Master internship proposal: quantified reachability properties for provably explainable AI

October 2, 2024

Keywords: formal methods, neural networks, reachability analysis, explainability

Institution: LIX, Turing building, campus of Ecole Polytechnique

Advisors: Eric Goubault and Sylvie Putot ({goubault,putot}@lix.polytechnique.fr)

General presentation: Artificial Intelligence is now embedded into a number of everyday life applications. More and more, we depend on neural networks even in critical situations, such as control and motion planning for autonomous cars, and it is of primary importance to be able to verify their correct behavior.

Abstraction-based safety verification for neural networks has received considerable attention recently, with in particular reachability analysis of neural networks using polyhedric abstractions such as [9, 10], with application in particular to local robustness analysis. The context of this work is to develop sound abstractions for addressing more general robustness properties. More specifically, the objective is to propose provably correct explanations of neural networks behavior while most existing techniques are heuristic [8].

Objectives: The tractable inner and outer-approximations of ranges of functions proposed in [3] are a building block for proving very general quantified reachability problems [4]. These constitute a basis from which the student is expected to design new set-based methods to tackle properties of neural networks that can be expressed as quantified reachability problems. The objectives are to identify some properties of interest that can be expressed in this framework, and design and experiment reachability analyzes inspired from the techniques of [3, 4] to rigorously assess these properties. As a starting point, we can explore fairness properties, in the line of e.g. [6, 2, 7]. Another axis consists in providing rigorous explainability properties of neural networks such as abductive explanation [5, 1]: a minimum subset of input features, which by themselves determine the classification produced by the DNN. We can also imagine using such approaches to guide the sparsification of neural networks.

Context: LIX (Laboratoire d'Informatique de l'Ecole Polytechnique) is a joint research unit with two supervisory institutions, École Polytechnique, a member of the Institut Polytechnique de Paris (cluster of universities composed of Ecole Polytechnique, Télécom Paris, ENSTA Paris, Télécom Sud Paris, ENSAE), and the Centre National de la Recherche Scientifique (CNRS), and one partner, Inria Saclay, with shared buildings and mixed teams.

LIX is organized in four poles: "Computer Mathematics", "Data Analytics and Machine Learning", "Efficient and Secure Communications", "Modeling, Simulation and Learning" and "Proofs and Algorithms". The intern will be part of the Cosynus team in the "Proofs and Algorithms" pole. The members of the Cosynus team work on the semantics and static analysis of software systems, sequential, concurrent or distributed, and hybrid/control systems and cyber-physical systems.

This internship is proposed in the context of the SAIF project (Safe Artificial Intelligence through Formal Methods), of the French National Research Programme on Artificial Intelligence PEPR IA. A successful internship can be followed by a PhD funded by the SAIF project.

Expected Knowledge of the Student: background in maths and computer science, eagerness to both develop theoretical ideas and implement and experiment them practically

References

[1] Shahaf Bassan and Guy Katz. Towards formal xai: Formally approximate minimal explanations of neural networks. In *Tools and Algorithms for the Construction and Analysis of Systems: 29th International Conference, TACAS 2023, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2023, Paris, France, April 22–27, 2023, Proceedings, Part I*, page 187–207, Berlin, Heidelberg, 2023. Springer-Verlag.

[2] Sumon Biswas and Hridesh Rajan. Fairify: Fairness verification of neural networks. In *Proceedings of the 45th International Conference on Software Engineering*, ICSE '23, page 1546–1558. IEEE Press, 2023.

[3] Eric Goubault and Sylvie Putot. Robust under-approximations and application to reachability of non-linear control systems with disturbances. *IEEE Control Systems Letters*, 4(4):928–933, 2020.

[4] Eric Goubault and Sylvie Putot. Inner and outer approximate quantifier elimination for general reachability problems. In *Proceedings of the 27th ACM International Conference on Hybrid Systems: Computation and Control*, HSCC '24, New York, NY, USA, 2024. Association for Computing Machinery.

[5] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

[6] Haitham Khedr and Yasser Shoukry. Certifair: A framework for certified global fairness of neural networks, 2022.

[7] Brian Hyeongseok Kim, Jingbo Wang, and Chao Wang. Fairquant: Certifying and quantifying fairness of deep neural networks, 2024.

[8] Rabia Saleem, Bo Yuan, Fatih Kurugollu, Ashiq Anjum, and Lu Liu. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*, 513:165–180, 2022.

[9] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 10825–10836, Red Hook, NY, USA, 2018. Curran Associates Inc.

[10] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL), January 2019.