Using Capsule Networks to Disarm Adversarial Attacks

Jordan Alexander, Sahaj Garg, Tanay Kothari

Stanford University

Introduction & Motivation

Adversarial attacks have been shown to construct examples that drastically reduce the performance of classification models. We attempt to construct a model that is robust to adversarial examples by reconstructing an image that detects and removes adversarial perturbations. This is accomplished using a capsule network, a novel computer vision architecture recently published by *Sabour et. al.*

Approach

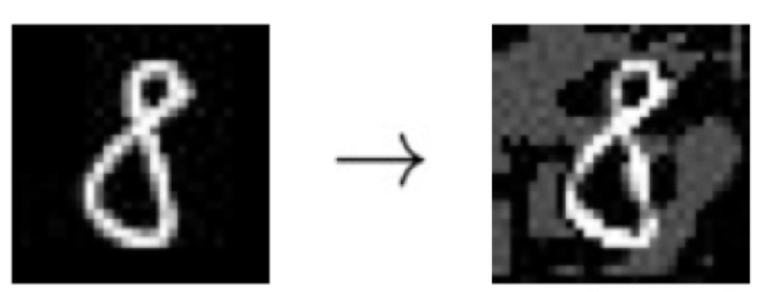
Dataset: MNIST (55,000 Train, 5,000 Dev, 10,000 Test)

Baseline: We trained a CNN with and without adversarial training and tested it with adversarial attacks.

Our model: We use a Capsule Network with FGSM to create adversarial examples and set its reconstruction target to the original image. It has an accuracy of 95% on adversarial examples after adversarial training. It is also able to filter out adversarially generated noise when reconstructing images.

Adversarial Attacks

Fast Gradient Sign Method



x "8" $\rightarrow$ $x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$ "0"

FGSM uses the parameters of a model, its input, and its target to find a small perturbation that maximizes the error of the model.

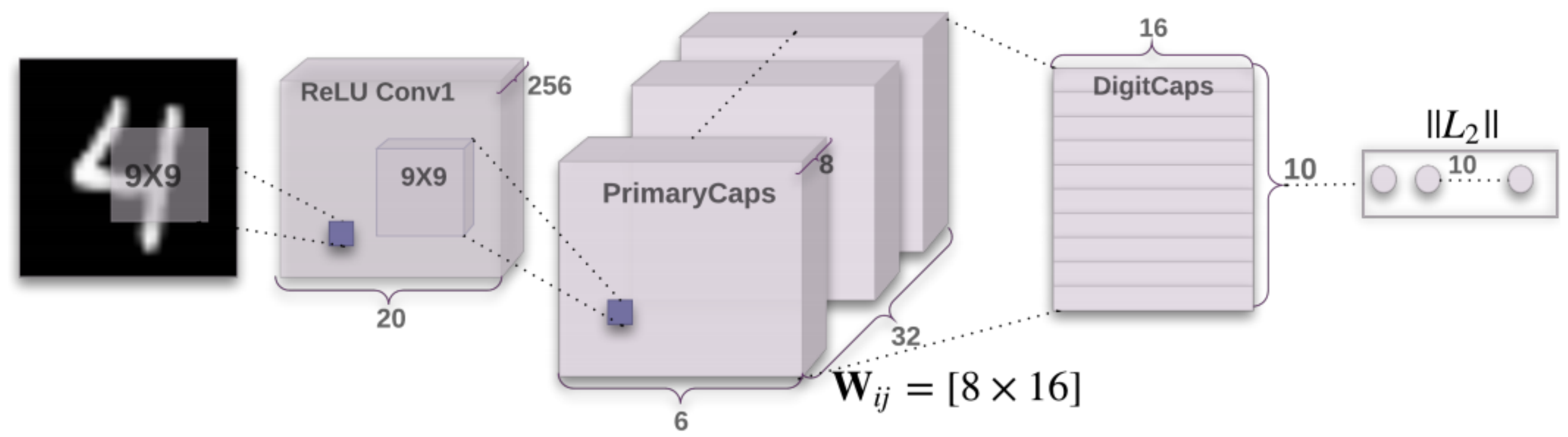
Although the perturbation appears like noise to humans, it is specially targeted to minimize the accuracy of the model.

Adversarial Training: We defend against attacks by training over a mini-batch of adversarial examples at the end of every mini-batch.

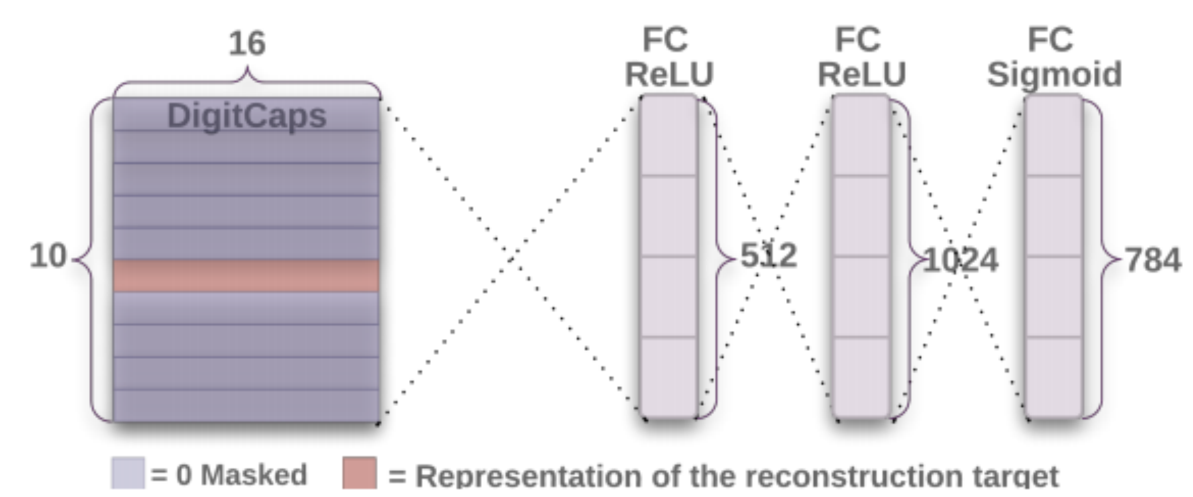
CNN Test Accuracy	Clean Trained	Adversarial Trained
Clean Test Set	99.08%	99.14%
Adversarial Test Set	79.92%	95.44%

Model and Architecture

Capsule Network: 1 conv layer, 1 capsule layer, 1 digit capsule layer.

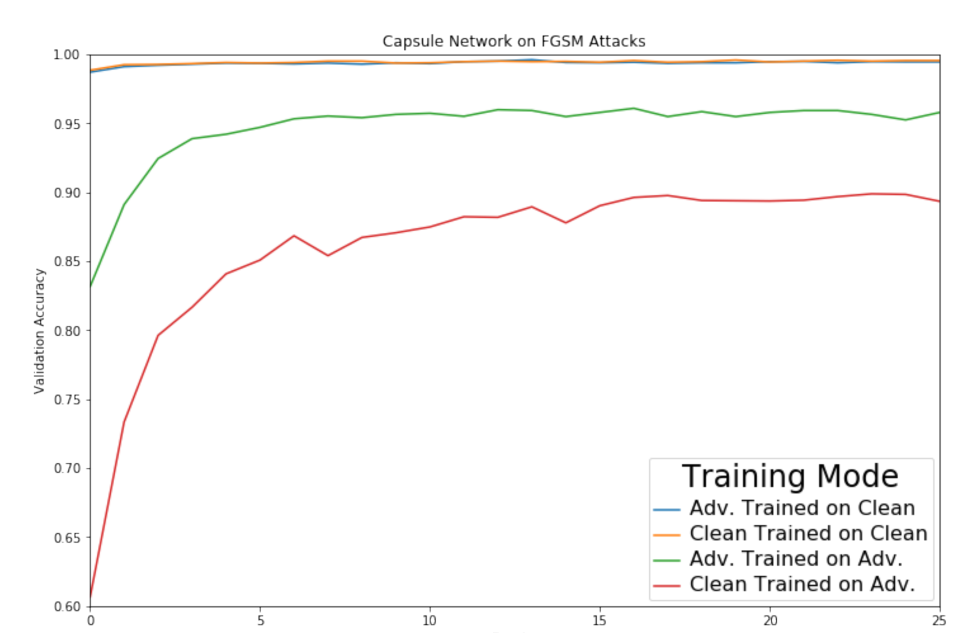


- Consists of groups of neurons called **capsules** whose outputs represent different characteristics of the same feature
- Outputs are routed from one layer to the next using dynamic routing instead of max pooling. This capture the relationship between **part and whole** in images.
- The outputs of the final capsules are decoded and reconstructed into an image



Results

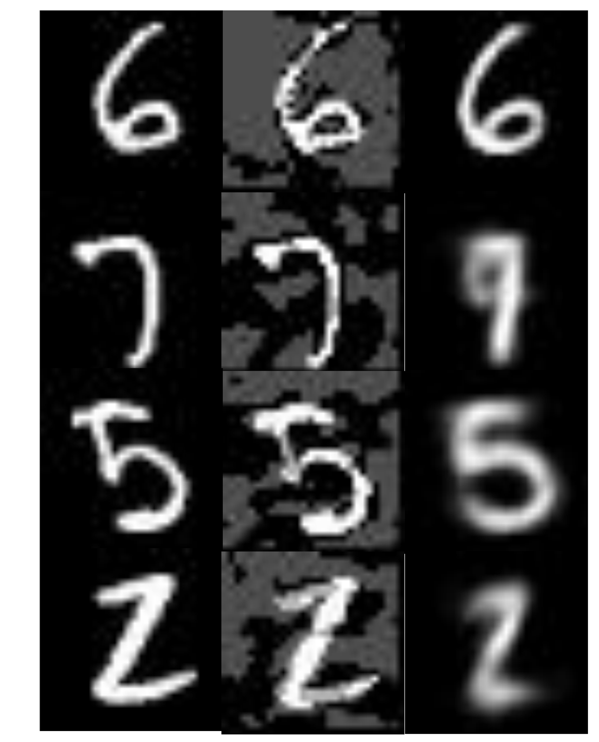
Clean-trained capsule networks have half the error rate of clean-trained CNNs. Both models perform similarly when trained on adversarial examples. Below are learning curves for the CNN (left) and Capsule Network (right) over epochs.



Analysis and Discussion

- To the right is a diagram of sample test images, corresponding FGSM attacks, and reconstructions by the capsule network.
- The reconstructions are quite similar to the original samples.
- When reconstructing a misclassified sample, some hints of the correct class are visible (second row)
- The capsule network learns to ignore adversarial perturbations when making predictions on test examples.

(Original, FGSM, Recon.)



Caps Test Accuracy	Clean Trained	Adversarial Trained
Clean Test Set	99.31%	99.27%
Adversarial Test Set	76.83%	95.18%

Future Work & Discussion

- Thoroughly test the transferability of adversarial attacks between models.
- Fully test and report details on the additional thermometer-encoded capsule network model
- Use Church-Window plots to examine nonlinearities in the decision boundaries for the predictions of the Capsule network and how they change after adversarial training.
- Implement a capsule network with multiple levels of dynamic routing and evaluate its performance.
- Testing our model against other datasets like ImageNet, SVHN, and CIFAR10.

References

- I. Goodfellow, **Explaining and Harnessing Adversarial Examples**. Mar 2015.
- S. Sabour, G. E. Hinton, **Dynamic Routing Between Capsules.** Nov 2017.
- Anonymous, **Thermometer Encoding: One Hot Way To Resist Adversarial Examples**. Nov 2017. Blind submission for ICLR 2018.