Master internship proposal: Verification of probabilistic properties for AI

October 2, 2024

Keywords: formal methods, neural networks, imprecise probabilities

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. But for some applications, e.g. perception in autonomous systems, through classifiers, we can only hope for probabilistic safety.

Moreover, in real-world systems, precise models representative of the data are not always available. For instance, several probabilistic models may be plausible for describing of a problem, or a probabilistic model may be known but with uncertain parameters. Therefore, we need to consider both probabilistic information and epistemic uncertainty. Recently was introduced in [1] an approach based on imprecise probabilities or probability boxes, which generalize probabilistic and epistemic uncertainties by defining sets of probability distributions, for the quantitative verification of neural networks. This approach provides qualitative measures of how likely the outputs of a neural network are to exhibit a certain behavior, given some assumptions on the inputs specified as imprecise probabilities. Based on an abstraction and propagation of sets of probability distributions in neural networks, the probability of a property satisfaction can be bounded, and the regions of the input space the more likely to lead to property violation identified.

Objectives: The approach of [1] proved to be both more general and more computationally efficient than the state of the art. However many challenges remain, the internship aims at investigating some of them:

- the current abstraction of probability boxes relies on a constant stepsize staircase discretization which will hardly scale with the input dimension of the networks; we would like to investigate other initial abstractions. Less importantly, distributions with bounded support are assumed, we believe this assumption could be lifted.

- independence is currently assumed between the inputs of the network; we would like to handle also multivariate input distributions, for instance using copulas [3] as in e.g. [2].

A longer-term objective is to study the extension of the approach to the analysis of Bayesian neural networks, where the weights and bias of the network are also defined by multivariate (imprecise) probability distributions. This extends the case of multivariate input distributions, but in much higher dimension, probably requiring a fresh view on the approach.

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] Eric Goubault and Sylvie Putot. A zonotopic dempster-shafer approach to the quantitative verification of neural networks. In *Formal Methods: 26th International Symposium, FM 2024, Milan, Italy, September 9–13, 2024, Proceedings, Part I*, page 324–342, Berlin, Heidelberg, 2024. Springer-Verlag.

[2] Ander Gray, Marcelo Forets, Christian Schilling, Scott Ferson, and Luis Benet. Verified propagation of imprecise probabilities in non-linear odes. *International Journal of Approximate Reasoning*, 164, 2024. Publisher Copyright: © 2023.

[3] Bernhard Schmelzer. Random sets, copulas and related sets of probability measures. *International Journal of Approximate Reasoning*, 160:108952, 2023.