UIUC Secure Learning Lab

TSS
Transformation-Specific Smoothing
for Robustness Certification

Linyi Li*, Maurice Weber*, Xiaojun Xu, Luka Rimanic,
Bhavya Kailkhura, Tao Xie, Ce Zhang, Bo Li

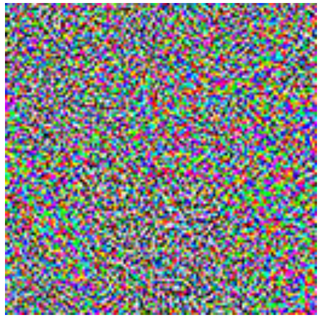
ILLINOIS ETHzürich NSF

Neural Networks are Vulnerable to Adversarial Attacks

- *W.l.o.g, consider image classification problem*
- Given an image as input, ML model predicts a class label
- However, attacker can usually craft adversarial input:
  - Indistinguishable from original input
  - But fool NN to make wrong prediction

Predicted as "cat"

+ 0.001 *



Small Perturbation

=

Predicted as "dog"

Szegedy, Christian, et al. "Intriguing properties of neural networks." ArXiv:1312.6199.

Adversarial Attack via Semantic Transformations

- Certifying and improving robustness for ML models against ℓ_p bounded perturbations is well-studied
  - *Clean input* $= x_0$
  - *Attacker needs to input* x *s.t.* $\|x - x_0\|_p \leq \epsilon$
- However, in the real-world, attacker can also apply semantic transformations (e.g., brightness, rotation, scaling) to fool ML models

(a) Input 1 **(b) Input 2 (darker version of 1)**

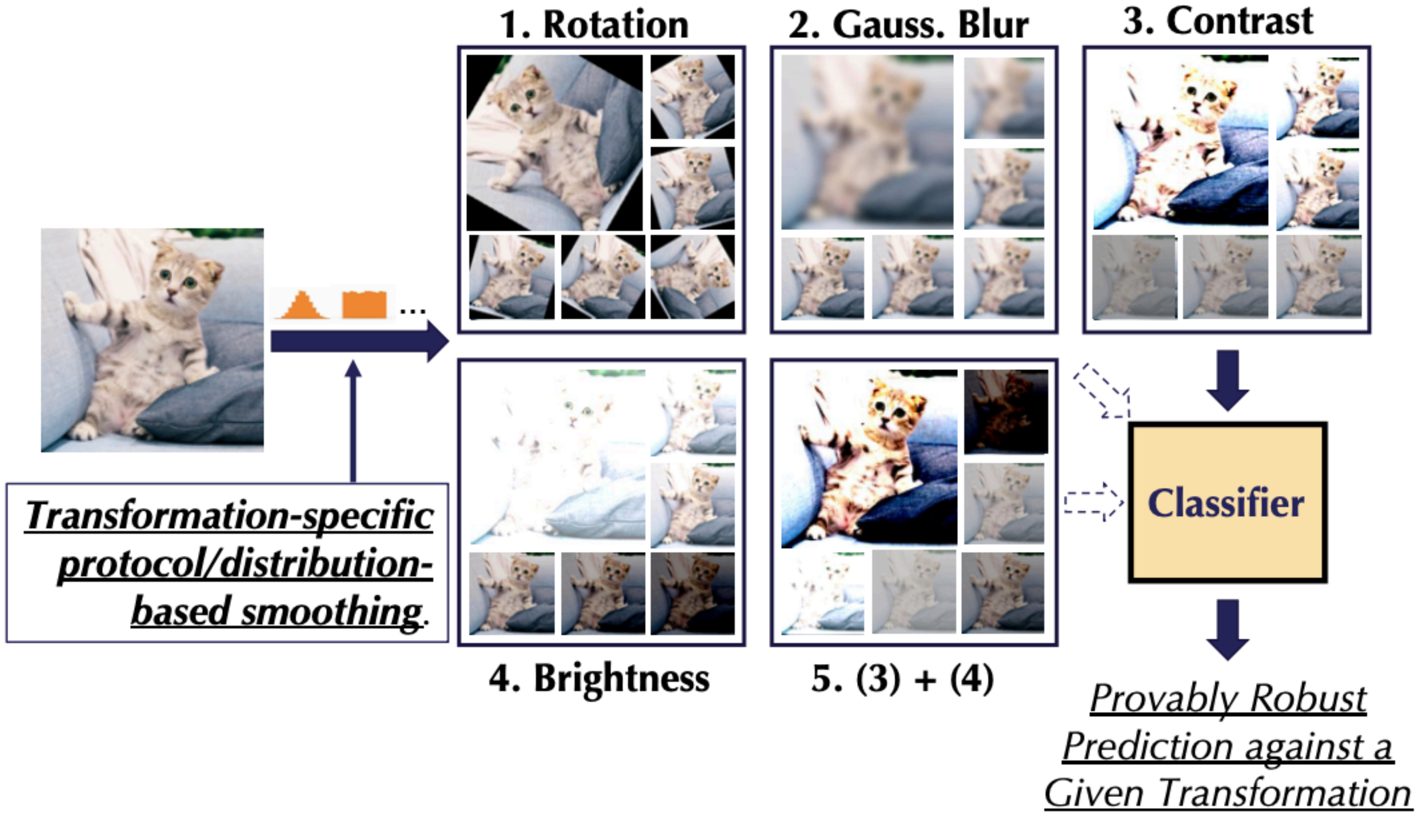
Adversarial examples found on Nvidia DAVE-2 self-driving car platform by DeepXplore

Pei, Kexin, et al. "Deepxplore: Automated whitebox testing of deep learning systems." *SOSP 2017.*

Can we get ML models that are certifiably robust to various semantic transformations?

Certify Robustness against Semantic Transformations

- We propose a **framework for certifying ML robustness against semantic transformations: TSS**



Compared with Existing Work

- Existing certified robustness methods:
  - Too **loose** on small models
  - Too **slow** for large models
  - Too **specific** for certain transformations
- Our work:
  - **Tight**: achieves state-of-the-art certified accuracy
  - **Scalable**: for the first time, achieve certified robustness on ImageNet
    - 30.4% certified accuracy against arbitrary rotation within 30°
  - **General**: general methodology for analyzing and certifying against transformations
    - Support > 10 common transformations:
      - rotation, scaling, brightness, contrast, blur, …

- ➢ Threat Model & Certification Goal
- Challenges
- Our Framework: TSS
- Experimental Evaluation

Threat Model

- Image classification task:
  - Input space: $\mathcal{X} \subseteq \mathbb{R}^d$
  - Output space: $\mathcal{Y} = \{1, \dots, C\}$
- Semantic transformation as a function

$$\phi: \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$$

  - Parameter space: $\mathcal{Z} \subseteq \mathbb{R}^m$
- Attacker can:
  1. arbitrarity choose parameter $\alpha \in \mathcal{Z}$
  2. transform x to $\phi(x, \alpha)$
  3. input $\phi(x, \alpha)$ to the classifier

Example:
- $\phi_R(x, \alpha)$ rotates input image x by α degree clockwise
- Define $\mathcal{Z} = [-30°, 30°]$

➢ Attacker can arbitrarily rotate the image within 30°

Certification Goal

- For our classifier $h: \mathcal{X} \rightarrow \mathcal{Y} = \{1, \dots, C\}$
- Given clean input $x \in \mathcal{X}$
- Wish to find a set $\mathcal{S} \subseteq \mathcal{Z}$ such that we can guarantee

$$h(x) = h\big(\phi(x, \alpha)\big), \forall \alpha \in \mathcal{S}$$

Threat Model & Certification Goal

➢ Challenges

Our Framework: TSS

Experimental Evaluation

Real-Valued Parameter Space

- The parameter space is real-valued
- The input image space is real-valued

➢ Infinite possible inputs after transformation
➢ Cannot certify via enumeration

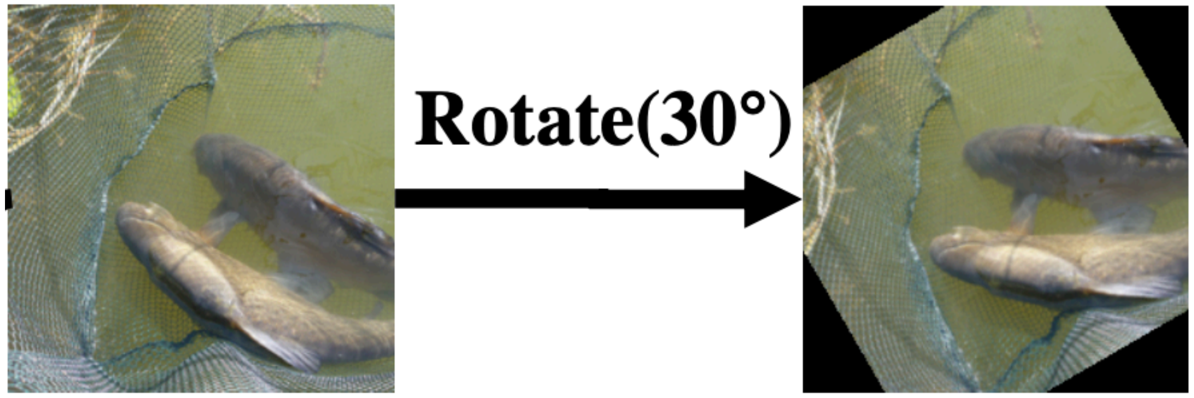
Large ℓ_p Difference

- Semantic transformation incurs large ℓ_p difference
  - Brightness +10% incurs ℓ_2 difference $0.1\times\sqrt{\#\ pixels} \approx 38.7$ on ImageNet

➢ Cannot certify with existing ℓ_p based methods

Interpolation

- Some transformations like rotation and scaling uses bilinear interpolation
- Certification needs to take complex interpolation effects into account



Threat Model & Certification Goal

Challenges

➢ Our Framework: TSS

- Generalized Randomized Smoothing
- TSS-R: Certifying Resolvable Transformations
- TSS-DR: Certifying Differentially Resolvable Transformations

Experimental Evaluation

Generalized Randomized Smoothing

- Given an arbitrary base classifier $h: \mathcal{X} \rightarrow \mathcal{Y} = \{1,2,\dots,C\}$
- Let $\phi(x, b) = x + b \cdot (1, \dots, 1)^{\mathrm{T}}$ be the brightness transformation
- Let $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ be the **smoothing distribution**
- Define $q(y|x;\varepsilon) = \Pr_{\varepsilon}\left(h\left(\phi(x,\varepsilon)\right) = y\right)$
    - ➢ q is probability of predicting class y under noise in parameter space
- We construct **smoothed classifier** $g: \mathcal{X} \rightarrow \mathcal{Y} = \{1,2,\dots,C\}$:

$$g(x;\varepsilon) = \operatorname{argmax}_{y \in \mathcal{Y}} q(y|x;\varepsilon)$$

➢ Returns the class with highest q

Smoothness Brings Robustness

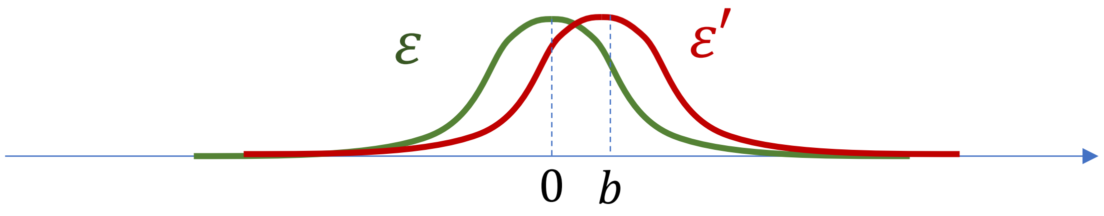
Recall $g(x;\varepsilon) = \text{argmax}_{y \in \mathcal{Y}}\, q(y|x;\varepsilon) = \text{argmax}_{y \in \mathcal{Y}} \Pr_{\varepsilon}\big(h\big(\phi(x,\varepsilon)\big) = y\big)$

- If for the clean input x_0, $q(\{\text{panda}, \text{monkey}, \text{cat}\}|x_0, \varepsilon) = \{0.80, 0.15, 0.05\}$
- Slightly change the brightness by b:

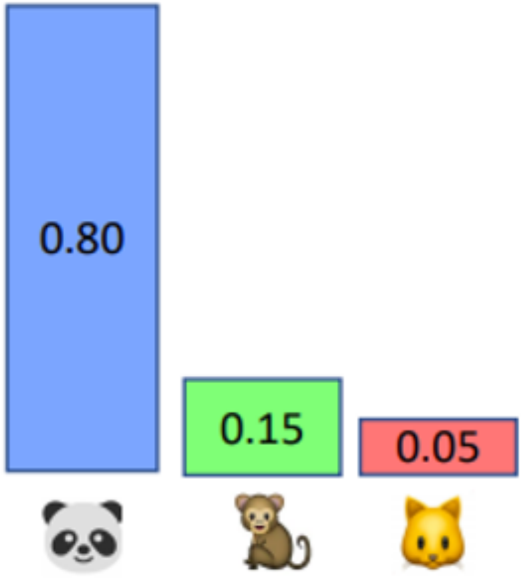
$$\varepsilon \sim \mathcal{N}(0, \sigma^2) \text{ becomes } \varepsilon' \sim \mathcal{N}(b, \sigma^2)$$

- Slightly shifting ε mean,

$q(\text{panda}|x_0, \varepsilon')$ is still **guaranteed** to be the largest



Prob. Density over Brightness Change



Credit to Cohen, Jeremy et al. Certified Adversarial Robustness via Randomized Smoothing. ICML 2019

Robustness Guarantee

- p_A: probability of top class (panda)
- p_B: probability of runner-up class (monkey)
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$: smoothing distribution:

g probably returns the top-class panda as long as brightness change

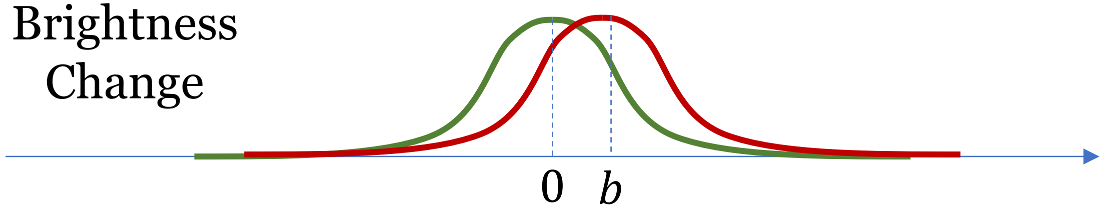
$$b \leq \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right),$$

where Φ^{-1} is the inverse standard Gaussian CDF

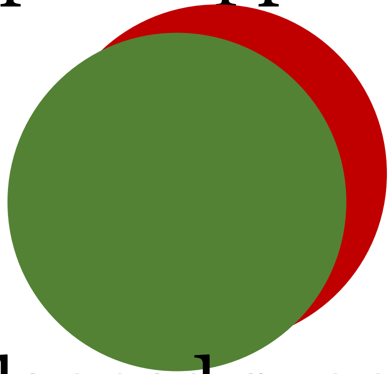
However…

- Guaranteed robustness relies on **overlapped supports** between original and transformed input
- For some transformations, there are overlapped supports

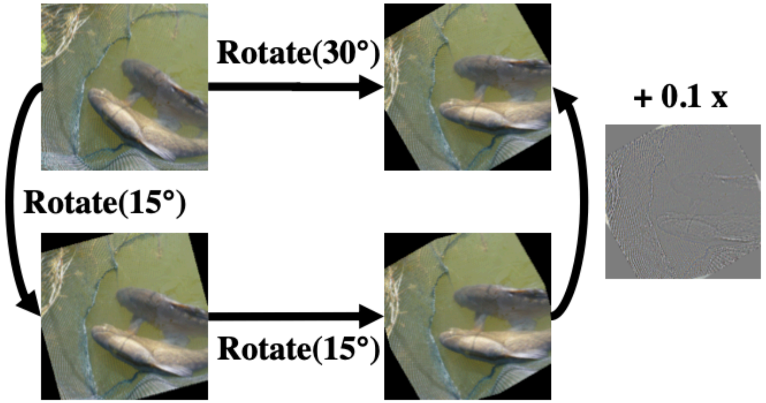
Brightness Change



Additive Perturbation



- For some transformations, hard to find overlapped supports
  - Smoothing over rotated input = Rotating two times
  - Rotate 15° + rotate 15° ≠ rotate 30°
    - Due to interpolation



Resolvable Transformations vs. Differentially Resolvable Transformations

- Transformation with overlapped supports = resolvable
  - Formally, for any $\alpha \in \mathcal{Z}$, there exists function $\gamma_\alpha : \mathcal{Z} \to \mathcal{Z}$,
$$\phi(\phi(x, \alpha), \beta) = \phi\big(x, \gamma_\alpha(\beta)\big)$$
- *(*informal)* Transformation without overlapped supports but continuous = differentially resolvable

TSS-R: Certifying Resolvable Transformations

- For resolvable transformations, use our generalized randomized smoothing to smooth and provide robustness certification
  - Brightness, contrast, translation, Gaussian blur, …

Interesting findings:

- Although Gaussian and uniform smoothing distribution shown best for ℓ_p bounded additive perturbations
- For these low-dimensional transformations, **Exponential distribution** usually performs the best
- Some transformations have constrained parameter space, customized smoothing distributions lead to higher certified robustness for them
  - *E.g.*, Gaussian blur's radius cannot be negative, use exponential or folded Gaussian as smoothing distributions

TSS-DR: Certifying Differentially Resolvable Transformations

- Differentially resolvable transformations may not have overlapped supports → **cannot** directly apply generalized randomized smoothing
- Luckily, we find
  - Transformations have low-dimensional parameter space
    - *E.g.*, one-dimensional rotation angle

  ➢ Moderate number of samples lead to an ϵ-cover of parameter space
  - (*informal) By definition, they are continuous w.r.t. parameter change
    - *E.g.*, rotated image w.r.t the rotation angle is continuous
    - *Preprocessing masks out pixels outside of inscribed circle to improve continuity*

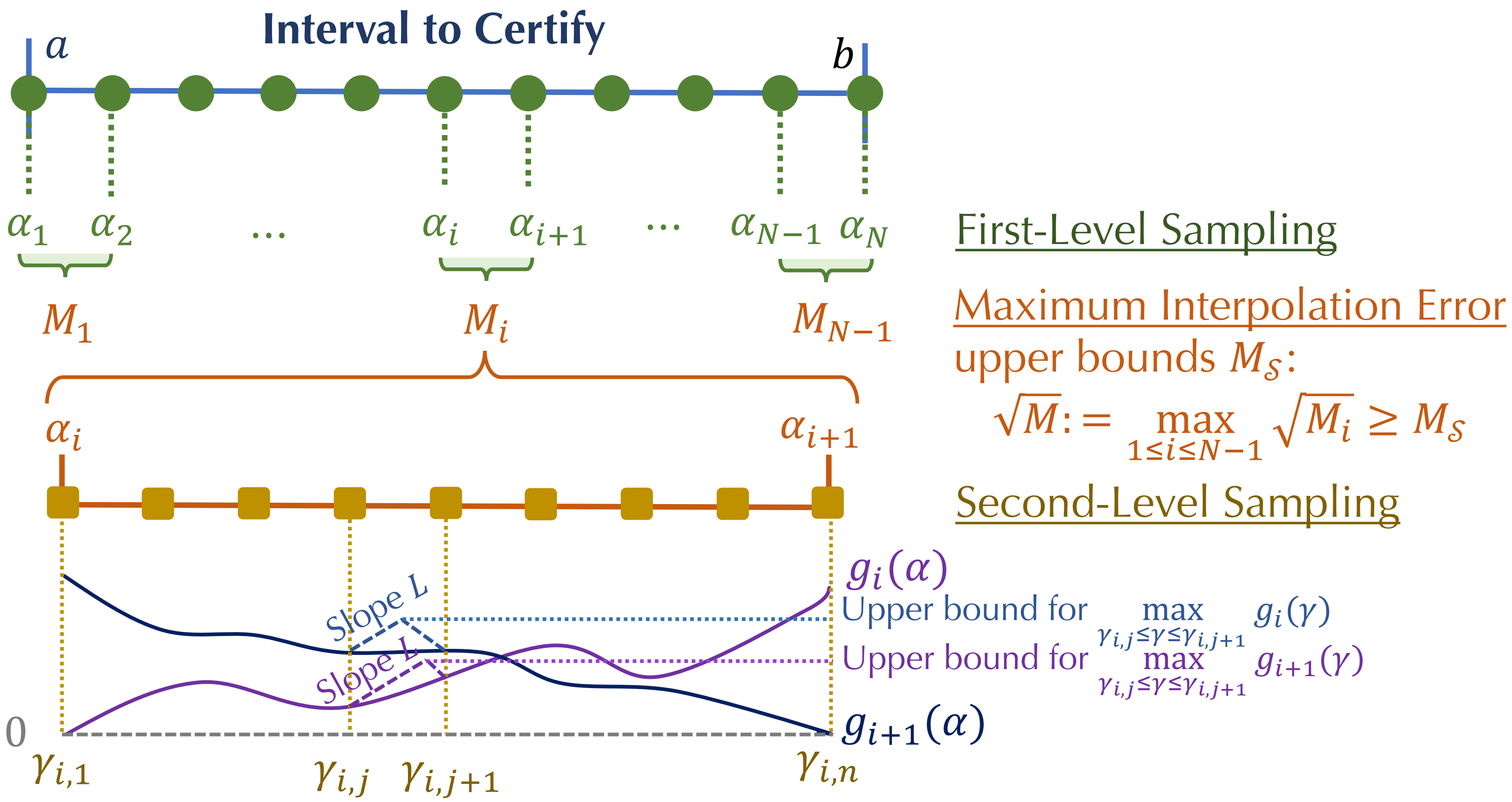
  ➢ Given Lipschitz L, maximum ℓ_2 difference from the nearest sample in ϵ-cover is ϵL

Reduction to Certifying ℓ_2 Robustness

- ➢ Moderate number of samples lead to an ϵ-cover of parameter space
- ➢ Given Lipschitz L, maximum ℓ_2 difference from the nearest sample in ϵ-cover is ϵL
- If for any sample in ϵ-cover, we can certify an ℓ_2 robust radius $\geq \epsilon L$, then we are done ☺
  - Certify an ℓ_2 robust radius?
  - Apply additive transformation suffices
- Problem to solve: compute the maximum ℓ_2 difference

Interpolation Error

- Given these samples, we now need to figure out the maximum interpolation error
  - i.e., maximum ℓ_2 difference from any transformed image to their nearest samples
- We combine stratified sampling and efficient Lipschitz computation to upper bound such difference



Bounding M_i from second-level sampling and Lipschitz constant:

$$M_i = \max_{1 \le j \le n-1} \min\{\text{Upper bound for} \max_{\gamma_{i,j} \le \gamma \le \gamma_{i,j+1}} g_i(\gamma), \text{Upper bound for} \max_{\gamma_{i,j} \le \gamma \le \gamma_{i,j+1}} g_{i+1}(\gamma)\}$$

Threat Model & Certification Goal

Challenges

Our Framework: TSS

➢ Experimental Evaluation

Experimental Setup

- Base Classifier Training:
  - We combined consistency-enhanced training [1] with transformation-specific data augmentation to obtain base classifier for smoothing
- Metric: **Certified Robust Accuracy**
  - The fraction of samples (within the test subset) that are
  - both certified robust and classified correctly
  - under any attack whose parameter is within predefined range

[1] Jongheon Jeong and Jinwoo Shin. Consistency Regularization for Certified Robustness of Smoothed Classifiers. NeurIPS 2020

Set-of-the-art Certified Robustness

Transformation	Type	Dataset	Attack Radius	Certified Robust Accuracy: TSS	DeepG [2]	Interval [47]	VeriVis [39]	Semanify-NN [35]	DistSPT [13]
Gaussian Blur	Resolvable	MNIST	Squared Radius $\alpha \leq 36$	90.6%	-	-	-	-	-
		CIFAR-10	Squared Radius $\alpha \leq 16$	63.6%	-	-	-	-	-
		ImageNet	Squared Radius $\alpha \leq 36$	51.6%	-	-	-	-	-
Translation (Reflection Pad.)	Resolvable, Discrete	MNIST	$\sqrt{\Delta x^2 + \Delta y^2} \leq 8$	99.6%	-	-	98.8%	98.8%	-
		CIFAR-10	$\sqrt{\Delta x^2 + \Delta y^2} \leq 20$	80.8%	-	-	65.0%	65.0%	-
		ImageNet	$\sqrt{\Delta x^2 + \Delta y^2} \leq 100$	50.0%	-	-	43.2%	43.2%	-
Brightness	Resolvable	MNIST	$b \pm 50\%$	98.2%	-	-	-	-	-
		CIFAR-10	$b \pm 40\%$	87.0%	-	-	-	-	-
		ImageNet	$b \pm 40\%$	70.0%	-	-	-	-	-
Contrast and Brightness	Resolvable, Composition	MNIST	$c \pm 50\%, b \pm 50\%$	97.6%	$\leq 0.4\%$ $(c, b \pm 30\%)$	0.0% $(c, b \pm 30\%)$	-	$\leq 74\%$ $(c \pm 5\%, b \pm 50\%)$	-
		CIFAR-10	$c \pm 40\%, b \pm 40\%$	82.4%	0.0% $(c, b \pm 30\%)$	0.0% $(c, b \pm 30\%)$	-	-	-
		ImageNet	$c \pm 40\%, b \pm 40\%$	61.4%	-	-	-	-	-
Gaussian Blur, Translation, Brightness, and Contrast	Resolvable, Composition	MNIST	$\alpha \leq 1, \sqrt{\Delta x^2 + \Delta y^2} \leq 5, c, b \pm 10\%$	90.2%	-	-	-	-	-
		CIFAR-10	$\alpha \leq 1, \sqrt{\Delta x^2 + \Delta y^2} \leq 5, c, b \pm 10\%$	58.2%	-	-	-	-	-
		ImageNet	$\alpha \leq 10, \sqrt{\Delta x^2 + \Delta y^2} \leq 10, c, b \pm 20\%$	32.8%	-	-	-	-	-
Rotation	Differentially Resolvable	MNIST	$r \pm 50°$	97.4%	$\leq 85.8\%$ $(r \pm 30°)$	$\leq 6.0\%$ $(r \pm 30°)$	-	$\leq 92.48\%$	82%
		CIFAR-10	$r \pm 10°$	70.6%	62.5%	20.2%	-	-	37%
			$r \pm 30°$	63.6%	10.6%	0.0%	-	$\leq 49.37\%$	22%
		ImageNet	$r \pm 30°$	30.4%	-	-	-	-	16% (rand. attack)
Scaling	Differentially Resolvable	MNIST	$s \pm 30\%$	97.2%	85.0%	16.4%	-	-	-
		CIFAR-10	$s \pm 30\%$	58.8%	0.0%	0.0%	-	-	-
		ImageNet	$s \pm 30\%$	26.4%	-	-	-	-	-
Rotation and Brightness	Differentially Resolvable, Composition	MNIST	$r \pm 50°, b \pm 20\%$	97.0%	-	-	-	-	-
		CIFAR-10	$r \pm 10°, b \pm 10\%$	70.2%	-	-	-	-	-
			$r \pm 30°, b \pm 20\%$	61.4%	-	-	-	-	-
		ImageNet	$r \pm 30°, b \pm 20\%$	26.8%	-	-	-	-	-
Scaling and Brightness	Differentially Resolvable, Composition	MNIST	$s \pm 50\%, b \pm 50\%$	96.6%	-	-	-	-	-
		CIFAR-10	$s \pm 30\%, b \pm 30\%$	54.2%	-	-	-	-	-
		ImageNet	$s \pm 30\%, b \pm 30\%$	23.4%	-	-	-	-	-
Rotation, Brightness, and ℓ_2	Differentially Resolvable, Composition	MNIST	$r \pm 50°, b \pm 20\%, \|\|\delta\|\|_2 \leq .05$	96.6%	-	-	-	-	-
		CIFAR-10	$r \pm 10°, b \pm 10\%, \|\|\delta\|\|_2 \leq .05$	64.2%	-	-	-	-	-
			$r \pm 30°, b \pm 20\%, \|\|\delta\|\|_2 \leq .05$	55.2%	-	-	-	-	-
		ImageNet	$r \pm 30°, b \pm 20\%, \|\|\delta\|\|_2 \leq .05$	26.6%	-	-	-	-	-
Scaling, Brightness, and ℓ_2	Differentially Resolvable, Composition	MNIST	$s \pm 50\%, b \pm 50\%, \|\|\delta\|\|_2 \leq .05$	96.4%	-	-	-	-	-
		CIFAR-10	$s \pm 30\%, b \pm 30\%, \|\|\delta\|\|_2 \leq .05$	51.2%	-	-	-	-	-
		ImageNet	$s \pm 30\%, b \pm 30\%, \|\|\delta\|\|_2 \leq .05$	22.6%	-	-	-	-	-

Robustness under Existing Attacks

- We study actual robustness under a random attack and an adaptive attack
  - TSS accuracy under attack > TSS certified robust accuracy
    - ➢ TSS certification is correct
  - TSS certified robust accuracy >> Standard models' accuracy under attack
    - ➢ TSS certification is meaningful in practice
  - Adaptive attack reduces standard models' accuracy more
    - ➢ TSS models provides strong robustness against adaptive attacks
  - The gap between accuracy under attack and certified robust accuracy is larger for larger dataset (e.g., ImageNet)
    - ➢ Improvement rooms exist

Other Findings

There are many more transformations in the wild world

- Evaluated on natural corruption datasets CIFAR-10-C and ImageNet-C:
  - TSS models are still better than standard models
  - Sometimes even better than SOTA on CIFAR-10-C and ImageNet-C

    * Evaluated on the highest level of corruptions
  - Provides strong robustness guarantees against transformation compositions, even on large-scale ImageNet

	CIFAR-10			ImageNet		
	Vanilla	AugMix [21]	**TSS**	Vanilla	AugMix [21]	**TSS**
Empirical Accuracy on CIFAR-10-C and ImageNet-C	53.9%	65.6%	**67.4%**	18.3%	**25.7%**	21.9%
Certified Accuracy against Composition of Gaussian Blur, Translation, Brightness, and Contrast	0.0%	0.4%	**58.2%**	0.0%	0.0%	**32.8%**

Other Findings (Cont.d)

- If the attack's perturbation radius (i.e., rotation angle) beyond the predefined radius used in training…
  - TSS still preserves high certified robust accuracy
  - For model defending 40% brightness change on ImageNet,
    - Certified accuracy against 40% change is 70.4%
    - Certified accuracy against 50% change is 70.0%
- Smoothing variance is a tunable hyperparameter
  - Small smoothing variance → high clean accuracy, small certified radius
  - Large smoothing variance → low clean accuracy, large certified radius
  - For highest certified accuracy under a given radius, an optimal smoothing variance exists

Conclusion

- **TSS**: a framework for certifying ML robustness against semantic transformations
- Categorize semantic transformations into resolvable (R) and differentiable resolvable (DR)
- Apply TSS-R and TSS-DR respectively
- Achieve significantly higher certified robustness than state-of-the-arts
- **First** work that achieves nontrivial certified robustness on ImageNet
- Achieve high empirical robustness against adaptive attacks and unforseen transformations

Full paper	arxiv.org/abs/2002.12398
Slides	linyil.com/res/pub/TSS-CCS21-slides.pdf
Code	github.com/AI-secure/semantic-randomized-smoothing