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

Original

Adversarial (random targets)

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

Original

Adversarial (random targets)

References


