AI Safety: Measuring uncertainty in ReLU neural networks by studying activation patterns

Keywords: Neural Network, ReLU, Out-of-Distribution detection

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.

Objectives

In recent years, neural networks (NNs) have become the backbone of most decision-making systems processing high dimensional inputs (such as images, videos, sounds and texts), mostly due to their ability to capture complex decisions boundaries through the use of linear operations and non-linear activation functions (sigmoid, tanh, ReLU). In particular, ReLUs (Rectified Linear Units) have contributed to the development of deep neural networks by reducing the likelyhood of an effect known as « vanishing gradient ». However, one of the main drawbacks of modern neural network relates to the difficulty of evaluating the level of confidence associated with their decision, which raises multiples questions regarding their adoption in critical systems (such as autonomous driving): in particular, most NN-based systems lack the ability to detect Out-of-Distribution (OoD) inputs (i.e. inputs which largely differs from examples in the training set) and may provide a decision with a false confidence measure. Evaluating the confidence of the system is a task known as network calibration, for which an extensive literature is already available (see the thorough review of [1]). In this internship, we want to study the activation patterns of ReLU gates across the training set in order to build a set of known behaviors that we hope will help detect OoD and ultimately provide a measure of uncertainty[2]. Indeed, a previous work on ReLU networks [3] has shown that far from being completely random, the number of possible activation patterns in such networks is actually quite limited. The goal of the internship is therefore to:

- Implement a method for extracting ReLU activation patterns from a given neural network
- Extract a set of common patterns from the training set and establish a measure of uncertainty of the decision based on the similarity with previously encountered examples
- Study the ability of this measure to detect Out-of-distribution inputs
- [Optional] Study the possibility of predicting activation patterns of the last layers of the network based on the behavior of the first layers.

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

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:

- Romain Xu-Darme (romain.xu-darme@cea.fr)
- Fabio-Alejandro Arnez Yagualca (fabio.arnez@cea.fr)
- Zakaria Chihani (zakaria.chihani@cea.fr)