Defenses against Adversarial Examples

Keller Jordan[1], Rene Gutierrez[2], Brett Gohre[3]

[1]Department of Computer Science
UCSC

[2]Department of Applied Mathematics & Statistics
UCSC

[3]Department of Physical & Biological Sciences
UCSC

CMPS290, Winter 2018

Optimizer Robustness

We studied:

- Vanilla gradient descent
- EG plus/minus

Procedure

- Train same model with GD and EG$\pm$ on MNIST
- Model is fully-connected 784-100-10
- Run non-targeted adversarial attack until fooled on subset
- Attacks were gradient ascent (GA) and fast gradient sign method (FGS)
- Average added noise for each class
- Compare results between models

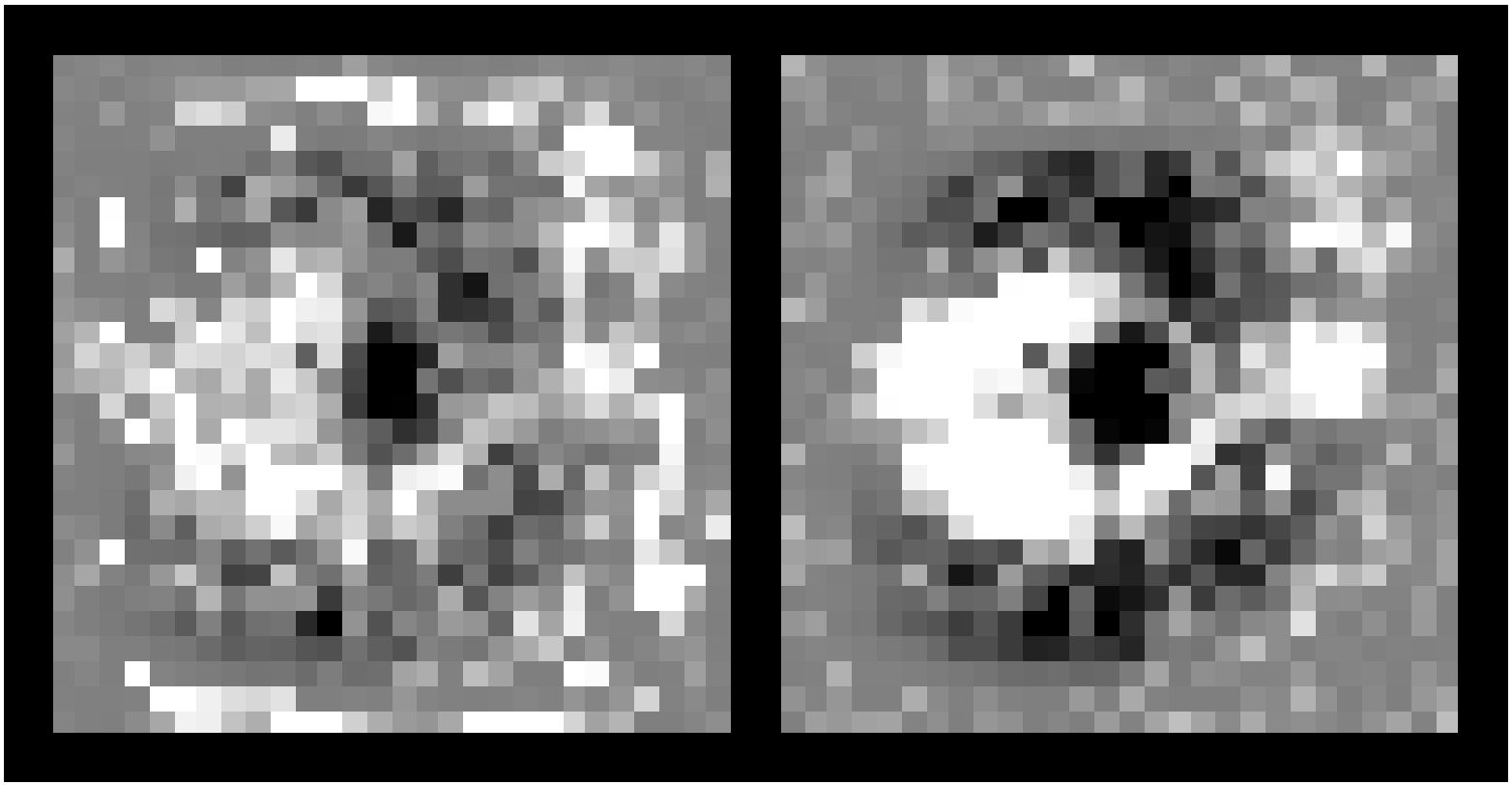
Attack Difficulty

Number of iterations to fool network

Method / Optimizer	SGD	EG
Gradient Ascent	60.9($\pm$32.3)	85.1($\pm$40.5)
Fast Gradient Sign	52.0($\pm$26.1)	91.0($\pm$43.5)

A network is "fooled" when its prediction changes (untargeted attack)

Average Perturbation



SGD (left), EG± (right)

Transferability Results

Transferability of attacks between optimizers

Method / Src→Dst	SGD→EG	EG→SGD
Gradient Ascent	67.4%	99.0%
Fast Gradient Sign	88.2%	99.8%

Iterations held constant at 200

Notes

- FGS looks better to humans, worse for MSE
- GA better at revealing structure of model since cares about strength of change
- Next step: should try L1 norm weights to see difference

Reconstruction as a Defense

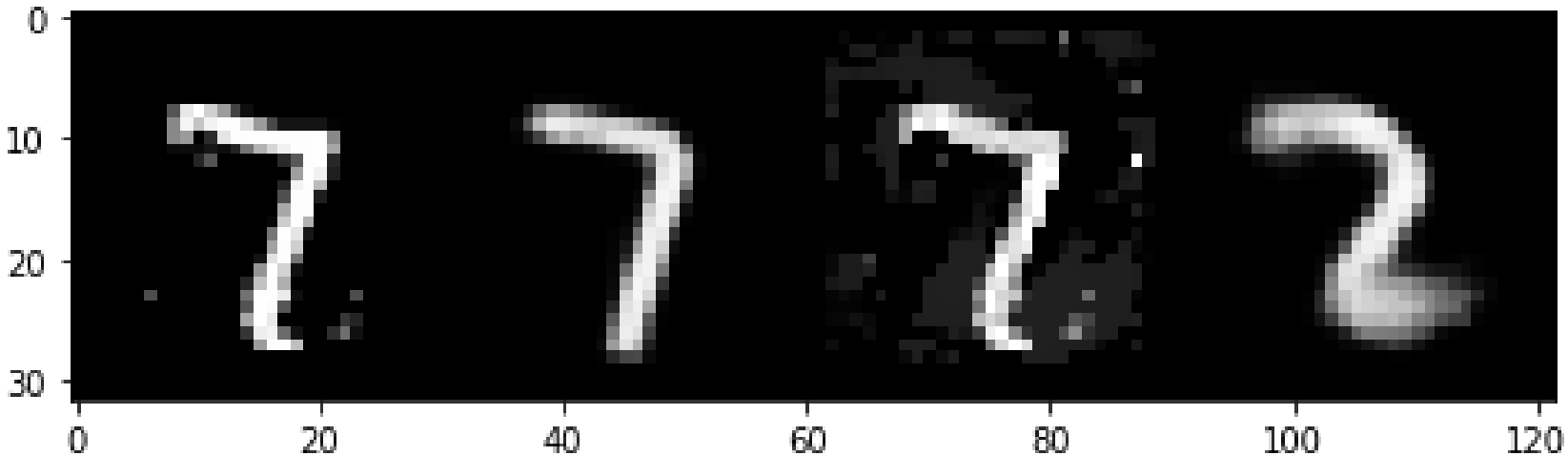
How to get recon err?

Need some way to reconstruct image
Could use encoder/decoder
I went with capsule network

Capsule Network Refresher

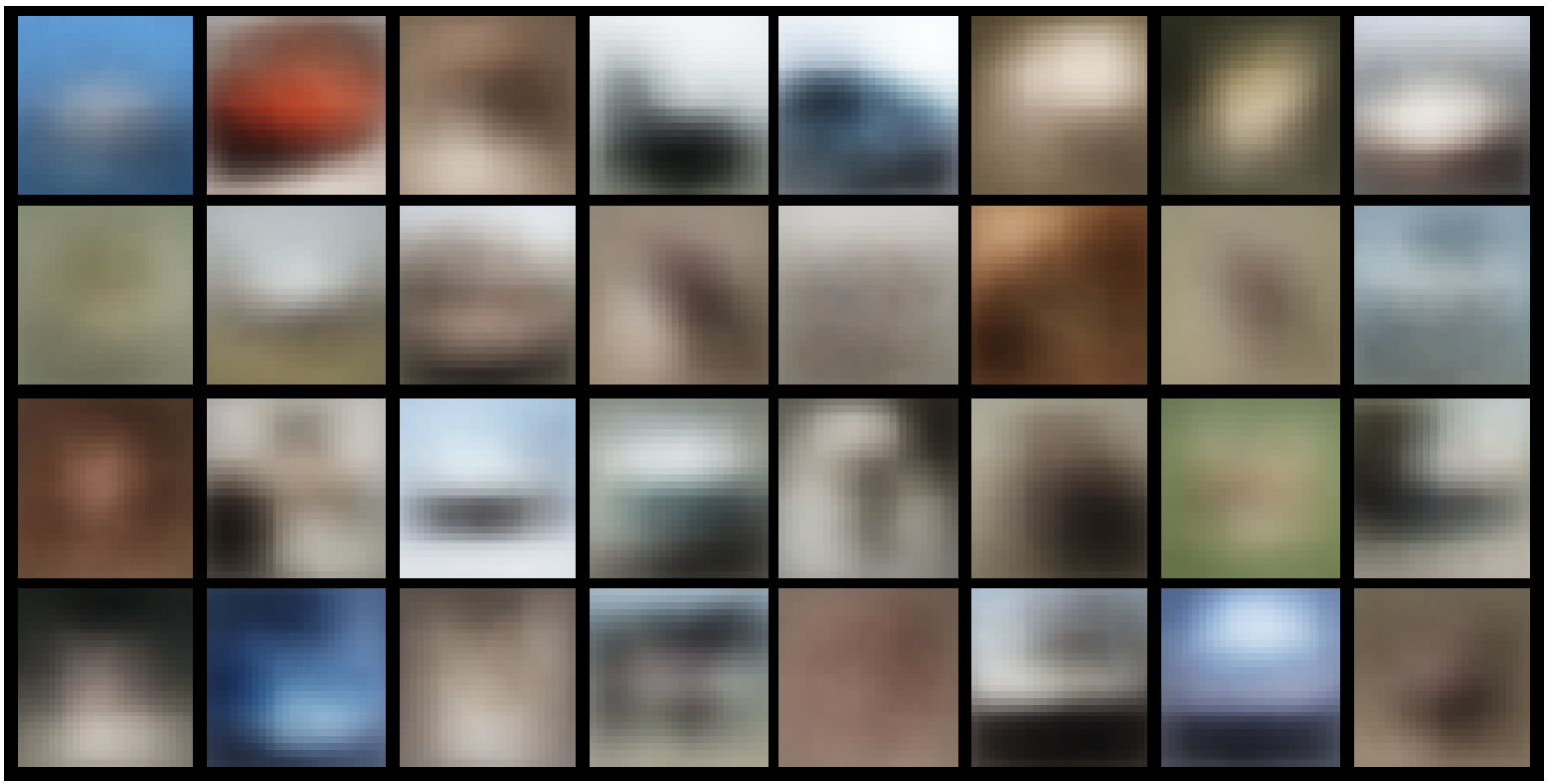
details about reconstruction network and capsnet

Results

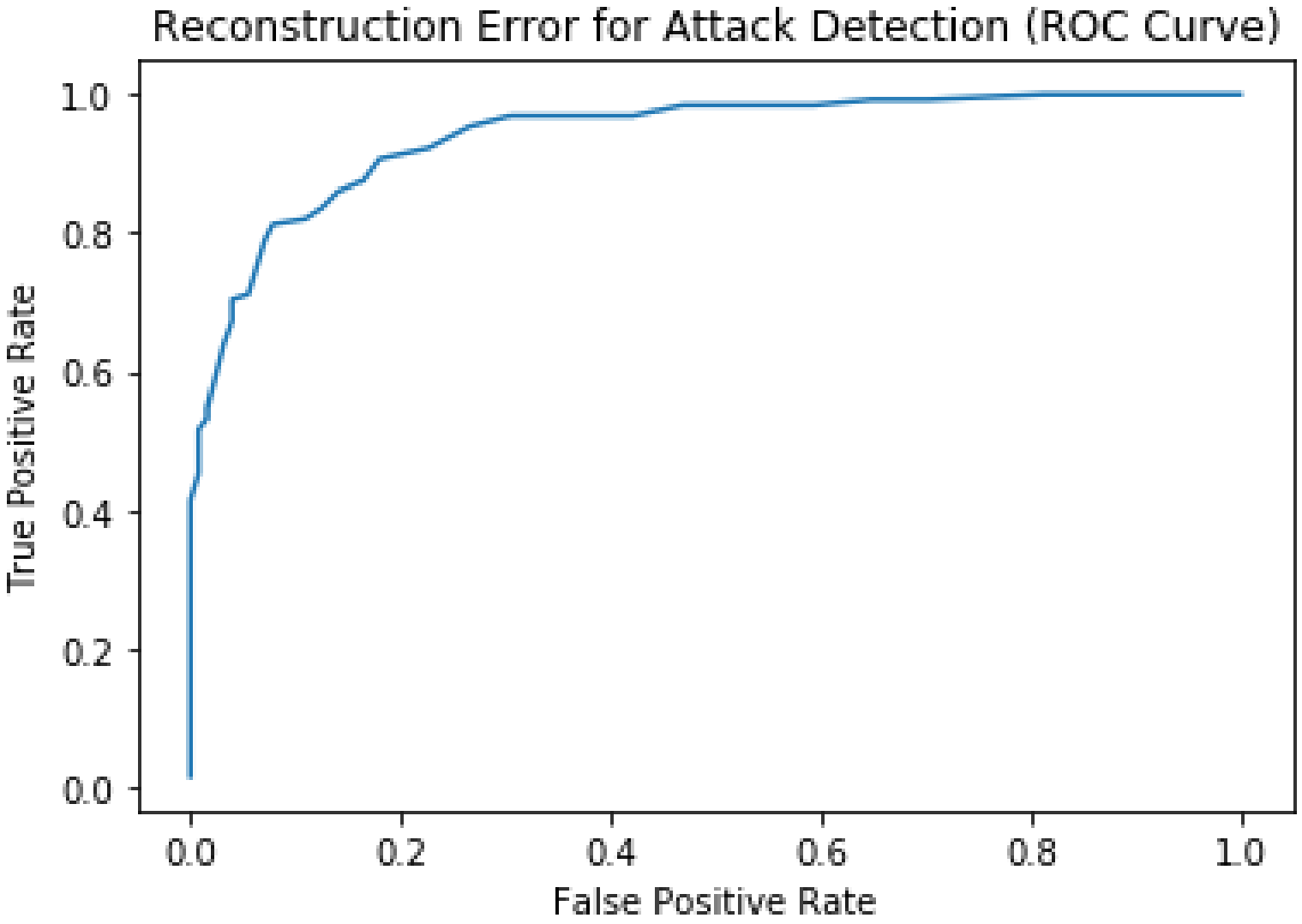


CIFAR10

Capsules are not there yet.



ROC Curve


Reconstruction Error for Attack Detection (ROC Curve)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0