On the Robustness of the CVPR 2018 White-box Adversarial Example Defenses

Anish Athalye* (aathalye@mit.edu) and Nicholas Carlini* (npc@cs.berkeley.edu)

Pixel Deflection

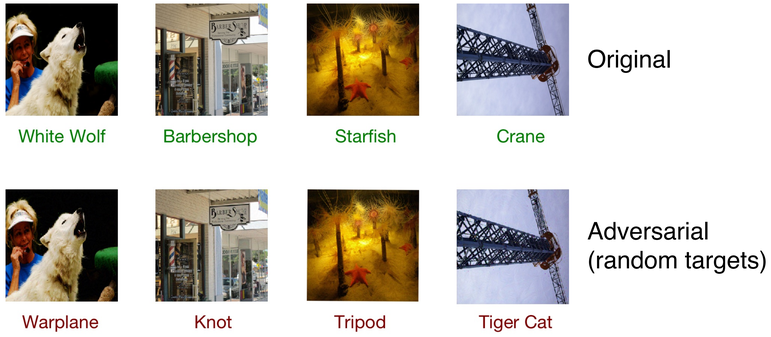
Defense Overview

- Pixel Deflection is a technique that defends against adversarial examples by randomly replacing pixels with nearby neighbors.
- This defense is non-differentiable and randomized, making existing attack algorithms fail at generating adversarial examples.

Attack

- We apply BPDA and EOT to generate adversarial examples.
- Evaluated over 1000 randomly chosen ImageNet images, with an l-infinity perturbation bound of 4/255, our attack reduces the accuracy of the defended classifier to 3%.

Demonstration



github.com/carlini/pixel-deflection

Background

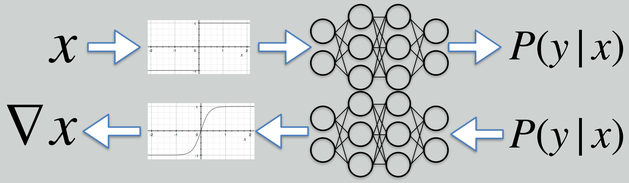
- Adversarial Examples are inputs specifically crafted to fool a neural network.

Evaluation

- We generate adversarial examples in the *white box* threat model, where the adversary knows the model architecture and parameters.

Backward Pass Differentiable Approximation

- BPDA is a generic attack framework introduced by Athalye *et al.* 2018a to generate adversarial examples for non-differentiable defenses



- Run the forward pass as usual to obtain the input Y given the network. On the backward pass, replace the non-differentiable pre-processing function with an approximation (e.g., by replacing hard thresholding with smoothed version).

Expectation Over Transformation

- EOT is an attack approach introduced by Athalye *et al.* 2018b to produce adversarial examples that are robust to randomized transformations sampled from a distribution T by optimizing the expected classification:

$$\mathbb{E}_{t\sim T}\left[\log P(y \mid t(x))\right]$$

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. ICML, 2018a.

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. Synthesizing robust adversarial examples. ICML, 2018b.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A. Towards deep learning models resistant to adversarial attacks. ICLR, 2018.

arxiv.org/abs/1804.03286

High-level Guided Denoiser

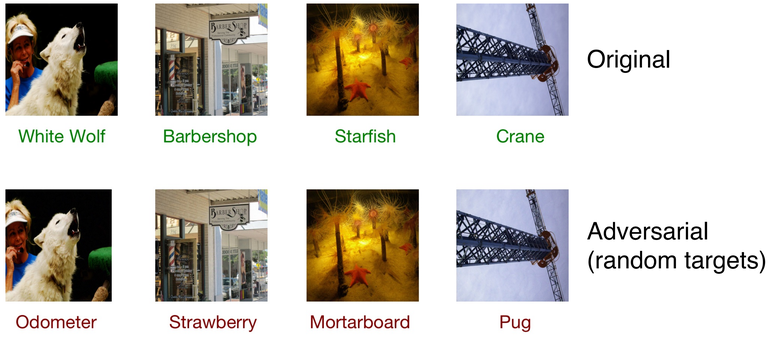
Defense Overview

- High-level representation Guided Denoiser is a technique that defends against adversarial examples by denoising inputs using a trained neural network before passing them to a standard classifier.
- The denoiser is a differentiable, non-randomized neural network.

Attack

- We apply Projected Gradient Descent (*Madry* et al. 2018), differentiating end-to-end through both the denoiser and the classifier.
- Evaluated over 1000 randomly chosen ImageNet images, with an l-infinity perturbation bound of 4/255, our attack reduces the accuracy of the defended classifier to 0%.

Demonstration



github.com/anishathalye/guided-denoise