Adversarial Example Attack

Andrew Jacobson

March 16, 2025

1 Computation Time

The victim model took 30.714485 seconds to train at 10 epochs. The attack generation took 0.0750 seconds.

2 Victim Model

The victim model is a Convolutional Neural Network and has the following architecture:

- Input: 28x28 images
- Convolutional Layer 1: 32 filters, 3x3 Kernel, and ReLU activation
- MaxPooling Layer: 2x2 pool size
- Convolutional Layer 2: 64 filters, 3x3 Kernel, and ReLU activation
- Flatten Layer: Flattens the layers into a 1D vector
- Fully Connected Layer: 128 neurons, ReLU activation
- Output Layer: 10 neurons, softmax activation

Hyperparameters:

- Optimizer: Adam
- Loss Function: Categorical cross entropy
- Batch Size: 32
- Epochs: 10

3 Victim Model Performance

```
Epoch 1/10
1875/1875 [==============================] - 3s 2ms/step - loss: 0.1276 - accuracy: 0.9616 - val_loss: 0.0445 - val_accuracy: 0.9859
Epoch 2/10
1875/1875 [==============================] - 3s 2ms/step - loss: 0.0413 - accuracy: 0.9871 - val_loss: 0.0331 - val_accuracy: 0.9883
Epoch 3/10
1875/1875 [==============================] - 3s 2ms/step - loss: 0.0288 - accuracy: 0.9911 - val_loss: 0.0345 - val_accuracy: 0.9888
Epoch 4/10
1875/1875 [==============================] - 3s 2ms/step - loss: 0.0212 - accuracy: 0.9932 - val_loss: 0.0375 - val_accuracy: 0.9889
Epoch 5/10
1875/1875 [==============================] - 3s 2ms/step - loss: 0.0155 - accuracy: 0.9947 - val_loss: 0.0277 - val_accuracy: 0.9911
Epoch 6/10
1875/1875 [==============================] - 3s 1ms/step - loss: 0.0125 - accuracy: 0.9959 - val_loss: 0.0286 - val_accuracy: 0.9907
Epoch 7/10
1875/1875 [==============================] - 3s 2ms/step - loss: 0.0102 - accuracy: 0.9966 - val_loss: 0.0298 - val_accuracy: 0.9922
Epoch 8/10
1875/1875 [==============================] - 3s 1ms/step - loss: 0.0073 - accuracy: 0.9977 - val_loss: 0.0409 - val_accuracy: 0.9900
Epoch 9/10
1875/1875 [==============================] - 3s 1ms/step - loss: 0.0076 - accuracy: 0.9975 - val_loss: 0.0370 - val_accuracy: 0.9909
Epoch 10/10
1875/1875 [==============================] - 3s 2ms/step - loss: 0.0060 - accuracy: 0.9980 - val_loss: 0.0343 - val_accuracy: 0.9912
313/313 [==============================] - 0s 670us/step - loss: 0.0343 - accuracy: 0.9912
Model Accuracy: 99.12%
The training time is: 30.714485 seconds
The current time is: 2025-03-21 15:02:03.502772
```

Figure 1: Training Process Complete

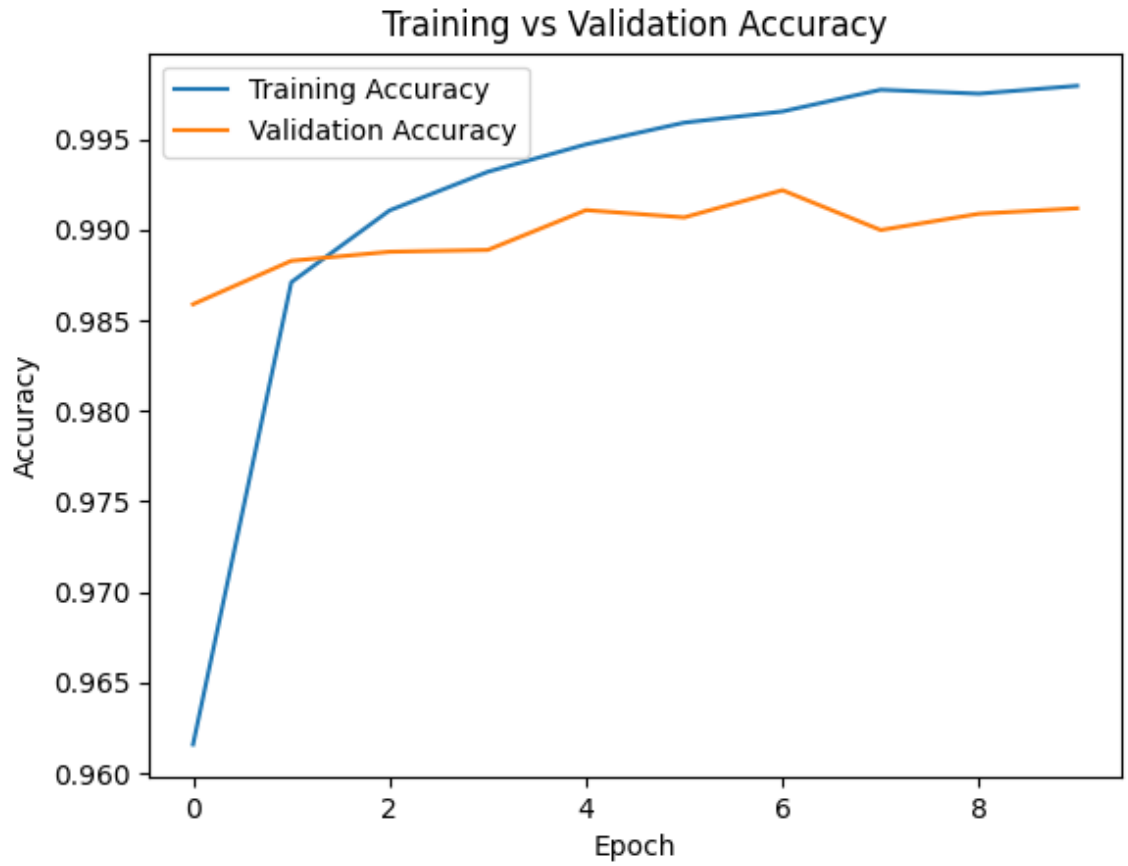

Figure 2: Testing Accuracy

4 Adversarial Example Attack

I used the Fast Gradient Sign Method. This attack method adds a small perturbation to an input image, attempting to mislead the victim model. The formula is:

$$adv_x = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

where:

- adv_x: adversarial example.
- x: original image.
- ϵ: attack strength for the perturbations.
- ∇_x: gradient of the loss function.
- $sign()$: applies the sign function for the direction of the perturbation.
- θ: The CNN model and its parameters.
- y: original label of the image.
- J: loss function.

5 Attack Results / Performance

The FGSM attack had a success rate of 70%.

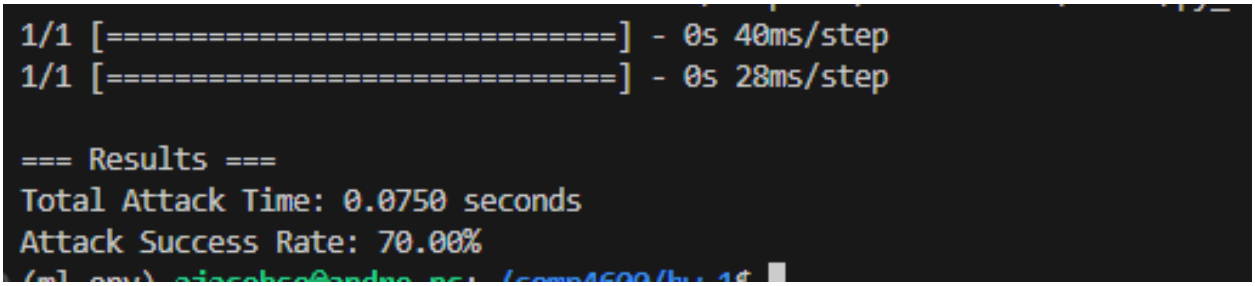

Figure 3: Attack Process Complete

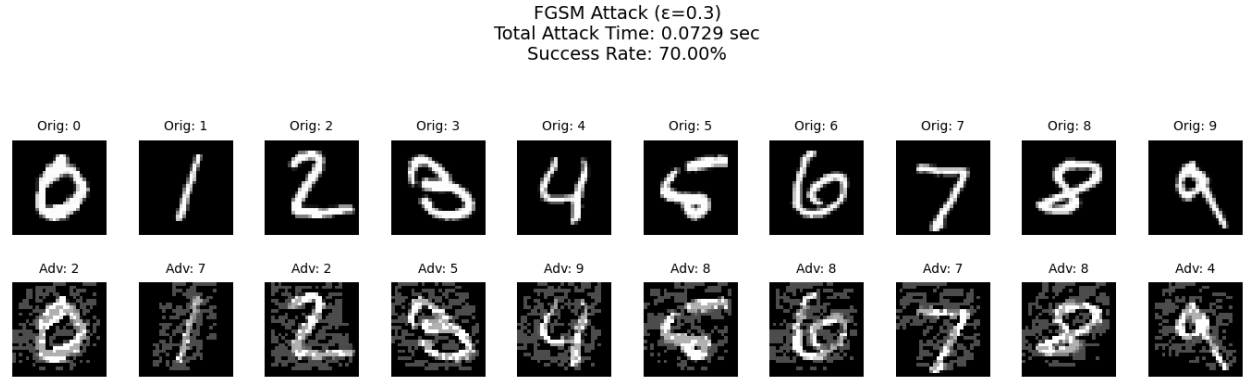

Figure 4: Attack Results