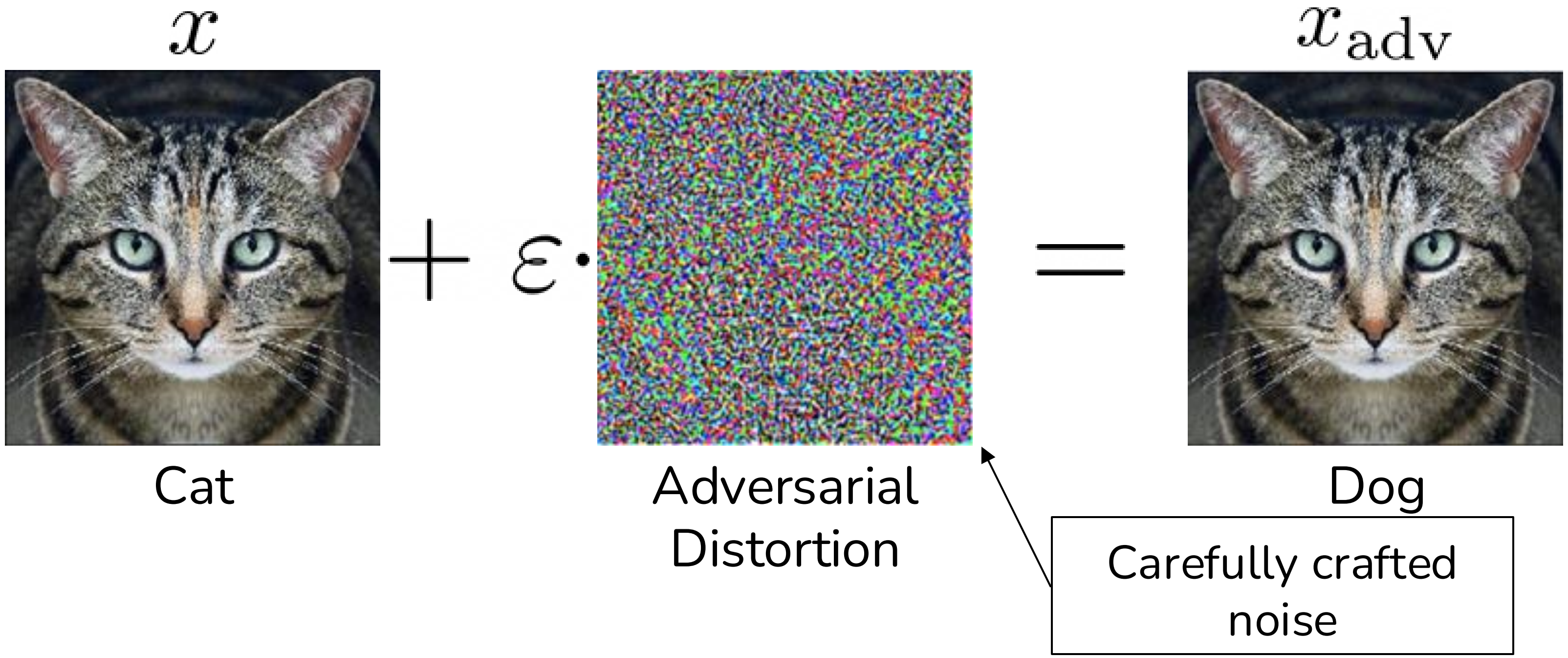
SoK
Certified Robustness for Deep Neural Networks

Linyi Li (UIUC), Tao Xie (Peking University), Bo Li (UIUC)

UIUC Secure Learning Lab

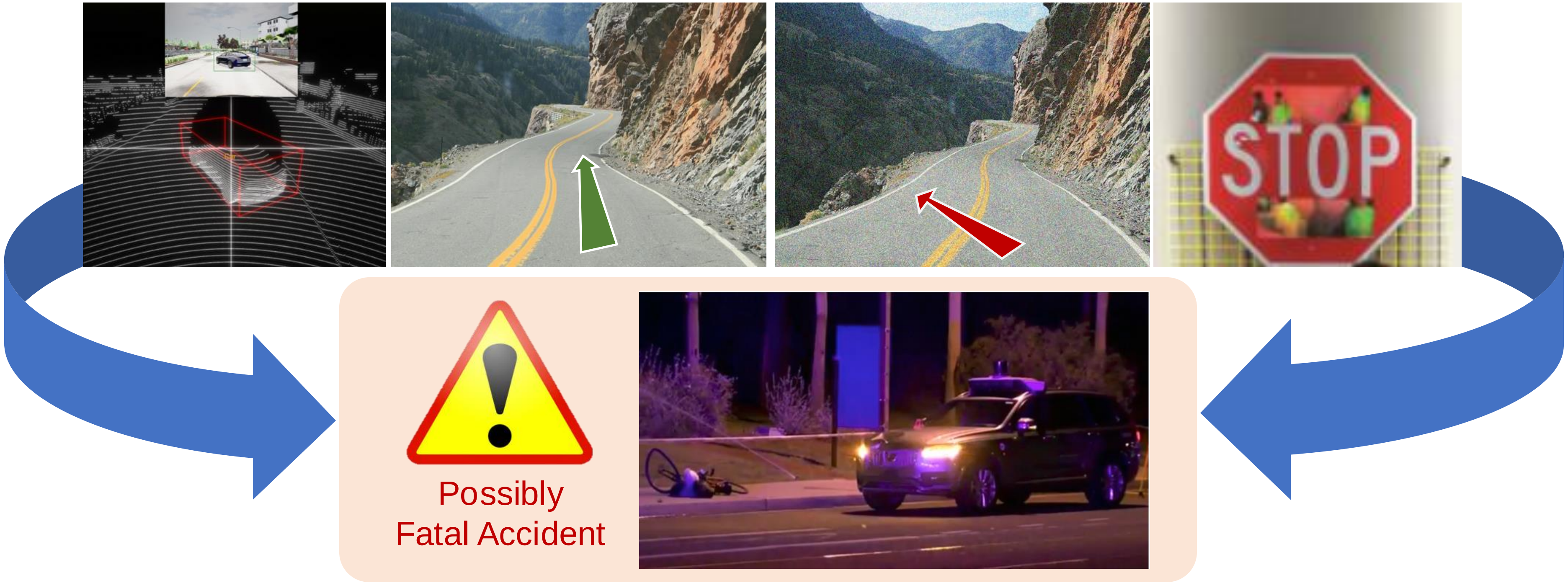
Adversarial Robustness – A Lasting Threat

- Deep neural networks (DNNs) can be easily fooled by adversarial examples
  - **Tiny** crafted perturbations can make DNNs give **wrong** predictions



Severe Safety Threats

Example: autonomous driving



Possibly
Fatal Accident

Arm Race

Defenses bypassed by follow-up attacks

discovery

[Szegedy et al. '14]

Ending Arm Race? Certified Robustness!

- Prove adversarial example doesn’t exist

Guarantee of Safety

Ending Arm Race? Certified Robustness!

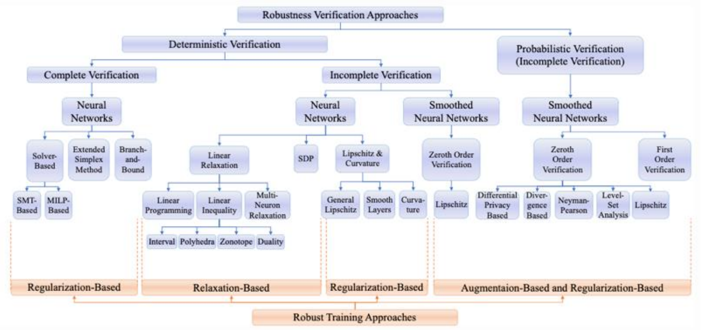
- Prove adversarial example doesn’t exist

Guarantee of Safety

First Systematization of Knowledge on Certified Robustness for Deep Neural Networks!

Content

TAXONOMY



SUMMARY

- Characteristics
- Strengths
- Limitations
- Connections
- Generalization
- …

DISCUSSION

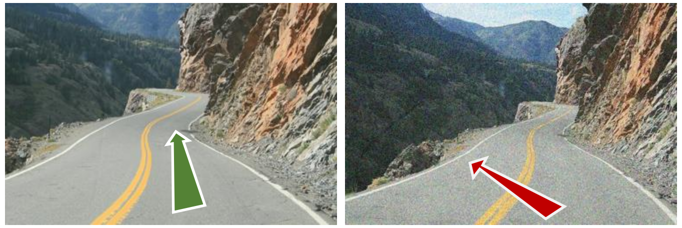
- Current Research
- Theoretical Barriers
- Main Challenges
- Future Directions
- …

BENCHMARK

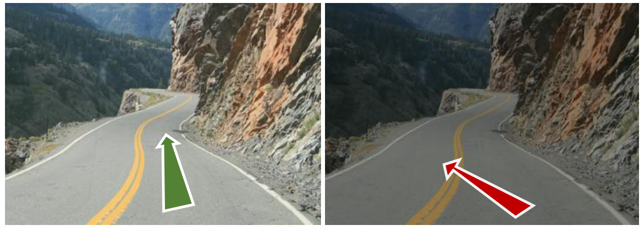
VeriGauge
Open-source platform for 20+ approaches

Threat Model

- Various types of adversarial examples exist



Tiny Perturbations



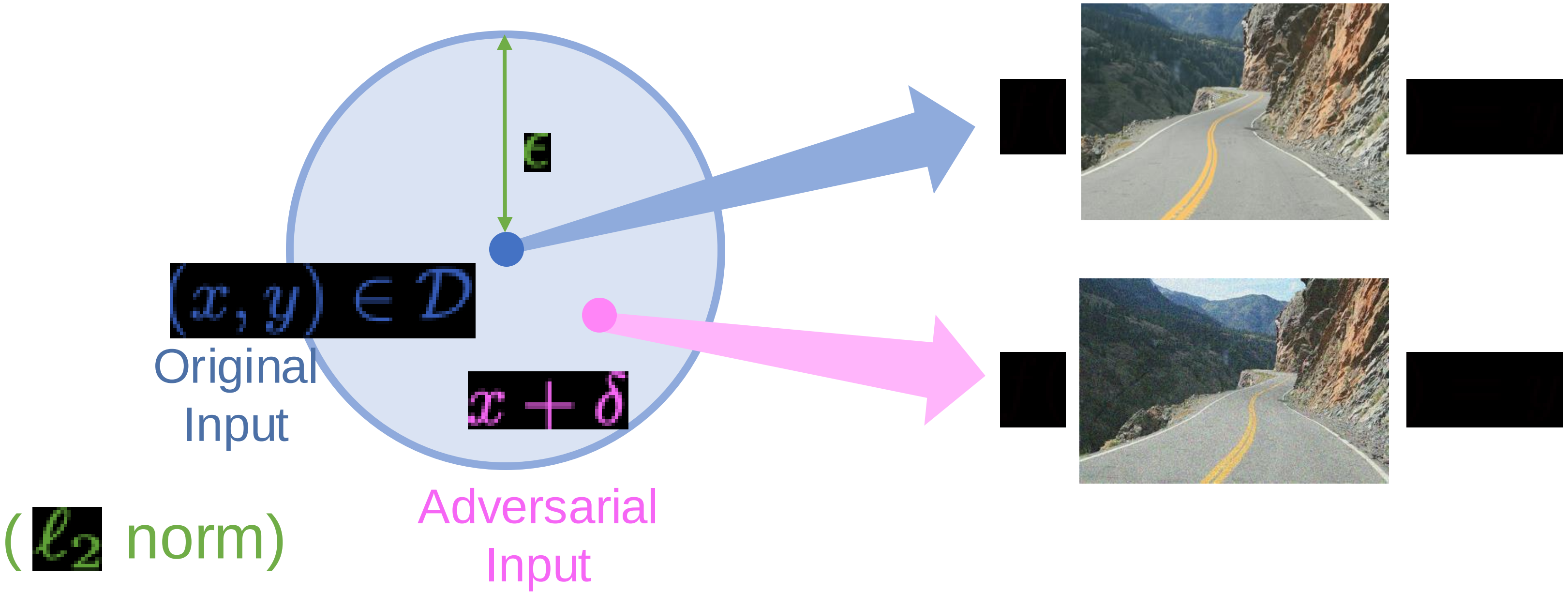
Semantic Transformations

- Focus on (p-norm constrained) perturbations
  - Widely studied
  - Techniques generalizable to other types of adversarial examples

Robustness against (p-Norm Constrained) Perturbations

Given a DL model [redacted], finite test dataset [redacted]

Check: $\forall \delta, \|\delta\|_p \leq \epsilon, f(x+\delta) = y$

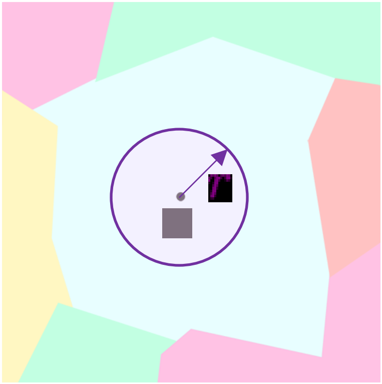


Formal Definition of Certification

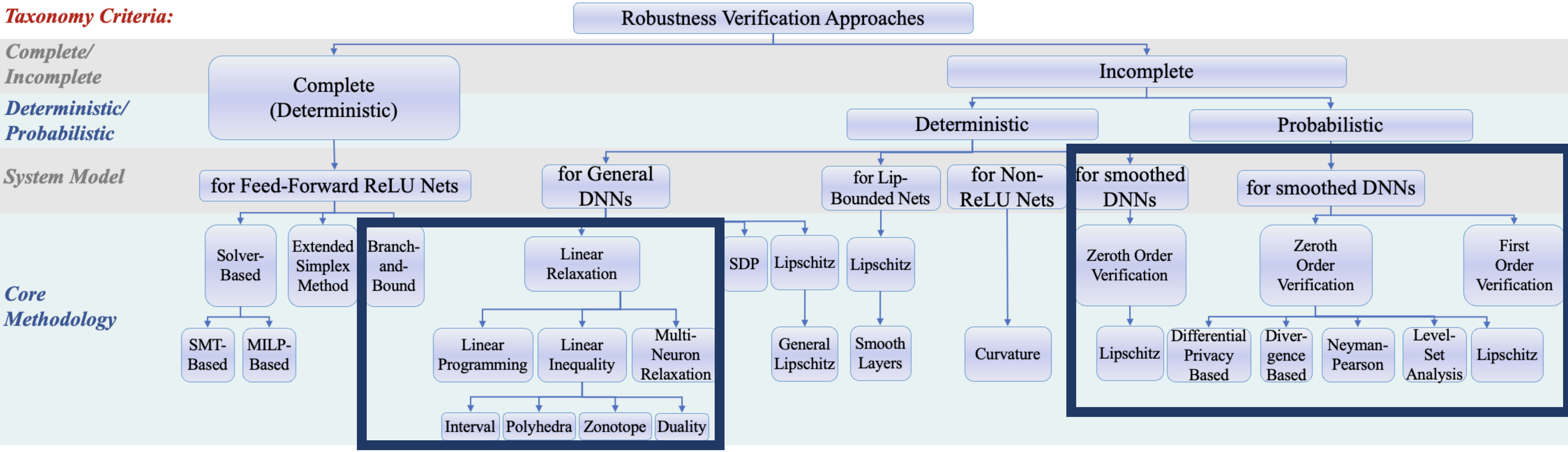
For given system [redacted] and data instance [redacted] with true label [redacted], compute larger [redacted], such that

[redacted]

Larger certified radius
= Tighter certification
= Better certified robustness



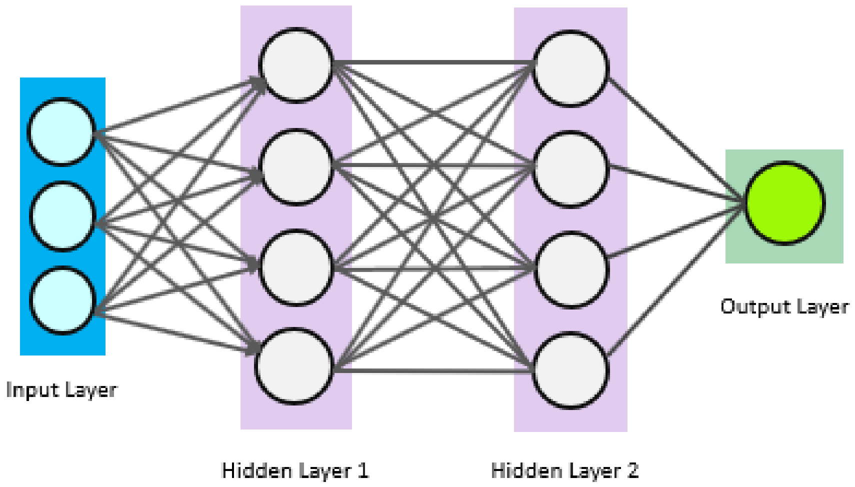
Taxonomy of Verification Method



SOTA determinisitc certified robustness for general DNNs

SOTA probabilistic certified robustness
Only method supporting large models

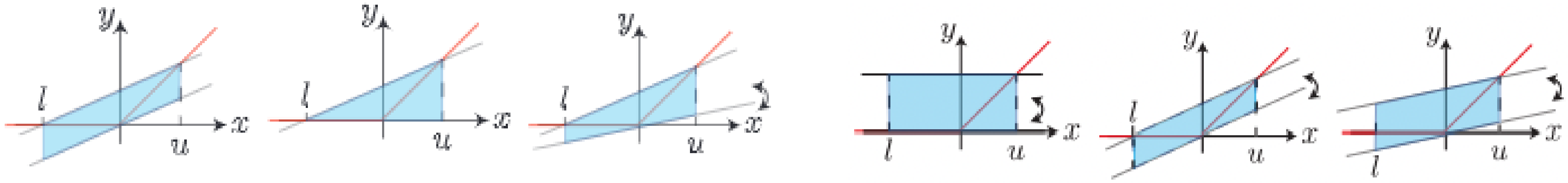
DNN Architecture



- **Input layer**: vector x_0
- **Weights**: $(W_0, b_0), (W_1, b_1), \dots (W_{L-1}, b_{L-1})$.
- **Activation function**:
  - $\text{ReLU}(x) = \max\{x, 0\}$
- **Computation**:
  - $x_1 = \text{ReLU}(W_0 x_0 + b_0)$,
  - $x_2 = \text{ReLU}(W_1 x_1 + b_1)$,
  - $\dots$
  - $x_L = W_{L-1} x_{L-1} + b_{L-1}$
- **Output**: x_L - confidence score for each class

https://www.educba.com/deep-learning-technique/

Linear Relaxation of ReLU

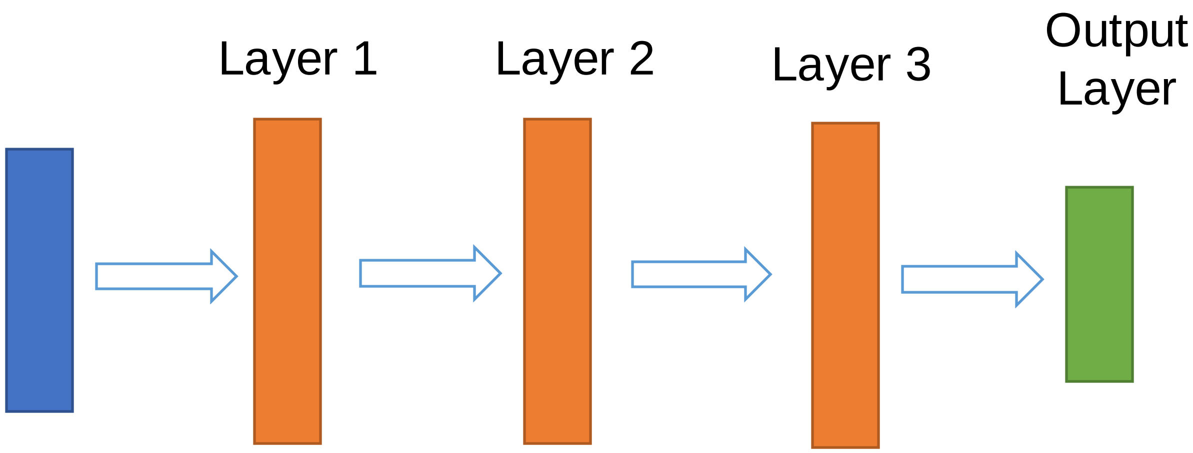


Different Linear Relaxations for ReLU$(x) = \max\{x, 0\}$

Weng, Lily, et al. “Towards fast computation of certified robustness for relu networks.” ICML 2018
Wong, Eric, and Zico Kolter. “Provable defenses against adversarial examples via the convex outer adversarial polytope.” ICML 2018
Singh, Gagandeep, et al. “Fast and Effective Robustness Certification.” NIPS 2018

Linear Relaxation Induces Linear Inequality

Propagate linear inequalities that bound possible output region



Linear Bound Propagation	$[x, x]$	$[L_1 x + l_1,$ $U_1 x + u_1]$	$[L_2 x + l_2,$ $U_2 x + u_2]$	$[L_3 x + l_3,$ $U_3 x + u_3]$	$[L_4 x + l_4,$ $U_4 x + u_4]$

Combat Over-Relaxation with Branch-and-Bound

Conditioned on two branches: $x \leq 0$ and $x > 0$,

each ReLU neuron is reduced to **linear constraints**: $y = 0$ or $y = x$

Select some neurons to condition on, and solve two subproblems

- If `x<=0, y=0` (linearized subproblem)
- If `x> 0, y=x` (linearized subproblem)
- Relax other neurons by linear relaxation and bound propagation

- **Most scalable verification method so far**

Randomized Smoothing

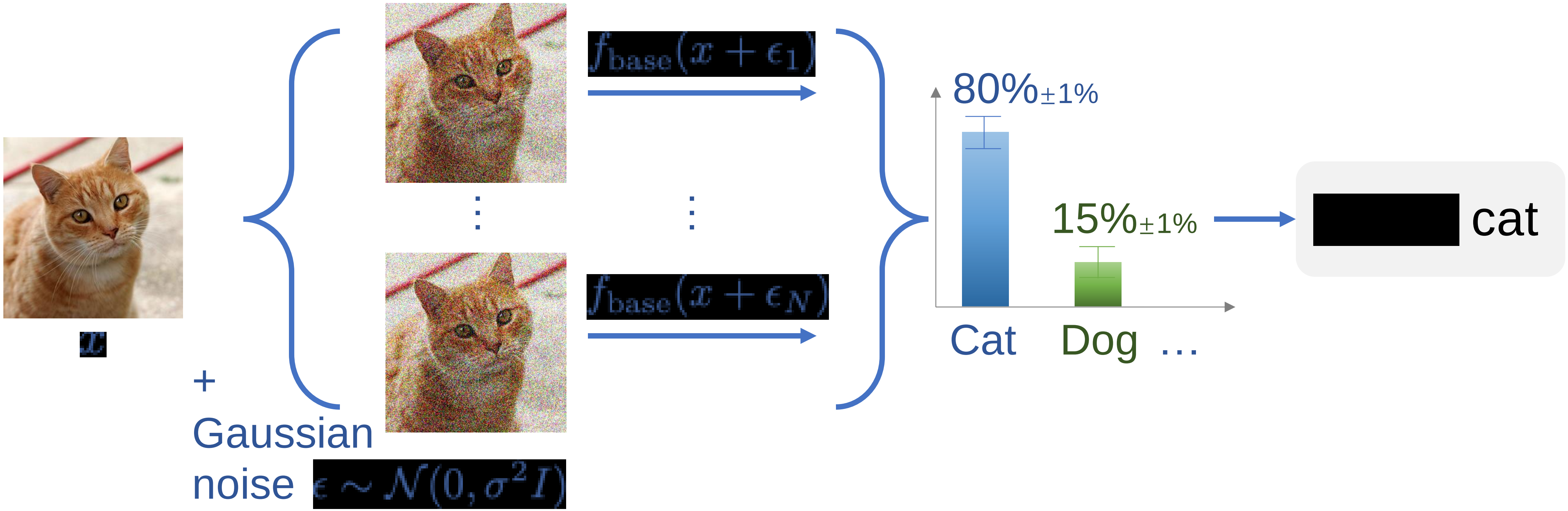
1. Train a model f_{base} (“base classifier”) under some known noise
2. Smooth f_{base} into a new classifier f (“smoothed classifier”), such that

   $f(x)$ = the most probable prediction by f_{base} under noised corruptions of ■

In deployment, use smoothed classifier f

Illustration of Randomized Smoothing

[redacted]



x

+ Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

$f_{\text{base}}(x + \epsilon_1)$

⋮

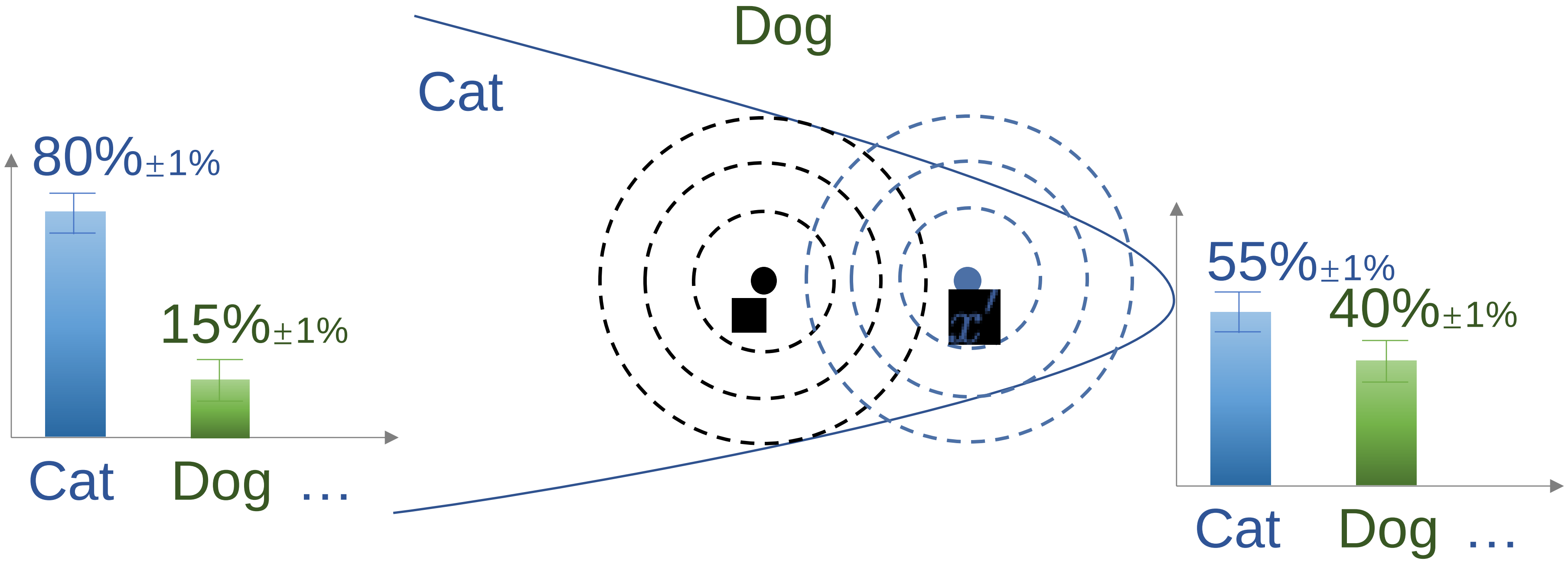
$f_{\text{base}}(x + \epsilon_N)$

[redacted] cat

Randomized Smoothing Enables Certified Robustness

Shift center of the distribution cannot change probability much

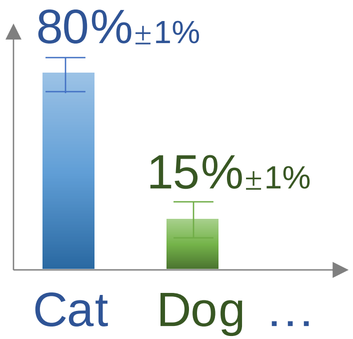
- If order doesn’t change, then consistent prediction guaranteed



$f(x') =$ cat

Closed-form Robustness Guarantee

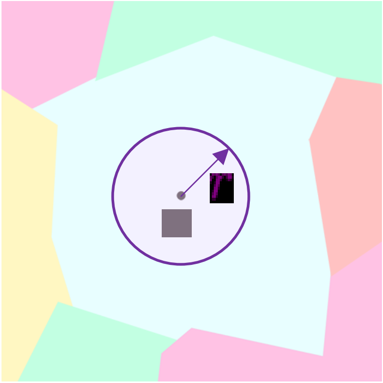
- [redacted] ’s probability of the top class (cat)
- [redacted] ’s probability of the runner-up class (dog)



f certifiably returns top class within an ℓ_2 ball around x of radius

$$r = \frac{\sigma}{2}\left(\Phi^{-1}(P_A) - \Phi^{-1}(P_B)\right)$$

- σ : variance of Gaussian smoothing noise
- Φ^{-1}: the inverse standard Gaussian CDF



Jeremy M Cohen, Elan Rosenfeld, J. Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. ICML 2019

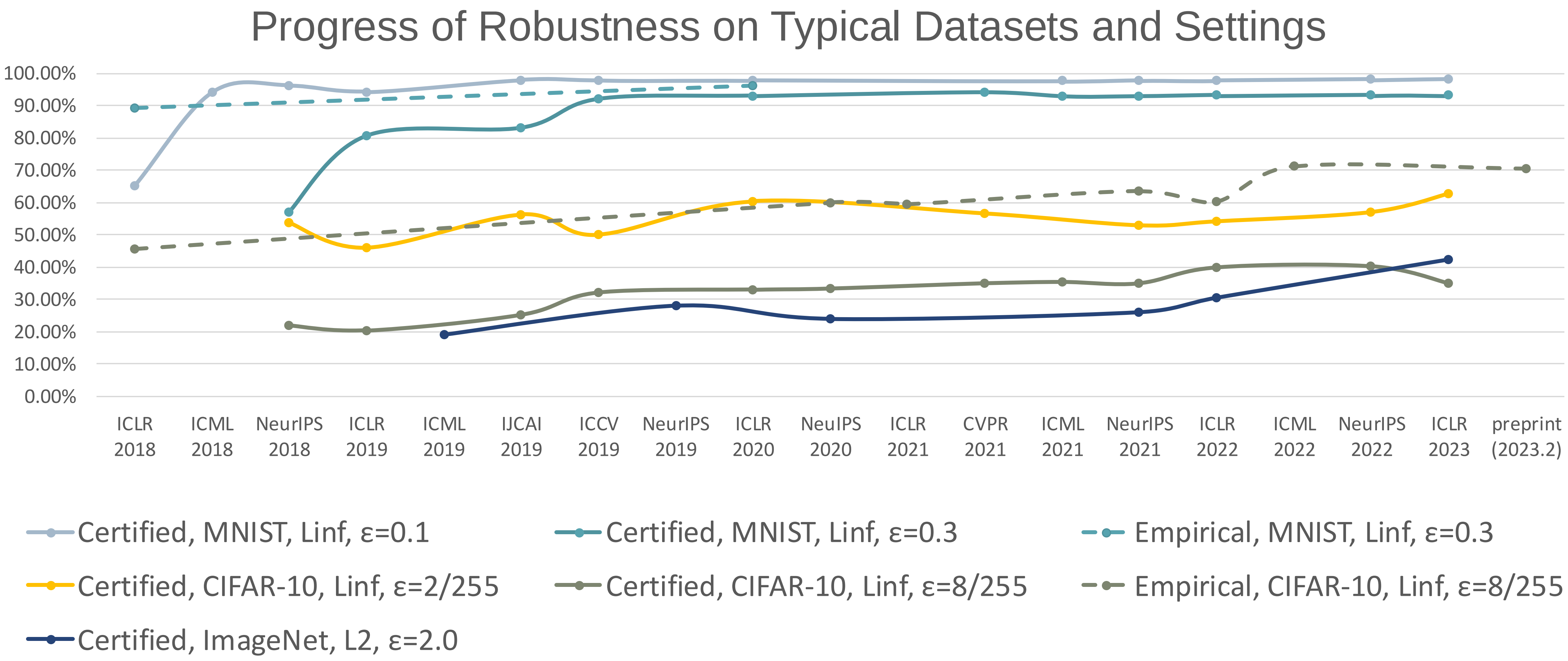
Certification Induces Robust Training

- Training DNN in specific ways can improve certified robustness

For linear relaxation + branch-and-bound:
training to **reduce upper bound of loss function** computed from over-approximation

For randomized smoothing:
training to **predict correctly for noised inputs**

How Far Are We on Real-World Datasets?



On MNIST

ℓ_∞ norm, $r = 0.3$

- SOTA Certified Robust Accuracy: **94.02%**
  - *[CVPR 2021] Towards Evaluating and Training Verifiably Robust Neural Networks*
- SOTA Empirical Robust Accuracy (against existing attacks): **96.34%**
  - https://github.com/MadryLab/mnist_challenge
  - *Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples*
  - *ArXiv: 2010.03593*

➢Not much difference

On CIFAR-10

ℓ_∞ norm, $r = 8/255$:

- SOTA Certified Robust Accuracy: **40.39%**
  - *[NeurIPS 2022] Rethinking Lipschitz Neural Networks and Certified Robustness: A Boolean Function Perspective*
- SOTA Empirical Robust Accuracy (against existing attacks): **71.29%**
  - *[ICML 2022] Diffusion Models for Adversarial Purification*

➢Still a gap

On ImageNet

ℓ_2 norm, $r = 2.0$

- SOTA Certified Robust Accuracy: **30.4%**
  - *Our paper at [ICLR 2022] On the Certified Robustness for Ensemble Models and Beyond*
- SOTA empirical robustness accuracy: **43.18%**
  - Against ℓ_∞ norm, $r = \frac{4}{255}$
  - *[ICML 2022] Diffusion Models for Adversarial Purification*

- Hard to achieve robustness

Key Messages

- Since 2017, **many methods proposed** to provide & improve DNN certified robustness
  - Linear relaxation
  - Branch-and-bound
  - Randomized smoothing
  - …
- **Remarkable certified robustness** achieved on small datasets, **but still challenging** on large ones
  - Good on MNIST
  - To be improved on CIFAR-10 and ImageNet

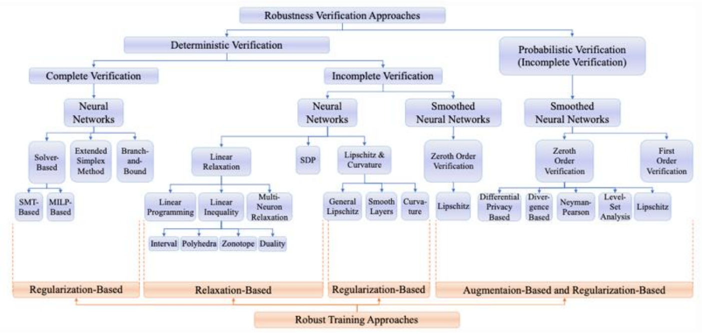
- Certification for p-norm bounded adversary **generalizable** for other threat models
  - Semantic adversary
  - Patch adversary
  - Word substitution adversary
  - Control state perturbation
  - Poisoning attack
  - …

☆ 96 10
☆ 73 6

sokcertifiedrobustness.github.io

SoK: Certified Robustness for Deep Neural Networks

TAXONOMY



SUMMARY

- Characteristics
- Strengths
- Limitations
- Connections
- Generalization
- …

DISCUSSION

- Current Research
- Theoretical Barriers
- Main Challenges
- Future Directions
- …

BENCHMARK

VeriGauge

Open-source platform for 20+ approaches