Semantic perturbations for Neural Network verifications

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.

Such works can, for example, rely on formal methods to give mathematical guarantees on the safety of an AI. However, they only consider simple robustness approaches against classical adversarial attacks. Using an infinity norm, they represent various kind of perturbation on the AI input. Nevertheless, they often are too generic and can fail to capture more precise perturbations. Recent publications (such as this one, or that one, have suggested to limit the perturbation to more meaningful things, such as rotation, brightness, saturation or biasfield.

Internship

The goal of this internship is implement those approaches so that they can be used in a follow up safety verification (with our lab tools, the CAISAR platform and the PyRAT analyser. This work will build on a small existing library which already includes simple perturbations (brightness, contrast, saturation and hue).

The internship goals are:

- Implementing the rotation and the bias field perturbation in the tools
- Directly modify the ONNX networks to include the perturbations
- Benchmark these perturbations on real dataset with formal verification through PyRAT
- If time allows, some retraining of network can be envisioned to robustify networks against these perturbations.

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:**
  - €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;

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)