Boosting neural network analysis with reinforcement learning and GNN

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. This problem has been proven NP-complete and as often with NP-problems there is a trade-off between precision and cost (computation time, memory etc) induced by the different neural network verification techniques. A simple technique to improve precision is to divide the problem into many sub-problems recursively until all sub-problems are solved. Nevertheless, naive approaches tend to simply exponentially increase the number of sub-problems but more careful approaches can drastically reduce their number. Exploring these new approaches will be the subject of the internship.

Internship

This internship builds on already existing work in our lab on the subject. Firstly, the paper ReCIPH which provides incentives on upon which axis to divide the current problems. Then it also builds on a previous work in determining the number of sub-problems that should be created at each step. The internship will more specifically improve on this second question and the mechanism called boosting (i.e. splitting in more than 2 at each step to boost the analysis).

This previous work has aimed to determine the best number of division by using an external neural network during our analysis. This was tackled with classical NN but also with a reinforcement learning model. This model, while unfinished, brought some results but more importantly some interesting perspectives. Further studies will be led to see how to improve on it or how to use different types of models such as RNN or GNN to further improve the performances.

As such this internship will consist in: - Analysing the previous results with reinforcement learning to get statistics and see where it performs well and where it does not. - Exploring ways to improve the reinforcement learning model to increase its performances. - Exploring new architectures to answer the question of the best number of division for a given problem.

Finally, all of this will be implemented in our tool PyRAT to see how the analysis can be guided along the problem and sub-problems and solve more efficiently.

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, ...)
  - notions of reinforcement learning
  - notions of Graph Neural Network and/or Recurrent Neural Network
  - 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)