Adversarial Example Free Zones for Specific Inputs and Neural Networks

Tibor Csendes[1], Nándor Balogh[2], Balázs Bánhelyi[1], Dániel Zombori[1], Richárd Tóth[1], and István Megyeri[1]

[1] University of Szeged, Szeged, Hungary
[2] Redink Ltd., Szeged, Hungary
csendes@inf.szte.hu

Keywords: artificial neural networks, adversarial example, interval arithmetic, inscribed interval

Recent machine learning models are highly sensitive to adversarial input perturbation. That is, an attacker may easily mislead a well-performing image classification system by altering some pixels. However, proving that a network will have correct output when changing some regions of the images, is quite challenging – mostly due to the high dimensionality and/or the nonlinearity of the problem. Because of this, only a few works targeted this problem, and some of these verification tools are not reliable [3]. Although there are an increasing number of studies on this field, reliable robustness evaluation is still an open issue. We will present interval arithmetic based algorithms to provide adversarial example free image patches for trained artificial neural networks [2]. The method is based on an earlier interval technique to bound level sets of parameter estimation problems [1].

The obtained results are illustrated on Figure 1 for some of the studied images from the MNIST dataset. The calculated number of pixels to be changed arbitrarily were between 88 and 190 (compare it with the $28 \times 28 = 784$ pixels in the images). The combined running time for the second round of 10 test images was 1971.87 second, i.e. closely half an hour.

We are still in the phase when we explore the capabilities of interval arithmetic based algorithms for describing the sensitivity of trained natural neural networks to changes in object to be classified, but we find our present results encouraging enough to continue our research project.

Acknowledgements. This research was supported by the project "Extending the activities of the HU-MATHS-IN Hungarian Industrial and Innovation Mathematical Service Network" EFOP3.6.2-16-2017-00015, and 2018-1.3.1-VKE-2018-00033.

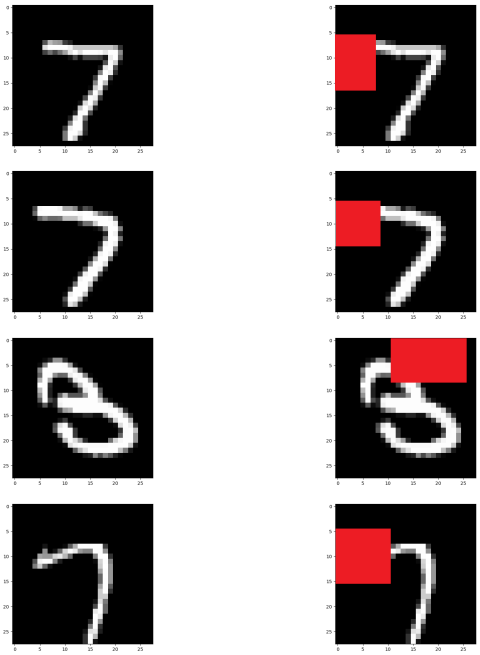

Figure 1: Original pictures and proven rectangles where we can change *everything* without having an adversarial example.

References

[1] T. Csendes: An interval method for bounding level sets of parameter estimation problems, *Computing*, 41 (1989), 75–86.

[2] T. Csendes, N. Balogh, B. Bánhelyi, D. Zombori, R. Tóth, I. Megyeri: Adversarial Example Free Zones for Specific Inputs and Neural Networks. *Proc. ICAI*, Eger, Hungary, 2020, 76–84, http://ceur-ws.org/Vol-2650/paper9.pdf

[3] D. Zombori, B. Bánhelyi, T. Csendes, I. Megyeri, M. Jelasity: Fooling a Complete Neural Network Verifier. *Proc. ICLR*, 2021, https://openreview.net/pdf?id=4IwieFS44l