Certifying Neural Networks with Abstract Interpretations

INF575 - Final Report

LAO Quentin

December 12, 2022

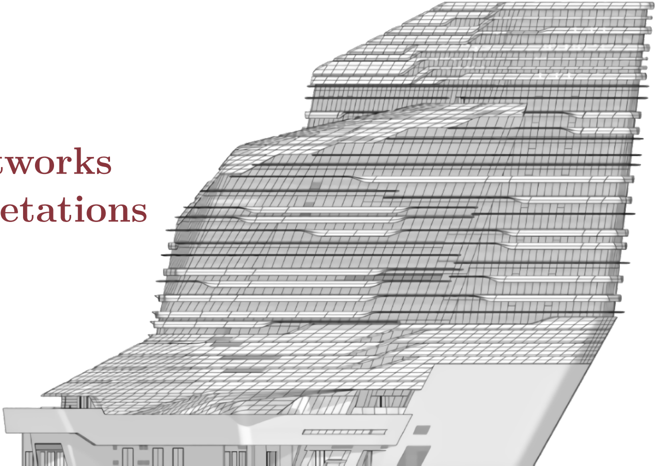
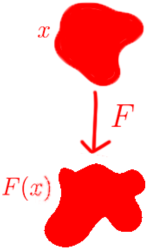


Table of Contents

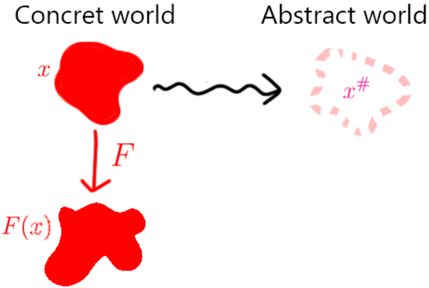
1 Classical Abstract interpretations

- Classical Abstract interpretations
- DeepPoly relaxation

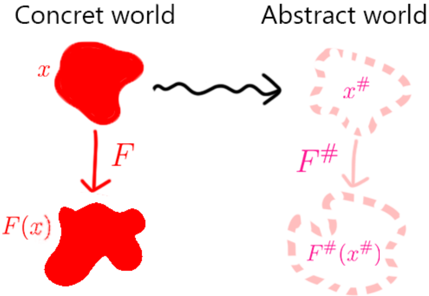
What do we expect from an abstract interpretation ?



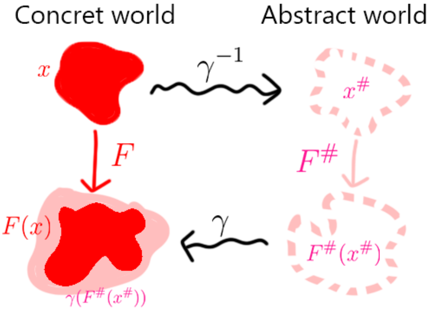
What do we expect from an abstract interpretation ?



What do we expect from an abstract interpretation ?

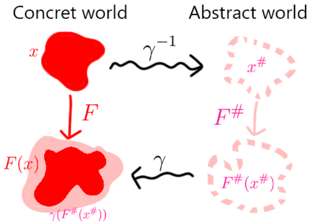


What do we expect from an abstract interpretation ?



What do we expect from an abstract interpretation ?

- **Robustness** : if $x \in I$, is $F(x) \in I'$?
- **Soundness** : $F(x) \subset \gamma(F^{\#}(x^{\#}))$
- Find an abstract form (similar for every neuron) that conveys the possible values with fewer approximations.



Concret world

Abstract world

Box relaxation

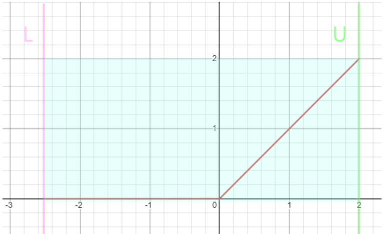
Abstraction : $x \in [\ell, u]$

Affine Transformation

- New neuron : $x_{\text{new}} = b + \sum_i w_i x_i$
- Abstraction for new neuron :
  - $\ell_{\text{new}} := b + \sum_i w_i (\mathbb{1}_{[w_i \geq 0]} \ell_i + \mathbb{1}_{[w_i < 0]} u_i)$
  - $u_{\text{new}} := b + \sum_i w_i (\mathbb{1}_{[w_i \geq 0]} u_i + \mathbb{1}_{[w_i < 0]} \ell_i)$

ReLU Transformation

- Abstraction for new neuron :
  - $\ell_{\text{new}} := \max(0, \ell)$
  - $u_{\text{new}} := \max(0, u)$



Zonotope relaxation

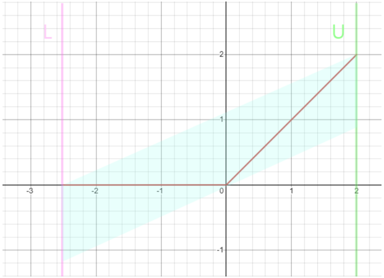
Abstraction : $x = b + \sum_k \varepsilon_k a_k$, with $\varepsilon_k \in [-1, 1]$

Affine Transformation

- Abstraction for new neuron :
  - $x_{\text{new}} := (b + \sum_i w_i b_i) + \sum_{i,k} \varepsilon_k^i (a_k^i w_i)$

ReLU Transformation

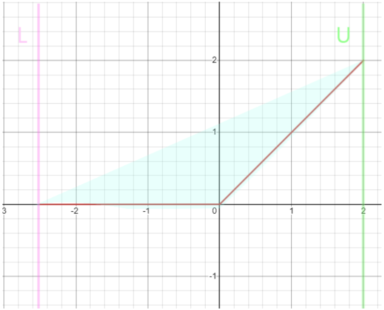
- Abstraction for new neuron :
  - $\text{ReLU}(x) = \lambda x + \varepsilon_{\text{new}} \frac{\mu}{2} + \frac{\mu}{2}$



Polyhedra relaxation

Abstraction : many $\sum_i a_{k,i} x_i \leq b_k$ constraints

ReLU Transformation



Summary

Soundness OK

Abstract name	+	-
Box Relaxation	memory friendly	not exact (Affine **and** ReLU)
Zonotone Relaxation	exact (Affine)	not exact (ReLU) + new uncertainties
Polyhedra Relaxation	more precised	computationally expensive

Summary

Soundness OK

Abstract name	+	**-**
Box Relaxation	memory friendly	not exact (Affine **and** ReLU)
Zonotone Relaxation	exact (Affine)	not exact (ReLU) + new uncertainties
DeepPoly?	...	...
Polyhedra Relaxation	more precised	computationally expensive

Table of Contents

2 DeepPoly relaxation

- Classical Abstract interpretations

- DeepPoly relaxation

DeepPoly relaxation

Abstraction :

- a lower bound ℓ_i and a upper bound u_i (Interval constraints) : $\ell_i \leq x_i \leq u_i$
- $a_i^{\leq}$ and $a_i^{\geq}$ both of the form $\sum_j w_j x_j + v$ (Relation constraints) : $a_i^{\leq} \leq x_i \leq a_i^{\geq}$.

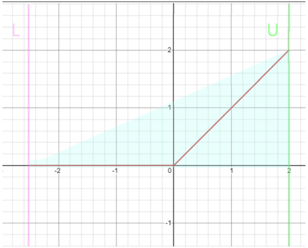
DeepPoly relaxation

Abstraction :

- a lower bound ℓ_i and a upper bound u_i (Interval constraints) : $\ell_i \leq x_i \leq u_i$
- $a_i^{\leq}$ and $a_i^{\geq}$ both of the form $\sum_j w_j x_j + v$ (Relation constraints) : $a_i^{\leq} \leq x_i \leq a_i^{\geq}$.

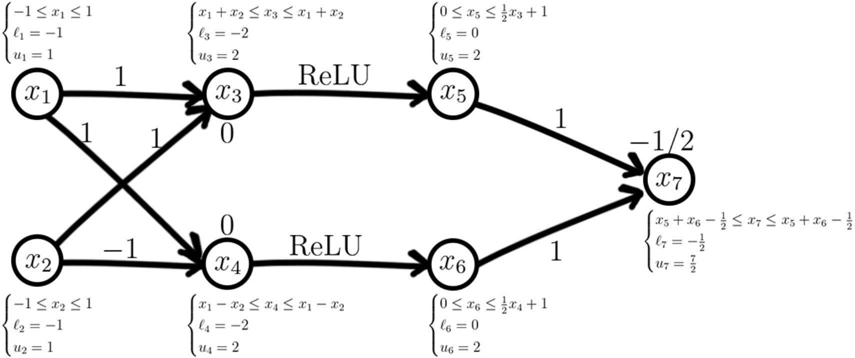
ReLU Transformation

- Abstraction for new neuron :
  - $a_{\text{new}}^{\leq} =: 0 \leq x_{\text{new}} \leq \lambda x + \mu =: a_{\text{new}}^{\geq}$
  - $\ell_{\text{new}} = 0$
  - $u_{\text{new}} = \lambda u + \mu$



DeepPoly relaxation

Backsubstitution



$$\begin{cases} -1 \leq x_1 \leq 1 \\ \ell_1 = -1 \\ u_1 = 1 \end{cases}$$

$$\begin{cases} x_1 + x_2 \leq x_3 \leq x_1 + x_2 \\ \ell_3 = -2 \\ u_3 = 2 \end{cases}$$

$$\begin{cases} 0 \leq x_5 \leq \frac{1}{2}x_3 + 1 \\ \ell_5 = 0 \\ u_5 = 2 \end{cases}$$

$$\begin{cases} x_5 + x_6 - \frac{1}{2} \leq x_7 \leq x_5 + x_6 - \frac{1}{2} \\ \ell_7 = -\frac{1}{2} \\ u_7 = \frac{7}{2} \end{cases}$$

$$\begin{cases} -1 \leq x_2 \leq 1 \\ \ell_2 = -1 \\ u_2 = 1 \end{cases}$$

$$\begin{cases} x_1 - x_2 \leq x_4 \leq x_1 - x_2 \\ \ell_4 = -2 \\ u_4 = 2 \end{cases}$$

$$\begin{cases} 0 \leq x_6 \leq \frac{1}{2}x_4 + 1 \\ \ell_6 = 0 \\ u_6 = 2 \end{cases}$$

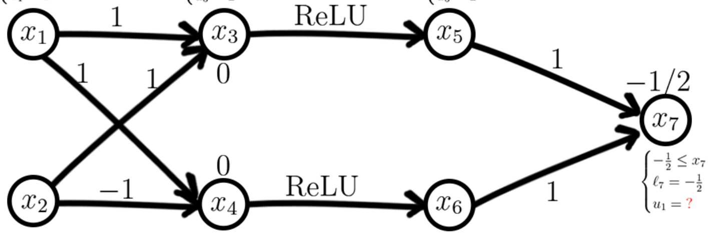
DeepPoly relaxation

Backsubstitution

$$\begin{cases} -1 \leq x_1 \leq 1 \\ \ell_1 = -1 \\ u_1 = 1 \end{cases}$$

$$\begin{cases} x_1 + x_2 \leq x_3 \leq x_1 + x_2 \\ \ell_3 = -2 \\ u_3 = 2 \end{cases}$$

$$\begin{cases} 0 \leq x_5 \leq \frac{1}{2}x_3 + 1 \\ \ell_5 = 0 \\ u_5 = 2 \end{cases}$$



$$\begin{cases} -\frac{1}{2} \leq x_7 \leq \frac{1}{2}x_3 + \frac{1}{2}x_4 + \frac{3}{2} \\ \ell_7 = -\frac{1}{2} \\ u_1 = ? \end{cases}$$

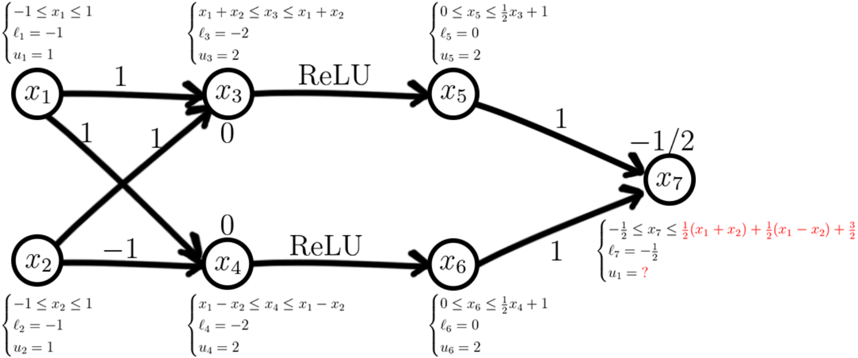
$$\begin{cases} -1 \leq x_2 \leq 1 \\ \ell_2 = -1 \\ u_2 = 1 \end{cases}$$

$$\begin{cases} x_1 - x_2 \leq x_4 \leq x_1 - x_2 \\ \ell_4 = -2 \\ u_4 = 2 \end{cases}$$

$$\begin{cases} 0 \leq x_6 \leq \frac{1}{2}x_4 + 1 \\ \ell_6 = 0 \\ u_6 = 2 \end{cases}$$

DeepPoly relaxation

Backsubstitution



$$\begin{cases} -1 \le x_1 \le 1 \\ \ell_1 = -1 \\ u_1 = 1 \end{cases}$$

$$\begin{cases} x_1 + x_2 \le x_3 \le x_1 + x_2 \\ \ell_3 = -2 \\ u_3 = 2 \end{cases}$$

$$\begin{cases} 0 \le x_5 \le \frac{1}{2}x_3 + 1 \\ \ell_5 = 0 \\ u_5 = 2 \end{cases}$$

$$\begin{cases} -\frac{1}{2} \le x_7 \le \frac{1}{2}(x_1 + x_2) + \frac{1}{2}(x_1 - x_2) + \frac{3}{2} \\ \ell_7 = -\frac{1}{2} \\ u_1 = ? \end{cases}$$

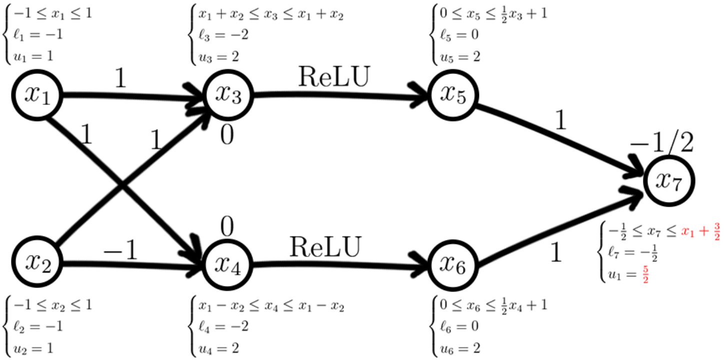
$$\begin{cases} -1 \le x_2 \le 1 \\ \ell_2 = -1 \\ u_2 = 1 \end{cases}$$

$$\begin{cases} x_1 - x_2 \le x_4 \le x_1 - x_2 \\ \ell_4 = -2 \\ u_4 = 2 \end{cases}$$

$$\begin{cases} 0 \le x_6 \le \frac{1}{2}x_4 + 1 \\ \ell_6 = 0 \\ u_6 = 2 \end{cases}$$

DeepPoly relaxation

Backsubstitution



$$\begin{cases} -1 \leq x_1 \leq 1 \\ \ell_1 = -1 \\ u_1 = 1 \end{cases}$$

$$\begin{cases} -1 \leq x_2 \leq 1 \\ \ell_2 = -1 \\ u_2 = 1 \end{cases}$$

$$\begin{cases} x_1 + x_2 \leq x_3 \leq x_1 + x_2 \\ \ell_3 = -2 \\ u_3 = 2 \end{cases}$$

$$\begin{cases} x_1 - x_2 \leq x_4 \leq x_1 - x_2 \\ \ell_4 = -2 \\ u_4 = 2 \end{cases}$$

$$\begin{cases} 0 \leq x_5 \leq \frac{1}{2}x_3 + 1 \\ \ell_5 = 0 \\ u_5 = 2 \end{cases}$$

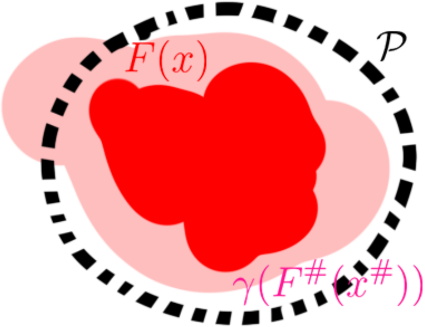
$$\begin{cases} 0 \leq x_6 \leq \frac{1}{2}x_4 + 1 \\ \ell_6 = 0 \\ u_6 = 2 \end{cases}$$

$$\begin{cases} -\frac{1}{2} \leq x_7 \leq x_1 + \frac{3}{2} \\ \ell_7 = -\frac{1}{2} \\ u_1 = \frac{5}{2} \end{cases}$$

DeepPoly relaxation

DeepSRGR to improve DeepPoly

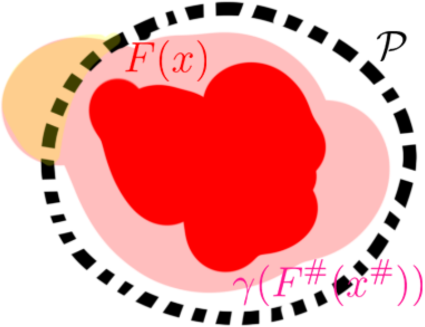
Main idea : remove *spurious* regions



DeepPoly relaxation

DeepSRGR to improve DeepPoly

Main idea : remove *spurious* regions



References

- Gagandeep Singh, Timon Gehr, Markus Püschel, Martin Vechev 2019. *An Abstract Domain for Certifying Neural Networks*

- Matthew Mirman, Timon Gehr, and Martin Vechev. 2018. *Differentiable Abstract Interpretation for Provably Robust Neural Networks. In Proc. International Conference on Machine Learning (ICML)*. 3575-3583.

- Pengfei Yang, Renjue Li, Jianlin Li, Cheng-Chao Huang, Jingyi Wang, Jun Sun, Bai Xue, and Lijun Zhang, *Improving Neural Network Verification through Spurious Region Guided Refinement*, 2021