New Neural Network Architecture for Verifiably Robust and Fair AI – Application to LLMs

The School of Physical and Mathematical Sciences (SPMS) and the School of Computer Science and Engineering (SCSE) at Nanyang Technological University (NTU) in Singapore, is seeking highly motivated candidates for several Ph.D. student positions in the areas of machine learning and/or cryptography. Interested applicants should send their detailed CVs to Prof. Thomas Peyrin (thomas.peyrin@ntu.edu.sg) preferably as soon as possible and before end of March 2024 (for August 2024 intake). The scholarship provides support for up to 4 years of PhD studies including tuition fees and a monthly stipend.

Candidates are expected to hold a Bachelor or a Master's degree in Computer Science or Mathematics, and to have a strong experience with machine learning and/or cryptography. Exposure to computer security is a plus.

More general information about graduate admissions at NTU can be found here: https://www.ntu.edu.sg/admissions/graduate/radmissionguide

Research context.

We have developed in NTU a new neural network architecture, so-called “Truth Table Deep Convolutional Neural Networks” or TTnets (https://arxiv.org/abs/2208.08609), that is very promising for many real-life scenarios. They can be seen as compressed neural networks (NN) based on small lookup tables and they allow the transformation of the inference into a small system of SAT equations or into a compact collection of small circuits after learning. TTnet are very simple NNs (easier to interpret and work with, essentially very generalized decision trees), while still providing a good accuracy and scaling to large datasets. Thus, they are very well suited for example in constrained environments such as embedded systems, mobile phones, etc., and present other interesting properties like formal complete verifiability. They are, in addition, the first global and exact explainable NNs that scale to large datasets, such as CIFAR-10 or ImageNET.

PhD research topic.

The goal of this PhD proposal is to exploit and extend the capabilities of TTnets with regards to AI verification, with applications to robustness and fairness. Deployment of verifiable AI models is crucial for many industries, such as military, healthcare, critical systems, etc. In addition, it is expected that strong government regulations towards increased AI explainability and verification will be imposed worldwide in the coming years. Providing a fast and complete verifiable NN remains an open problem as of today. Yet, with their ability to be easily transformed into a system of SAT equations, TTnets already improved state-of-the-art results for verifiable accuracy against adversarial attacks on simple/medium tabular or image datasets. More complex datasets or other types of datasets (graphs, text) remain a challenge today. In addition, despite their importance, robustness or fairness of Large Language Models (LLMs) on given tasks remains to be studied.

This verifiably robust/fair AI research can be conducted from different and non-exclusive directions:

- new robustness and fairness training methods, dedicated to TTnet.
- improving the inner building blocks of TTnet, but also considering the training and benchmarking of fundamental variations in the TTnet global architecture and parameters.
- participation to the annual worldwide VNN-Comp (Verification of Neural Networks Competition) on the verification of neural networks. Study the possibility for a global robustness testing framework ().
- scaling to larger verification tasks: formal proofs for watermarking, verification for ImageNet, etc.
- application to LLMs: thanks to their good zero-shot or few-shot performances, LLMs are increasingly used for very specialized tasks. Yet, it is extremely hard for the user to get any knowledge or guarantee regarding the robustness and fairness of the LLM's inferences. The student will apply a distillation (with good matching rate) of the LLM capabilities on a given task into a TTnet model. From the TTnet model, the candidate will be able to study and improve the robustness and fairness of the extracted model.

Of course, the candidate will have the freedom to explore other directions if willing to do so.

About NTU.

Nanyang Technological University (NTU) is a research-intensive university with globally acknowledged strengths in science and engineering. The university provides a high-quality global education to more than 33,500 undergraduate and postgraduate students. Hailing from more than 66 countries, the university's 3,600-strong teaching and research staff bring dynamic international perspectives and years of solid industry experience. NTU is ranked 19th in the world (QS World University Rankings 2022) and 1st among the world's best young universities (QS World University Rankings and Times Higher Education World University Rankings 2023). Notably, its computer science and engineering schools have been ranked in top 15 in the world (QS World University Rankings and Times Higher Education World University Rankings 2023).

Contact:

Prof. Thomas Peyrin
Nanyang Technological University
School of Physical & Mathematical Sciences / School of Computer Science and Engineering
Email: thomas.peyrin@ntu.edu.sg