Internship position: Confidence-based safety properties in CAISAR

Keywords: Neural Networks, Confidence-based Safety Properties, Formal Verification

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, the use of models produced by machine learning has become widespread, even in industrial settings. However, studies are flourishing showing the dangers that such models can bring, in terms of safety, privacy or even 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 and improve trust in AI, research works are constantly emerging, resulting into several tools like ERAN (Gehr et al. 2018), CROWN (Wang et al. 2021), Reluplex (Katz et al. 2017) and its successor Marabou (Katz et al. 2019).

In the past couple of years, the AISER team has developed CAISAR (Alberti et al. 2022), an open-source platform that focuses on the specification and verification of AI systems' robustness and safety. In particular, CAISAR provides a language for formally specifying

AI properties, and integrates various state-of-the-art tools for establishing their validity in a guaranteed way.

Internship objectives

The concept of a *confidence-based safety property* has been recently introduced by Athavale et al. (Athavale et al. 2024) to recast robustness and fairness properties in terms of the confidence score with which a neural network generates its outcomes.

The main objective of this internship will be to investigate, design, and implement a support for confidence-based safety properties in the CAISAR, the AISER's open-source platform for characterizing AI systems' safety and robustness.

The broad internship goals are:

- familiarization with the state-of-the-art on formal approaches to properties for AI safety (Casadio et al. 2022)
- familiarization with the work on confidence-based safety properties (Athavale et al. 2024)
- getting started with the CAISAR platform
- design and implementation of confidence-based safety properties in CAISAR
- identification and evaluation against benchmarks

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. We strive to provide an inclusive and enjoyable workplace. We are aware of discriminations based on gender (especially prevalent on our fields), race or disability, we are doing our best to fight them. One of our team member is formally trained against psychological harassment and sexual abuse.

Minimal

- master student or equivalent (2nd/3rd engineering school year) in computer science
- some knowledge of the OCaml programming language (or any other functional language)
- some knowledge of the Python programming language
- ability to work in a team

Preferred

- notions in formal methods (*i.e.* abstract interpretation, SAT/SMT solving, etc.)
- notions in machine learning and, in particular, neural networks

Characteristics

The candidate will be monitored by two research engineers of the team, with at least one weekly meeting.

Duration

5 to 6 months from early 2025

Locations

CEA Nano-INNOV, Paris-Saclay Campus, France

Compensation and funding

- 1400€ monthly stipend
- possible allowance for housing and travel expense (in case a relocation is needed) with a limit of €229 per month
- 75% refund of transit pass
- subsidized lunches
- 3 days of remote work per week

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:

- Michele Alberti (michele.alberti@cea.fr)
- François Bobot (françois.bobot@cea.fr)

References

Alberti, Michele, François Bobot, Zakaria Chihani, Julien Girard-Satabin, and Augustin Lemesle. 2022. "CAISAR: A platform for Characterizing Artificial Intelligence Safety and Robustness." In *AISafety*. CEUR-Workshop Proceedings. Vienne, Austria. https://hal.archives-ouvertes.fr/hal-03687211.

Athavale, Anagha, Ezio Bartocci, Maria Christakis, Matteo Maffei, Dejan Nickovic, and Georg Weissenbacher. 2024. "Verifying Global Two-Safety Properties in Neural Networks with Confidence." In *Computer Aided Verification - 36th International Conference, CAV 2024, Montreal, QC, Canada, July 24-27, 2024, Proceedings, Part II*, edited by Arie Gurfinkel and Vijay Ganesh, 14682:329–51. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/978-3-031-65630-9/_17.

Casadio, Marco, Ekaterina Komendantskaya, Matthew L Daggitt, Wen Kokke, Guy Katz, Guy Amir, and Idan Refaeli. 2022. "Neural Network Robustness as a Verification Property: A Principled Case Study." In *International Conference on Computer Aided Verification*, 219–31. Springer.

Gehr, Timon, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. 2018. "AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation." In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE. https://doi.org/10.1109/sp.2018.00058.

Katz, Guy, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. 2017. "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks." *arXiv Preprint arXiv:1702.01135*. https://doi.org/10.1007/978-3-319-63387-9_5.

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. https://doi.org/10.1007/978-3-030-25540-4_26.

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.