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

Optimizers Studied

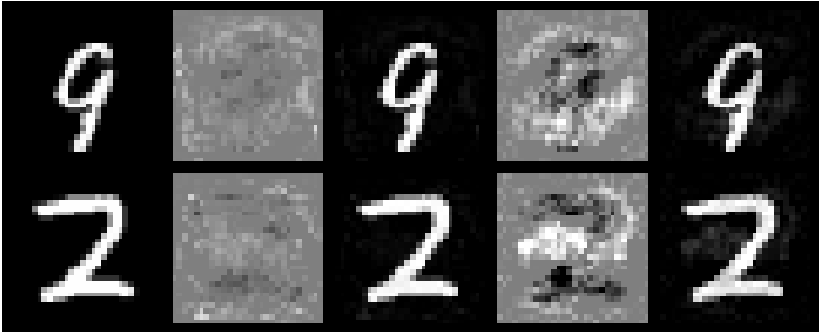
Method	SGD	EG
Rule	$w_{t+1} \doteq w_t - \eta \nabla L(w_t)$	$w_{t+1} \doteq w_t \, e^{-\eta \nabla L(w_t)}$

- Used extension of EG to +/- weights case for training

Procedure

- Train FC 784-100-10 using GD and EG± on MNIST
- Run untargeted adversarial attack methods
  - Gradient Ascent (GA)
  - Fast Gradient Sign (FGS)
- Compare resulting models and adversarial examples
  - Number of iters to fool
  - Transferability of strong attacks
  - Average perturbation

Examples



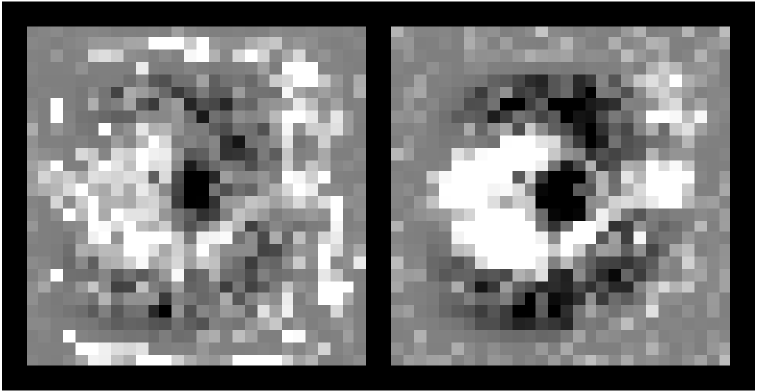
Attack Difficulty

Number of iterations to fool network

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

- A network is defined as “fooled” when its prediction changes

Average Perturbation



SGD (left), EG± (right)

Transferability Results

Probability of success on other optimizer

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 (should be comparably strong)

Results

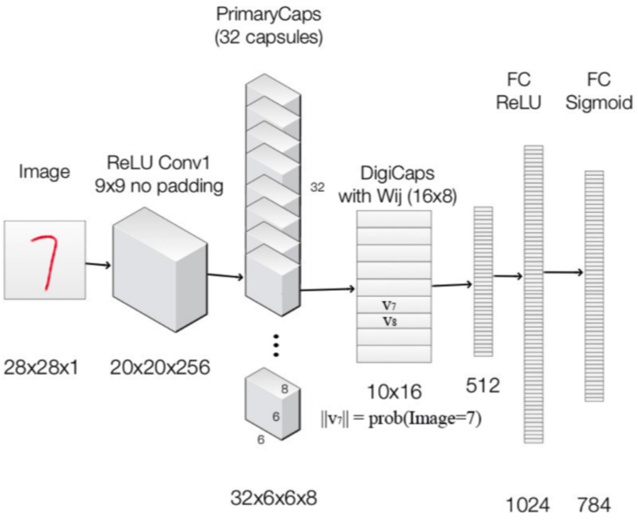
- Requires 1.5× stronger attacks to fool EG-trained model
- EG shows some robustness to attacks transferred from SGD
- SGD is not robust to attacks transferred to EG
- Attacks against EG make more sense w.r.t. expected structure of digit space

Defending using Reconstruction Error

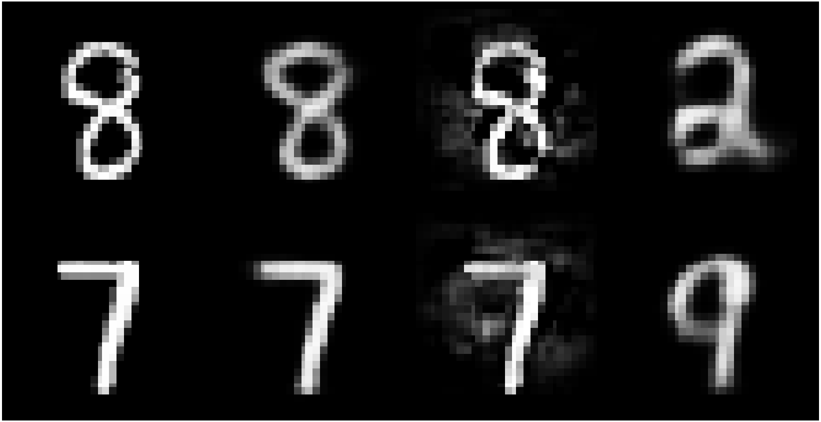
Basic idea

- Use an architecture that reconstructs input images (CapsNet)
- Model will reconstruct some element of decoder-space for fooled class
- Adversarial images are unlikely to be in this space
- Expect high reconstruction error (MSE)

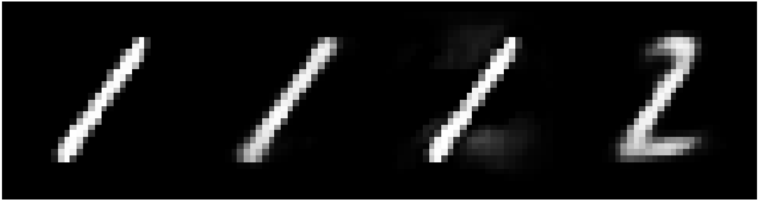
Capsule Network Refresher

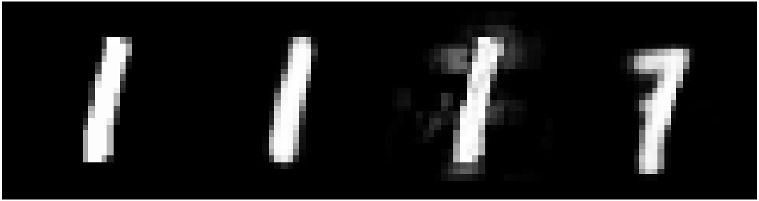


Reconstructions

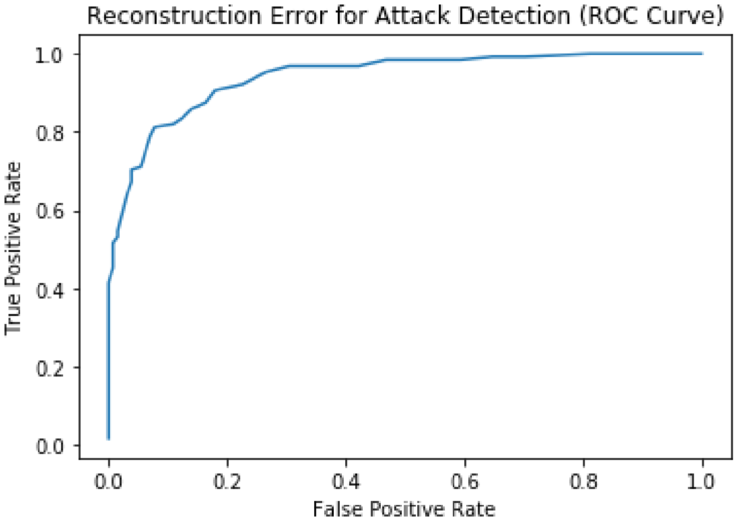


Reconstructions





Results


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
0.0
0.2
0.4
0.6
0.8
1.0

Results

- Method successfully detects ∼70% of of attacks with 5% false-positive rate
- Could be improved by better loss function
- Unknown vulnerability to white-box attacks
- Expect good black-box performance due to variability of decoders and loss functions