment with mathematical proofs or their boolean relevance [3]. Limitations of SHAP include the absence of contextual information in the explanation process and of causality information w.r.t. contributions of features to ML model predictions.

Objective. This internship is a follow-up to an earlier project on the application of XAI methods to ML-based design of processor microarchitectures [4]. It aims to explore the broader XAI landscape. It will focus specifically on investigating the soundness of SHAP explanations, including their underlying mathematical foundations. Typical questions of interest are: to what extent can SHAP explanations be soundly proven to accurately address model behavior? How could one assess the reliability of SHAP explanations and their alignment with well-defined notions of relevancy, e.g. Boolean relevancy?

Tentative working methodology.

- Literature review: The intern will first conduct an extensive literature review to understand the current state of SHAP explanations and their application to machine learning explainability. This review will also explore existing approaches to promoting explainable AI, including symbolic and hybrid AI approaches. It will also cover existing Boolean reasoning.
- Theoretical framework: Based on the literature review, the intern will work on elaborating a theoretical framework that connects SHAP explanations with a carefully selected reasoning approach from the literature review. This framework should enable us to evaluate SHAP explanations' soundness.
- Empirical evaluation: To validate the theoretical framework, the intern will apply it to machine learning models dedicated to computer system design.

Expected outcomes.

- A theoretical framework for evaluating SHAP explanations' soundness with respect to well-established reasoning approaches.
- Insights into the alignment (or misalignment) of SHAP explanations with these approaches.
- A report detailing the methodology, findings, and recommendations for improving SHAP explanation reliability.

Qualification.

Prospective interns should possess the following qualifications:

- A solid background in symbolic AI and/or machine learning.
- Strong background in logic/Boolean reasoning, and more generally in formal/mathematical reasoning
- Proficiency in Python and relevant libraries for machine learning and data analysis.
- Familiarity with SHAP and other interpretability methods is a plus.
- Excellent analytical and problem-solving skills.
- Effective communication skills for documenting and presenting findings.

Keywords: explainable artificial intelligence, SHAP method, machine learning

Contact information.

Applications are to be sent to Abdoulaye Gamatié (Abdoulaye.Gamatie@lirmm.fr). The internship will be hosted in LIRMM (*Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier*), which is a cross-faculty research laboratory of the University of Montpellier and the National Center for Scientific Research (CNRS).

Bibliography.

1. A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai," Information Fusion, vol. 58, pp. 82–115, 2020
2. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
3. Joao Marques-Silva, Xuan Xiang Huang. "Explainability is NOT a Game". 2023 hal-04154767v2.enw
4. M. Rapp et al., "MLCAD: A survey of research in machine learning for CAD keynote paper," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 41, no. 10, pp. 3162–3181, 2021.