Noise symbols reduction and explainability for PyRAT's neural network analysis

Keywords: PyRAT, Neural Network, Abstract Interpretation, Zonotopes

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, even in 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 adversarial attacks, small perturbations invisible to naked eyes which can drastically change the output of our AI. 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 PyRAT, a Python tool based on abstract interpretation techniques to assess the robustness of a neural network in face of perturbations. It propagates abstract domains representing all possible inputs through the network to find all reachable outputs and thus decide on their safety or not. One of the abstract domains used in PyRAT is based on Zonotopes and introduces something called noise symbols at each non linearity. Unfortunately such non-linearity are numerous especially in big networks. As such, it is necessary to find heuristics in order to reduce (in a safe way) the number of noise symbols during an analysis.

Internship

The aim of this internship is to envisioned and implement noise symbols reduction techniques in PyRAT which would soundly overapproximate the analysis. Approaches such as the one proposed here: https://arxiv.org/pdf/2305.01932.pdf will be tested. The advantages and drawbacks of each heuristics will be assessed and benchmarked on various datasets.

An additional aspect of the internship will be to link these noise symbols reduction techniques to the explainability of the network. Indeed, noise symbols can be good indicators of which pixels have the most influence on the network. Reducing and merging certain noise symbols may impact the quality of these explanations. On the other hand, merging noise symbols together can also help explainability by showing their combined influence on the outputs (i.e. groups of input pixels might influence the network more than individual pixels). The second part of this stage will thus be dedicated to explore this link with explainability.

Qualifications

- **Minimal**
  - Master student or 2nd or 3rd year of engineering school
  - knowledge of Python
  - notions of AI and neural networks, and of AI frameworks (TF, Keras, Pytorch, . . .)
  - ability to work in a team
- **Preferred**
  - some knowledge of abstract interpretation or formal method

Characteristics

- **Duration:** 5 to 6 months from early 2024
- **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

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:

- Augustin Lemesle (augustin.lemesle@cea.fr)
- Zakaria Chihani (zakaria.chihani@cea.fr)