

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.031	4.38	0.031
Accuracy	98%	12%	98%

Example Misclassifications:

Number of misclassified images for Clean: 1

Attack: fgsm\_cw\_attack

Dataset: MNIST

Training Epochs: 10

Trained Clean Images: 64

Test Images: 140

Accuracy: 0.98

Precision: 0.99

Recall: 0.99

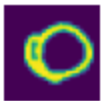
F1-score: 0.99

ROC AUC Score: 1.00

Misclassifications:

0 -> 6: 1

0 -> 6



Number of misclassified images for No Defense Attack: 56

Attack: fgsm\_cw\_attack

Dataset: MNIST

Training Epochs: 10

Adversarial Training Images: 60

Test Images: 140

Accuracy: 0.12

Precision: 0.23

Recall: 0.20

F1-score: 0.22

ROC AUC Score: 1.00

Misclassifications:

0 -> 9: 9

0 -> 7: 7

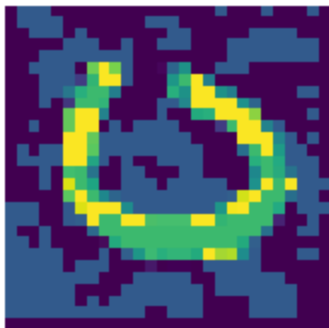
0 -> 8: 3

0 -> 6: 21

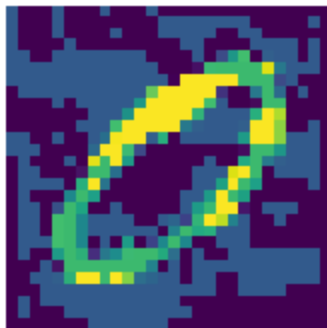
0 -> 5: 14

0 -> 4: 2

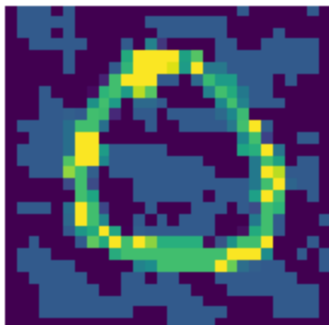
0 -&gt; 9



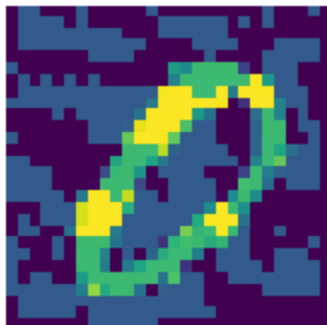
0 -&gt; 7



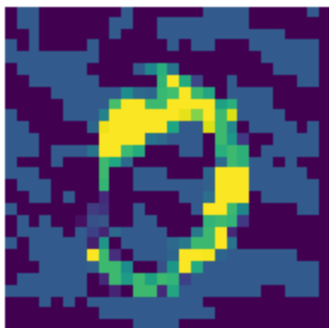
0 -&gt; 9



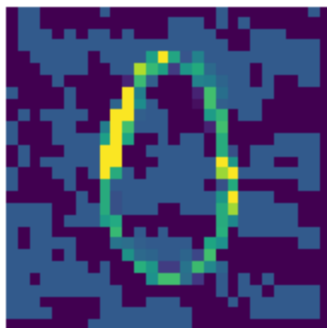
0 -&gt; 8



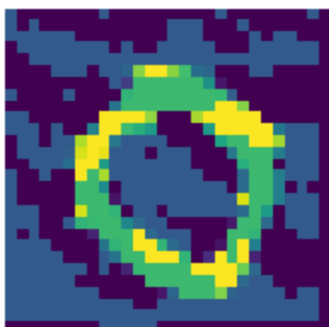
0 -&gt; 7



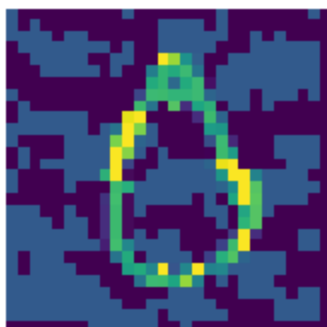
0 -&gt; 6



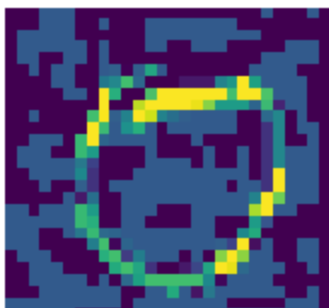
0 -&gt; 6



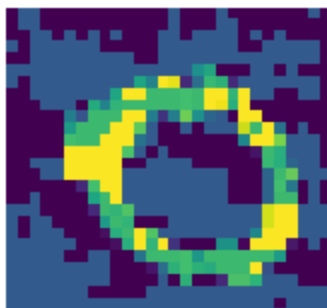
0 -&gt; 5



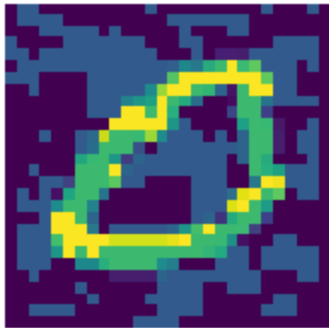
0 -&gt; 5



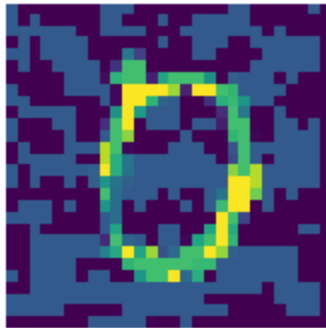
0 -&gt; 5



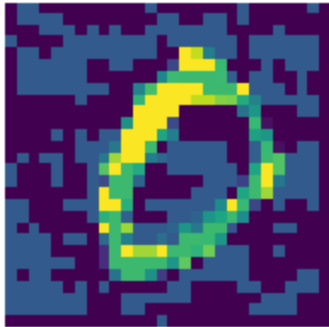
0 -&gt; 5



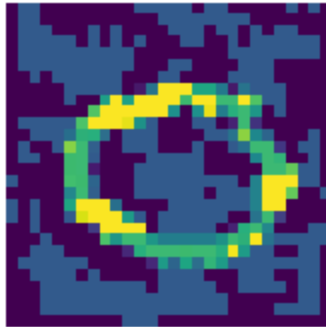
0 -&gt; 9



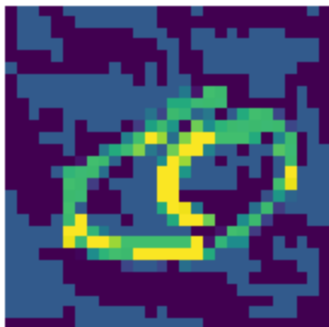
0 -&gt; 6



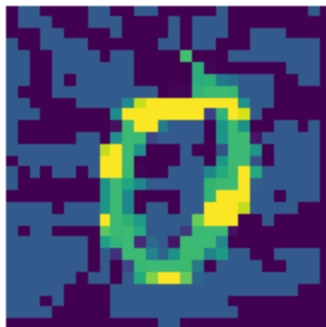
0 -&gt; 9



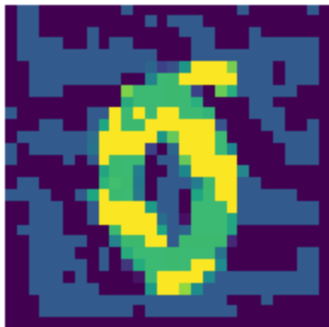
0 -&gt; 5



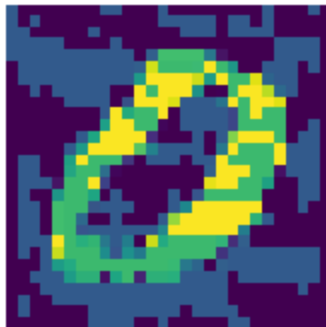
0 -&gt; 7



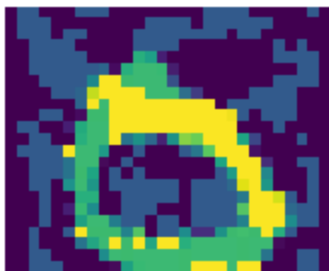
0 -&gt; 5



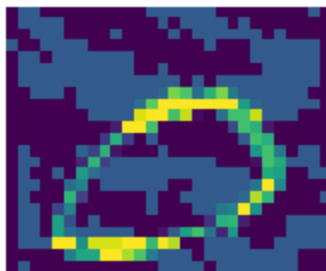
0 -&gt; 7



0 -&gt; 5

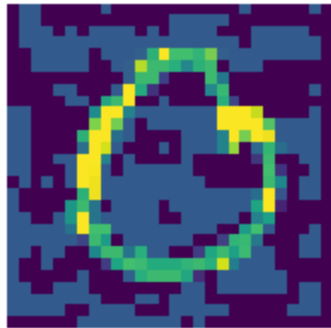


0 -&gt; 5

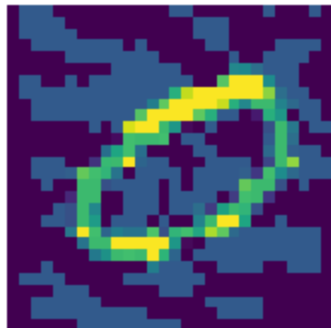




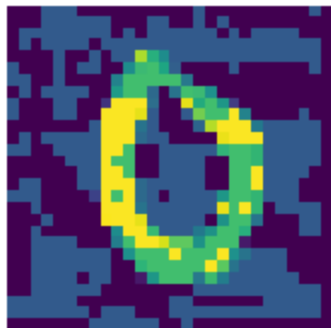
0 -&gt; 6



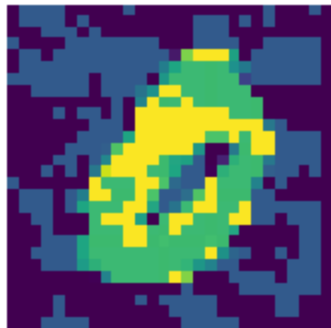
0 -&gt; 5



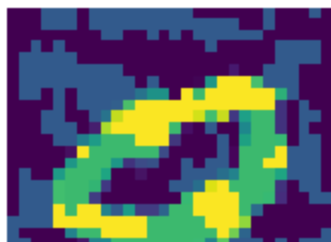
0 -&gt; 6



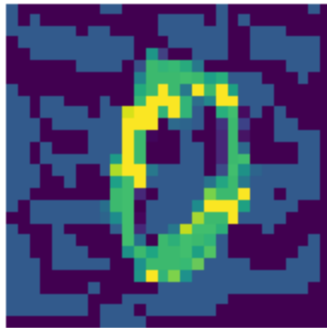
0 -&gt; 6



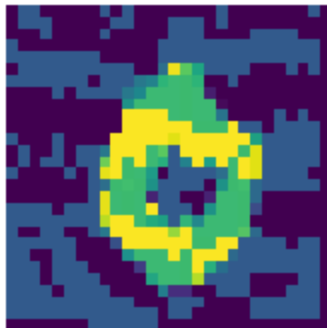
0 -&gt; 9



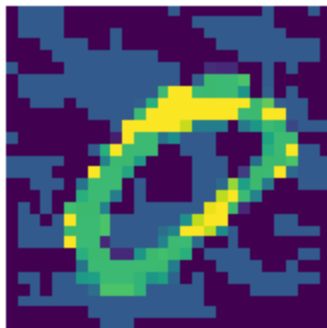
0 -&gt; 7



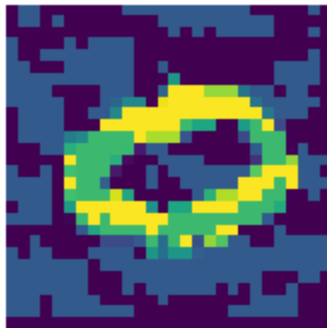
0 -&gt; 6



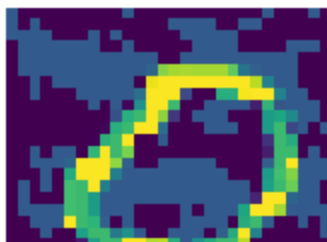
0 -&gt; 7



0 -&gt; 9

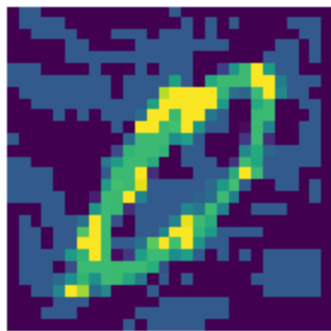


0 -&gt; 6

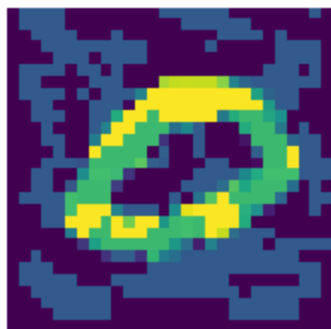




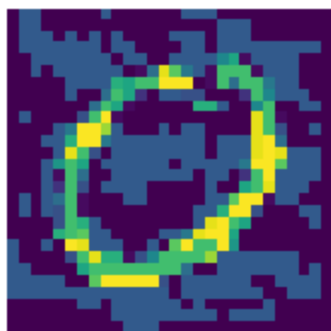
0 -&gt; 8



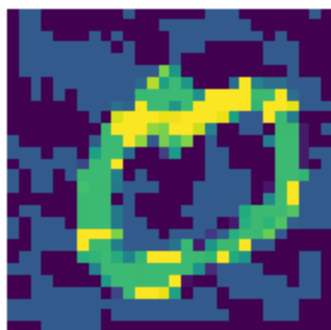
0 -&gt; 9



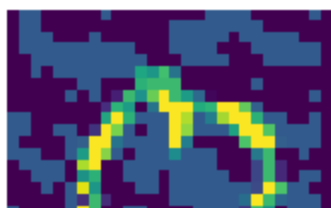
0 -&gt; 9



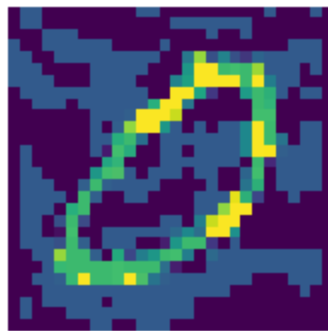
0 -&gt; 5



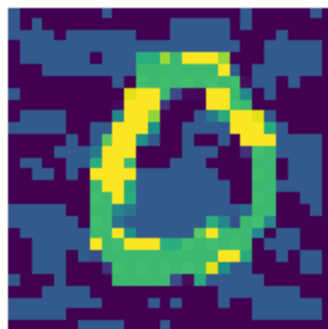
0 -&gt; 7



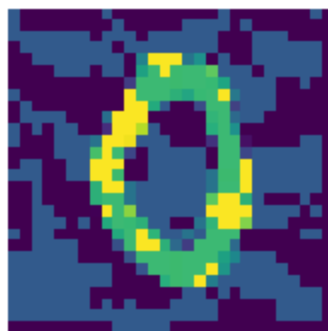
0 -&gt; 5



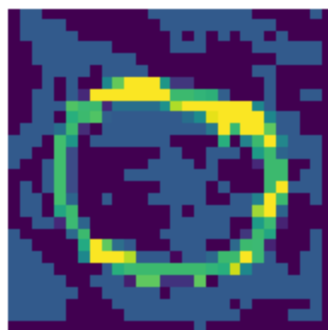
0 -&gt; 6



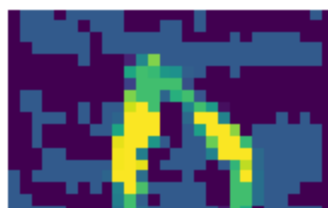
0 -&gt; 6



0 -&gt; 5

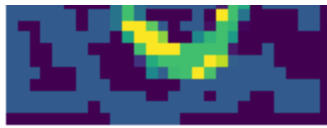


0 -&gt; 6

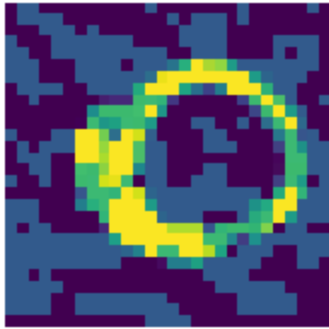




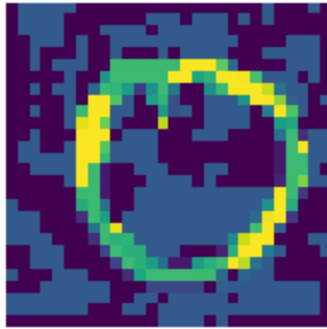
0 -&gt; 6



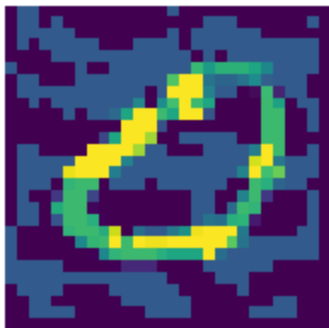
0 -&gt; 6



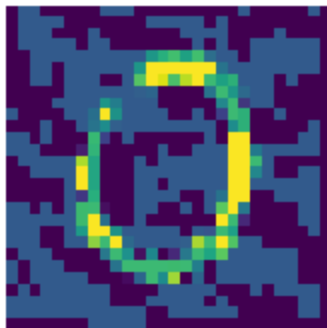
0 -&gt; 4



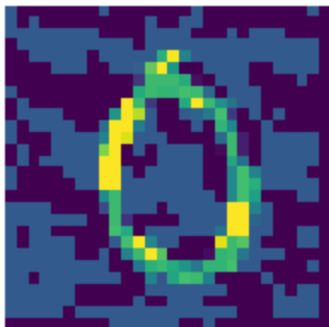
0 -&gt; 9



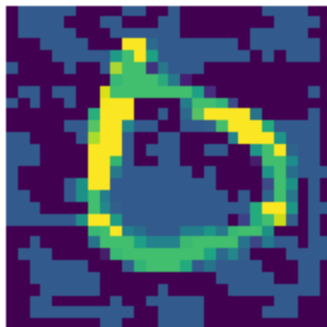
0 -&gt; 6



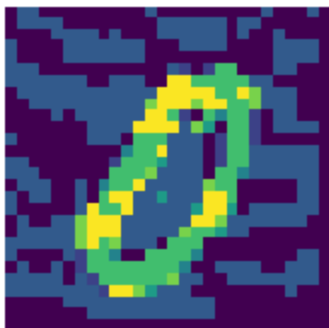
0 -&gt; 6



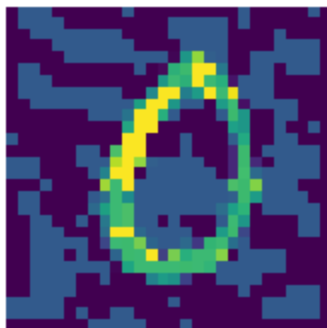
0 -&gt; 8



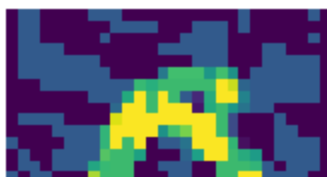
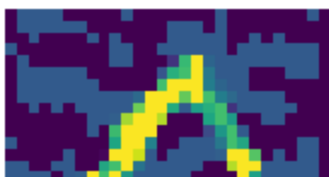
0 -&gt; 6

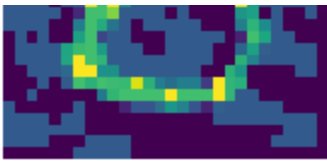


0 -&gt; 6

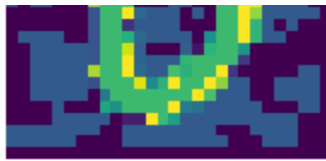


0 -&gt; 5

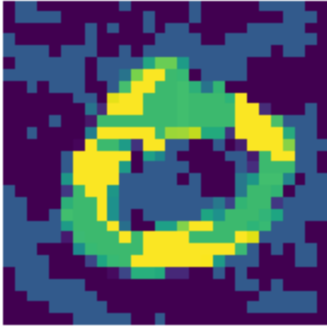




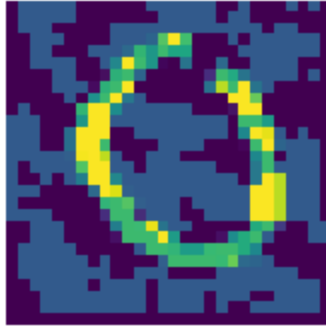
0 -&gt; 6



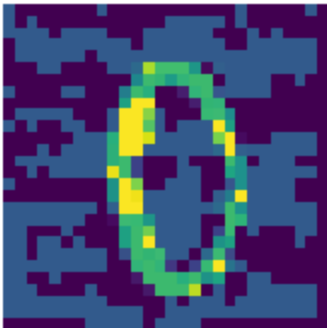
0 -&gt; 6



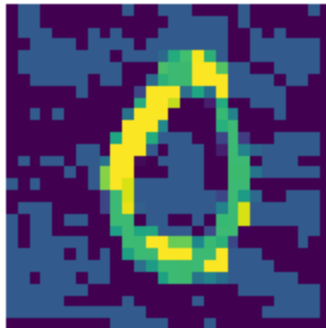
0 -&gt; 6



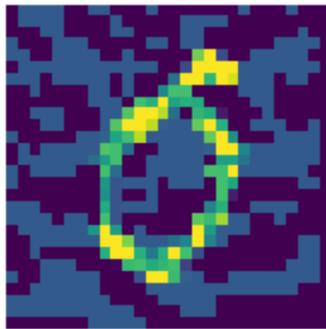
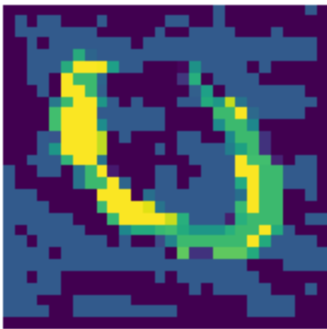
0 -&gt; 6



0 -&gt; 4



0 -&gt; 5



-----  
Number of misclassified images for w/ Defense Attack: 1

Attack: fgsm\_cw\_attack

Dataset: MNIST

Training Epochs: 10

Retrained Clean and Adversarial Images: 124

Test Images: 140

Accuracy: 0.98

Precision: 0.99

Recall: 0.99

F1-score: 0.99

ROC AUC Score: 1.00  
-----

Misclassifications:

0 -> 6: 1

0 -> 6

