A reliable algorithm to determine adversary example free zones for artificial neural networks

Tibor Csendes, Nándor Balogh, Balázs Bánhelyi, Dániel Zombori, Richárd Tóth, and István Megyeri

University of Szeged, Redink Ltd.

Adversarial examples in artificial neural networks

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.

Illustration

A single page introduction to interval calculation

$$[a, b] + [c, d] = [a + c, b + d],$$

$$[a, b] - [c, d] = [a - d, b - c],$$

$$[a, b] \cdot [c, d] = [\min(ac, ad, bc, bd), \max(ac, ad, bc, bd)],$$

$$[a, b]/[c, d] = [a, b] \cdot [1/d, 1/c] \text{ if } 0 \notin [c, d].$$

The inclusion of the function

$$f(x) = x^2 - x$$

obtained for the interval $[0, 1]$ is $[-1, 1]$, while the range of it is here just $[-0.25, 0.0]$.

Using more sophisticated techniques the problem of the too loose enclosure can be overcome – at the cost of higher computing times.

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.

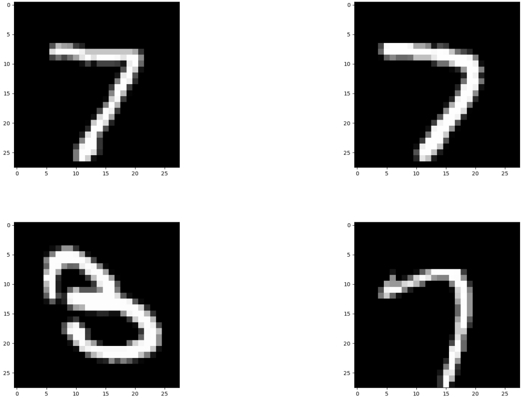
According to our experiences, benevolent problems show much better complexity numbers compared to theoretically possible pessimistic convergence rates.

The pseudo code of the algorithm

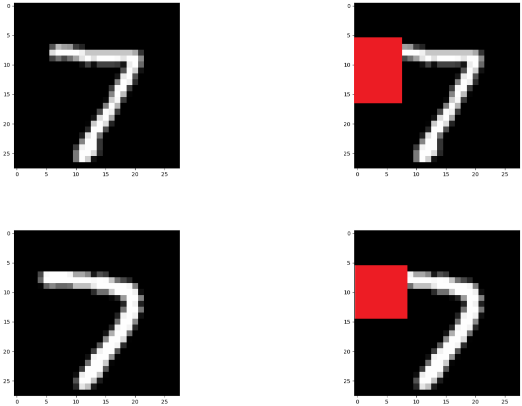
0. If $F(p0) > 0.5$ then $greater = true$, otherwise $greater = false$
1. Iterate until $percent <= 100$
2. Let P be an n dimensional interval
3. For $i = 1$ to n do
   1. If $p_i = 0$, then $P_i = [0, 2 * percent/100]$
   2. Otherwise, if $p_i = 1$, then $P_i = [1 - 2 * percent/100, 1]$
   3. Otherwise $P_i = [p_i - percent/100, p_i + percent/100]$, and check the end points: if the lower one is negative, then set it to zero, if the upper one is larger than 1, then set it to 1.
4. If $greater = true$ and $F(P) \geq 0.5$, or $greater = false$ and $F(P) < 0.5$ then do:
   1. If $percent < 1$, then $maxpercent = percent$, and break the main cycle, Stop.
   2. Otherwise $maxpercent = percent$, and $percent = percent + 1$
5. Otherwise if $percent < 1$, then set $percent = percent - 0.1$
   1. If now $percent = 0$, then set $maxpercent = 0$ and STOP
   2. Otherwise break the outer loop
6. End of the cycle started in the first step

Proven amount of changes on the gray scale *everywhere* on the picture without having an adversarial example

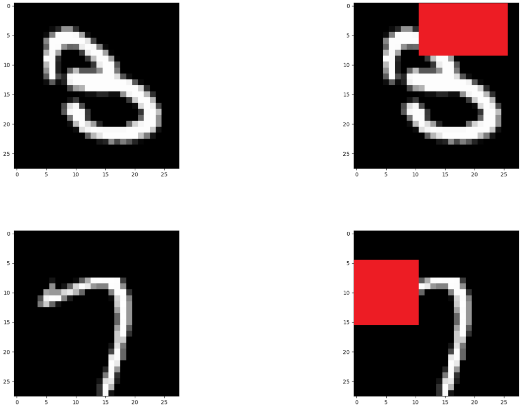
In the order of appearance: 2%, 4%, 8%, and 3%, respectively.



Original pictures & proven rectangles where we can change *everything* without having an adversarial example



Original pictures & proven rectangles where we can change *everything* without having an adversarial example # 2



Conclusion and future research

We could demonstrate that our interval based algorithm is capable to verify simple artificial neural networks on small real life picture recognition problems.

Next steps:

- Test larger realistic networks.
- Try Julia to speed up the algorithm.
- Implement the so-called "interval propagation" trick to fight the dependency problem.
- Design heuristic greedy search methods to have an efficient technique.
- Check how our method scales up with increasing problem size and with more complex networks.
- Which activation function fits our procedure best?

References

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

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry: Adversarial Examples Are Not Bugs, They Are Features. arXiv:1905.02175

Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi: One pixel attack for fooling deep neural networks. arXiv:1710.08864

Michal Zaj, Konrad Zolna, Negar Rostamzadeh, and Pedro O. Pinheiro: Adversarial Framing for Image and Video Classification. The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

References 2

Tibor Csendes, Nándor Balogh, Balázs Bánhelyi, Dániel Zombori, Richárd Tóth, and István Megyeri: Adversarial Example Free Zones for Specific Inputs and Neural Networks. ICAI Proceedings, 2020, 76-84.

Dániel Zombori, Balázs Bánhelyi, Tibor Csendes, István Megyeri, Márk Jelasity: Fooling a Complete Neural Network Verifier. Int. Conf. on Learning Representations (ICLR 2021), https://openreview.net/forum?id=4IwieFS44l.

Scholar rank

Publication forum, h5, h5 median

1. Nature, 414, 607
2. The New England Journal of Medicine, 410, 704
3. Science, 391, 564
4. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 356, 583
5. The Lancet, 345, 600
6. Advanced Materials, 294, 406
7. Cell, 288, 459
8. Nature Communications, 287, 389
9. Chemical Reviews, 270, 434
10. **International Conf. on Learning Representations, 253, 470**

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.

The related papers are available on the page

`www.inf.u-szeged.hu/~csendes`