Adversary example free zones for neural networks

Tibor Csendes, Nándor Balogh, Richárd Tóth, and István Megyeri

University of Szeged

One of the hottest topics in present artificial intelligence research is to understand the phenomenon of adversarial examples for machine learning technics applying artificial neural networks.

The typical problem is that in many practical cases, e.g. in image recognition, after the proper training of the network, surprisingly close pictures to the actual ones result in a denial decision.

We developed an interval arithmetic based algorithm that is capable to describe the level sets of an artificial neural network around a feasible positive sample.

In this way, we could ensure with mathematical rigor that adversarial samples cannot exist within the found bounds. The key question is how the algorithm that was published earlier by T. Csendes scales up with increasing dimension.

According to our experiences, benevolent problems show much better complexity numbers compared to theoretically possible pessimistic convergence rates.

Acknowledgement: This research was supported by the project ”Extending the activities of the HU-MATHS-IN Hungarian Industrial and Innovation Mathematical Service Network” EFOP-3.6.2-16-2017-00015, 2018-1.3.1-VKE-2018-00033.

Reference:

Csendes, T.: An interval method for bounding level sets of parameter estimation problems. Computing 41(1989) 75-86.