# BART: BARRAGE OF RANDOM TRANSFORMS

## FOR ADVERSARIALLY ROBUST DEFENSE

Booz | Allen | Hamilton

*Edward Raff* [1,2,4] *Jared Sylvester* [1,2,4] *Steven Forsyth* [3] *Mark McLean* [1]

1 Laboratory for Physical Sciences 2 Booz Allen Hamilton 3 NVIDIA 4 U.M.B.C.

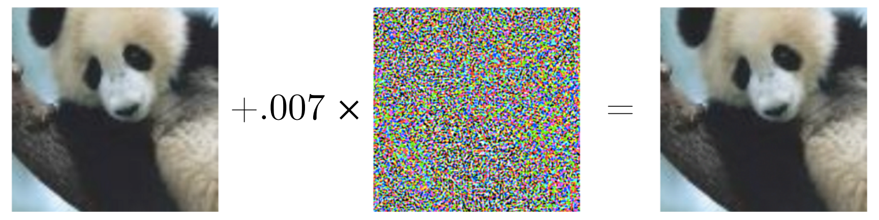
IEEE Conference on Computer Vision & Pattern Recognition (CVPR), 16–20 June 2019, Long Beach

## Abstract

*Defenses against adversarial examples are historically easy to defeat. The common understanding is that a combination of simple image transformations and other various defenses are insufficient to provide the necessary protection when the obfuscated gradient is taken into account. In this paper, we explore the idea of stochastically combining a large number of individually weak defenses into a single barrage of randomized transformations to build a strong defense against adversarial attacks. We show that, even after accounting for obfuscated gradients, the Barrage of Random Transforms (BaRT) is a resilient defense against even the most difficult attacks, such as PGD. BaRT achieves up to a 24x improvement in accuracy compared to previous work, and has even extended effectiveness to a previously untested maximum adversarial perturbation of* $\varepsilon = 32$.

## Motivation & Approach

*Problem:* An attacker can make small perturbations to inputs that are numerically significant, but semantically & perceptually meaningless.



panda, 57.7% attack perturbation "gibbon," 99.3%

*Solution: Make our own perturbations to inputs.*

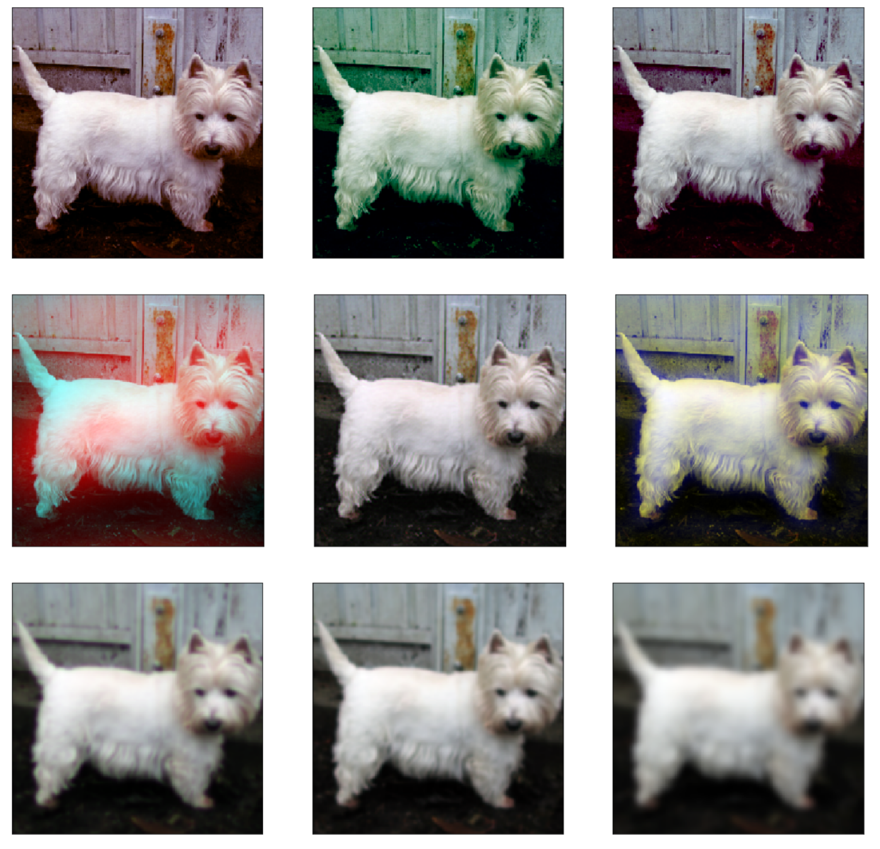
(Image credit: Goodfellow, Shlens & Szegedy. "Explaining and Harnessing Adversarial Examples." ICLR, 2015.)

## Method

- Modify the image at inference time (e.g. by blurring, adding noise, desaturating, etc.).
  - This should interfere with the adversary's ability to find a successful attack perturbation.
- This has been tried before … and it didn't work.
  - Input transforms make following the gradient between the original and attacked image only trivially harder.
- Input transformations haven't worked before, *so what makes BaRT different?*
- Our approach:
  1. Take a large set of transformations.
  2. Parameterize each one randomly.
  3. Select a random subset to apply for each input.
  4. Apply them in randomized, serial order.
- *Individually, all of our 25 transforms can be easily defeated, but together they form a strong ensemble.*

## Randomized Transforms

### Example outputs from single, randomized transforms



Alter XYZ: Convert to CIE XYZ color space, perturb w/ random offset, convert back to RGB.

FFT Alteration: Randomly scale and/or zero out some coefficients in frequency space.

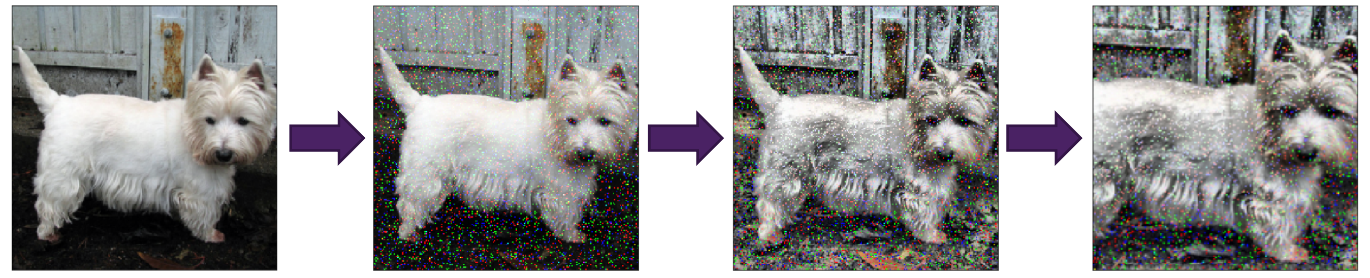
Gaussian Blur: Blur using a Gaussian with randomly chosen standard deviation.

Example output #1 Example output #2 Example output #3

### Transforms are applied in series

Transforms are randomly selected and then applied in a random order, using random parameters.

Original Image Transform 1: Noise Injection & Transform 2: Histogram Eq. & Transform 3: Partial Zoom



### "Randomness on top of randomness"

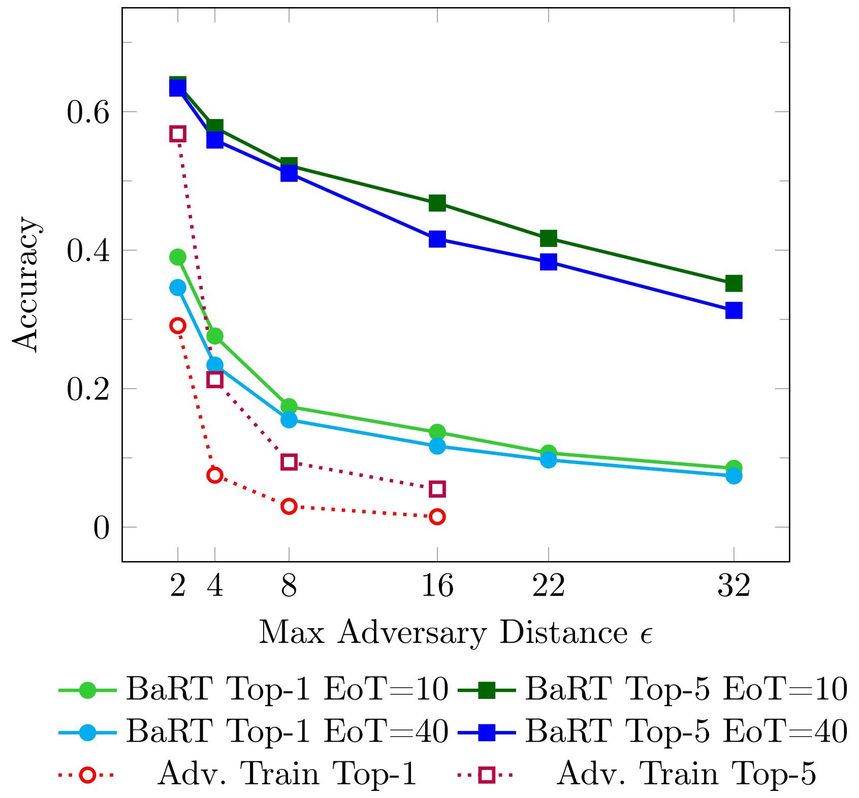
Randomization removes information from the expectation of the gradient. You don't get to attack an single input, you have to attack the expectation over an infinite set of inputs like this:

Example outputs with five random transforms applied.

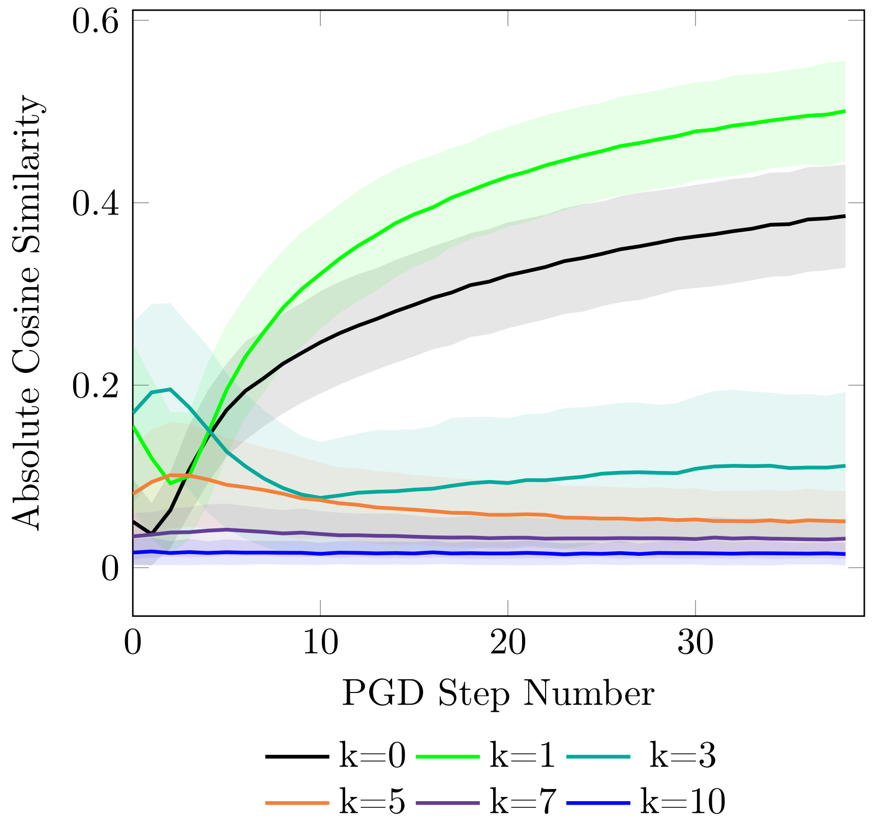
### Creating a strong adversary

- Used PGD to construct the strongest adversary we could.
- Implemented BPDA and EoT to allow the adversary to approximate each transform and avoid reliance on obfuscated gradients.
- Allowed the attacker to know the randomly chosen parameters of each defense.
- Allowed *adversarial distance of up to* $\varepsilon = 32$.
- Thoroughly tested for vanishing & exploding gradients.
- Created a new attack we thought might be better able to defeat BaRT: Medoid-over-Transform.

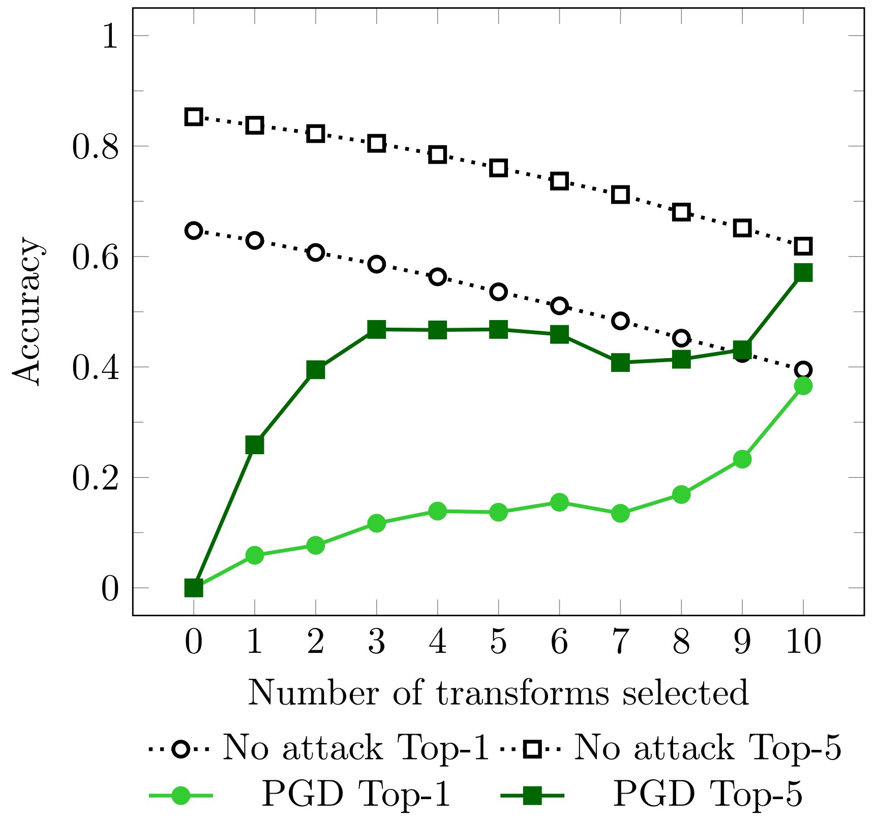
## Results



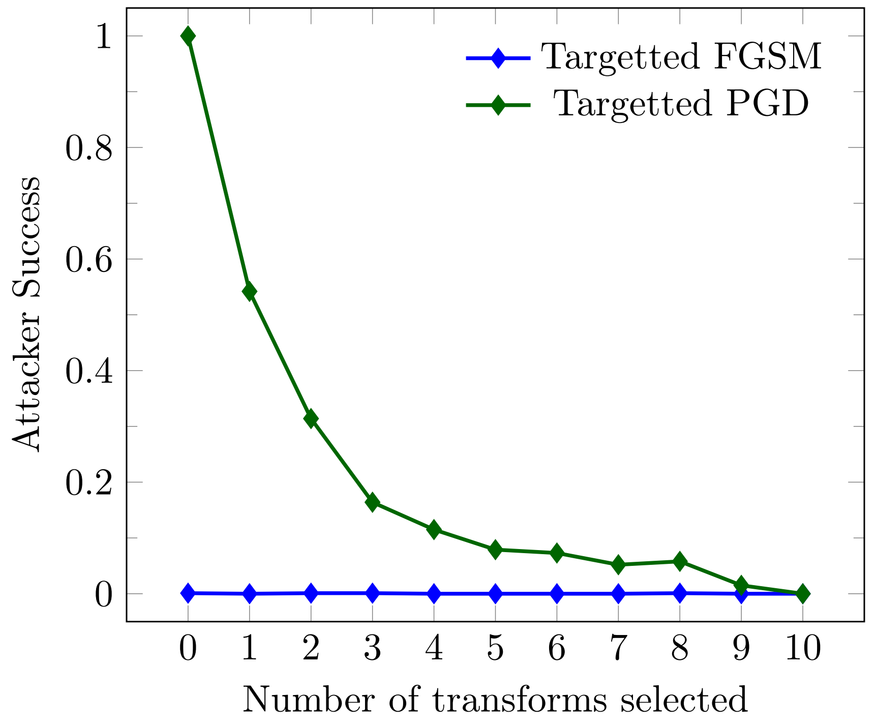
- *BaRT surpasses the previous state-of-the-art defense for ImageNet.* (Adversarial Training; Kurakin et al. 2017).
- Top-5 accuracy >57% when attacked.
- Higher Top-1 accuracy than the Top-5 accuracy of Adversarial Training when ε ≥ 4.



- The cosine similarity between successive steps taken by the PGD attacker is low when the number of transforms k > 1.
- This indicates that the attacker is hindered in finding a path to an adversarial example.



- Adding more transforms to the ensemble costs accuracy when not under attack, but increases it when being attacked.



- For targeted attacks with no defensive transforms, the PGD attacker wins 100% of the time.
- With 10 transforms, *the attacker's success falls to 0%.*