ICAI 2023
12th International Conference on Applied Informatics
Eger, Hungary, 2023

Interval Based Verification of Adversarial Example Free Zones for Neural Networks

Tibor Csendes[a], Nóra Büki[a], Soma Timer[a]

[a]Institute of Informatics, University of Szeged
csendes@inf.szte.hu

Recent machine learning models are sensitive to adversarial input perturbation. That is, an attacker may easily mislead an otherwise well-performing image classification system by altering some pixels. It is quite challenging to prove that a network will have correct output when changing slightly some regions of the images. This is why only a few works targeted this problem. Although there are an increasing number of studies on this field, really reliable robustness evaluation is still an open issue. In this work, we will attempt to contribute in this direction. We will present an improved interval arithmetic based algorithm to provide adversarial example free image patches for trained artificial neural networks.

Szegedy et al. [2] showed first the phenomenon of adversarial examples. Since then the efforts for verification algorithms were concentrated around optimized models of the trained neural networks [3]. We were able to prove that these verification algorithms are unfortunately not reliable [4]. In the last ICAI conference we reported our first results with an interval based verification algorithm [1]. This approach applied simple natural interval extension for the calculation of inclusion functions, and we were able to produce realistic size adversarial example free zones for simple networks, that had good accuracy for the MNIST picture database distinguishing the hand written figures for the digits 3 and 7.

In the present talk we report on our new approaches. We implemented our algorithm in the Julia language. The main difficulty in applying interval arithmetic in the evaluation of trained neural network lies in the so-called dependency problem. In spite of the fact that addition and multiplication can be implemented for intervals in a precise way, the hidden dependencies of input variables pose a substantial problem in terms of overestimation of the bounded ranges. Affine arithmetic and the interval propagation technique can help. We report our computational test results on the full MNIST database of 10 different hand written digits with a modest but realistic size network. We give details on the tricky handling of the ReLU

activation function as well. We give some empirical values on the probabilities with which the ReLU activation functions obtain an argument very close to zero – causing trouble in the verification algorithm.

References

[1] T. Csendes, N. Balogh, B. Bánhelyi, D. Zombori, R. Tóth, I. Megyeri: *Adversarial Example Free Zones for Specific Inputs and Neural Networks*, in: Proceedings of the 2020 ICAI, Eger, Hungary, URL: https://ceur-ws.org/Vol-2650/paper9.pdf.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus: *Intriguing properties of neural networks*, in: Proceedings of the 2014 International Conference on Learning Representations, DOI: https://doi.org/10.48550/arXiv.1312.6199.

[3] V. Tjeng, K. Xiao, R. Tedrake: *Evaluating Robustness of Neural Networks with Mixed Integer Programming*, in: Proceedings of the 2019 International Conference on Learning Representations, DOI: https://doi.org/10.48550/arXiv.1711.07356.

[4] D. Zombori, B. Bánhelyi, T. Csendes, I. Megyeri, M. Jelasity: *Fooling a complete neural network verifier*, in: Proceedings of the 2021 International Conference on Learning Representations, URL: https://openreview.net/forum?id=4IwieFS44l.