Implementing new types of neural networks for analysis in PyRAT

Keywords: PyRAT, Neural Network, Abstract Interpretation, Recurent Neural Network, Transformers

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 a large number of 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. Recently the field of activity of the laboratory has been extended to artificial intelligence safety and security verification. In particular, PyRAT and CAISAR are two tools developed in that context to verify safety properties on neural networks using abstract interpretation techniques.

Objectives

As mentioned, PyRAT leverages the principles of abstract interpretation to propagate abstract domains (input) through abstract operations representing the layers of the neural network and in order to assess the reachable states (output). In comparison to classical software verification, PyRAT works directly on the weights, biases, and parameters of a neural network model thus making PyRAT lighter and faster to use for neural network analyses. PyRAT is developed in Python as it is a widely used language for neural network frameworks such as Keras, Pytorch or Tensorflow. As of now, the primary use of PyRAT is to assess robustness w.r.t. some perturbation around inputs on small neural networks. However, on larger neural networks or on larger inputs a simple pass with PyRAT will lack precision.

To compensate the loss of precision in such cases, we introduced a domain splitting approach inside PyRAT. This iterative approach splits the input domain into smaller ones until PyRAT can prove the property on them or find a counterexample. This also allows to use PyRAT to also prove some safety properties, i.e. bigger intervals of input on certain neural networks.

In this internship, we aim to focus on exploring and understanding new types of neural network that are not currently supported by PyRAT such as Recurrent Neural Networks or Transformers. The former are of especially great interest as they are often found in use cases focusing on time series by our industrial partners. The layers and function of the network should be already available for the most part in PyRAT but the crux of the problem will be to modify the propagation in PyRAT to suit this kind of networks. https://arxiv.org/pdf/2005.13300.pdf, https://arxiv.org/abs/2007.10135, https://arxiv.org/pdf/1905.07387.pdf

Qualifications

- **Minimal**
- Master student or 2nd or 3rd year of engineering school
- knowledge of Python
- some knowledge 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 2022
- **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)
- Michele Alberti michele.alberti@cea.fr
- Zakaria Chihani (zakaria.chihani@cea.fr)