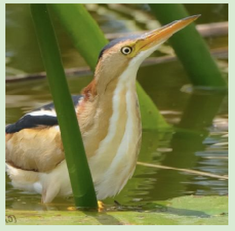
A New Defense Against Adversarial Images: Turning a Weakness Into a Strength

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 x is robust to Gaussian noises due to the low density of adversarial perturbations.

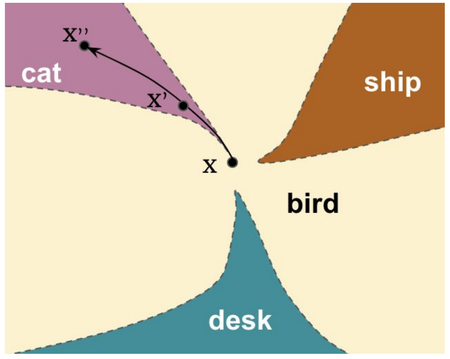
This could be bypassed by moving from x to x''.

2) Close proximity to decision boundary:
A real image x can be easily attacked to a different class within several steps, which we adopt to measure the proximity to decision boundary.

This could be bypassed by moving from x to x''.

However, 1) and 2) are contradictory!

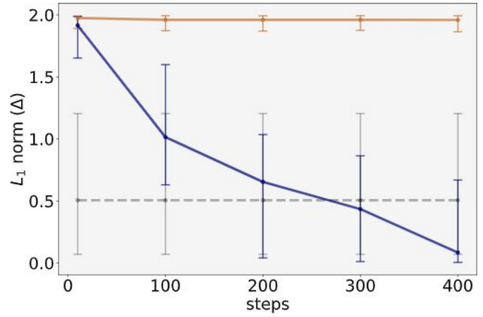
Image Distribution



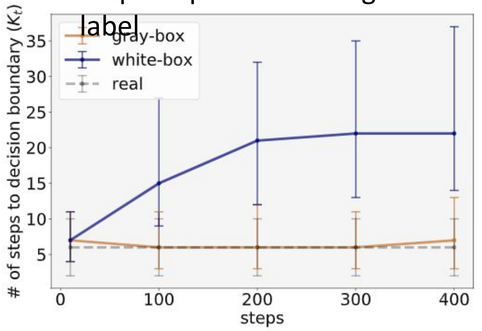
Detection Strategy:

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



Results on ImageNet:

Adversarial radius and some settings

Detection rate