Deconstructive Defense against Adversaries

Chin-Wei Huang*, Nan Rosemary Ke*, Chris Pal

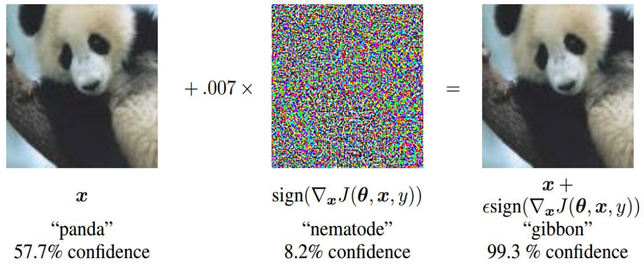
Université de Montréal
POLYTECHNIQUE MONTRÉAL
MILA

Introduction

- Neural Network classifiers are vulnerable to adversarial attacks and model inversion attacks.
- Adversarial examples [1]: adding carefully chosen noise imperceivable by human eye to fool a classifier.
- Model inversion attacks [2]: reconstructing data samples from trained model.
- We propose to train a classifier with an auxiliary autoencoder. Decoder takes real gradients, and encoder takes negative gradients.

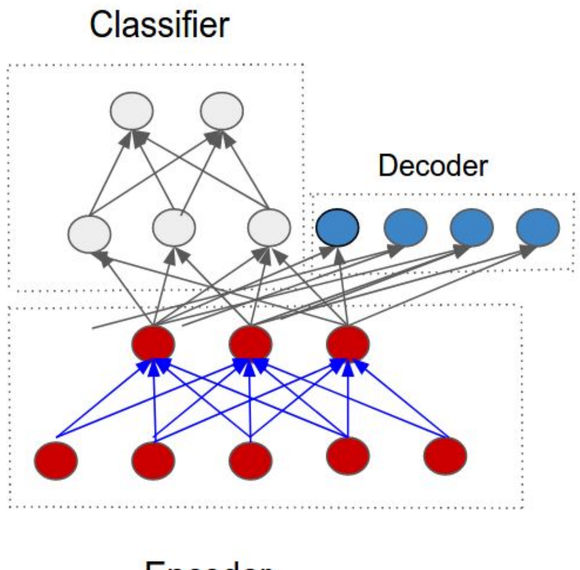
Adversaries

- Adversarial Attacks:
  - Adding small perturbation that humans cannot recognize but enough to hurt classification performance



- Model Inversion Attacks
  - Recover original image given trained model

Destruction defense model



Reversal Gradient

- 2 training objectives (shared encoder):
  Classification & Reconstruction
- Decoder takes real gradient
- Encoder gets 2 gradients
  - Gradient from reconstruction loss
    - (positive or negative)
  - Gradient from classification loss
- Shuffle 50% of reconstruction targets for adversarial attack experiments
  - If real digit, take real gradient
  - If fake digit, take negative gradient

Table 1: Adversarial attacks on MNIST dataset

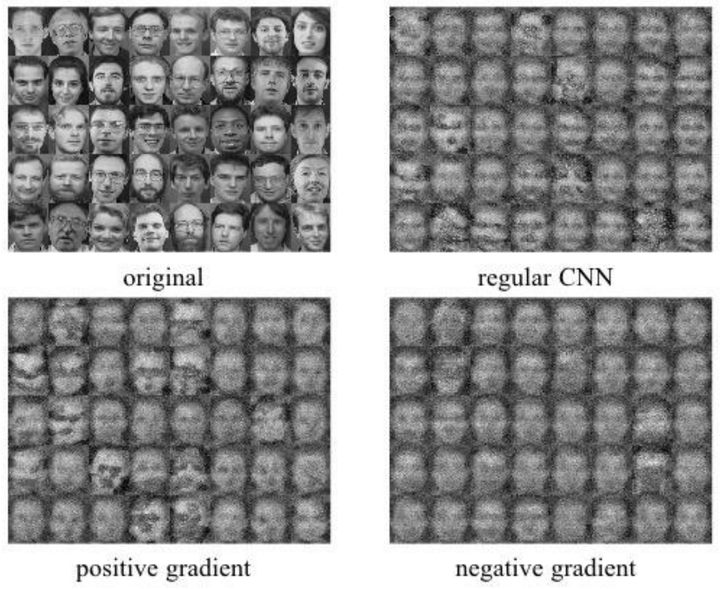
Experiment	Perturbation, %	Accuracy,
Baseline	0%	96.8%
Baseline	5%	95.6%
Baseline	10%	95.6%
Baseline	25%	95.6%
Positive gradient	0%	96.3%
Positive gradient	5%	95.0%
Positive gradient	10%	95.0%
Positive gradient	25%	95.0%
Our model	0%	95.2%
Our model	5%	94.9%
Our model	10%	94.9%
Our model	25%	94.9%

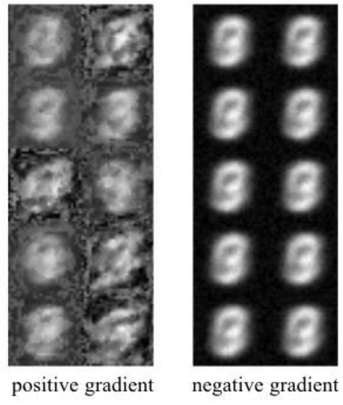
Experiments

- For Adversarial Attacks, we evaluated our model on MNIST dataset, using the Fast Gradient Sign method.
- For Model Inversion Attacks, we evaluated on MNIST and AT&T faces.
  - We warm-start the attack with an **average image**.
  - For different models we fix the number of iterations for the attacker.

Results

- Samples for Model Inversion Attacks





References

[1] Goodfellow I.-J., Shlens J., and Szegedy C. Explaining and Harnessing Adversarial Examples.

[2] Fredrikson, M., Jha, S., and Ristenpart, T. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures.