Demystifying the Adversarial Robustness of Random Transformation Defenses

Chawin Sitawarin   Zachary Golan-Strieb   David Wagner

UC Berkeley

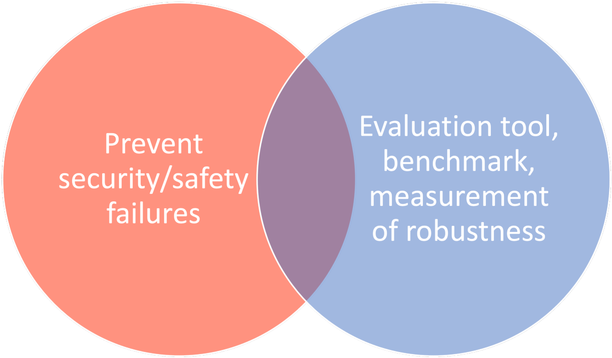
AAAI Workshop on Adversarial Machine Learning 2022

Contact: `chawins@berkeley.edu`

Outline

- Introduction
- Part 1: Pitfalls of BPDA Attack
- Part 2: Our Best Attack

Why Study Adversarial Examples?

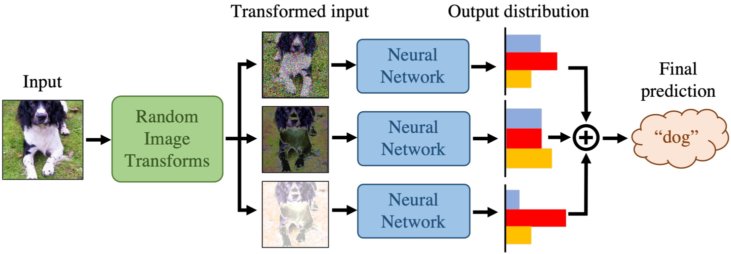


- To achieve both goals, we need an accurate tool for measuring the robustness of machine learning models in diverse settings.

Random Transformations as a Defense

- Many works have proposed noise / random transforms as a way to improve the robustness of neural networks, e.g.,
  - **Random image transforms**: Dhillon et al. [2018], Xie et al. [2018], He et al. [2019], Zhang and Liang [2019], Bender et al. [2020]
  - **Randomized smoothing**: Liu et al. [2018], Lecuyer et al. [2019], Cohen et al. [2019].
  - **Random weights**: Liu et al. [2019]
- However, stochastic defenses are poorly understood, and we still lack reliable tools for measuring their robustness.
- This work tries to address this problem and particularly focuses on *Barrage of Random Transforms or BaRT* [Raff et al., 2019] (CVPR 2019) which claims a significant robustness result on ImageNet.

Notation: Random Transform Defense



- Average softmax output over a distribution of random transformations.
- Expectation is approximated by Monte Carlo sampling.
- BaRT sequentially applies k different transformations for each Monte Carlo sample (n samples for one input).

Original Evaluation of BaRT

- Raff et al. [2019] use Backward-Pass Differentiable Approximation (BPDA) to "approximate" gradients for non-differentiable transforms by substituting with a surrogate neural network.
- Use PGD attack with Expectation over Transformations (EoT).
- Find a large robustness improvement compared to adversarial training:

	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

Ref: Raff et al. [2019]

Outline

- Introduction
- Part 1: Pitfalls of BPDA Attack
- Part 2: Our Best Attack

BPDA Attack is NOT Sufficiently Strong

Table: Comparison of attacks with different gradient approximations on BaRT with all transformations and only differentiable ones. Lower = better attack.

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.**

Pitfalls of BPDA

- BPDA cannot approximate some transforms because the architecture has limited expressivity, e.g., small “receptive field” = cannot approximate large geometric transforms.

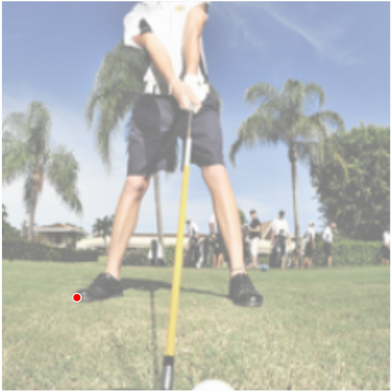
Original

Real zoom transform

Pitfalls of BPDA

- BPDA cannot approximate some transforms because the architecture has limited expressivity, e.g., small "receptive field" = cannot approximate large geometric transforms.

Original



Real zoom transform

Pitfalls of BPDA

- BPDA cannot approximate some transforms because the architecture has limited expressivity, e.g., small "receptive field" $=$ cannot approximate large geometric transforms.

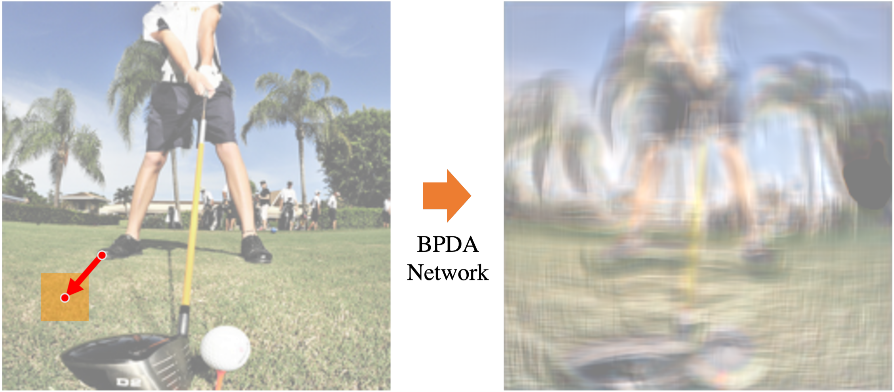
BPDA Network

Original

Real zoom transform

Pitfalls of BPDA

- BPDA cannot approximate some transforms because the architecture has limited expressivity, e.g., small "receptive field" = cannot approximate large geometric transforms.



Original

BPDA zoom transform

Pitfalls of BPDA

- The BPDA network overfits to training images.
- During the attack, the trained BPDA networks are given partially transformed images, yet the BPDA networks are only trained with untransformed inputs.
- Since we are backpropagating through several transforms, one poor transform gradient approximation could ruin the entire estimate.

Focus on Differentiable Transforms

- We suggest that future works focus only on differentiable transformations as part of a stochastic defense (until there is a reliable black-box or gradient approximation technique).
- Separate studies on stochastic and non-differentiable models
- Benefits of using only differentiable transforms: (i) more accurate and efficient evaluation, (ii) adversarial training.
- From this point on, we only consider differentiable transforms and use Bayesian optimization to tune the transforms' hyperparameters.

Outline

- Introduction
- Part 1: Pitfalls of BPDA Attack
- Part 2: Our Best Attack

Our Best Attack: Overview

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: stochastic optimization (non-convex, constrained)
- Design principle: variance reduction

Our Best Attack: Objective Function

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**

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

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

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

  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

- Linear loss
- Improve transferability with SGM [Wu et al., 2020]

Our Best Attack: Gradient Clipping

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

- Clipped gradients remove outliers and reduce variance

Our Best Attack: Optimizer

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

- Aggregated Momentum (AggMo) as optimizer: momentum that is not very sensitive to hyperparameters
- Signed updates for ℓ_∞-norm constraint

Robustness Results and Attack Comparison

Table: Comparison between the baseline EoT attack, AutoAttack, and our attack on the differentiable RT defense.

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

- Our attack beats EoT attack and AutoAttack in both randomized and standard modes by a large margin.

Summary & Open Problems

- We show that even an adaptive technique for circumventing non-differentiability (i.e.,BPDA) is not effective against existing RT defenses and reveal that these defenses are likely non-robust.
- We propose a new state-of-the-art attack for random transform defenses, improving the baseline EoT attack and explaining its effectiveness through variance of the gradient estimates.

Future improvements:

- Study other defenses that we have not considered, but our findings may apply (e.g., randomized smoothing, weight perturbation).
- These defenses are interesting settings to study stochastic optimization methods (e.g., variance reduction, acceleration).
- Black-box and standardized attacks for stochastic defenses.

Thank You!

References I

C. Bender, Y. Li, Y. Shi, M. K. Reiter, and J. Oliva. Defense through diverse directions. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 756–766. PMLR, July 2020.

J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, June 2019.

G. S. Dhillon, K. Azizzadenesheli, J. D. Bernstein, J. Kossaifi, A. Khanna, Z. C. Lipton, and A. Anandkumar. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations*, 2018.

References II

Z. He, A. S. Rakin, and D. Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–597, 2019.

M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672, 2019. doi: 10.1109/SP.2019.00044.

X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh. Towards robust neural networks via random self-ensemble. In *ECCV (7)*, pages 381–397, 2018.

X. Liu, Y. Li, C. Wu, and C.-J. Hsieh. Adv-BNN: Improved adversarial defense through robust bayesian neural network. In *International Conference on Learning Representations*, 2019.

References III

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.

C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

Y. Zhang and P. Liang. Defending against whitebox adversarial attacks via randomized discretization. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 684–693. PMLR, Apr. 2019.