Unsupervised Neural Network Verification

Benedikt Böing [1] **Emmanuel Müller** [1]

Neural networks are at the forefront of machine learning being responsible for achieve- ments such as AlphaGo. As they are being deployed in more and more environments - even in safety-critical ones such as health care - we are naturally interested in assu- ring their reliability. However, the discovery of so-called adversarial attacks for supervised neural networks demonstrated that tiny distortions in the input space can lead to misclassifications and thus, to potentially catas- trophic errors: Patients could be dia- gnosed wrongly, or a car might confuse stop signs and traffic lights. Thus, ideally, we would like to guarantee that these types of attacks cannot occur. In this thesis we extend the research on reliable neural networks to the realm of unsuper- vised learning. This includes defining proper notions of reliability, as well as analyzing and adapting unsupervised neural networks with respect to this notion. Our definitions of reliability depend on the underlying neural networks and the problems they are meant to solve. However, in all our cases, we aim for guarantees on a continuous input space containing infinitely many points. Therefore we extend the traditional setting of testing against a finite dataset such that we require specialized tools to actually check a given network for reliability. We will demonstrate how we can leverage neural network veri- fication for these purposes. Using neural network verification, however, entails a major challenge: It does not scale up to large networks. To overcome this limitation, we design a novel training procedure yielding networks that are both more reliable according to our definition as well as more amenable for neural network verification. By exploiting the piecewise affine structure of our networks, we can locally simplify them and thus decrease verification runtime significantly. We also take a perspective that complements a neural network's training by exploring how we can repair non-reliable neural network ensembles. Making use of this particular model structure, we circumvent the need to verify large scale neural networks and allow for a post-processing step that ensures reliability. Our method yields a model which provably cannot be attacked by adversarial attacks and keeps the original predictive capability. With this thesis, we paradigmatically show the necessity and the complications of unsu- pervised neural network verification. It aims to pave the way for more research to come and towards a safe usage of these simple-to-build yet difficult-to-understand models gi- ven by unsupervised neural networks.

[1]Chair of Data Science and Data Engineering, Technical University of Dortmund, Dortmund, Germany. Correspondence to: Benedikt Böing <benedikt.boeing@cs.tu-dortmund.de>.

References

Böing, B. and Müller, E. On training and verifying robust autoencoders. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10. IEEE, 2022.

Böing, B., Roy, R., Müller, E., and Neider, D. Quality guarantees for autoencoders via unsupervised adversarial attacks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part II*, pp. 206–222. Springer, 2021.

Böing, B., Howar, F., Hüntelmann, J., Müller, E., and Stewing, R. Neural network verification with DSE. In *OVERLAY@AI*IA 2022*, 2022a.

Böing, B., Klüttermann, S., and Müller, E. Post-robustifying deep anomaly detection ensembles by model selection. In *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 861–866. IEEE, 2022b.