Demystifying the Adversarial Robustness of Random Transformation Defenses

Chawin Sitawarin Zachary Golan-Strieb David Wagner

UC Berkeley

ICML | 2022

Thirty-ninth International Conference on Machine Learning

Contact: chawins@berkeley.edu

Random Transform Defense against Adversarial Examples

- Many works have proposed random input transformation to improve the adversarial robustness of neural networks.

Random Transform Defense against Adversarial Examples

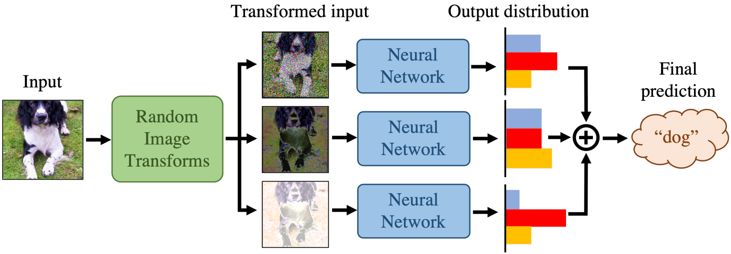
- Many works have proposed random input transformation to improve the adversarial robustness of neural networks.
- However, stochastic defenses are poorly understood, and **we still lack reliable tools for measuring their robustness.**

Random Transform Defense against Adversarial Examples

- Many works have proposed random input transformation to improve the adversarial robustness of neural networks.
- However, stochastic defenses are poorly understood, and **we still lack reliable tools for measuring their robustness.**
- We focus on **Barrage of Random Transforms (BaRT)** [Raff et al., 2019] which sequentially applies multiple image transforms to its inputs in random order and with random parameters. Transforms are sampled from a pool of over 20, both differentiable and not.

Random Transform Defense against Adversarial Examples

- Many works have proposed random input transformation to improve the adversarial robustness of neural networks.
- However, stochastic defenses are poorly understood, and **we still lack reliable tools for measuring their robustness.**
- We focus on **Barrage of Random Transforms (BaRT)** [Raff et al., 2019] which sequentially applies multiple image transforms to its inputs in random order and with random parameters. Transforms are sampled from a pool of over 20, both differentiable and not.



Original Evaluation of BaRT [Raff et al., 2019]

- Raff et al. [2019] used PGD Attack on steroid:

Original Evaluation of BaRT [Raff et al., 2019]

- Raff et al. [2019] used PGD Attack on steroid:
  - **BPDA** (Backward-Pass Differentiable Approximation): use neural networks to approximate gradients of non-differentiable transforms.

Original Evaluation of BaRT [Raff et al., 2019]

- Raff et al. [2019] used PGD Attack on steroid:
    - **BPDA** (Backward-Pass Differentiable Approximation): use neural networks to approximate gradients of non-differentiable transforms.
    - **EoT** (Expectation over Transformations): deal with randomness.

Original Evaluation of BaRT [Raff et al., 2019]

- Raff et al. [2019] used PGD Attack on steroid:
  - **BPDA** (Backward-Pass Differentiable Approximation): use neural networks to approximate gradients of non-differentiable transforms.
  - **EoT** (Expectation over Transformations): deal with randomness.
- This was state-of-the-art attack at the time.

Original Evaluation of BaRT [Raff et al., 2019]

- Raff et al. [2019] used PGD Attack on steroid:
  - **BPDA** (Backward-Pass Differentiable Approximation): use neural networks to approximate gradients of non-differentiable transforms.
  - **EoT** (Expectation over Transformations): deal with randomness.
- This was state-of-the-art attack at the time.
- A large improvement compared to adversarial training on deterministic models. Increases adversarial accuracy from 1.5% to 36%.

Original Evaluation of BaRT [Raff et al., 2019]

- Raff et al. [2019] used PGD Attack on steroid:
  - **BPDA** (Backward-Pass Differentiable Approximation): use neural networks to approximate gradients of non-differentiable transforms.
  - **EoT** (Expectation over Transformations): deal with randomness.
- This was state-of-the-art attack at the time.
- A large improvement compared to adversarial training on deterministic models. Increases adversarial accuracy from 1.5% to 36%.

Accuracy of multiple models trained ImageNet Raff et al. [2019].

	Clean Images		Attacked	
Model	Top-1	Top-5	Top-1	Top-5
Inception v3	78	94	0.7	4.4
Inception v3 w/Adv. Train	78	94	1.5	5.5
ResNet50	76	93	0.0	0.0
ResNet50-BaRT, $k = 5$	65	85	16	51
ResNet50-BaRT, $k = 10$	65	85	36	57

BPDA Attack is NOT Sufficiently Strong

Table 1: BaRT replicate on a 10-class subset of ImageNet dataset.

Transforms used in BaRT	Adversarial accuracy		
	Exact	BPDA	Identity
All	n/a	52	**36**
Only differentiable	**26**	65	41

- *Exact*: PGD attack with exact gradients.
- *Identity*: PGD attack with the transforms ignored in the backward pass (treated as an identity function).

BPDA Attack is NOT Sufficiently Strong

Table 1: BaRT replicate on a 10-class subset of ImageNet dataset.

Transforms used in BaRT	Adversarial accuracy		
	Exact	BPDA	Identity
All	n/a	52	**36**
Only differentiable	**26**	65	41

- *Exact*: PGD attack with exact gradients.
- *Identity*: PGD attack with the transforms ignored in the backward pass (treated as an identity function).
- **We found that BPDA attack is much weaker than Exact and is surprisingly weaker than Identity.**

Takeaway 1: Focus on Differentiable Transforms

- We suggest that future works **focus on differentiable transformations** only as part of a stochastic defense (until there is a reliable black-box attack or gradient approximation technique).

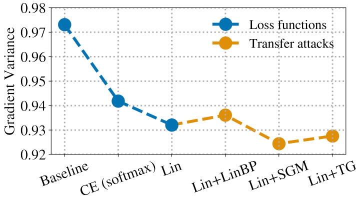
Takeaway 1: Focus on Differentiable Transforms

- We suggest that future works **focus on differentiable transformations** only as part of a stochastic defense (until there is a reliable black-box attack or gradient approximation technique).
- Separate studies on stochastic and non-differentiable models

Takeaway 1: Focus on Differentiable Transforms

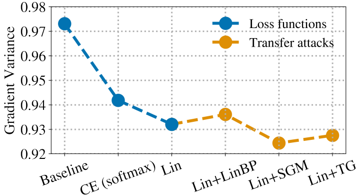
- We suggest that future works **focus on differentiable transformations** only as part of a stochastic defense (until there is a reliable black-box attack or gradient approximation technique).
- Separate studies on stochastic and non-differentiable models
- Benefits of using only differentiable transforms:
  - More accurate and efficient evaluation
  - Compatible with adversarial training

Better Attack on (Differentiable) Transform Defense



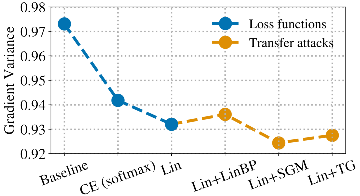
- Setting: non-convex, constrained SGD

Better Attack on (Differentiable) Transform Defense



- Setting: non-convex, constrained SGD
- Attack success rate is highly correlated with variance of gradient estimates.

Better Attack on (Differentiable) Transform Defense



- Setting: non-convex, constrained SGD
- Attack success rate is highly correlated with variance of gradient estimates.
- Key is **variance reduction**.

Better Attack on (Differentiable) Transform Defense

Algorithm 1 Our best attack on RT defenses

Input: Perturbation size ϵ, max. PGD steps T, step size $\{\gamma_t\}_{t=1}^{T}$, and AggMo's damping constants $\{\mu_b\}_{b=1}^{B}$.
Output: Adversarial examples x_{adv}
Data: Test input x and its ground-truth label y
$u \sim \mathcal{U}[-\epsilon, \epsilon], \quad x_{\text{adv}} \leftarrow x + u, \quad \{v_b\}_{b=1}^{B} \leftarrow \mathbf{0}$
for $t = 1$ **to** T **do**
  $\{\theta_i\}_{i=1}^{n} \sim p(\theta)$
  $G_n \leftarrow \nabla \mathcal{L}_{\text{Linear}} \left(\frac{1}{n} \sum_{i=1}^{n} f(t(x_{\text{adv}}; \theta_i)), y \right)$
  $\hat{G}_n \leftarrow \text{Clip}(G_n, \frac{-1}{\sqrt{d}}, \frac{1}{\sqrt{d}})$
  for $b = 1$ **to** B **do**
    $v_b \leftarrow \mu_b \cdot v_b + \hat{G}_n$
  end for
  $x_{\text{adv}} \leftarrow x_{\text{adv}} + \frac{\gamma_t}{B} \cdot \text{Sign} \left(\sum_{b=1}^{B} v_b \right)$
end for

- Setting: non-convex, constrained SGD
- Attack success rate is highly correlated with variance of gradient estimates.
- Key is **variance reduction**.
- Linear loss on logits

Better Attack on (Differentiable) Transform Defense

Algorithm 1 Our best attack on RT defenses

Input: Perturbation size ϵ, max. PGD steps T, step size $\{\gamma_t\}_{t=1}^{T}$, and AggMo's damping constants $\{\mu_b\}_{b=1}^{B}$.
Output: Adversarial examples x_{adv}
Data: Test input x and its ground-truth label y
$u \sim \mathcal{U}[-\epsilon, \epsilon], \quad x_{\text{adv}} \leftarrow x + u, \quad \{v_b\}_{b=1}^{B} \leftarrow \mathbf{0}$
for $t = 1$ **to** T **do**
  $\{\theta_i\}_{i=1}^{n} \sim p(\theta)$
  $G_n \leftarrow \nabla \mathcal{L}_{\text{Linear}} \left(\frac{1}{n} \sum_{i=1}^{n} f(t(x_{\text{adv}}; \theta_i)), y \right)$
  $\hat{G}_n \leftarrow \text{Clip}(G_n, \frac{-1}{\sqrt{d}}, \frac{1}{\sqrt{d}})$
  for $b = 1$ **to** B **do**
    $v_b \leftarrow \mu_b \cdot v_b + \hat{G}_n$
  end for
  $x_{\text{adv}} \leftarrow x_{\text{adv}} + \frac{\gamma_t}{B} \cdot \text{Sign}\left(\sum_{b=1}^{B} v_b \right)$
end for

- Setting: non-convex, constrained SGD
- Attack success rate is highly correlated with variance of gradient estimates.
- Key is **variance reduction**.
- Linear loss on logits
- Signed gradients and momentum

Better Attack on (Differentiable) Transform Defense

Algorithm 1 Our best attack on RT defenses

Input: Perturbation size ϵ, max. PGD steps T, step size $\{\gamma_t\}_{t=1}^{T}$, and AggMo's damping constants $\{\mu_b\}_{b=1}^{B}$.
Output: Adversarial examples x_{adv}
Data: Test input x and its ground-truth label y
$u \sim \mathcal{U}[-\epsilon, \epsilon], \quad x_{\text{adv}} \leftarrow x + u, \quad \{v_b\}_{b=1}^{B} \leftarrow \mathbf{0}$
for $t = 1$ **to** T **do**
  $\{\theta_i\}_{i=1}^{n} \sim p(\theta)$
  $G_n \leftarrow \nabla \mathcal{L}_{\text{Linear}}\left(\frac{1}{n}\sum_{i=1}^{n} f(t(x_{\text{adv}};\theta_i)), y\right)$
  $\hat{G}_n \leftarrow \text{Clip}(G_n, \frac{-1}{\sqrt{d}}, \frac{1}{\sqrt{d}})$
  for $b = 1$ **to** B **do**
    $v_b \leftarrow \mu_b \cdot v_b + \hat{G}_n$
  end for
  $x_{\text{adv}} \leftarrow x_{\text{adv}} + \frac{\gamma_t}{B} \cdot \text{Sign}\left(\sum_{b=1}^{B} v_b\right)$
end for

- Setting: non-convex, constrained SGD
- Attack success rate is highly correlated with variance of gradient estimates.
- Key is **variance reduction**.
- Linear loss on logits
- Signed gradients and momentum
- AggMo optimizer [Lucas et al., 2019]

Better Attack on (Differentiable) Transform Defense

Algorithm 1 Our best attack on RT defenses

Input: Perturbation size ϵ, max. PGD steps T, step size $\{\gamma_t\}_{t=1}^{T}$, and AggMo's damping constants $\{\mu_b\}_{b=1}^{B}$.
Output: Adversarial examples x_{adv}
Data: Test input x and its ground-truth label y

$u \sim \mathcal{U}[-\epsilon, \epsilon], \quad x_{\text{adv}} \leftarrow x + u, \quad \{v_b\}_{b=1}^{B} \leftarrow \mathbf{0}$

for $t = 1$ **to** T **do**

  $\{\theta_i\}_{i=1}^{n} \sim p(\theta)$

  $G_n \leftarrow \nabla \mathcal{L}_{\text{Linear}} \left(\frac{1}{n} \sum_{i=1}^{n} f(t(x_{\text{adv}}; \theta_i)), y \right)$

  $\hat{G}_n \leftarrow \text{Clip}(G_n, \frac{-1}{\sqrt{d}}, \frac{1}{\sqrt{d}})$

  for $b = 1$ **to** B **do**

    $v_b \leftarrow \mu_b \cdot v_b + \hat{G}_n$

  end for

  $x_{\text{adv}} \leftarrow x_{\text{adv}} + \frac{\gamma_t}{B} \cdot \text{Sign} \left(\sum_{b=1}^{B} v_b \right)$

end for

- Setting: non-convex, constrained SGD
- Attack success rate is highly correlated with variance of gradient estimates.
- Key is **variance reduction**.
- Linear loss on logits
- Signed gradients and momentum
- AggMo optimizer [Lucas et al., 2019]
- Improve transferability (SGM [Wu et al., 2020])

Robustness Results and Attack Comparison

Table 2: Comparison between the baseline attack, AutoAttack (standard version + EoT), and our attack on differentiable Random Transform Defense.

Attack	Accuracy	
	CIFAR-10	Imagenette
No attack	81	89
Baseline	33	70
AutoAttack	61	85
Our attack	**29**	**6**

- Our attack beats the baseline (PGD+EoT) and AutoAttack by a large margin. Even a carefully tuned BaRT is not robust.

Robustness Results and Attack Comparison

Table 2: Comparison between the baseline attack, AutoAttack (standard version + EoT), and our attack on differentiable Random Transform Defense.

Attack	Accuracy	
	CIFAR-10	Imagenette
No attack	81	89
Baseline	33	70
AutoAttack	61	85
Our attack	**29**	**6**

- Our attack beats the baseline (PGD+EoT) and AutoAttack by a large margin. Even a carefully tuned BaRT is not robust.
- We also use our attack to adversarially trained BaRT, but it is still not as robust as adversarial training on a deterministic network.

Takeaway 2: Better Attacks

- Attacks on Random Transform Defense is much less efficient compared to deterministic models.

Takeaway 2: Better Attacks

- Attacks on Random Transform Defense is much less efficient compared to deterministic models.
- For better attacks, try
  - Reducing variance of gradient estimates.
  - Using a lot of steps (at least a few thousands).
  - Using momentum and accelerated gradient methods when possible.

Thank You!

Come see our poster at **Hall E #215** (Poster Session 1)!

References I

J. Lucas, S. Sun, R. Zemel, and R. Grosse. Aggregated momentum: Stability through passive damping. In *International Conference on Learning Representations*, 2019.

E. Raff, J. Sylvester, S. Forsyth, and M. McLean. Barrage of random transforms for adversarially robust defense. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6521–6530, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00669.

D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma. Skip connections matter: On the transferability of adversarial examples generated with ResNets. In *International Conference on Learning Representations*, 2020.