ARENA: Enhancing Abstract Refinement for Neural Network Verification @SPLASH SRC 2022

Yuyi Zhong, Quang-Trung Ta and Siau-Cheng Khoo

{yuyizhong, taqt, khoosc}@comp.nus.edu.sg

National University of Singapore

Introduction

Robustness verification of neural network requires a balance between precision and efficiency.

We enhance the existing effectiveness of DeepSRGR [1] and propose an **abstract refinement** process that:

- Encode the network **more precisely** for precision improvement
- Eliminate **multiple** adversarial labels simultaneously for efficiency improvement

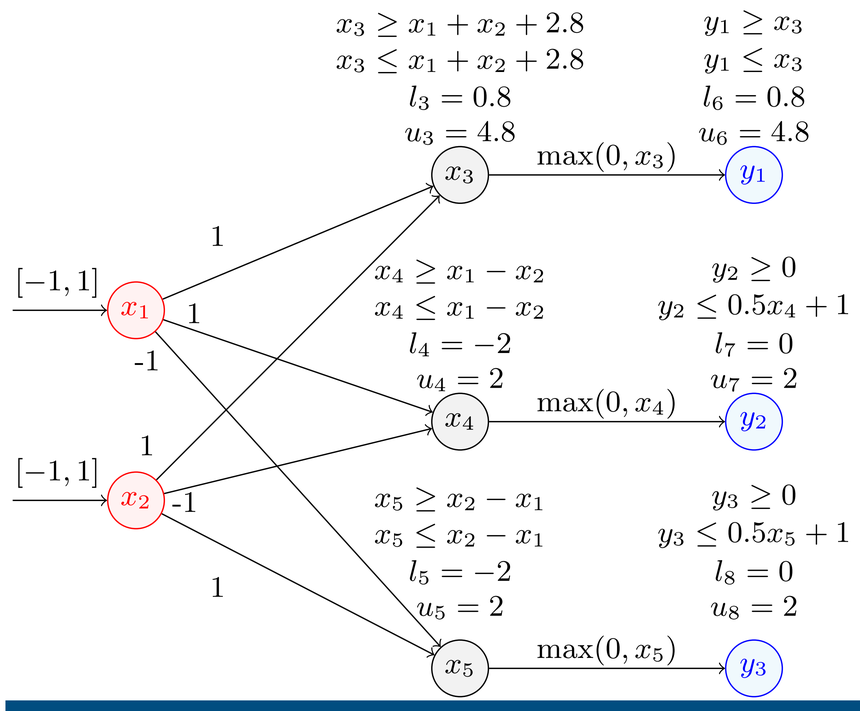
Contributions

To our best knowledge, our system outperforms existing approximated approaches running on CPU.

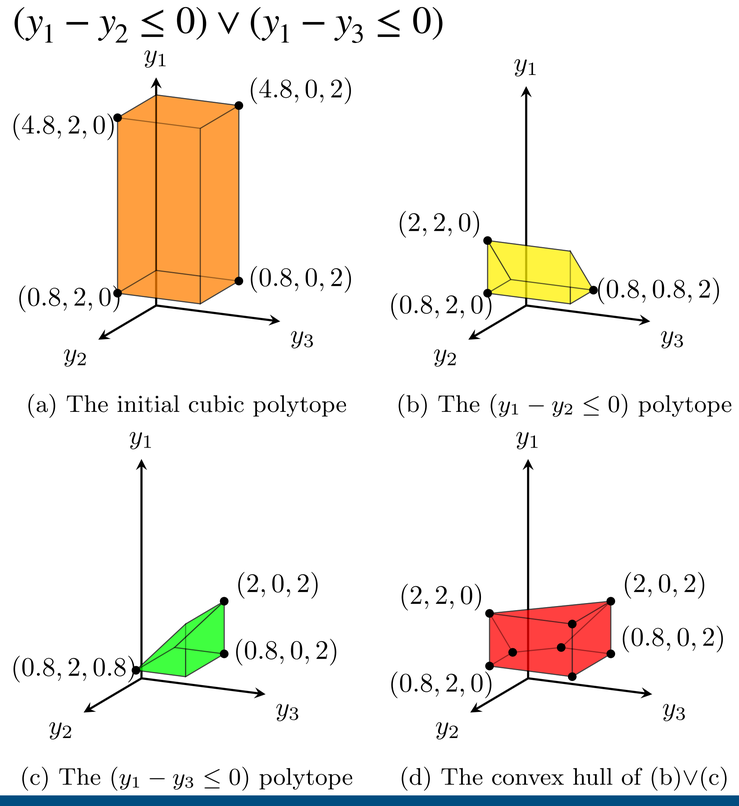
- It adapts double description method [2] to solve disjuncts of constraints in LP encoding, for system speed-up.
- It leverages more precise ReLU encoding in PRIMA [3] to increase precision
- It utilizes the solutions returned by the LP solver to detect adversarial examples

Illustrative Example

1. Abstract interpretation, DeepPoly [3]. Each neuron is represented by linear constraints.
Goal: $(y_1 - y_2 > 0) \wedge (y_1 - y_3 > 0)$



2. Add negated goals as constraints
$(y_1 - y_2 \leq 0) \vee (y_1 - y_3 \leq 0)$



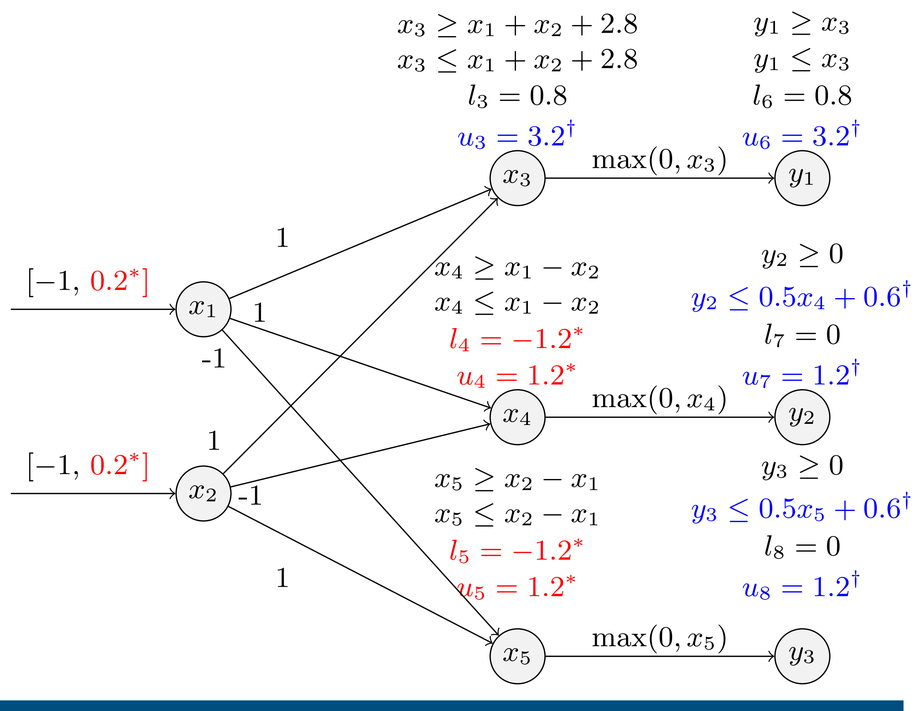
(a) The initial cubic polytope

(b) The $(y_1 - y_2 \leq 0)$ polytope

(c) The $(y_1 - y_3 \leq 0)$ polytope

(d) The convex hull of (b)∨(c)

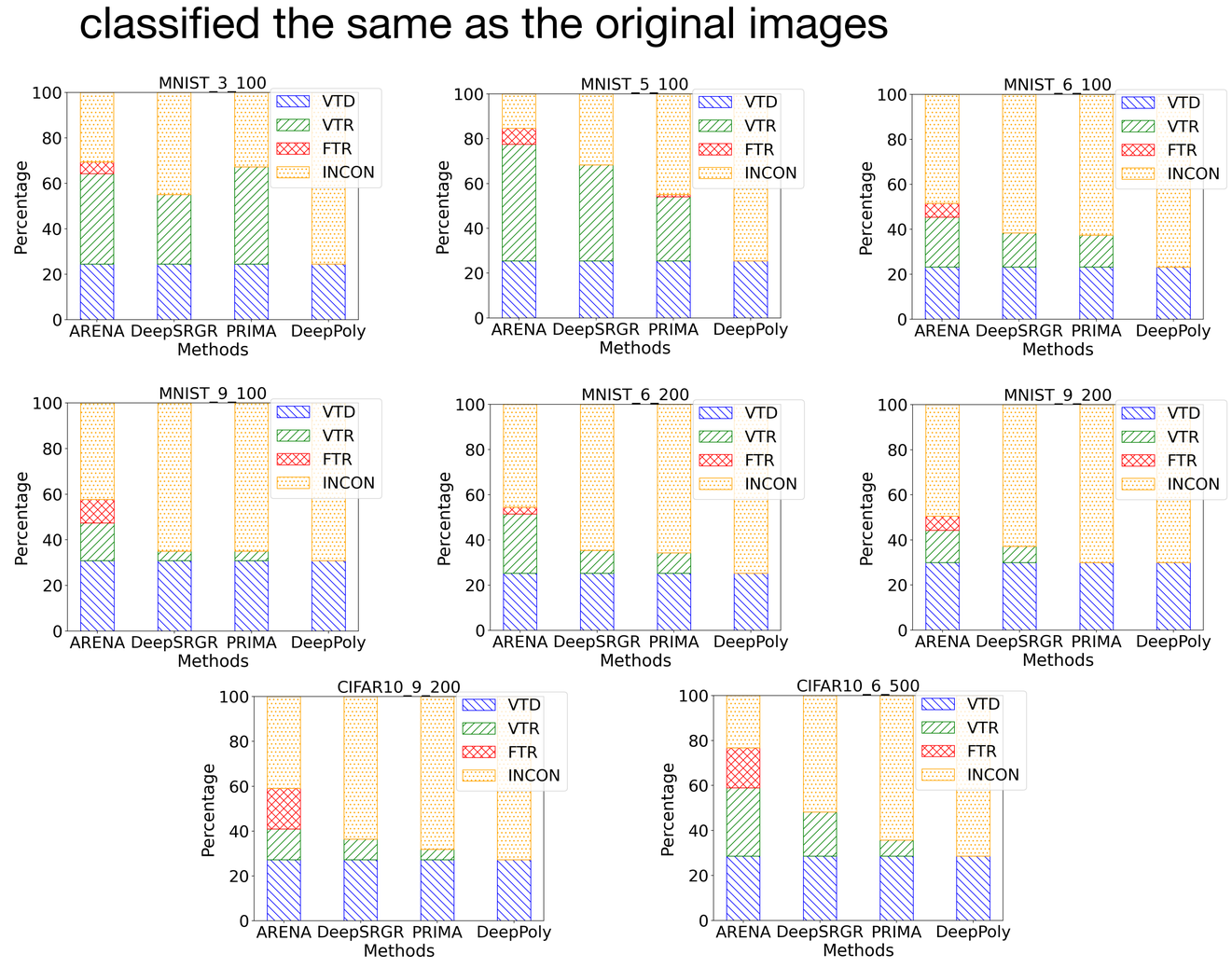
3. Resolve the neuron intervals with LP solver. The abstraction of the network is refined. Now $y_1 - y_2 \leq 0$ and $y_1 - y_3 \leq 0$ are infeasible



Experiments

- Our prototypical verifier ARENA is available online at https: //github.com/arena-verifier/ARENA
- The report is online at https://jacksonzyy.github.io/homepage/files/VMCAI_tech_report.pdf
- We compare ARENA with state-of-the-art verifiers including DeepSRGR [1], PRIMA [3], and DeepPoly [4]
- Robustness: all images perturbed within ϵ will be classified the same as the original images
- We eliminate 4 labels (batch size δ) at the same time
- On average, ARENA returns **18.7%** more conclusive images than DeepSRGR; **22.1%** more than PRIMA

Neural Net	ϵ	ARENA			DeepSRGR		PRIMA		DeepPoly	
		Verify	Falsify	Time	Verify	Time	Verify	Time	Verify	Time
MNIST_3_100	0.028	63	5	88.6	54	87.2	66	99.8	24	0.1
MNIST_5_100	0.08	76	7	227.7	67	203.2	53	13.0	25	0.9
MNIST_6_100	0.025	45	6	814.0	38	454.5	37	172.1	23	0.2
MNIST_9_100	0.023	46	10	2725.6	34	1248.7	34	158.1	30	0.8
MNIST_6_200	0.016	51	3	2430.0	35	1685.6	34	238.5	25	0.9
MNIST_9_200	0.015	43	6	6284.5	36	4383.8	29	271.6	29	2.1
CIFAR10_9_200	0.0011	9	4	6893.9	8	8192.6	7	478.9	6	10.6
CIFAR10_6_500	0.0032	33	10	4190.7	27	6531.3	20	410.2	16	26.5



Future Work

- **Accelerate** by handling output constraints by GPU, instead of costly LP solver
- **Explore** dynamic assignment of batch size δ

References

[1] Pengfei et al. "Improving Neural Network Verification through Spurious Region Guided Refinement." *TACAS* 2021.

[2] Komei et al. "Double Description Method Revisited." *Combinatorics and Computer Science 1995*

[3] Mark et al. "PRIMA: general and precise neural network certification via scalable convex hull approximations." *POPL* 2022

[4] Gagandeep et al. "An abstract domain for certifying neural networks." *POPL* 2019.