Some Recent Perspectives on Ensuring Neural Networks Safety

Jan Křetínský

Abstract. In this talk we advocate several underdeveloped research directions for safety verification of neural networks. The talk builds upon recent work [AHKM20,CKM23,HKMS21,HKRS23].
We focus at several classical tools of software verification, which have proven crucial for its practical success, yet remain largely unexplored in the context of neural networks. In particular, we firstly discuss the role of *abstraction* as a separate instrument not to be confused with heuristics used in various verification procedures. Abstraction is a key verification technique to improve scalability, currently essentially inapplicable to neural networks. This illustrates the lack of basic understanding how to scalably approach the verification problem. Secondly, we discuss *runtime verification and monitoring* as a more practical compromise for ensuring safety. Again, while confidence in the outcomes of neural networks has been studied in machine learning, actual monitoring has not been systematically approached and utilized, despite its appeal to the industry. In the talk, we sketch some of the issues and difficulties as well as suggestions and arguments for examining the directions in more detail.

References

[AHKM20] Pranav Ashok, Vahid Hashemi, Jan Křetínský, and Stefanie Mohr. Deepabstract: Neural network abstraction for accelerating verification. In *ATVA*, volume 12302 of *Lecture Notes in Computer Science*, pages 92–107. Springer, 2020.

[CKM23] Calvin Chau Jan Křetínský, and Stefanie Mohr. Syntactic vs semantic linear abstraction and refinement of neural networks. In *ATVA*, Lecture Notes in Computer Science. Springer, 2023. To appear.

[HKMS21] Vahid Hashemi, Jan Křetínský, Stefanie Mohr, and Emmanouil Seferis. Gaussian-based runtime detection of out-of-distribution inputs for neural networks. In *RV*, volume 12974 of *Lecture Notes in Computer Science*, pages 254–264. Springer, 2021.

[HKRS23] Vahid Hashemi, Jan Křetínský, Sabine Rieder, and Jessica Schmidt. Runtime monitoring for out-of-distribution detection in object detection neural networks. In *FM*, volume 14000 of *Lecture Notes in Computer Science*, pages 622–634. Springer, 2023.