On the Limitations of Stochastic Pre-processing Defenses

Yue Gao\textsuperscript{1}, Ilia Shumailov\textsuperscript{2,4}, Kassem Fawaz\textsuperscript{1}, Nicolas Papernot\textsuperscript{3,4}

\textsuperscript{1}University of Wisconsin–Madison, \textsuperscript{2}University of Cambridge, \textsuperscript{3}University of Toronto, \textsuperscript{4}Vector Institute
Background

Adversarial examples, defenses, and evaluations.
Adversarial Examples

$x$

“panda”
57.7% confidence

$\times .007\times$

(sign($\nabla_x J(\theta, x, y)$))

“nematode”
8.2% confidence

$= x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$

“gibbon”
99.3% confidence

Figure from Goodfellow et al. Explaining and harnessing adversarial examples. ICLR 2015.
Stochastic Pre-processing Defenses

**Intuition:** Adversarial examples must generalize to all transformations.

*Does this strategy make the attack any harder?*
Evaluations Rely on Adaptive Attacks

  - Random Cropping, Random Rescaling, …
  - Obfuscated Gradients Give a False Sense of Security.

- **Round 2 (2019 – 2020)**
  - MixUp, Random Pixel Dropping, …
  - On Adaptive Attacks to Adversarial Example Defenses.

- **Round 3 (2019 – 2022)**
  - Barrage of Random Transformations (BaRT).
  - Demystifying the Adversarial Robustness of Random Transformation Defenses.

- **Round 4 (2022 – ?)**
  - Diffusion Models for Adversarial Purification (DiffPure).
  - ?
Lessons (not) Learned from Adaptive Attacks

- Adaptive attacks become hard to design & evaluate.
  - **BaRT**: broken after 3 years on a smaller-scale dataset (ImageNet → ImageNette).
  - **DiffPure**: requires “1-4 high-end NVIDIA GPUs with 32 GB of memory.”

- Fundamental weaknesses remain unknown.
  - Why doesn’t randomness provide robustness as we expected?
  - How could future defenses avoid the pitfalls of existing stochastic defenses?

*We should look for fundamental limitations.*
Lack of Sufficient Randomness

Limitation 1 of 2
Formulations

- Stochastic Classifier

\[ f_\theta : \mathcal{X} \rightarrow \mathbb{R}^C, \quad \mathcal{X} = [0, 1]^d, \quad \theta \sim \Theta \]

- Prediction (majority vote)

\[ F_{\text{vote}}(x) := \arg \max_{c \in [C]} \sum_{i=1}^n \mathbb{1}\left\{ \arg \max_{j \in [C]} f_{\theta_{i,j}}(x) = c \right\} \]

number of classes

input dimension

random parameter

randomization space

number of votes

output for class j

sampled parameter \( \theta_i \sim \Theta \)
Core Attack Techniques: PGD + EOT

- Projected Gradient Descent (PGD)

\[ x^{i+1} \leftarrow x^i + \alpha \cdot \text{sgn}\{\nabla \mathcal{L}(f_\theta(x^i), y)\} \]

- Expectation over Transformation (EOT)

\[ x^{i+1} \leftarrow x^i + \alpha \cdot \text{sgn}\left\{\mathbb{E}_{\Theta \sim \Theta} \left[\nabla \mathcal{L}(f_\theta(x^i), y)\right]\right\} \approx x^i + \alpha \cdot \text{sgn}\left\{\frac{1}{m} \sum_{j=1}^{m} \nabla \mathcal{L}(f_{\theta_j}(x^i), y)\right\} \]
Literature’s (Rightful) View of EOT

Initially proposed for “synthesizing examples that are adversarial over a chosen distribution of transformations.” (Athalye et al.)

Adopted to “correctly compute the gradient over the expected transformation to the input.” (Athalye et al.)

Became “standard technique for computing gradients of models with randomized components” (Tramèr et al.)

Finally, evaluations explicitly detect randomized components and enforce the application of EOT. (Croce et al.)
Blind Spot: Unclear Security under Weaker Attacks

- Case Study: Random Rotation

\[ t_\theta(x) := \text{rotate}(x, \theta), \quad \theta \sim \mathcal{U}(-90^\circ, 90^\circ) \]

- Attacking with PGD-\(k\) and EOT-\(m\)

<table>
<thead>
<tr>
<th>Attacks</th>
<th>(k)</th>
<th>(m)</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Untargeted</td>
<td>10</td>
<td>5</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>Targeted</td>
<td>10</td>
<td>5</td>
<td>99.0%</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1</td>
<td>99.0%</td>
</tr>
</tbody>
</table>

Randomness can be insecure even under standard attacks (w/o handling randomness)
Most Stochastic Defenses Lack Sufficient Randomness

- Revisit previously broken defenses w/o EOT

Notations: attack iterations $k$, EOT samples $m$, learning rate $\alpha$, number of gradient queries $k \times m$.

<table>
<thead>
<tr>
<th>Defenses</th>
<th>Original Adaptive Evaluation (w/ EOT)</th>
<th></th>
<th></th>
<th>Success Rate</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k$</td>
<td>$m$</td>
<td>$\alpha$</td>
<td>$k \times m$</td>
<td></td>
<td>$k$</td>
<td>$m$</td>
<td>$\alpha$</td>
<td>$k \times m$</td>
</tr>
<tr>
<td>Guo et al. [11]</td>
<td>1,000</td>
<td>30</td>
<td>0.1</td>
<td>30,000</td>
<td>100%</td>
<td>1,000</td>
<td>1</td>
<td>0.001</td>
<td>1,000</td>
</tr>
<tr>
<td>Xie et al. [40]</td>
<td>1,000</td>
<td>30</td>
<td>0.1</td>
<td>30,000</td>
<td>100%</td>
<td>200</td>
<td>1</td>
<td>0.1</td>
<td>200</td>
</tr>
<tr>
<td>Dhillon et al. [8]</td>
<td>500</td>
<td>10</td>
<td>0.1</td>
<td>5,000</td>
<td>100%</td>
<td>500</td>
<td>1</td>
<td>0.1</td>
<td>500</td>
</tr>
<tr>
<td>Xiao et al. [39]</td>
<td>100</td>
<td>1,000</td>
<td>0.01</td>
<td>100,000</td>
<td>100%</td>
<td>40,000</td>
<td>1</td>
<td>0.1/255</td>
<td>40,000</td>
</tr>
<tr>
<td>Roth et al. [28]</td>
<td>100</td>
<td>40</td>
<td>0.2/255</td>
<td>4,000</td>
<td>100%</td>
<td>4,000</td>
<td>1</td>
<td>0.1/255</td>
<td>4,000</td>
</tr>
</tbody>
</table>

- Standard attacks already perform well ...

... as long as they run for more iterations with a smaller learning rate
EOT is Only Beneficial for Sufficient Randomness

- Targeted Attacks on Randomized Smoothing

<table>
<thead>
<tr>
<th>EOT Samples</th>
<th>Attack Success Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24.3 43.2 66.7 80.9 87.5 93.1 94.9</td>
</tr>
<tr>
<td>5</td>
<td>67.1 81.4 91.1 93.8 95.7 96.7 97.2</td>
</tr>
<tr>
<td>10</td>
<td>80.1 89.5 94.7 96.4 96.8 97.2 97.5</td>
</tr>
<tr>
<td>20</td>
<td>86.1 93.6 96 96.8 97.4 97.9 98.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EOT Samples</th>
<th>Attack Success Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.81 9.02 16.7 21.9 28 33.5 34.8</td>
</tr>
<tr>
<td>5</td>
<td>17.3 24.4 31.5 36.2 39 42 43.9</td>
</tr>
<tr>
<td>10</td>
<td>24.1 31.2 36.7 40.3 42.7 45.2 45.6</td>
</tr>
<tr>
<td>20</td>
<td>30.9 37.5 41.9 43.7 45.3 46.4 46.1</td>
</tr>
</tbody>
</table>

Lower Randomness ($\sigma = 0.25$)  Higher Randomness ($\sigma = 0.50$)

Randomization’s contribution to robustness is overestimated.
Renewed Understanding of Randomization

Why could we break stochastic defenses?

- **Before**: Because we used EOT.
- **Now**: Because they did not have sufficient randomness.

I want to apply random rotation, am I secure?

- **Before**: Maybe, as long as the attack does not apply EOT.
- **Now**: No, not even under standard attacks.

*Next: What if the defenses do have sufficient randomness?*
Trade-off: Robustness vs. Invariance

Limitation 2 of 2
Stochastic Defenses & Model Invariance

- What does it mean for a model to be invariant?

\[ F_\theta(x) := F(t_\theta(x)) = F(x), \quad \forall \theta \in \Theta, x \in \mathcal{X} \]

- If the defended model is invariant to the defense …

\[ \text{Attack}(F_\theta, x) = \text{Attack}(F, x) \]

- Attacking the **defended model** is the same as attacking the **original model**!

*Stochastic pre-processing defenses are not expected to work.*
Theoretical Setting: Binary Classification

- **Settings**
  - Label: \( y \in \{-1, +1\} \)
  - Input: \( x|y \sim \mathcal{N}(y, 1) \)
  - Adversary: \( \|\delta\|_\infty \leq \epsilon \)

- **Robust Accuracy**

  \[ \text{Rob} := \frac{\text{dotted area}}{\text{shadowed area}} \]
Theoretical Setting: Binary Classification

- Undefended Classification
  - Bayesian Optimal Classifier

\[ F(x) = \text{sgn}(x) \]

- Robust Accuracy

\[
\text{Rob} = \frac{\text{dotted area}}{\text{shadowed area}} = \frac{\Phi(1 - \epsilon)}{\Phi(1)}
\]

\( \Phi(\cdot) \) is the CDF of \( \mathcal{N}(0, 1) \)
Theoretical Setting: Binary Classification

- Defended Classification
  - Introduce the Defense
    \[ t_\theta := x + \theta, \quad \theta \sim \mathcal{N}(1, \sigma^2) \]
  - Processed Input Distribution
    \[ t_\theta(x) \sim \mathcal{N}(y + 1, 1 + \sigma^2) \]
  - Higher Robust Accuracy
    \[ \text{Rob} = \frac{\text{dotted area}}{\text{shadowed area}} = \frac{\Phi'(-\epsilon) + \Phi'(2 - \epsilon)}{\Phi'(0) + \Phi'(2)} \]
    \[ \Phi'(\cdot) \text{ is the CDF of } \mathcal{N}(0, \sigma^2) \]
Theoretical Setting: Binary Classification

- Defended Classification (w/ Trained Invariance)
  - Processed Input Distribution
    \[ t_\theta(x) \sim \mathcal{N}(y + 1, 1 + \sigma^2) \]
  - New Bayesian Optimal Classifier
    \[ F_\theta^+(x) = \text{sgn}(x + \theta - 1) \]
  - Reduced Robust Accuracy
    \[ \text{Rob} = \frac{\text{dotted area}}{\text{shadowed area}} = \frac{\Phi'(1 - \epsilon)}{\Phi'(1)} \]
    \( \Phi'(\cdot) \) is the CDF of \( \mathcal{N}(0, \sigma^2) \)
Theoretical Setting: Binary Classification

- Defended Classification (w/ Perfect Invariance)
  - New Bayesian Optimal Classifier
    \[ F_{\theta}^+(x) = \text{sgn}(x + \theta - 1) \]
  - Majority Vote
    \[ F_{\theta}^*(x) \rightarrow \mathbb{E}_{\theta \sim \Theta}[F_{\theta}^+(x)] \]
    \[ = \mathbb{E}_{\theta \sim \Theta}[\text{sgn}(x + \theta - 1)] \]
    \[ = \text{sgn}(x) \]
    \[ = F(x) \]
Formalized Robustness vs. Invariance Trade-off

- Rate of Invariance
  \[ R(k) := \Pr[F_\theta(x) = F(x)] \]
  defended model
  defended model’s boundary

- Theorem
  “When the defended classifier achieves higher invariance to preserve utility, the adversarial robustness provided by the defense strictly decreases.”

Stochastic pre-processing defenses explicitly control invariance
Fine-tuning Makes Defenses Less Robust

- The same attack on randomized smoothing before & after fine-tuning.

Untargeted Attacks

![Graph showing the effect of noise level on benign accuracy and attack success rate before and after fine-tuning for untargeted attacks.]

Targeted Attacks

![Graph showing the effect of noise level on benign accuracy and attack success rate before and after fine-tuning for targeted attacks.]

High Invariance: 50% more attacks

Low Invariance: ineffective attacks
Discussions

What can we learn from these two limitations?
What Do Stochastic Defenses Really Do?

- They do not provide “inherent robustness” to the model.
  - Currently, only adversarial training can improve the model’s robustness.

- They shift the input distribution through randomness and transformations.
  - This is an explicit control of the model’s invariance.
  - The observed “robustness” is a result of introduced errors.
Implications for Future Research

- Should we abandon stochastic defenses?
  - No, they still make black-box attacks harder.

- How do we improve stochastic defenses?
  - Look for new ways of using randomness.
  - Decouple robustness and invariance.
  - Force the attack to target non-transferable subproblems.

Orthogonal Models
Independent Patches
Different Modalities
Summary & Questions

● Motivation
  - Adaptive attacks become extremely hard to design & evaluate.
  - We need to understand the defense’s fundamental limitations.

● Our Findings
  - Most stochastic defenses are insecure even under standard attacks.
  - Trade-off between robustness and invariance.

● Takeaways
  - Stochastic pre-processing defenses are not promising.
  - Look for new ways of using randomness.
Thank You

Yue Gao
Ph.D. Student, University of Wisconsin–Madison
Research Interests: Trustworthy Machine Learning, Security and Privacy
Contact: gy@cs.wisc.edu
Homepage: https://pages.cs.wisc.edu/~gy