ICML | 2022
Thirty-ninth International Conference on Machine Learning

Demystifying the Adversarial Robustness of Random Transformation Defenses

Chawin Sitawarin (*chawins@berkeley.edu*) Zachary Golan-Strieb David Wagner

Berkeley UNIVERSITY OF CALIFORNIA
BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Motivations

- Many works have proposed random input transformation to improve the adversarial robustness of neural networks.
- Unlike deterministic models, **stochastic defenses are poorly understood, and reliable tools for measuring their robustness are lacking.**
- We address this problem, focusing on **Barrage of Random Transforms** or **BaRT** [Raff et al., 2019] (CVPR 2019). BaRT applies multiple transforms sequentially to its inputs in random order and with random parameters.

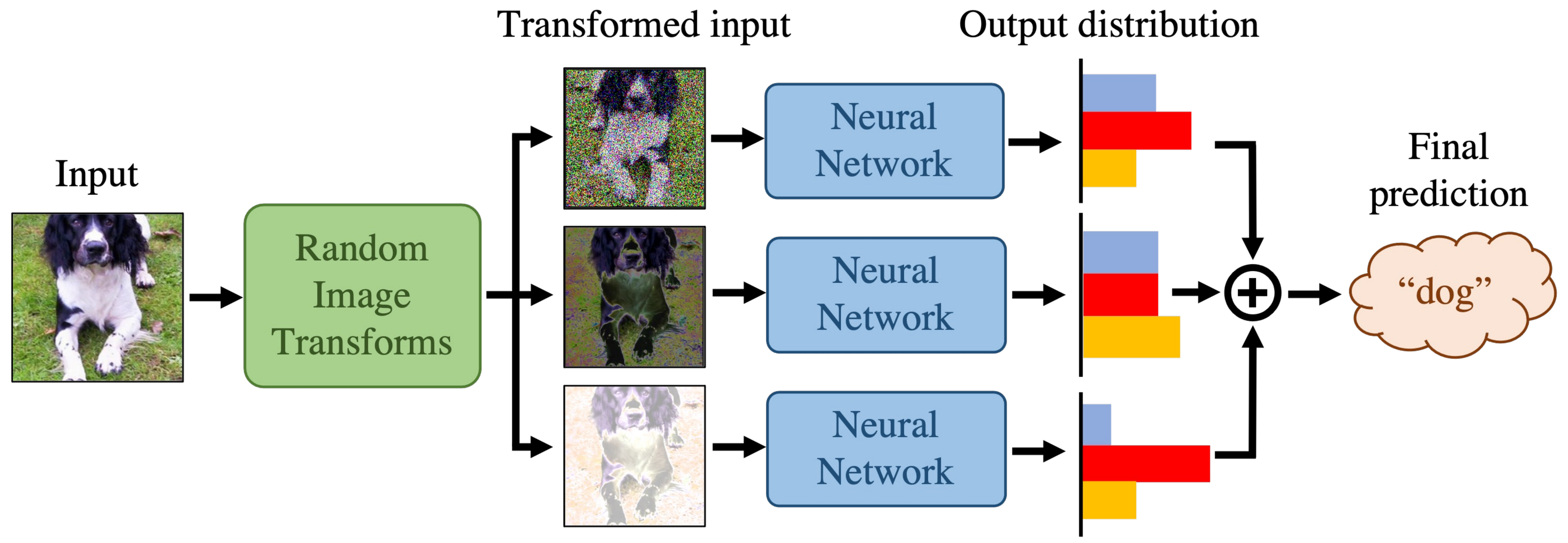


Fig 1: Diagram of a generic random transform defense

- BaRT was evaluated with the state-of-the-art method at the time:
  PGD + **EoT** (Expectation over Transformation)
  \+ **BPDA** (Backward-Pass Differentiable Approximation)
- EoT [Athalye et al., 2017] deals with the randomness
- BPDA [Athalye et al., 2018] deals with the non-differentiable transforms by using a trained neural network to approximate each of them and backprop through the networks as a proxy.
- They claim a huge robustness improvement on ImageNet. Increases adversarial accuracy from 1.5% to 36%.

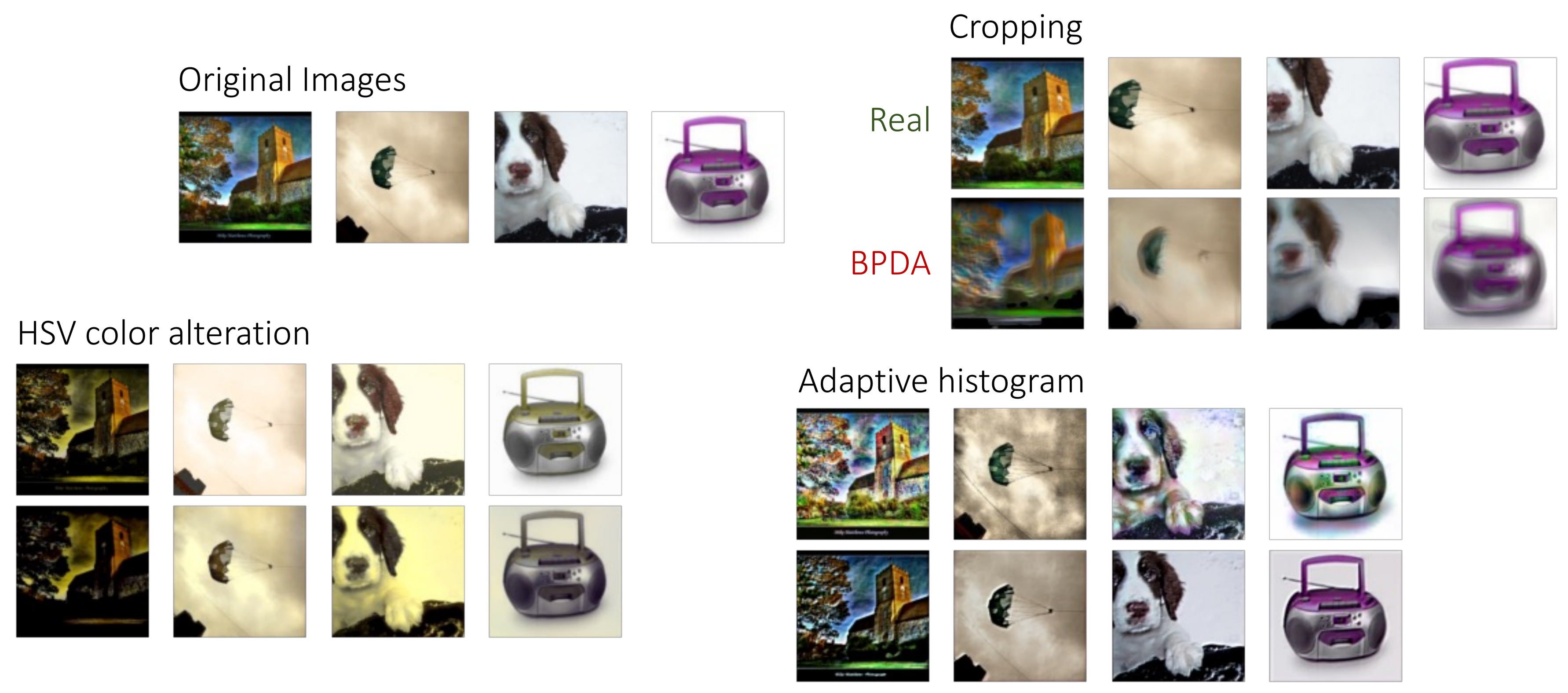
Model	Clean Images Top-1	Clean Images Top-5	Attacked Top-1	Attacked Top-5
Inception v3	78	94	0.7	4.4
Inception v3 w/Adv. Train	78	94	1.5	5.5
ResNet50	76	93	0.0	0.0
ResNet50-BaRT, $k = 5$	65	85	16	51
ResNet50-BaRT, $k = 10$	65	85	36	57

BPDA is Not Sufficiently Strong

Table 1: Effectiveness of PGD attack with different gradient approximation method on Imagenette dataset (10-class subset of ImageNet). $\epsilon_\infty = 16/255$ and 40 steps.

Transforms Used in BaRT	Adversarial Accuracy with Different Gradient Approximations: Exact	BPDA	Identity	Combo
Full	n/a	52.32	36.49	**25.24**
Only Differentiable	**26.06**	65.28	41.25	n/a

- *Exact*: PGD attack with exact gradients. *Identity*: ignore transform in the backward pass. *Combo*: BPDA (non-diff.) + Exact (diff.)
- BPDA attack is much weaker than any other gradient approximation.
- Why does BPDA fail?
  - Cannot approximate the transforms well enough
  - Overfits to training images which are all clean
  - Error amplifies with more transforms

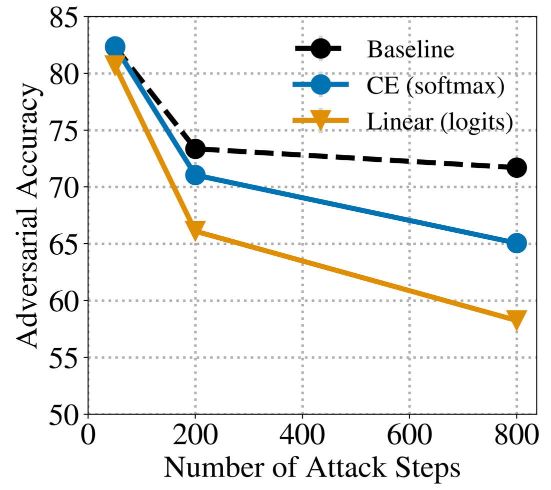


Takeaway 1

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

Stronger Attack on (Differentiable) Random Transform Defense

- Even with differentiable transforms alone, current attack is sub-optimal.
- Requires thousands of steps but does not converge to good local optima.
- Attack on Random Transform Defense = SGD.
- **Our attack combines baseline (PGD+EoT) with multiple techniques:**
  - Variance reduction
  - Signed gradients and momentum
  - Improved transferability with SGM [Wu et al., 2020]
  - Linear loss on logits
  - AggMo optimizer (acceleration & less tuning) [Lucas et al., 2019]



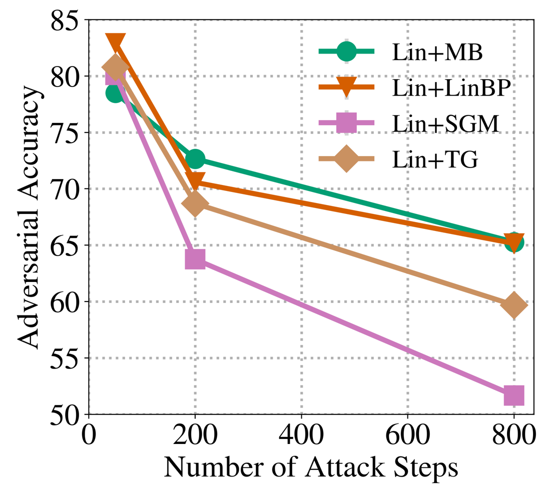
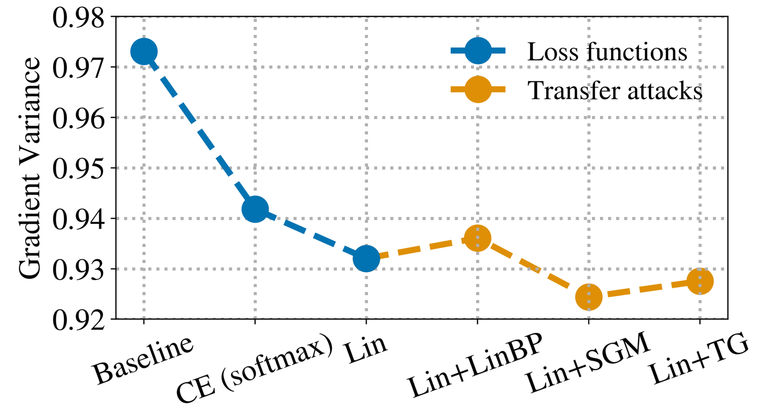


Table 2: Attack comparison on Random Transform defense. AutoAttack uses standard version combined with EoT. For Imagenette, $\epsilon = 16/255$, and for CIFAR-10, $\epsilon = 8/255$.

Attacks	Accuracy: CIFAR-10	Imagenette
No attack	81.12 ± 0.54	89.04 ± 0.34
Baseline	33.83 ± 0.44	70.79 ± 0.53
AutoAttack	61.13 ± 0.85	85.46 ± 0.43
Our attack	$\mathbf{29.91} \pm 0.35$	$\mathbf{6.34} \pm 0.35$

- Attack effectiveness is strongly correlated to variance of the gradient estimates.



Defenses	Imagenette Clean Accuracy	Imagenette Adv. Accuracy	CIFAR-10 Clean Accuracy	CIFAR-10 Adv. Accuracy
Normal model	**95.41**	0.00	**95.10**	0.00
Madry et al. (2018)	78.25	**37.10**	81.90	45.30
Zhang et al. (2019)	87.43	33.19	81.26	**46.89**
RT defense	89.04 ± 0.34	6.34 ± 0.35	81.12 ± 0.54	29.91 ± 0.35
AdvRT defense	88.83 ± 0.26	8.68 ± 0.52	80.69 ± 0.66	41.30 ± 0.49

Takeaway 2

- Randomness makes attacks a lot less efficient.
- For better attacks, try (1) reducing variance of the gradients, (2) using accelerated methods, (3) running the attack with lots of steps.
- Combining the defense with adversarial training helps but is *not as good as* adversarial training on normal deterministic neural networks.