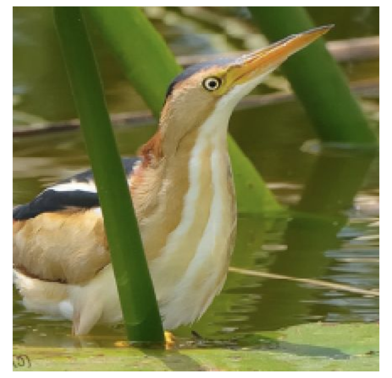
A New Defense Against Adversarial Images: Turning a Weakness Into a Strength

Tao Yu*[1], Shengyuan Hu*[1], Chuan Guo[1], Wei-Lun Chao[2], Kilian Q. Weinberger[1]
[1] Department of Computer Science, Cornell University.
[2] 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.

- **Common and inevitable:**
  - Gradient-guided search can perturb real images to any other class.
  - Any classifier has a fundamental limit on the robustness that it can achieve for adversarial examples.



"bittern"
99.99% confidence

- **Hard to defend against:**
  - Defenses such as adversarial training are very slow and not suitable for large datasets like ImageNet.
  - State-of-the-art defense (TRADES) achieves only 56% accuracy against strongest attack with a drop of 11% clean accuracy on CIFAR10.

"canoe"
29.53% confidence

Vulnerability of Existing Defenses

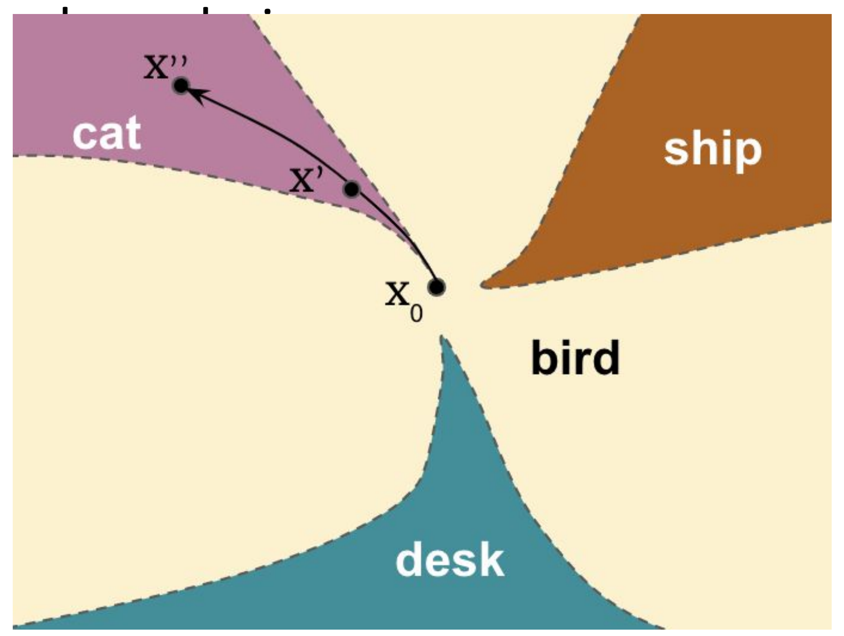
Attackers can easily bypass defenses by optimizing them in black-box and white-box settings.

- **Black-box:**
  Attackers can access model and defense structure, but not any parameter, one can use decision-based methods or natural gradients to optimize the defense criterion.
- **White-box:**
  Attackers have full knowledge of model and defense, one can use gradient-guided methods for differentiable functions or approximate non-differentiable functions with identity (BPDA) to optimize defense criterion.

Methodology

We propose a set of seemingly contradictory criteria to detect adversarial examples.

Illustration of class



- **Robustness to random noise (C1):**
  Prediction of a real image $\mathbf{x}_0$ is robust to random noise, i.e., low density of adversarial perturbations
  ----> Input **x'** can be detected effectively.
  ----> However, input **x''** can bypass this detection.
- **Existence of nearby adversarial examples (C2):**
  Gradient-guided attacks can easily find an adversarial example of $\mathbf{x}_0$. However, the optimization against **C1** makes it hard to find an "adversarial example" of the adversarial example **x''**.
  ----> Adversarial example **x''** can be detected effectively.

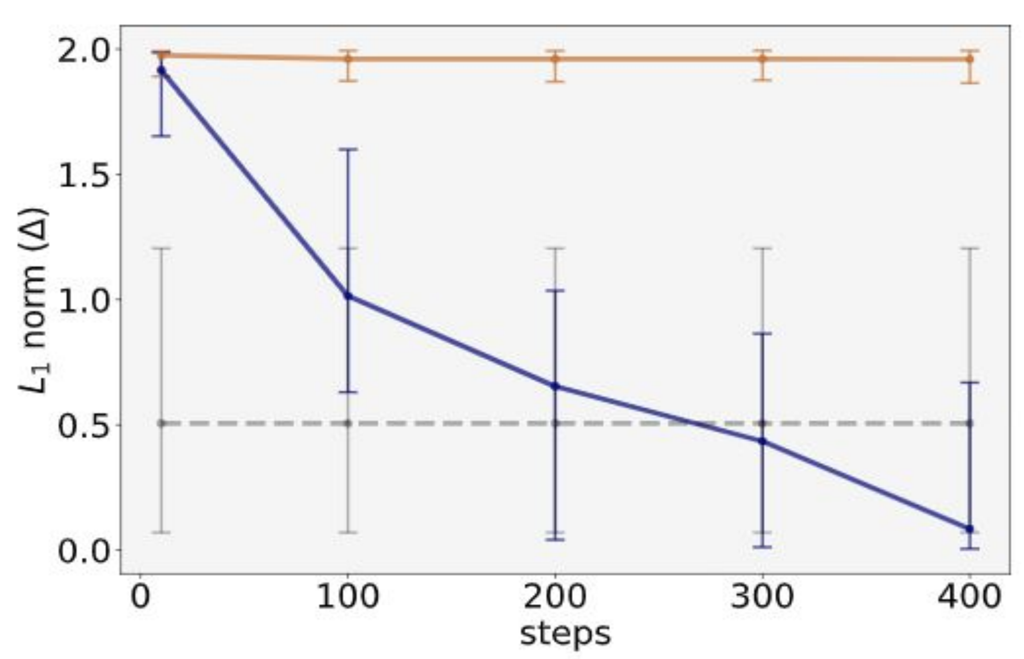
----> **C1+C2** can detect both adversaries **x', x''**.

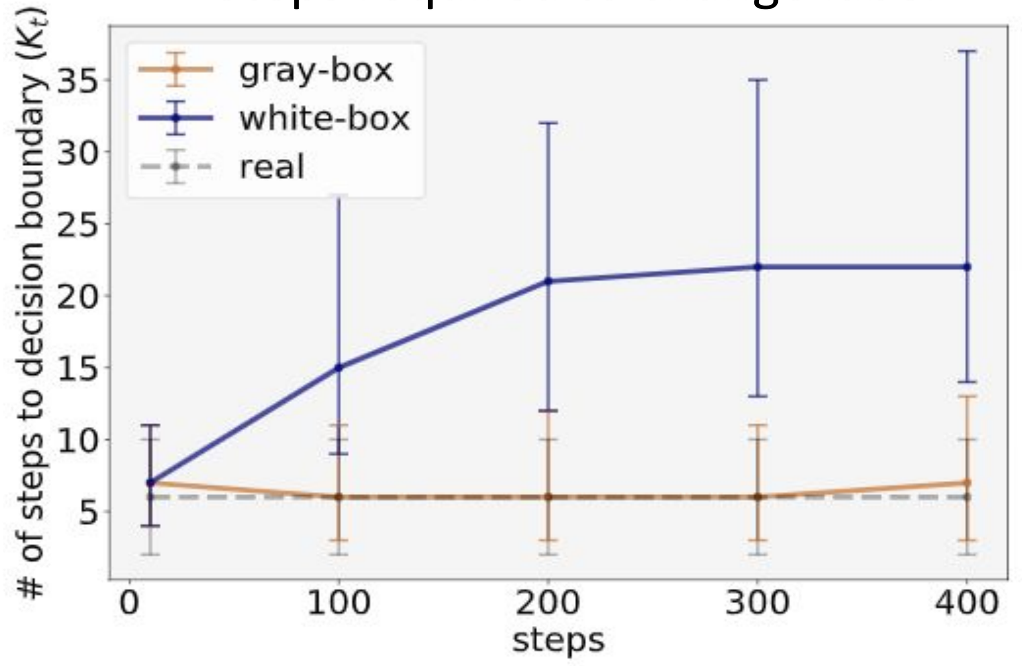
- **Contradictory optimization for attackers:**
  - Optimizing **C1** pulls the adversarial example away from the boundary (towards **x''**).
  - Optimizing **C2** pulls the adversarial example close to the boundary (towards **x'**).

  Optimizing **C1+C2** during attack leads to competing objectives!

Detection Strategy

- **C1:** Robustness to Gaussian noise ($||\cdot||_1$ between predictions of $\mathbf{x}_0$ w/ & w/o noises).
- **C2t:** Susceptibility to targeted iterative adversarial attack (# of iterations until success).
- **C2u:** Susceptibility to untargeted iterative adversarial attack.

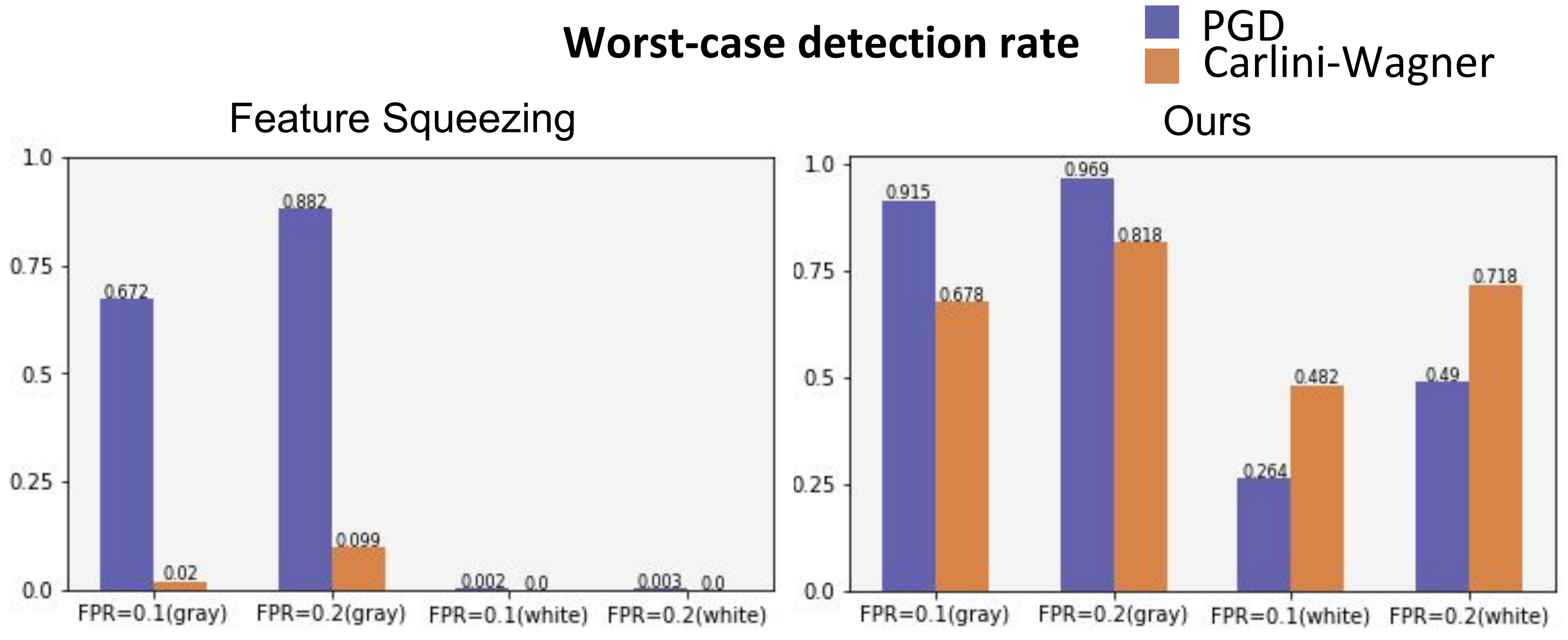




Experiment on ImageNet

- White and gray box:
  - Attack: 50 steps PGD and CW with different learning rates.
  - Model: pretrained ResNet-101 (Top-1 error: 22.63).
  - Adversarial radius: $||\cdot||_\infty$ = 0.1.

  Attacker optimizes loss function designed to bypass C1 and C2(t/u)



- Ablation study:
  Variations of white-box attack can still be detected.

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

- Black box:
  - Attack: boundary attack against different detectors
  - Model: pretrained ResNet-101.
  - Adversarial radius: MSE 0.01 (equivalent to $||\cdot||_\infty$ = 0.1).