ARENA: Enhancing Abstract Refinement for Neural Network Verification

Yuyi Zhong Quang-Trung Ta Siau-Cheng Khoo
National University of Singapore

NUS
National University of Singapore

Contributions

- **Eliminate multiple adversarial labels for efficiency**
- **Multi-ReLU convex abstraction for precision**
- **Adversarial example detection for falsification**
- **A Tool - ARENA**
  - A CPU-based prototypical analyzer named ARENA (Abstract Refinement Enhancer for Neural network verificAtion)
- **Evaluation**
  - Improved precision vs. the state-of-the-art approximation based analyzers

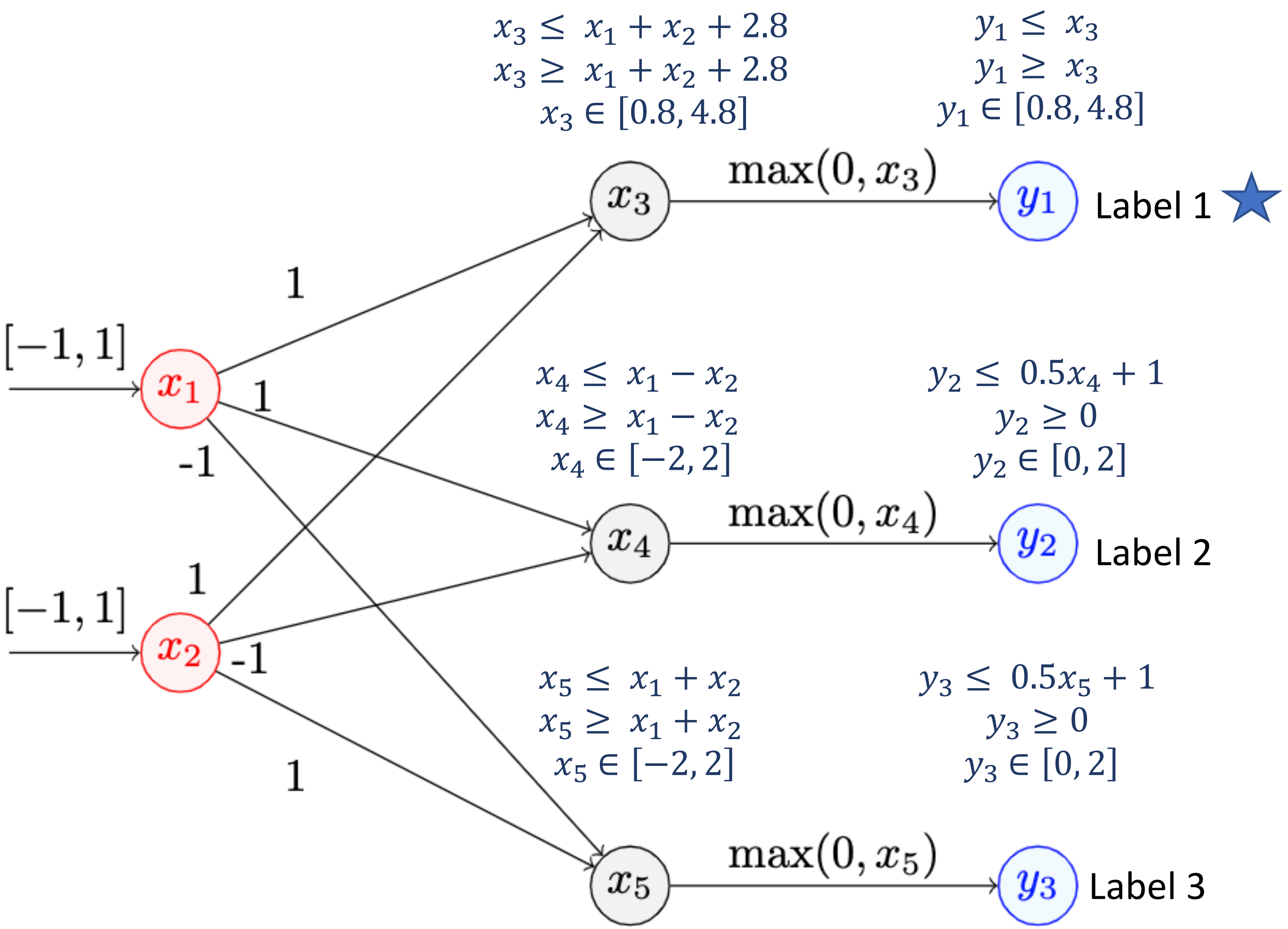
Robustness Analysis

- **Robustness analysis with perturbation ϵ**
  - The input space region $\mathbb{R}$: $\bigtimes_{i=1}^{n}[p_i - \epsilon, p_i + \epsilon]$
  - All inputs $\in \mathbb{R}$, all classified as ground truth label
- **Input**
  - Network to be verified, input space
- **Output**
  - Verified / Falsified / Inconclusive

Image A

Image B

Illustrative Example



- Initial abstract interpretation (abstract domain designed by DeepPoly)

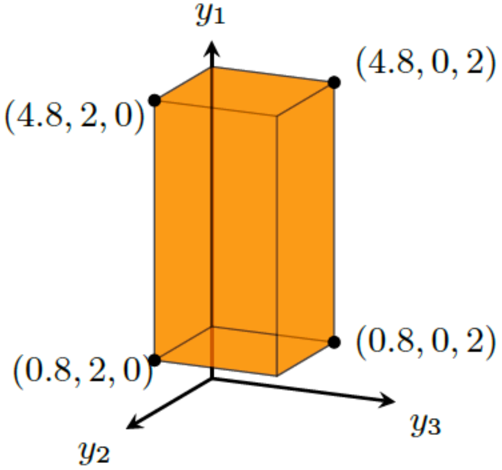
Verify Robustness

- To prove that label 1 is the dominant label
  - Goal: $y_1 - y_2 > 0 \wedge y_1 - y_3 > 0$
- From previous abstract values, we only obtain
  - $y_1 - y_2 \geq -0.2$ and $y_1 - y_3 \geq -0.2$
  - Fail to ascertain robustness, need to eliminate adversarial label 2 and 3
- Refine the abstraction using linear programming (LP)
  - Constraint set = symbolic constraints of all neurons (network encoding Π) + $(y_1 - y_2 \leq 0 \vee y_1 - y_3 \leq 0)$
  - $\Pi \wedge ((y_1 - y_2 \leq 0) \vee (y_1 - y_3 \leq 0))$ encodes the existence of adversarial examples
  - Objective function = maximize/minimize neurons
  - LP solver returns tighter bounds, leading to a better abstraction
  - The tighter abstraction shows the existence of adversarial examples to be spurious

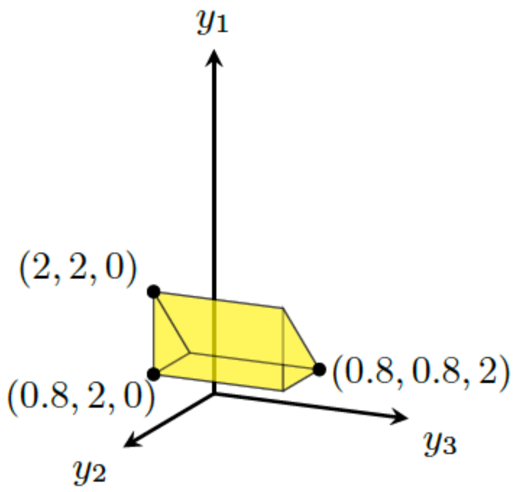
Handle Disjunction in LP

- Constraint set $\Pi \wedge ((y_1 - y_2 \leq 0) \vee (y_1 - y_3 \leq 0))$
- Linear programming does not naturally support the disjunction of linear inequalities
- To address the challenge, we compute the over-approximate convex hull P of $(y_1 - y_2 \leq 0) \vee (y_1 - y_3 \leq 0)$ under Π
- A convex hull is represented as a set of linear inequalities, LP is amenable to handle $\Pi \wedge P$
- Leverage double description method to compute the convex hull P

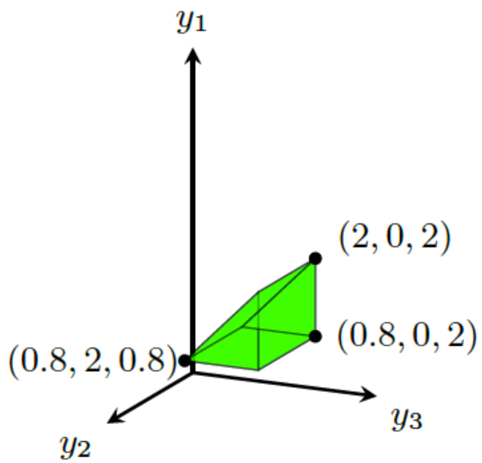
Convex Hull Computation



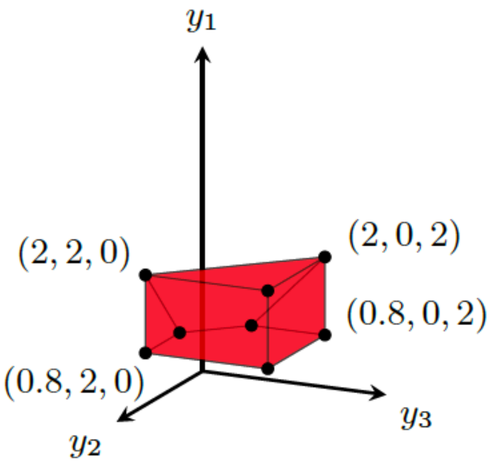
(a) The initial cubic polytope under Π



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



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

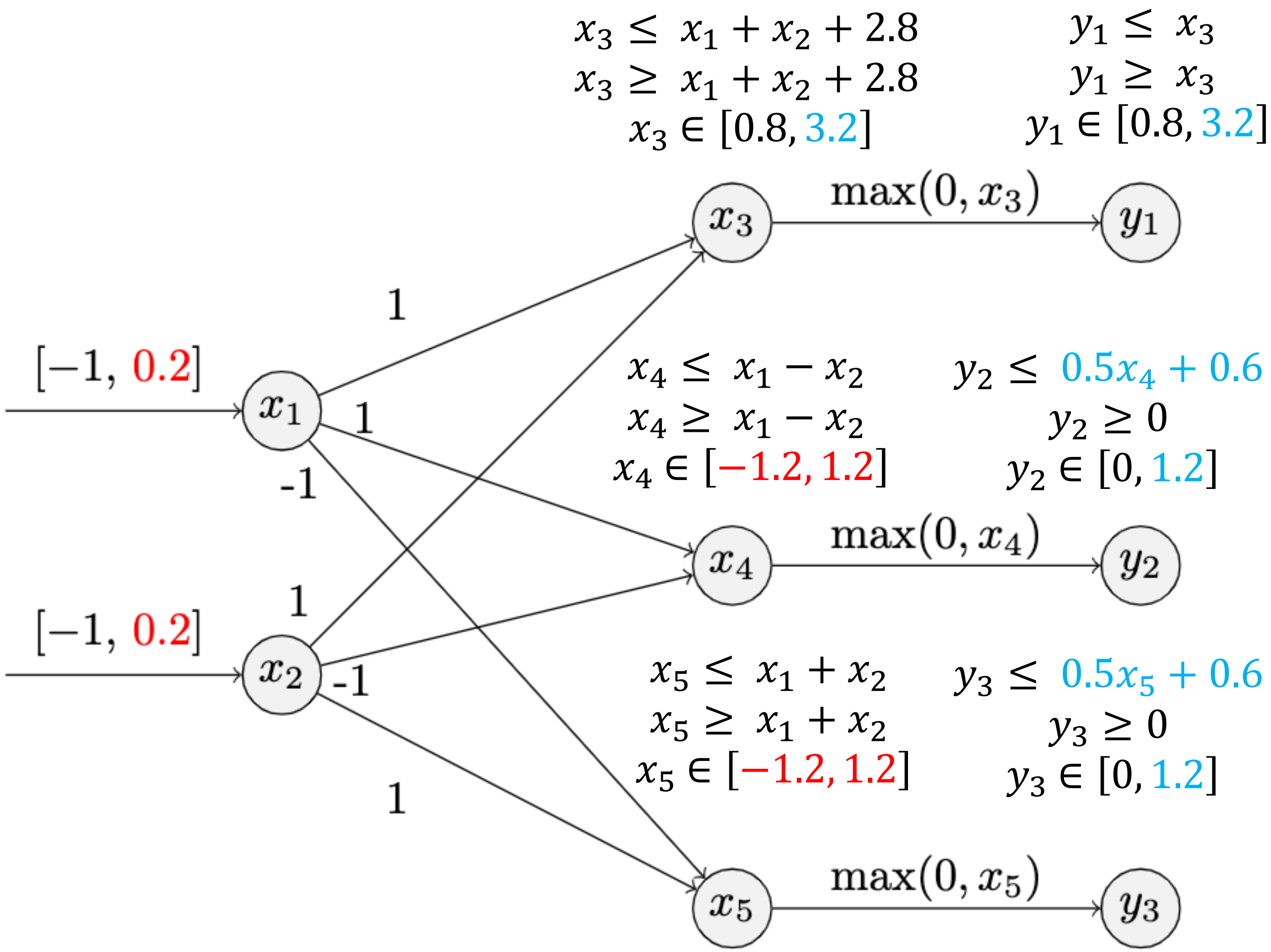


(d) The convex hull of union of (b),(c)

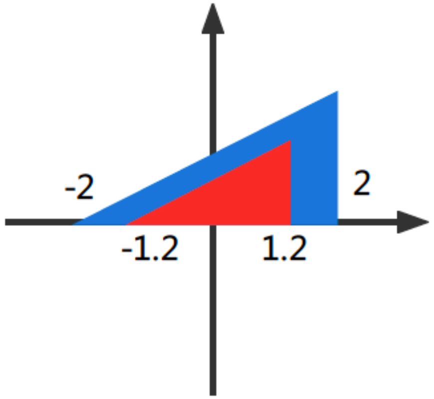
- The convex hull $(y_1 - y_2 \leq 0) \vee (y_1 - y_3 \leq 0)$ under Π
- The convex hull P is defined as:
  - $-y_1 + y_2 + y_3 \geq 0$
  - $y_2 \geq 0$
  - $y_3 \geq 0$
  - $-1 + 1.25y_1 \geq 0$
  - $2 - y_1 \geq 0$
  - $2 - y_2 \geq 0$
  - $2 - y_3 \geq 0$
- The constraint set $\Pi \wedge P$ is now fully conjunct

Refined Abstraction via LP Solving

- Resolve input or unstable ReLU input intervals via LP solver



- Based on the new interval, the abstract values of other neurons are updated
- The abstraction of ReLU neuron is refined



Multiple Adversarial Labels Elimination

- Given the refined abstraction
- $y_1 - y_2 \geq 0.2$ and $y_1 - y_3 \geq 0.2$
- Making adversarial label 2 and 3 infeasible
- Label 1 is the dominant label, and robustness verified

System ARENA

- **Multiple adversarial labels elimination**
  - Encode multiple adversarial labels in LP solver and resolve tighter abstraction
  - Tighter abstraction leads to infeasibility of adversarial labels
- **More precise ReLU encoding**
  - Adopt multi-ReLU convex abstraction in PRIMA, capturing the dependencies among ReLU neurons
- **Adversarial example detection**
  - A feasible constraint set indicates the possibility of a property violation.
  - Check if the optimal solution from the LP solver constitutes an adversarial example

Evaluation

- MNIST/CIFAR10 test set; Compare with PRIMA, DeepSRGR, DeepPoly

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

- On average, ARENA returns 18.7% more conclusive images than DeepSRGR; 22.1% more than PRIMA.

Verification Time

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

- Not competitive in time
- We use the costly LP solver on CPU
- Future work: the solving process can be implemented on GPU

Key Takeaways

- **ARENA: enhanced abstract refinement**
  - Eliminate multiple adversarial labels
  - Conduct counterexample detection
  - Achieve improved precision

- **Online resources**
  - GitHub repo:
    https://github.com/arena-verifier/ARENA
  - Full paper/report :
    https://jacksonzyy.github.io/homepage/files/VMCAI_tech_report.pdf

Thank you!

Q & A

Backup1, Single-ReLU Encode

- Independent ReLU encode without considering the relationship of neurons in the same layer

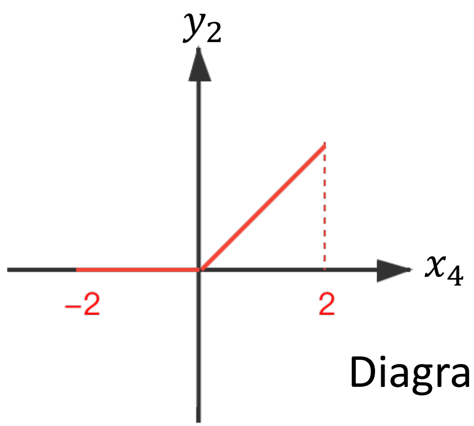


Diagram A

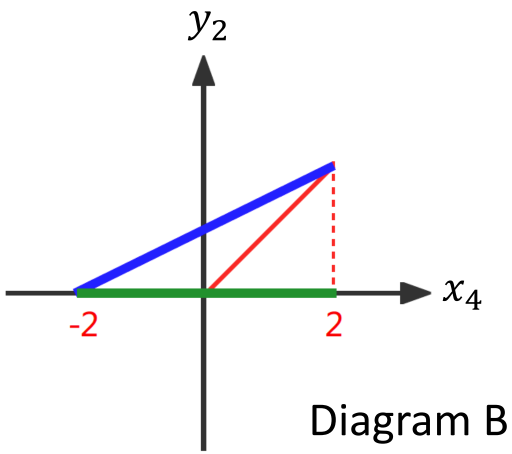
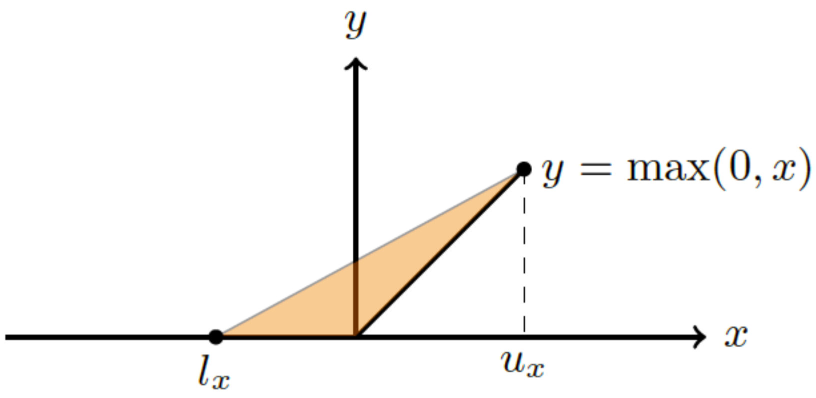


Diagram B

Backup1, Multi-ReLU Encode



- Multi-ReLU relaxation captures the dependency, and computes the convex abstraction of k-ReLU neurons via novel convex hull approximation algorithms.
- $k = 2$, ReLU neurons y_1, y_2 with inputs x_1, x_2, get a convex hull in $(y_1, y_2\ x_1, x_2)$ space
- $\{x_1 + \ x_2 - 2y_1 - 2y_2 \geq -2,\ \ y_1 \geq 0,\ \ \ y_2 \geq 0,\ \ -x_1 + y_1 \geq 0, -x_2 + y_2 \geq 0,$ $0.375x_2 - y_2 \geq -0.75\}$
- $x_1 + \ x_2 - 2y_1 - 2y_2 \geq -2$ correlates $(y_1, y_2\ x_1, x_2)$ all together, which is beyond the single ReLU encoding.