Abstract Interpretation for Explainable Artificial Intelligence (AI4XAI)

PhD Subject

The PhD student will work within the **ForML** project (https://www.irit.fr/ForML). The project is a collaboration between the Institut de recherche en informatique de Toulouse (IRIT) (Aurélie Hurault, Toulouse INP, and Martin Cooper, Toulouse III), Sorbonne Université (Antoine Miné, LIP6), and Inria Paris (Caterina Urban, ANTIQUE), and is led by the IRIT. ForML aims to develop new static analysis techniques based on abstract interpretation and new model checking techniques based on counterexample-guided abstraction refinement to verify robustness, fairness, and explainability properties of machine-learned software. The PhD student will be based in Paris and will be co-supervised by Caterina Urban and Antoine Miné. Research visits to Toulouse and collaborations with the IRIT members of the project are also expected.

Specifically, the PhD student will work on the *explainability* axis of the ForML project. A previous work by the IRIT members of the project [Marques-Silva et al. 2021] describes novel algorithms for computing formal explanations of (black-box) monotonic classifiers. In essence, these algorithms identify minimal subsets of the input features that are sufficient for the prediction (AXp) or for changing the prediction (CXp). Formally verified implementations of these algorithms are extracted from Coq proofs of their correctness [Hurault et Marques-Silva, 2023]. With these previous works as a starting point, we envision a number of avenues that can be investigated.

First, instead of assuming that classifiers are monotonic, we would require a white-box access to the classifier to design a static analysis by abstract interpretation to formally verify this hypothesis. A starting point for this analysis could be the sound proof system with judgments specifying whether a program is monotonic that has been introduced by ANTIQUE in recent work [Campion et al. 2024].

We would also like to investigate alternative algorithms to compute explanations that leverage numeric and symbolic abstraction for machine learning software [Urban and Miné 2021]. Based on the above mentioned recent work by ANTIQUE [Campion et al. 2024], we expect abstractions based on numerical value ranges to be sound as well as complete for monotonic (or stable) classifiers.

On the other hand, the hypothesis of monotonicity of the classifiers is rather strong. We would like to investigate possible avenues to relax such requirement, potentially obtaining more approximate explanations. Specifically, this will involve designing and implementing abstract interpretation-based forward analysis algorithms for computing (approximate but sound) AXp and CXp explanations of classifiers that are not necessarily monotonic.

These abstract interpretation-based algorithms would also allow generalizing explanation beyond a single classified entry to a neighborhood around these entry in the input space of the classifier.

Finally, another venue worth investigating comes from a recent work of ANTIQUE with the members of the project at Sorbonne Université [Moussaoui Remil et al. 2024], which proposed a backward analysis based on abstract interpretation for determining the sets of program variables that an attacker can control to ensure a certain program outcome. These sets can be seen as non-minimal AXp explanations of the program outcome. The analysis builds upon an inference of sufficient preconditions for Computation Tree Logic (CTL) program properties that was previously developed by ANTIQUE [Urban et al. 2018]. It would be interesting to port this work to the context of machine learning classifiers and formally establishing relationships with (approximate) AXp and CXp explanations [Marques-Silva et al. 2021]. Combinations of forward and backward static analysis would also be interesting to explore.

We expect these static analysis methods to be implemented and thoroughly evaluated experimentally. Existing infrastructure and prototypes developed by ANTIQUE in Python and

Ocaml can be built upon, if desired. We will leverage benchmarks from previous work done by ANTIQUE and IRIT [Urban et al. 2020, Hurault et Marques-Silva, 2023] for evaluation. Certified implementations of the developed abstract interpretation-based algorithms will leverage the proof assistant Coq.

[Marques-Silva et al. 2021] João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, Nina Narodytska. Explanations for Monotonic Classifiers (ICML 2021)
[Hurault et Marques-Silva, 2023] Aurélie Hurault, João Marques-Silva. Certified Logic-Based Explainable AI - The Case of Monotonic Classifiers (TAP 2023)
[Campion et al. 2024] Marco Campion, Mila Dalla Preda, Roberto Giacobazzi, Caterina Urban. Monotonicity and the Precision of Program Analysis (POPL 2024)
[Urban and Miné 2021] Caterina Urban, Antoine Miné. A Review of Formal Methods applied to Machine Learning (https://arxiv.org/abs/2104.02466, 2021)
[Moussaoui Remil et al. 2024] Naïm Moussaoui Remil, Caterina Urban, Antoine Miné. Automatic Detection of Vulnerable Variables for CTL Properties of Programs (LPAR 2024)
[Urban et al. 2018] Caterina Urban, Samuel Ueltschi, Peter Müller. Abstract Interpretation of CTL Propertie (SAS 2018)
[Urban et al. 2020] Caterina Urban, Maria Christakis, Valentin Wüstholz, Fuyuan Zhang. Perfectly Parallel Fairness Certification of Neural Networks (OOPSLA 2020)