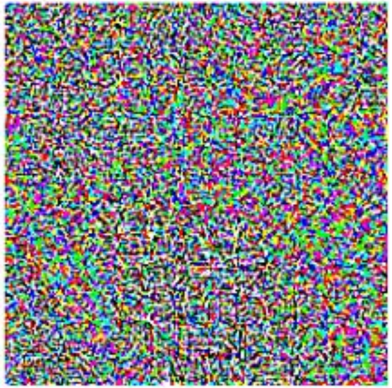
Universal Approximation with Certified Networks

Maximilian Baader, Matthew Mirman, Martin Vechev
Department of Computer Science
ETH Zurich, Switzerland

Adversarial Examples

+ 0.007



=

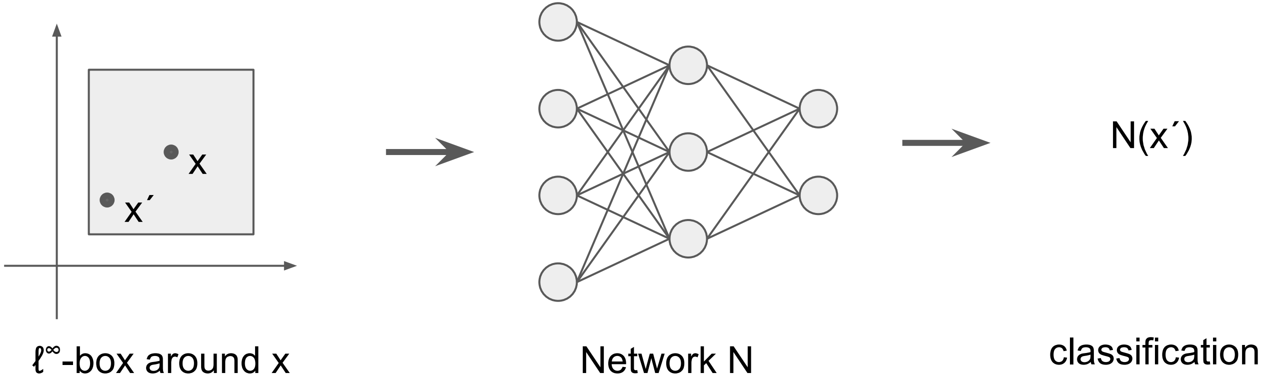
x | e | x + 0.007·e

N(x) = “panda” | noise | N(x + 0.007e) = “gibbon”

[1] Goodfellow et. al. Explaining and Harnessing Adversarial Examples. ICLR 2015

ℓ^∞-robustness

A neural network N is ε **ℓ^∞-robust** around an image x, if for all images x’ having ℓ^∞-distance to x of at most ε, it holds that N(x) = N(x´).



Goal: Prove N(x) = N(x´) for all x´

ℓ^∞-robustness certification via Interval analysis

A common method to prove ℓ^∞-robustes is linear relaxation to intervals ([2], [3], [4]). **Interval analysis** is the **fastest** non probabilistic certification method (~4x slower than classification time), can **scale to large networks** and when used for training **produces state-of-the-art results** ([3], [4]).

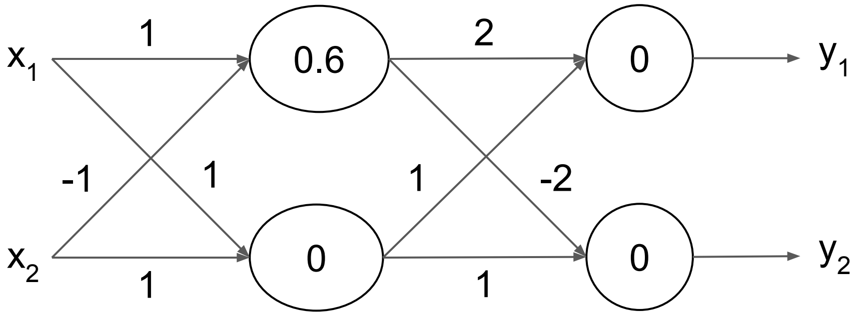
However, interval analysis loses precision -- it can induce too large of an over-approximation of the actual values.

[2] Gehr et al. AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation. IEEE S&P 2018.
[3] Mirman et al. Differentiable Abstract Interpretation for Provably Robust Neural Networks. ICML 2018.
[4] Gowal et al. On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models. arXiv 2018.

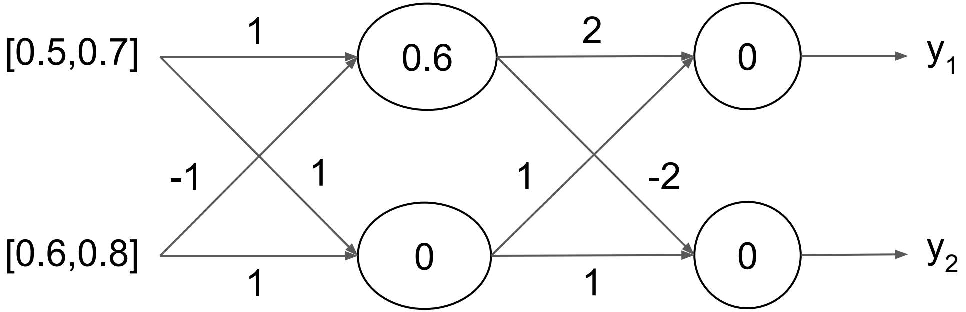
Certification with Interval Analysis

Example: Assume we have a 2 pixel image x = (0.6, 0.7), ε=0.1. The intervals for x_1 and x_2 are [0.6-ε, 0.6+ε] = [0.5,0.7] and [0.7-ε, 0.7+ε] = [0.6,0.8] respectively.

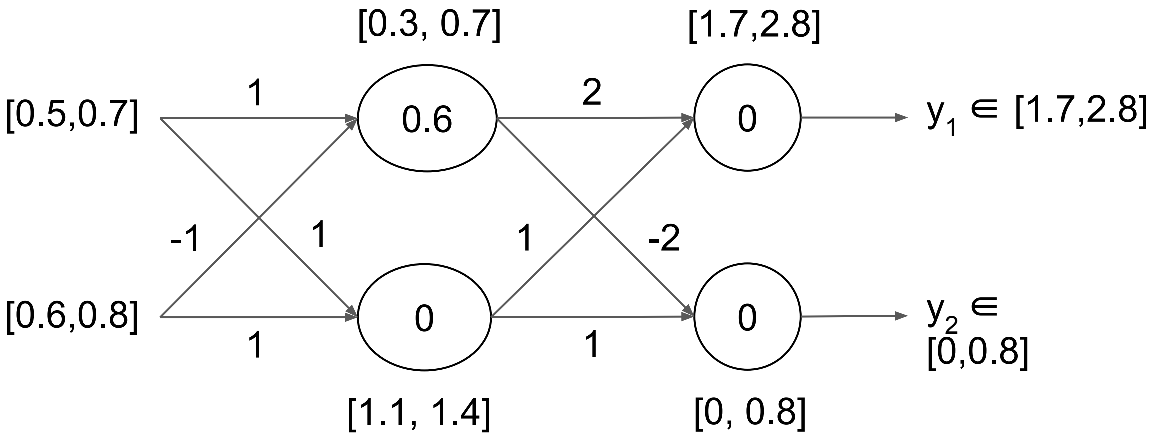
Let the network N be:



Certification with Interval Analysis



Certification with Interval Analysis



We can prove $y_1 > y_2$, thus classification is robust.

Key challenge

Neural networks trained to be amenable to certification with interval analysis often have unsatisfactory accuracy (< 60%) and certifiability (< 40%, ε=8/255) on standard datasets (CIFAR-10). This is unfortunate as intervals scale to large networks.

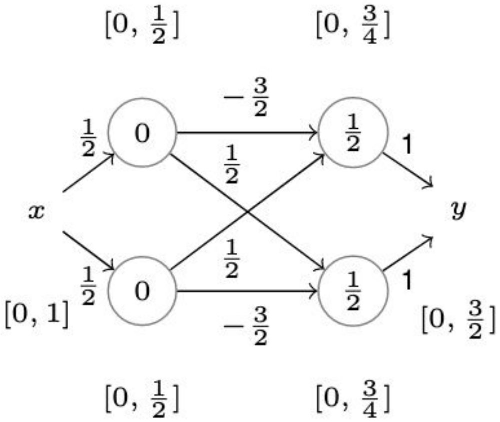
Fundamental Question

Do interval-certifiable networks **actually exist**, which approximate **any continuous function**?

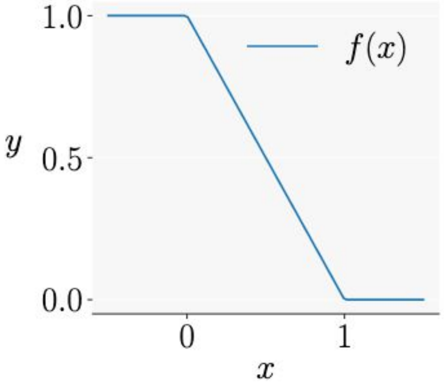
Implication: If yes, it can mean that there may actually be hope in using interval analysis for creating accurate and provable large neural networks!

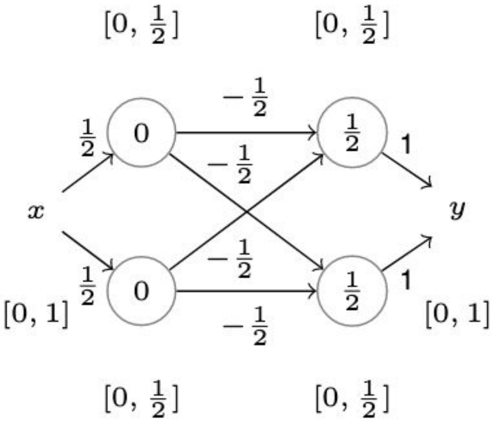
Classical Universal Approximation is insufficient:

Two networks can approximate the same function f, but they behave different under interval analysis:



N_1





N_2

Here, we cannot prove $N_1([0,1]) \subseteq [0,1]$, but $N_2([0,1]) \subseteq [0,1]$ although $N_1(x) = N_2(x)$ for all x in $\mathbb{R}$.

In this work we prove:

Theorem: Let $f : [0,1]^m \to \mathbb{R}$ be a continuous function.

For all $\delta > 0$ exists a ReLU network N such that for all $x \in [0,1]^m$ and for $\varepsilon > 0$ interval analysis can prove that N approximates f up to δ.

Specifically if $l = \min f([x-\varepsilon, x+\varepsilon])$ and $u = \max f([x-\varepsilon, x+\varepsilon])$ then $N^\#([x-\varepsilon, x+\varepsilon])$ satisfies

$$[l + \delta, u - \delta] \subseteq N^\#([x-\varepsilon, x+\varepsilon]) \subseteq [l - \delta, u + \delta].$$

ReLU networks can interval provably approximate continuous functions!

Future work: optimize the construction and study interval training in depth