Offer #2025-08776

PhD Position F/M Abstract Interpretation for Explainable Artificial Intelligence (AI4XAI)

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

Context

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 supervised by Caterina Urban. Research visits to Toulouse and collaborations with the IRIT members of the project are also expected.

Assignment

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 su?cient 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]. These previous works will be the starting point for the PhD thesis.

[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)

Main activities

A number of avenues will be investigated during the PhD.

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].

Second, we aim to extend the definitions of AXp and CXp to operate over the *latent space* of a model, e.g., on activation patterns in hidden layers of a neural network [Geng et al. 2023]. We aim to define algorithms to compute the minimal latent space explanations su?cient for preserving or altering a model prediction. Notably, latent space explanations will allow us to reason about non-convex regions over the model input space, generalizing explanation beyond a single input to a neighborhood in the input space. We also plan to infer relational explanations, e.g., establishing dependencies between neuron activations. We will leverage (combinations of) existing numeric and symbolic abstractions for machine learning software [Urban and Miné 2021] to practically compute such explanations. We will target ReLU-activated neural networks to start with, with the objective to later generalize the results to target graph neural networks —a particularly good fit since the size of their input space is not fixed —and language models — defining latent explanations over token embeddings, attention patterns, or context windows.

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 su?cient 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.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

General Information

- **Theme/Domain :** Proofs and Verification
  Software engineering (BAP E)
- **Town/city :** Paris
- **Inria Center :** Centre Inria de Paris
- **Starting date :** 2025-10-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2025-05-27

Contacts

- **Inria Team :** ANTIQUE
- **PhD Supervisor :**
  Urban Caterina / caterina.urban@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.