

# On the Limitations of Stochastic Pre-processing Defenses

**Yue Gao**<sup>1</sup>, Ilia Shumailov<sup>2,4</sup>, Kassem Fawaz<sup>1</sup>, Nicolas Papernot<sup>3,4</sup> <sup>1</sup>University of Wisconsin–Madison, <sup>2</sup>University of Cambridge, <sup>3</sup>University of Toronto, <sup>4</sup>Vector Institute



# Background

Adversarial examples, defenses, and evaluations.



### **Adversarial Examples**



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**

• Round 1 (2018 – 2019)

➢ 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  $\rightarrow$  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

6



### **Formulations**



Prediction (majority vote)

$$F_{\text{vote}}(\boldsymbol{x}) \coloneqq \arg \max_{c \in [C]} \sum_{i=1}^{n} \mathbb{1} \left\{ \arg \max_{j \in [C]} \frac{f_{\boldsymbol{\theta}_i, j}(\boldsymbol{x}) = c}{sampled \text{ parameter } \theta_i \overset{\text{i.i.d.}}{\sim} \Theta} \right\}$$



# **Core Attack Techniques: PGD + EOT**

• Projected Gradient Descent (PGD) sigma sinction  $x^{i+1} \leftarrow x^i + \alpha \cdot \operatorname{sgn}\{\nabla \mathcal{L}(f_{\theta}(x^i), y)\}$  learning rate (step size)

 Expectation over Transformation (EOT) actual EOT

$$\boldsymbol{x}^{i+1} \leftarrow \boldsymbol{x}^{i} + \alpha \cdot \operatorname{sgn} \left\{ \mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{\Theta}} \Big[ \nabla \mathcal{L} \big( f_{\boldsymbol{\theta}}(\boldsymbol{x}^{i}), \boldsymbol{y} \big) \Big] \right\} \approx \boldsymbol{x}^{i} + \alpha \cdot \operatorname{sgn} \left\{ \frac{1}{m} \sum_{j=1}^{m} \nabla \mathcal{L} \big( f_{\boldsymbol{\theta}_{j}}(\boldsymbol{x}^{i}), \boldsymbol{y} \big) \right\}$$
estimated EOT



# Literature's (Rightful) View of EOT

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

ICML 2018

**ICML 2018** 

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

NeurIPS 2020

**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

$$t_{ heta}(x) := \mathrm{rotate}(x, heta), \quad heta \sim \mathcal{U}(-90^\circ,90^\circ)$$

• Attacking with PGD-*k* and EOT-*m* 

Attacks	k	m	Success Rate
Untargeted	10	5	100%
Ontargeted	50	1	100%
Taratad	10	5	99.0%
Targeleu	50	1	99.0%

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  $k \times m$ .

Defenses	Original Adaptive Evaluation (w/ EOT)				Our Ablation Study (w/o EOT)					
	k	m	$\alpha$	k  imes m	Success Rate	k	m	lpha	k  imes m	Success Rate
Guo et al. [11]	1,000	30	0.1	30,000	100%	1,000	1	0.001	1,000	99.0%
Xie et al. [40]	1,000	30	0.1	30,000	100%	200	1	0.1	200	100%
Dhillon et al. [8]	500	10	0.1	5,000	100%	500	1	0.1	500	100%
Xiao et al. [39]	100	1,000	0.01	100,000	100%	40,000	1	0.1/255	40,000	98.4%
Roth et al. [28]	100	40	0.2/255	4,000	100%	4,000	1	0.1/255	4,000	96.1%

• 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



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

14



## **Stochastic Defenses & Model Invariance**

• What does it mean for a model to be invariant? randomization space

$$F_{ heta}(x) := F(t_{ heta}(x)) = F(x), \quad orall \ heta \in \Theta, x \in \mathcal{X}$$
 $ext{defended model}$  original model input space

• If the defended model is invariant to the defense ...

 $\operatorname{Attack}(F_{\theta}, x) = \operatorname{Attack}(F, x)$ 

• Attacking the defended model is the same as attacking the original model!

Stochastic pre-processing defenses are not expected to work.



### Settings

- $\succ$  Label  $y \in \{-1,+1\}$
- $\succ$  Input  $x|y \sim \mathcal{N}(y,1)$
- ightarrow Adversary  $\|\delta\|_{\infty} \leq \epsilon$
- Robust Accuracy

 $Rob := \frac{dotted area}{shadowed area}$ 





Undefended Classification
 > Bayesian Optimal Classifier

 $F(x) = \operatorname{sgn}(x)$ 

Robust Accuracy





- Defended Classification
  - Introduce the Defense

 $t_{ heta} := x + heta, \quad heta \sim \mathcal{N}(1, \sigma^2)$ 

Processed Input Distribution

 $t_{ heta}(x) \sim \mathcal{N}(y+1,1+\sigma^2)$ 

Higher Robust Accuracy

$$egin{aligned} \mathsf{Rob} &= rac{ ext{dotted area}}{ ext{shadowed area}} \ &= rac{\Phi'(-\epsilon) + \Phi'(2-\epsilon)}{\Phi'(0) + \Phi'(2)} \ \Phi'(\cdot) ext{ is the CDF of } \mathcal{N}(0,\sigma^2) \end{aligned}$$





Defended Classification (w/ Trained Invariance)
 Processed Input Distribution

 $t_{ heta}(x) \sim \mathcal{N}(y+1,1+\sigma^2)$ 

New Bayesian Optimal Classifier

 $F^+_ heta(x) = \mathrm{sgn}(x+ heta-1)$ 

 $\begin{array}{l} \blacktriangleright \mbox{Reduced Robust Accuracy} \\ \mbox{Rob} = \frac{\mbox{dotted area}}{\mbox{shadowed area}} \\ = \frac{\Phi'(1-\epsilon)}{\Phi'(1)} \\ \mbox{\Phi'}(\cdot) \mbox{ is the CDF of } \mathcal{N}(0,\sigma^2) \end{array}$ 

back to undefended -2  $0-\epsilon$   $+\epsilon^2$  4 $\mathcal{N}(0, 1 + \sigma^2)$  $\mathcal{N}(2, 1 + \sigma^2)$ Boundary



20

# **Theoretical Setting: Binary Classification**

- Defended Classification (w/ Perfect Invariance)
  - New Bayesian Optimal Classifier

 $F^+_ heta(x) = \mathrm{sgn}(x+ heta-1)$ 

Majority Vote

$$egin{aligned} F^*_ heta(x) & o \mathbb{E}_{ heta\sim\Theta}[F^+_ heta(x)] \ &= \mathbb{E}_{ heta\sim\Theta}[ ext{sgn}(x+ heta-1)] \ &= ext{sgn}(x) \ &= F(x) \end{aligned}$$





## Formalized Robustness vs. Invariance Trade-off



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.





# **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