Formal Explanations for Trustworthy Artificial Intelligence

Formal Explanations for Trustworthy Artificial Intelligence

Keywords: Explainability, Neural Network, Why3, CAISAR, PyRAT

Institution

The French Alternative Energies and Atomic Energy Commission (CEA) is a key player in research, development, and innovation. Drawing on the widely acknowledged expertise gained by its 16,000 staff spanned over 9 research centers with a budget of 4.1 billion Euros, CEA actively participates in more than 400 European collaborative projects with numerous academic (notably as a member of Paris-Saclay University) and industrial partners. Within the CEA Technological Research Division, the CEA List institute addresses the challenges coming from smart digital systems.

Among other activities, CEA List's Software Safety and Security Laboratory (LSL) research teams design and implement automated analysis in order to make software systems more trustworthy, to exhaustively detect their vulnerabilities, to guarantee conformity to their specifications, and to accelerate their certification. The lab recently extended its activities on the topic of AI trustworthiness and gave birth to a new research group: AISER (Artificial Intelligence Safety, Explainability and Robustness).

Scientific context

Systems incorporating artificial intelligence (AI) components have a considerable influence on society and the physical infrastructures on which they are based. The role played by these systems means that there is an unprecedented need for audit and trust. However, the scale of the data processed by these systems (sensory and temporal data) means that there is an unprecedented need for audit and trust. However, the scale of the data processed by these systems (sensory and temporal data) and the computer complexity of these programmes complicate their analysis. In addition, existing explicability techniques are

difficult to compare and their unreliability limits their applicability to realistic systems.

A recent line of work consist on formulating explanations for AI models as Satisfaction Modulo Theory queries (Marques-Silva and Ignatiev 2022), using feature importance. They thus provide sound basis for explanations as they directly operate on the model, compared to linear approximation-based approaches like LIME (Ribeiro, Singh, and Guestrin 2016). The scalability of this approach on realistic use cases is yet to be evaluated. Other approaches (Pal et al. 2022) are able to soundly bound the validity domain of overapproximations.

Internship

The aim of this internship is to explore the scalability of formal explainable AI techniques. In particular, the internship's goal is to identify the limits of existing techniques and propose new approaches, for instance inspired by (Bassan and Katz 2023). To do so, the intern will leverage two tools developped by the team: the CAISAR platform for formulating verification queries, and the PyRAT analyser to solve them.

The broad internship goals are:

- familiarization with the state-of-the-art on explainable AI ((Molnar 2022))
- implementation of contrastive explanation methods in the CAISAR platform
- benchmark the solution on deep neural networks on selected datasets
- if time allows, use the PyRAT analyzer to formulate a validity bound for overapproximation-based explanations

Qualifications

The candidate will work at the crossroads of formal verification and artificial intelligence. As it is not realistic to be expert in both fields, we encourage candidates that do not meet the full qualification requirements to apply nonetheless. We strive to provide an inclusive and enjoyable workplace. We are aware of discriminations based on gender (especially prevalent on our fields), race or disability, we are doing our best to fight them.

- **Minimal**
  - Master student or equivalent (2nd/3rd engineering school year) in computer science
  - knowledge of OCaml
  - ability to work in a team, some knowledge of version control
- **Preferred**
  - notions of AI and neural networks

  - knowledge of formal verification in general, of SMT solving in particular
  - knowledge of Why3

Characteristics

The candidate will be monitored by two research engineers of the team.

- **Duration:** 5 to 6 months from early 2023
- **Location:** CEA Nano-INNOV, Paris-Saclay Campus, France
- **Compensation:**
  - €700 to €1300 monthly stipend (determined by CEA compensation grids)
  - maximum €229 housing and travel expense monthly allowance (in case a relocation is needed)
  - CEA buses in Paris region and 75% refund of transit pass
  - subsidized lunches
  - 3 days of remote work

Application

If you are interested in this internship, please send to the contact persons an application containing:

- your resume;
- a cover letter indicating how your curriculum and experience match the qualifications expected and how you would plan to contribute to the project;
- your bachelor and master 1 transcripts;
- the contact details of two persons (at least one academic) who can be contacted to provide references.

Applications are welcomed until the position is filled. Please note that the administrative processing may take up to 3 months.

Contact persons

For further information or details about the internship before applying, please contact:

- Julien Girard-Satabin (julien.girard2@cea.fr) (also available on LinkedIn)
- Zakaria Chihani (zakaria.chihani@cea.fr)

References

Bassan, Shahaf, and Guy Katz. 2023. "Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks." https://arxiv.org/abs/2210.13915.

Marques-Silva, Joao, and Alexey Ignatiev. 2022. "Delivering Trustworthy AI Through Formal XAI." *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (11): 12342–50. https://doi.org/10.1609/aaai.v36i11.21499.

Molnar, Christoph. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. https://christophm.github.io/interpretable-ml-book.

Pal, Abhinandan, Francesco Ranzato, Caterina Urban, and Marco Zanella. 2022. "Abstract Interpretation-Based Feature Importance for SVMs." https://arxiv.org/abs/2210.12456.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. ""Why Should i Trust You?": Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.