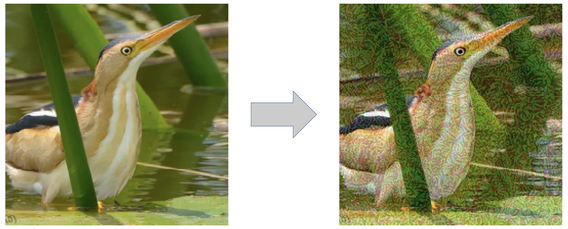
A New Defense Against Adversarial Images: Turning a Weakness Into a Strength

Tao Yu*, Shengyuan Hu*, Chuan Guo, Wei-Lun Chao, Kilian Q Weinberger
Department of Computer Science, Cornell University.
Department of Computer Science and Engineering, The Ohio State University.

Background:

Neural Networks are prone to imperceptible changes in the input -- adversarial perturbations -- that alter the model's decision entirely and can be efficiently discovered by gradient-guided search.



"bittern"
99.99% confidence

"canoe"
29.53% confidence

Motivation:

1) Robustness to random noise:
Classification of a real image $\mathbf{x}_0$ is robust to Gaussian noises due to the low density of adversarial perturbations. We measure this with L1 norm between probability vectors, where small values represent robust and real images.
This could be bypassed by moving from $\mathbf{x}_0$ to **x''**.

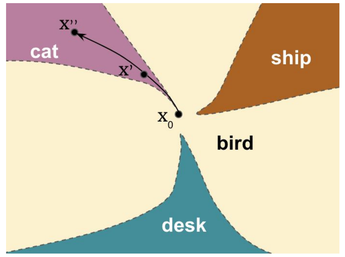
2) Close proximity to decision boundary:
A real image **x** can be easily attacked to a different class within several steps, which we consider to be the proximity to decision boundary. We measure this using least steps of PGD required to attack successfully.
This could be bypassed by moving from $\mathbf{x}_0$ to **x'**.

However, when combined together, optimization of 1) and 2) are contradictory!

Attack Terminology

Definition:
$\mathbf{x}$: image y: true label y_t: target label
$h(\mathbf{x})$: predicted class probability vector
$\mathbf{p}^{\text{adv}}$: target adversary probability distribution
$\mathcal{L}(\cdot,\cdot)$: cross-entropy loss
want image $\mathbf{x}$ misclassified as y_t:
$\mathcal{L}_1 = \mathcal{L}(h(\mathbf{x}), \mathbf{p}^{\text{adv}})$
want $\mathbf{x}$ being robust to random noise:
$\mathcal{L}_2 = \mathbb{E}_{\epsilon\sim N(0,\sigma^2 I)}\left[\|h(\mathbf{x}) - h(\mathbf{x}+\epsilon)\|_1\right]$

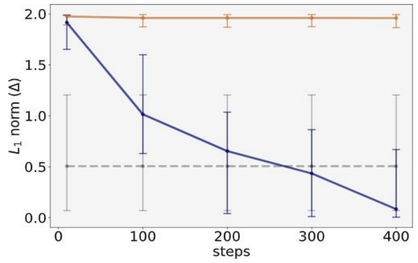
Image Distribution



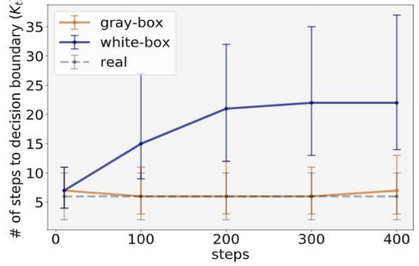
Detection Strategy:

C1: Robustness to Gaussian noise
C2t: Susceptibility to adversarial noise by targeted iterative attack
C2u: Susceptibility to adversarial noise by untargeted iterative attack

Robustness to Gaussian Noise



Steps required to change label



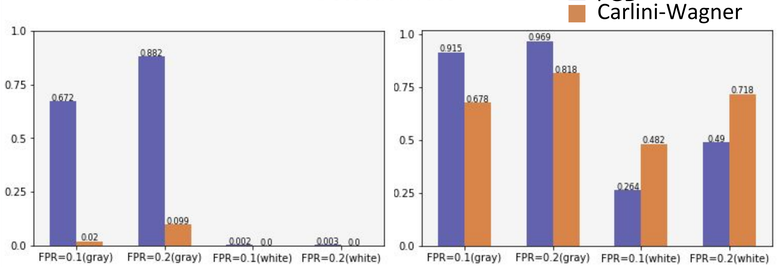
Best effort white-box adversary:

It is difficult to directly optimize the discrete C2t/u. We encourage the constructed adversarial image to change prediction to a different class from y_t after a single gradient step using targeted attack or untargeted attack.
Define $\delta_{y'} = \nabla_{\mathbf{x}}\mathcal{L}_{\mathcal{A}}(h(\mathbf{x}), y')$, and following two losses:
$\mathcal{L}_3 = \mathbb{E}_{y'\sim\text{Uniform}, y'\neq y_t}[\mathcal{L}(h(\mathbf{x}-\alpha\delta_{y'}), y')]$(bypass C2t) $\mathcal{L}_4 = -\mathcal{L}(h(\mathbf{x}+\alpha\delta_{y_t}), y_t)$(bypass C2u)
Loss for best effort white-box adversary is: $\mathcal{L}_* = \lambda\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4$

Results on ImageNet:

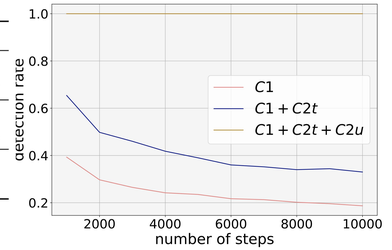
For white and gray box attack, we use PGD and CW to attack a pretrained ResNet101 for 50 steps under adversarial radius 0.1 (L-inf norm), with learning rates 0.01, 0.03, 0.1. Worst case detection results using Feature Squeezing and ours are reported below.

Detection rate



For black box attack, we use boundary attack and treat both the model and detection (in 3 different scenarios) as a black box. A successful adversary fools the detector and has MSE less than 0.01. We plot the detection rates as the number of steps changes.

Detector	FPR	PGD	CW
Small radius ($\tau = 0.03$)	0.2	0.571	0.713
Small radius ($\tau = 0.03$)	0.1	0.418	0.500
$\mathcal{L}^* = \mathcal{L}_1 + \mathcal{L}_2$	0.2	0.695	0.512
$\mathcal{L}^* = \mathcal{L}_1 + \mathcal{L}_2$	0.1	0.527	0.304
Untargeted Attack	0.2	0.994	0.538
Untargeted Attack	0.1	0.987	0.342



For ablation study, we evaluate our defense under different white box attack scenarios: small perceptibility(0.03); attacking criterion C1 only, and untargeted iterative attack. Worst case results are reported above.