Automata learning from Recurrent Networks via Clustering

Keywords: Artificial Inelligence, Classification, Neural Networks, Abstract Interpretation

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

Through the recent developments of AI, their use has become even more widespread, both in domestic and industrial settings. Nevertheless, studies are flourishing showing the dangers that such AI can bring, whether in terms of safety, privacy or fairness. We can for example cite the case of adversarial attacks, where small perturbations invisible to naked eyes, can drastically change the output of an AI model. To face these dangers, works and tools are constantly emerging to increase the trust one can have in AI systems.

One of the tool developed at CEA in the AISER team is CaiSAR, an open-source platform that focuses on the characterization of AI systems' Robustness and Safety. In order to ensure the safety of an AI system, this platform can call several provers including PyRAT, a Python tool based on Abstract Interpretation techniques also developed at CEA in the AISER team. Those two tools are under active development, as new features are added to improve their accuracy and the expressivity of their specification language. CaiSAR is written in Ocaml, PyRAT is written in Python3.

When assessing the Safety of an AI classifier, it is sometimes useful to qualify the judgement given by our tools by allowing the AI system to be abstract safe, instead of either Safe or Unsafe, when we can prove that the system's misclassifications respect the formal hierachy. For example, with a hierarchy that distinguish "animal" classes and "vehicule" classes, a system that misclassifies a dog as "a cat" but not as "a car" could be considered to be abstract safe.

Internship

The main objective of this internship is to introduce the notion of Hierarchical Classification in CAISAR, and use it to verify AI systems with PyRAT. The main steps are:

- to define a file format to represent the hierarchy of classes
- to modify how the provers are called in CAISAR and their outputs to determine whether the AI system is abstract safe
- to experiment on at least one case study

This work will have contributions to the field of automata learning and to neural networks verification. The internship will likely conclude by publishing a paper (workshop, conference) depending on the quality of the work to be carried.

Qualifications

The candidate will work at the crossroads of formal methods and machine learning. 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 or applied mathematics
  - knowledge of OCaml
  - ability to work in a team
  - fluent in French or English
- **Preferred**
  - knowledge of Python
  - notions of abstract interpretation and/or formal methods for computer science

Characteristics

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

- **Duration:** 3 to 6 months from early 2026
- **Location:** CEA Nano-INNOV, Paris-Saclay Campus, France
- **Compensation:**
  - €1300 if you are in M1/second year of engineering school, €1400 if you are in M2/third year of engineering school
  - 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
  - 2 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;

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:

- Tristan Le Gall (tristan.le-gall@cea.fr)