Specification from Generative Models for Machine Learning Verification

Keywords: Formal Verification, Machine Learning, Generative Models, Why3

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. Recently the field of activity of the laboratory has been extended to artificial intelligence safety and security verification.

Scientific context

Perceptive programs are to become more pervasive in a vast amount of fields, among those being healthcare, autonomous transportation or legal assistance. Trusting those programs as they will operate on multiple aspects of our lives is thus paramount, both for public acceptance and an easier development process. Among others, formal verification and testing are a set of techniques that help to build trust.

The specificities of perceptive programs makes them difficult to verify. Specifically, they lack a *formal specification of the inputs*. In the general setting of classification or detection, the only specification is a mapping between the input and output on a limited dataset (labelling). For instance, given a program that must detect pedestrians, there is no formal definition of what is an image with a pedestrian, apart from examples from a dataset. Since it is impossible to have a unambiguous definition on the inputs, it is furthermore impossible to formulate properties to verify. The high dimensionality of their inputs and the use of specific mathematical operations (activation functions) are also a problem.

Existing work on specifying perceptive programs: robustness against perturbation (Madry et al. 2017), differential privacy (Abadi et al. 2016), using various approaches such as reachability analysis (Müller et al. 2021), (Wang et al. 2021) and SMT solving (Katz et al. 2019).

Simulators can be used to constraint the inputs of a machine learning program to achieve some specification (Girard-Satabin et al. 2020). (Xie, Kersting, and Neider 2022) further refines this approach by including neural networks output into a first-order-logic reasoning procedure. Finally, program synthesis and generative models can be combined to create correct-by-construction programs (Fijalkow and Gupta 2019), (Toledo et al. 2021).

Internship goal

The student will work with CAISAR, an AI verification platform currently developped in the lab. The goal is to formalize the use of neural network as a *partial specification* for program verification.

This internship can be described by the following goals:

- familiarization with the CAISAR platform, as well as with the state-of-the-art on program synthesis and neurosymbolic verification
- formalize how to include prototypal neural networks as logical objects into CAISAR's reasoning mechanism
- implement instances of such prototypal neural networks, and test those instances

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.

- **Minimal**
  - Master student or equivalent (2nd/3rd engineering school year) in computer science
  - notions of AI and neural networks
  - ability to work in a team, some knowledge of version control
- **Preferred**
  - knowledge of OCaml
  - knowledge of formal verification
  - knowledge of Why3

Characteristics

The candidate will be monitored by two research engineers of the team.

- **Duration:** 5 to 6 months from early 2023
- **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 with daily bonus

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:

- Julien Girard-Satabin (julien.girard2@cea.fr) (also available on LinkedIn)
- Zakaria Chihani (zakaria.chihani@cea.fr)

Bibliography

Abadi, Martin, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. "Deep Learning with Differential Privacy." In *23rd ACM Conference on Computer and Communications Security (ACM CCS)*, 308–18.

Fijalkow, Nathanaël, and Mohit Kumar Gupta. 2019. "Verification of Neural Networks: Specifying Global Robustness Using Generative Models," October. https://arxiv.org/abs/1910.05018.

Girard-Satabin, Julien, Guillaume Charpiat, Zakaria Chihani, and Marc Schoenauer. 2020. "CAMUS: A Framework to Build Formal Specifications for Deep Perception Systems Using Simulators." In *ECAI 2020 - 24th European Conference on Artificial Intelligence*. Santiago de Compostela, Spain. https://hal.inria.fr/hal-02440520.

Katz, Guy, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, et al. 2019. "The Marabou Framework for Verification and Analysis of Deep Neural Networks." In *Computer Aided Verification*, edited by Isil Dillig and Serdar Tasiran, 11561:443–52. Cham: Springer International Publishing.

Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. "Towards Deep Learning Models Resistant to Adversarial Attacks." *arXiv:1706.06083 [Cs, Stat]*, June. https://arxiv.org/abs/1706.06083.

Müller, Christoph, François Serre, Gagandeep Singh, Markus Püschel, and Martin Vechev. 2021. "Scaling Polyhedral Neural Network Verification on GPUs." *Proceedings of Machine Learning and Systems* 3.

Toledo, Felipe, David Shriver, Sebastian Elbaum, and Matthew B Dwyer. 2021. "Distribution Models for Falsification and Verification of DNNs." In *The 36th IEEE/ACM International Conference on Automated Software Engineering*, 13.

Wang, Shiqi, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. 2021. "Beta-CROWN: Efficient Bound Propagation with Per-neuron Split Constraints for Complete and Incomplete Neural Network Robustness Verification." October 31, 2021. http://arxiv.org/abs/2103.06624.

Xie, Xuan, Kristian Kersting, and Daniel Neider. 2022. "Neuro-Symbolic Verification of Deep Neural Networks." In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, edited by Lud De Raedt, 3622–28. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2022/503.