TSS
Transformation-Specific Smoothing for Robustness Certification

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

\[ + 0.001 * \text{Small Perturbation} = \text{Predicted as “dog”} \]

Adversarial Attack via Semantic Transformations

• Certifying and improving robustness for ML models against $\ell_p$ 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

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

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

Transformation-specific protocol/distribution-based smoothing.

1. Rotation
2. Gauss. Blur
3. Contrast
4. Brightness
5. (3) + (4)

Provably Robust Prediction against a Given Transformation
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: \( \mathcal{Y} = \{1, \ldots, C\} \)

• Semantic transformation as a function \( \phi: \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{X} \)
  • Parameter space: \( \mathcal{Z} \subseteq \mathbb{R}^m \)

• Attacker can:
  1. arbitrarily 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 \( \alpha \) degree clockwise
• Define \( \mathcal{Z} = [-30^\circ, 30^\circ] \)

\[\text{Attacker can arbitrarily rotate the image within } 30^\circ\]
Certification Goal

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

$$h(x) = h(\phi(x, \alpha)), \forall \alpha \in S$$
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 $\ell_p$ Difference

- Semantic transformation incurs large $\ell_p$ difference
  - Brightness +10% incurs $\ell_2$ difference $0.1 \times \sqrt{\# \text{pixels}} \approx 38.7$ on ImageNet

➤ Cannot certify with existing $\ell_p$ 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 $h: \mathcal{X} \to \mathcal{Y} = \{1, 2, \ldots, C\}$
- Let $\phi(x, b) = x + b \cdot (1, \ldots, 1)^T$ be the brightness transformation
- Let $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ be the smoothing distribution
- Define $q(y|x; \varepsilon) = \Pr(h(\phi(x, \varepsilon)) = y)$
  - $q$ is probability of predicting class $y$ under noise in parameter space
- We construct smoothed classifier $g: \mathcal{X} \to \mathcal{Y} = \{1, 2, \ldots, C\}$:
  $$g(x; \varepsilon) = \arg\max_{y \in \mathcal{Y}} q(y|x; \varepsilon)$$
  - Returns the class with highest $q$
Smoothness Brings Robustness

Recall $g(x; \varepsilon) = \arg\max_{y \in Y} q(y|x; \varepsilon) = \arg\max_{y \in Y} \Pr_{\varepsilon}(h(\phi(x, \varepsilon)) = y)$

- If for the clean input $x_0$, $q(\{\text{panda, monkey, cat}\}|x_0, \varepsilon) = \{0.80, 0.15, 0.05\}$
- Slightly change the brightness by $b$:
  \[ \varepsilon \sim \mathcal{N}(0, \sigma^2) \text{ becomes } \varepsilon' \sim \mathcal{N}(b, \sigma^2) \]
- Slightly shifting $\varepsilon$ mean, $q(\text{panda}|x_0, \varepsilon')$ is still guaranteed to be the largest

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

\[
b \leq \frac{\sigma}{2} (\Phi^{-1}(p_A) - \Phi^{-1}(p_B)),
\]

where $\Phi^{-1}$ 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 😊

- For some transformations, hard to find overlapped supports 😞
  - Smoothing over rotated input = Rotating two times
  - Rotate $15^\circ$ + rotate $15^\circ \neq$ rotate $30^\circ$
    - Due to interpolation
Resolvable Transformations vs. Differentially Resolvable Transformations

- Transformation with overlapped supports = resolvable
  - Formally, for any $\alpha \in \mathcal{Z}$, there exists function $\gamma_{\alpha}: \mathcal{Z} \to \mathcal{Z}$, $\phi(\phi(x, \alpha), \beta) = \phi(x, \gamma_{\alpha}(\beta))$

- (*informal) Transformation without overlapped supports but continuous = differentially resolvable

### Differentially Resolvable Transformations

- **Resolvable Transformations**
  - Translation
  - Brightness
  - Contrast
  - Gaussian Blur

- **Brightness & Contrast**
- **Rotation & Brightness**
- **Rotation**
- **Scaling**
- **Other Compositions**
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 $\ell_p$ 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 $\epsilon$-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 $\ell_2$ difference from the nearest sample in $\epsilon$-cover is $\epsilon L$
Reduction to Certifying $\ell_2$ Robustness

- Moderate number of samples lead to an $\varepsilon$-cover of parameter space
- Given Lipschitz $L$, maximum $\ell_2$ difference from the nearest sample in $\varepsilon$-cover is $\varepsilon L$
  
  - If for any sample in $\varepsilon$-cover, we can certify an $\ell_2$ robust radius $\geq \varepsilon L$, then we are done
    
    - Certify an $\ell_2$ robust radius?
    - Apply additive transformation suffices
  
  - Problem to solve: compute the maximum $\ell_2$ difference
Interpolation Error

- Given these samples, we now need to figure out the maximum interpolation error.
  - i.e., maximum $\ell_2$ difference from any transformed image to their nearest samples.
- We combine stratified sampling and efficient Lipschitz computation to upper bound such difference.

First-Level Sampling

Maximum Interpolation Error upper bounds $M_S$:

$$\sqrt{M} = \max_{1 \leq i \leq N-1} \sqrt{M_i} \geq M_S$$

Second-Level Sampling

Bounding $M_i$ from second-level sampling and Lipschitz constant:

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

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Type</th>
<th>Dataset</th>
<th>Attack Radius</th>
<th>TSS</th>
<th>Certified Robust Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Blur</td>
<td>Resolvable</td>
<td>MNIST</td>
<td>Squared Radius $a \leq 36$</td>
<td>90.6%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>Squared Radius $a \leq 16$</td>
<td>63.6%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ImageNet</td>
<td>Squared Radius $a \leq 36$</td>
<td>51.6%</td>
<td>-</td>
</tr>
<tr>
<td>Translation (Reflection Pad.)</td>
<td>Resolvable, Discrete</td>
<td>MNIST</td>
<td>$\sqrt{\Delta x^2 + \Delta y^2} \leq 8$</td>
<td>99.6%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>$\sqrt{\Delta x^2 + \Delta y^2} \leq 20$</td>
<td>80.8%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ImageNet</td>
<td>$\sqrt{\Delta x^2 + \Delta y^2} \leq 100$</td>
<td>50.0%</td>
<td>-</td>
</tr>
<tr>
<td>Brightness</td>
<td>Resolvable</td>
<td>MNIST</td>
<td>$b \pm 50%$</td>
<td>98.2%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>$b \pm 40%$</td>
<td>87.0%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ImageNet</td>
<td>$b \pm 40%$</td>
<td>70.0%</td>
<td>-</td>
</tr>
<tr>
<td>Contrast and Brightness</td>
<td>Resolvable, Composition</td>
<td>MNIST</td>
<td>$c \pm 50%, b \pm 50%$</td>
<td>97.6%</td>
<td>$\leq 0.4%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>$c \pm 40%, b \pm 40%$</td>
<td>82.4%</td>
<td>(c, b $\pm 30%$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ImageNet</td>
<td>$c \pm 40%, b \pm 40%$</td>
<td>61.4%</td>
<td>(c, b $\pm 30%$)</td>
</tr>
<tr>
<td>Gaussian Blur, Translation, Bright-</td>
<td>Resolvable, Composition</td>
<td>MNIST</td>
<td>$\alpha \leq 1, \sqrt{\Delta x^2 + \Delta y^2} \leq 5, c, b \pm 10%$</td>
<td>90.2%</td>
<td>(c, b $\pm 30%$)</td>
</tr>
<tr>
<td>ness, and Contrast</td>
<td></td>
<td>CIFAR-10</td>
<td>$\alpha \leq 1, \sqrt{\Delta x^2 + \Delta y^2} \leq 5, c, b \pm 10%$</td>
<td>58.2%</td>
<td>(c, b $\pm 30%$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ImageNet</td>
<td>$\alpha \leq 1, \sqrt{\Delta x^2 + \Delta y^2} \leq 10, c, b \pm 20%$</td>
<td>32.8%</td>
<td>(c, b $\pm 30%$)</td>
</tr>
<tr>
<td>Rotation</td>
<td>Differentially Resolvable</td>
<td>MNIST</td>
<td>$r = 50^\circ$</td>
<td>97.4%</td>
<td>$\leq 85.8%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>$r = 10^\circ$</td>
<td>70.6%</td>
<td>(r $\pm 30^\circ$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>$r = 30^\circ$</td>
<td>63.6%</td>
<td>62.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ImageNet</td>
<td>$r = 30^\circ$</td>
<td>30.4%</td>
<td>10.5%</td>
</tr>
<tr>
<td>Scaling</td>
<td>Differentially Resolvable</td>
<td>MNIST</td>
<td>s $\pm 30%$</td>
<td>97.2%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>s $\pm 30%$</td>
<td>58.8%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ImageNet</td>
<td>s $\pm 30%$</td>
<td>26.4%</td>
<td>-</td>
</tr>
<tr>
<td>Rotation and Brightness</td>
<td>Differentially Resolvable</td>
<td>MNIST</td>
<td>$r \pm 50^\circ, b \pm 20%$</td>
<td>97.0%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>$r \pm 10^\circ, b \pm 10%$</td>
<td>70.2%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>$r \pm 30^\circ, b \pm 20%$</td>
<td>61.4%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ImageNet</td>
<td>$r \pm 30^\circ, b \pm 20%$</td>
<td>26.8%</td>
<td>-</td>
</tr>
<tr>
<td>Scaling and Brightness</td>
<td>Differentially Resolvable</td>
<td>MNIST</td>
<td>s $\pm 50%, b \pm 50%$</td>
<td>96.6%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>s $\pm 30%, b \pm 30%$</td>
<td>54.2%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ImageNet</td>
<td>s $\pm 30%, b \pm 30%$</td>
<td>23.4%</td>
<td>-</td>
</tr>
<tr>
<td>Rotation, Brightness, and $\ell_2$</td>
<td>Differentially Resolvable</td>
<td>MNIST</td>
<td>$r \pm 50^\circ, b \pm 20%, |\ell|_2 \leq 0.05$</td>
<td>96.6%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>$r \pm 10^\circ, b \pm 10%, |\ell|_2 \leq 0.05$</td>
<td>64.2%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>$r \pm 30^\circ, b \pm 20%, |\ell|_2 \leq 0.05$</td>
<td>55.2%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ImageNet</td>
<td>$r \pm 30^\circ, b \pm 20%, |\ell|_2 \leq 0.05$</td>
<td>26.6%</td>
<td>-</td>
</tr>
<tr>
<td>Scaling, Brightness, and $\ell_2$</td>
<td>Differentially Resolvable</td>
<td>MNIST</td>
<td>s $\pm 50%, b \pm 50%, |\ell|_2 \leq 0.05$</td>
<td>96.4%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIFAR-10</td>
<td>s $\pm 30%, b \pm 30%, |\ell|_2 \leq 0.05$</td>
<td>51.2%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ImageNet</td>
<td>s $\pm 30%, b \pm 30%, |\ell|_2 \leq 0.05$</td>
<td>22.6%</td>
<td>-</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>CIFAR-10</th>
<th>ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Empirical Accuracy</strong> on CIFAR-10-C and ImageNet-C</td>
<td>53.9%</td>
<td>65.6%</td>
</tr>
<tr>
<td><strong>Certified Accuracy</strong> against Composition of Gaussian Blur, Translation, Brightness, and Contrast</td>
<td>0.0%</td>
<td>0.4%</td>
</tr>
</tbody>
</table>
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 unforeseen transformations

Code: [github.com/AI-secure/semantic-randomized-smoothing](https://github.com/AI-secure/semantic-randomized-smoothing)