Counter example and adversarial attacks for neural network verification

Keywords: Verification, Rotation, biasfield, 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

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, among other the security and robustness of a neural network in face of these adversarial perturbations. Nevertheless, before proving the robustness, PyRAT tries to falsify the property by using adversarial attacks in order to avoid extensive analysis when simple counter examples exist. This aversarial search relies on classical methods such as FGSM or PGD attacks but lack the ability to adapt itself w.r.t. the neural network specificities or to capitalise on the analysis to reduce the search scope.

Internship

The goal of this internship is to improve the attack capacities of PyRAT to detect weakness in the model either by increasing the strength of the adversarial attacks used or by using additional information gathered from the analysis of PyRAT. Additionally, as finding attacks can be time consuming, compromises will be sought to find the best time ratio for attacks and for verification.

Multiple axis will be tackled during this internship:

- Implementing stronger and more adversarial attacks in PyRAT
- In case of multi step analysis, incorporate knowledge from previous analysis in the attacks search
- Automate and optimise the attack research in time or performance

The intern will also have the occasion to partake in the next edition of the VNNCOMP, the international neural network verification competition, in order to benchmark the new attacks capabilities of PyRAT.

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 Python
  - ability to work in a team, some knowledge of version control
- **Preferred**
  - notions of AI and neural networks
  - notions of Computer Vision

Characteristics

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

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

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)