PROVABLY BOUNDING NEURAL NETWORK PREIMAGES

Many safety-critical applications require the computation of preimages of neural networks for a given output.
We present INVPROP, an algorithm to compute a **convex overapproximation** of these preimages that **does not require LP solvers** and can be executed using **GPUs**. The experimental evaluation demonstrates that some overapproximations are over **2500× tighter** and computed **2.5× faster** than in prior work.

BENEFITS

- Intermediate layer bounds can be optimized
- No LP solver required
- Full GPU support
- Iterative SGD optimization

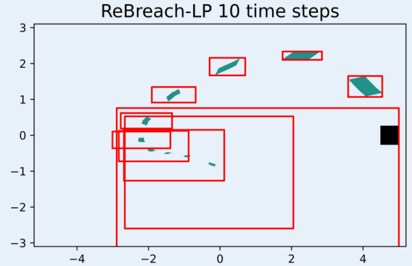
AUTHORS

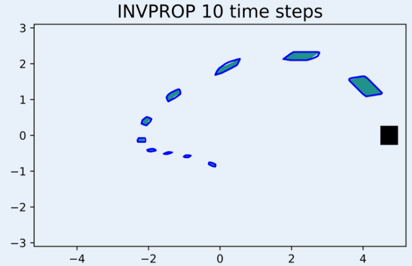
Suhas Kotha[1], Christopher Brix[2], Zico Kolter[1,4], Krishnamurthy (Dj) Dvijotham[3]*, Huan Zhang[1]*

[1] Carnegie Mellon University, Pittsburgh PA, 15213, USA.
[2] RWTH Aachen University, Aachen, 52056, Germany.
[3] Google Research, Brain Team.
[4] Bosch Center for AI.

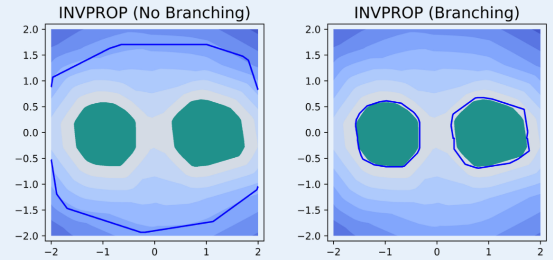
* These authors contributed equally to this work.

COMPARISON TO PRIOR SOTA





NON-CONVEX BOUNDS VIA INPUT BRANCHING



ORIGINAL PROBLEM

Which input bounds lead to outputs in the target area?

$$\min_{\mathbf{x}} \quad \mathbf{c}^\top \mathbf{x}$$
$$\text{s.t.} \quad \mathbf{x} \in \mathcal{X}; \quad \mathbf{H} f(\mathbf{x}) + \mathbf{d} \leq 0$$

DUALIZATION

The output constraint can be included in the objective
Inverting the order of min and max yields a lower bound

$$\max_{\boldsymbol{\gamma}} \min_{\boldsymbol{x}} \quad \mathbf{c}^\top \boldsymbol{x} + \boldsymbol{\gamma}^\top (\mathbf{H} f(\boldsymbol{x}) + \mathbf{d})$$
$$\text{s.t.} \quad \boldsymbol{x} \in \mathcal{X}; \quad \boldsymbol{\gamma} \geq 0$$

RELAXATION

The inner minimization is only constrained by x

$$\max_{\boldsymbol{\gamma}} \quad \text{AutoLiRPA}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$$
$$\text{s.t.} \quad \boldsymbol{x} \in \mathcal{X}; \quad \boldsymbol{\gamma} \geq 0; \quad 0 \leq \boldsymbol{\alpha} \leq 1$$

REFERENCES

Rober, N., Everett, M., Zhang, S., How, J.P.: A hybrid partitioning strategy for backward reachability of neural feedback loops. arXiv preprint arXiv:2210.07918 (2022)

Rober, N., Katz, S.M., Sidrane, C., Yel, E., Everett, M., Kochenderfer, M.J., How, J.P.: Backward reachability analysis of neural feedback loops: techniques for linear and nonlinear systems. arXiv preprint arXiv:2209.14076 (2022)

Xu, K., Shi, Z., Zhang, H., Wang, Y., Chang, K.-W., Huang, M., Kailkhura, B., Lin, X., Hsieh, C.-J.: Automatic perturbation analysis for scalable certified robustness and beyond. Advances in Neural Information Processing Systems 33 (2020)