Automatic recognition of fake news by linguistic and artificial intelligence tools

Csendes Tibor

Szegedi Tudományegyetem

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 techniques 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.

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 of describing the level sets of an artificial neural network around a feasible positive sample.

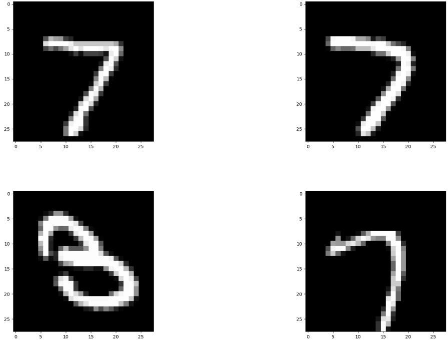
In this way, we could ensure with mathematical rigor that adversarial samples cannot exist within the found bounds. The key question is how the algorithm that was published earlier by T. Csendes scales up with increasing dimension.

The pseudo code of the algorithm on a single neuron

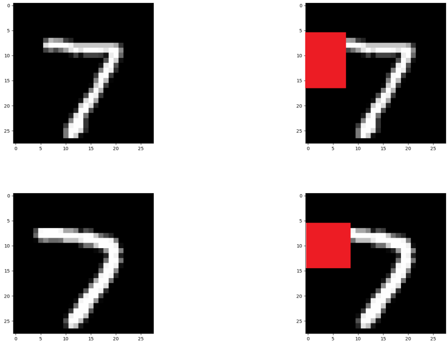
0. If $F(p0) > 0.5$ then $greater = true$, otherwise $greater = false$
1. Iterate until $percent <= 100$
2. Let P be an n dimensional interval
3. For $i = 1$ to n do
   1. If $p_i = 0$, then $P_i = [0, 2 * percent/100]$
   2. Otherwise, if $p_i = 1$, then $P_i = [1 - 2 * percent/100, 1]$
   3. Otherwise $P_i = [p_i - percent/100, p_i + percent/100]$, 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

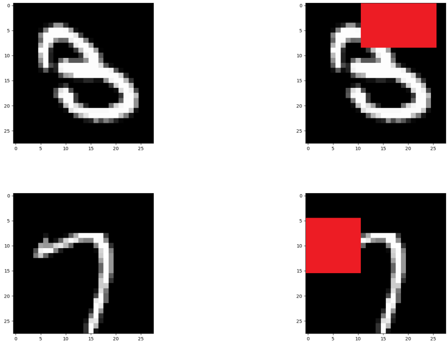
In the order of appearance: 2%, 4%, 8%, and 3%, respectively.



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



State of the art

On a single neuron on the 3-7 problem, interval arithmetic showcases its best properties. On a more realistic network, however, and with multiple output classes, there are problems to solve.

- The dependency problem of interval arithmetic blows up output widths unless we use more costly alternate representations.
- The alternate representations are ill-prepared to deal with the nonlinearity of a ReLU, leading to overestimations.
- The computer representation of floating point numbers results in overestimation due to outward rounding.
- Intervals are only partially ordered, so the certainty of the output cannot always be ensured.
- Real life greyscale stickers can be white or black, but interval stickers are both at the same time.

Overestimation

Assertion

For a fully connected feed forward standard artificial neural network the overestimation size $w(F(X)) - w(f(X))$ of the inclusion function can be zero only if at least one of the following conditions are fulfilled:

- *all input intervals are of zero width:* $w(x_i) = b - a = 0$,
- *for all input variables* x_i *in the computation of each of the outputs all weights of them are of the same sign: either all nonnegative, or all nonpositive, and*
- *all the final evaluation functions calculating the outputs of the network have negative arguments.*

These conditions are not only sufficient one by one, but a proper combination of them is also necessary.

Overestimation 2

Assertion

For a fully connected feed forward standard artificial neural network of k input intervals, m neurons in each of the even number of n hidden layers, and all weights w_i are bounded by $|w_i| \leq W$ the amount of overestimation $w(F(X)) - w(f(X))$ of the inclusion function of an output is not more than $2^{n/2} m^{n/2} W^n \sum_{i=1}^{k} w(X_i)$.

Corollary

A direct consequence of this Assertion is that we can have the same amount of overestimation due to the dependency problem with decreasing the number of hidden layers while increasing the number of neurons in a layer and vice versa.

Future plan

Our primary goal is to develop a verifier that guarantees mathematical certainty of an adversarial example-free zone. Our secondary goal is to ensure that this zone is of a non-negligible size. For this, we aim to derive several mathematical results and perform computational experiments.

Alternate interval representations such as affine forms and a so-called "symbolic propagation" method theoretically, with an appropriate extension of the rectifier function, provide the same result, but the practical implementations might differ because computers.

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.

Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana: Formal Security Analysis of Neural Networks using Symbolic Intervals arXiv:1809.08098

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, and by the Subprogramme for Linguistic Identification of Fake News and Pseudo-scientific Views, part of the Science for the Hungarian Language National Programme of the Hungarian Academy of Sciences (MTA).