Karlsruher Institut für Technologie

Application-oriented Formal Verification Research Group
Institute of Information Security and Dependability (KASTEL)
Am Fasanengarten 5
76131 Karlsruhe
https://formal.kastel.kit.edu

Refining Verified Floating-Point Sound NN Verifiers via Interactive Theorem Provers

Research Project (*Praxis der Forschung*)

Nowadays, neural networks (NNs) see increasing use even in safety or security critical systems. As traditional statistical evaluation cannot provide hard guarantees about the absence of errors, efficient algorithms for formal verification of NNs have been developed [3].
Most of these algorithms however, assume that computations are carried out over real numbers. In reality, however, both the NN and the verifier are executed using floating point arithmetic. This mismatch has been shown to introduce soundness issues, where verifiers may incorrectly declare a network safe [1, 4].
Interactive theorem provers such as Isabelle/HOL, Rocq or Lean allow users to formalize systems and mechanically check proofs about them. Via refinement, e.g., Isabelle/HOL is also able to generate efficiently executable code from the specification of an algorithm [2].
Given the unintuitive nature and many pitfalls of floating point arithmetic, the rigorous nature of interactive theorem provers makes them a promising tool for building floating-point sound NN verifiers.

Idea. The goal of this project is to specify a NN verification algorithm within an interactive theorem prover and prove that it is sound with respect to floating-point arithmetic. The verified specification will then be refined into efficient executable code.

Task. The project involves

- familiarization with interactive theorem proving
- studying existing formalizations of floating-point arithmetic and simple NNs models
- specifying a NN verifier and
- using refinement to generate efficient code.

Keywords: Neural Network Verification, Interactive Theorem Provers, Refinement, Floating Point

References

[1] Kai Jia et al. “Exploiting Verified Neural Networks via Floating Point Numerical Error”. In: *Static Analysis - 28th International Symposium, SAS 2021, Chicago, IL, USA, October 17-19, 2021, Proceedings*. Ed. by Cezara Dragoi et al. Vol. 12913. Lecture Notes in Computer Science. Springer, 2021, pp. 191–205. DOI: 10.1007/978-3-030-88806-0_9.

[2] Peter Lammich et al. *Automatic Refinement to Efficient Data Structures: A Comparison of Two Approaches*. In: *J. Autom. Reason.* 63.1 (2019), pp. 53–94. DOI: 10.1007/S10817-018-9461-9.

[3] Gagandeep Singh et al. *An abstract domain for certifying neural networks*. In: *Proc. ACM Program. Lang.* 3.POPL (2019), 41:1–41:30. DOI: 10.1145/3290354.

[4] Dániel Zombori et al. “Fooling a Complete Neural Network Verifier”. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Supervisors:

Philipp Kern, philipp.kern@kit.edu, Room 203 (Bldg. 50.34)
Michael Kirsten, kirsten@kit.edu, Room 228 (Bldg. 50.34)