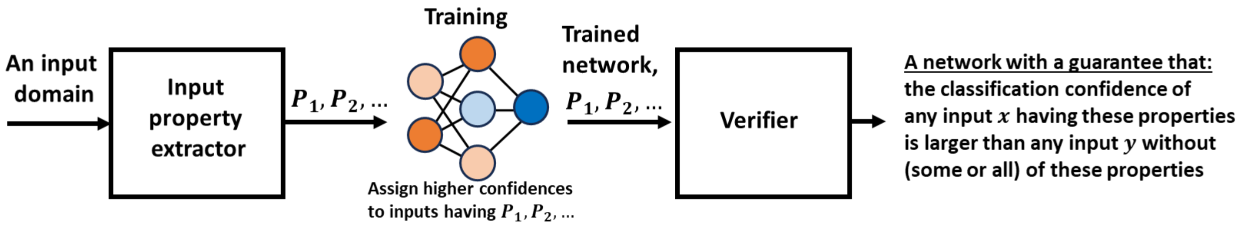
Towards Formal Guarantees for Networks' Overconfidence

Anan Kabaha, Dana Drachsler Cohen

Technion, Israel Institute of Technology

Although successful, neural networks are prone to making overconfident predictions when presented with out-of-distribution inputs. Even well-trained networks can show very high confidence for inputs that do not belong to the task they are trained for. This raises concerns about the network's reliability in real-world scenarios where it may encounter a wide range of inputs. For example, consider a network trained for traffic signs that upon encountering a tree – which was not part of the training set – classifies it as a stop sign with a high confidence. To cope, several works suggest training approaches aiming to decrease the confidence of out-of-distribution inputs [1,2,3,4], mainly by regularization terms. Each of these works considers a certain type of an out-of-distribution domain, samples from it, and incorporates the samples into the training process. These works demonstrate empirical success in reducing the network's overconfidence. However, they rely on a subset of inputs drawn from a specific distribution. As a result, they can only provide either empirical evidence [1,2,3] or asymptotic guarantees [4] to ensure that the network assigns low confidence to out-of-distribution inputs. However, these approaches cannot *formally guarantee* that the network assigns low confidence to *any out-of-distribution input*. In this work, we propose to leverage formal verification to identify and resolve the network's overconfidence. The key idea is to compute a set of representative properties (features), derived directly from the input domain, which enable to automatically identify whether a new input is in or out of the distribution. Then, similarly to prior works (with some differences), we propose a training approach that encourages the network to assign higher confidence to inputs with these properties. To provide formal guarantees for the trained network, we propose a verifier, relying on a MILP encoding, to prove that the network assigns higher confidence to inputs having these properties. Thereby, we propose a system that both mitigates the overconfidence issue as well as provides formal guarantees to inputs that cannot suffer from this issue. Figure 1 visualizes our system.



References
[1] Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In ICLR (2018).
[2] Hein, M., Andriushchenko, M., and Bitterwolf, J. Why ReLU networks yield high confidence predictions far away from the training data and how to mitigate the problem. In CVPR (2019).
[3] Ming, Y., Sun, Y. Dia, O. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection?. In ICLR (2023).
[4] Meinke, A., Bitterwolf, J., and Hein, M. Provably robust detection of out-of-distribution data (almost) for free. In NeurIPS (2022).