

Metric	Clean	No Defense Attack	w/ Defense Attack
Loss	0.105	5.14	0.134
Accuracy	97%	5%	92%

Example Misclassifications:

Number of misclassified images for Clean: 2

Attack: fgsm_bim_attack

Dataset: MNIST

Training Epochs: 10

Trained Clean Images: 64

Test Images: 140

Accuracy: 0.97

Precision: 0.99

Recall: 0.98

F1-score: 0.98

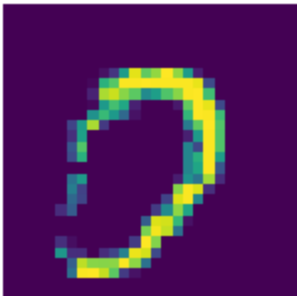
ROC AUC Score: 1.00

Misclassifications:

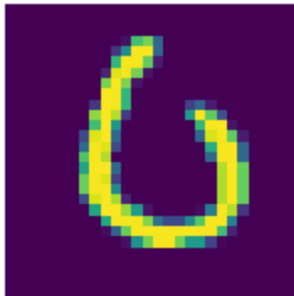
0 -> 7: 1

0 -> 6: 1

0 -> 7



0 -> 6



Number of misclassified images for No Defense Attack: 61

Attack: fgsm_bim_attack

Dataset: MNIST

Training Epochs: 10

Adversarial Training Images: 60

Test Images: 140

Accuracy: 0.05

Precision: 0.09

Recall: 0.08

F1-score: 0.08

ROC AUC Score: 1.00

Misclassifications:

0 -> 6: 14

0 -> 8: 4

0 -> 5: 14

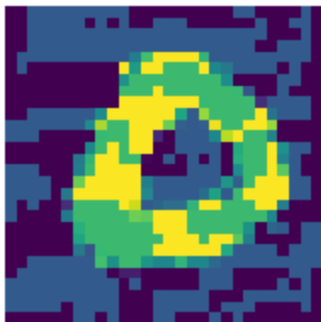
0 -> 9: 11

0 -> 7: 12

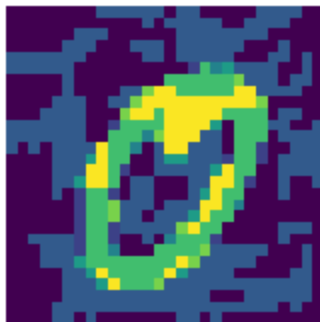
0 -> 4: 5

0 -> 2: 1

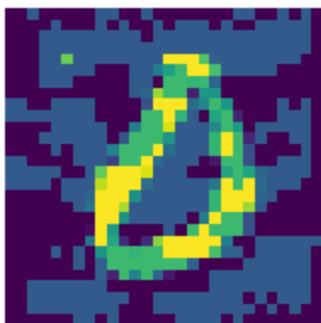
0 -> 6



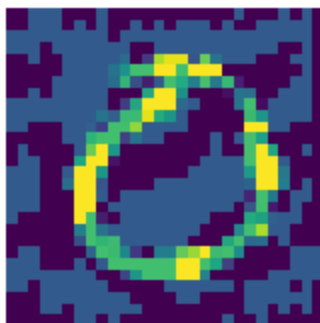
0 -> 8



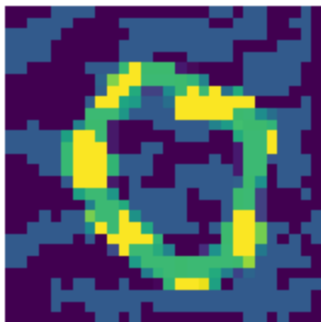
0 -> 6



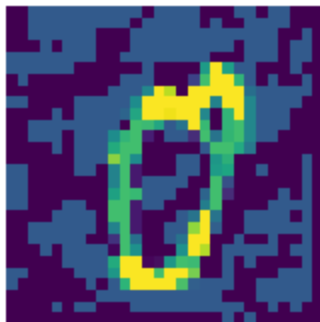
0 -> 6



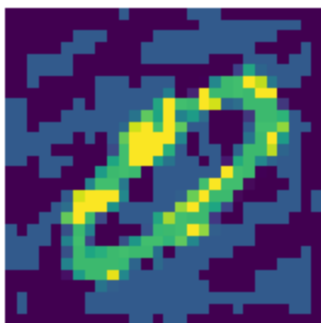
0 -> 5



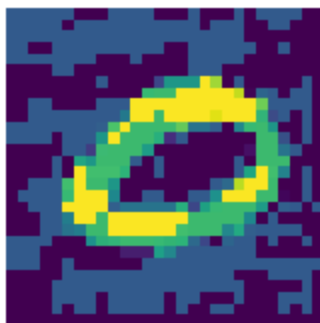
0 -> 5



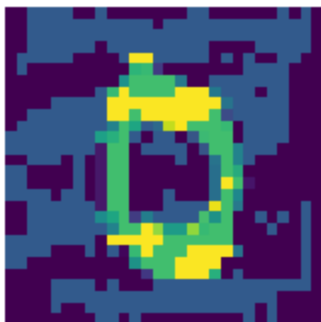
0 -> 8



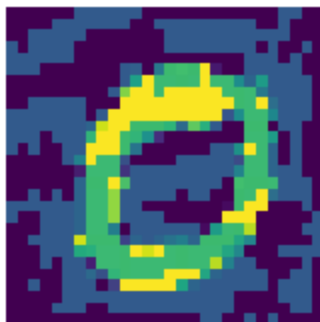
0 -> 9



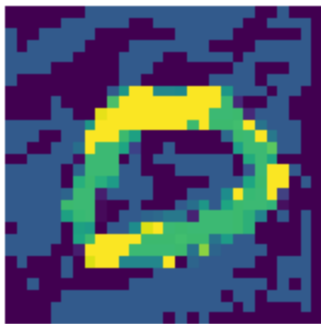
0 -> 5



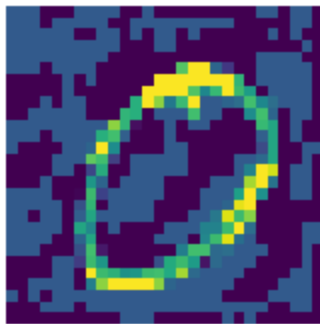
0 -> 5



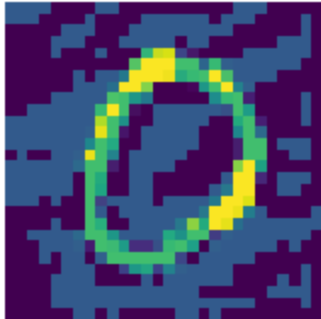
0 -> 5



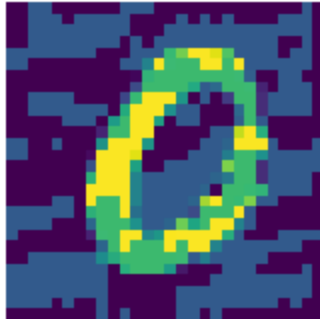
0 -> 9



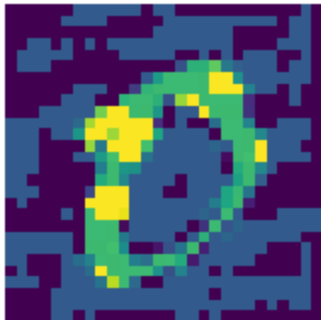
0 -> 9



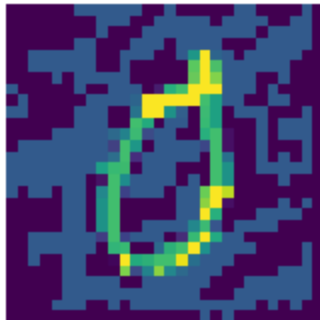
0 -> 6



0 -> 7



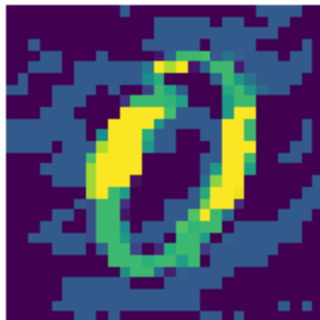
0 -> 7



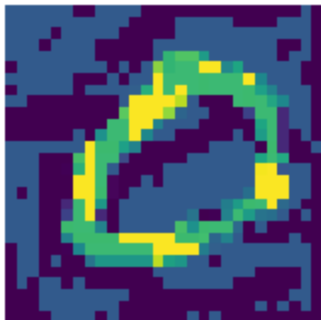
0 -> 4



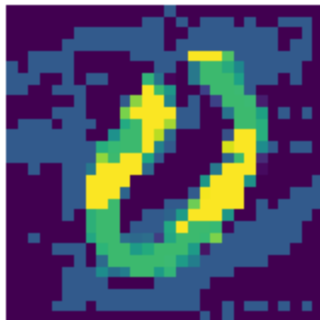
0 -> 7



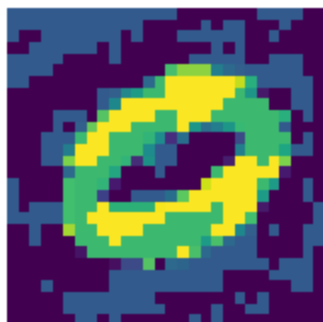
0 -> 6



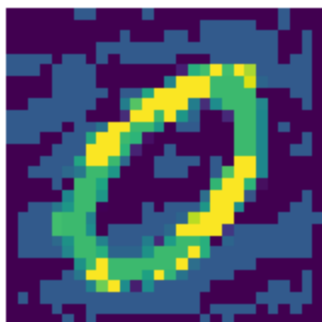
0 -> 4



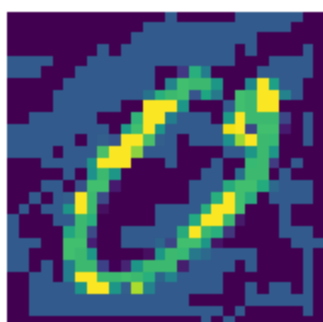
0 -> 9



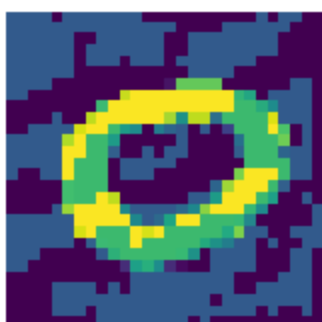
0 -> 7



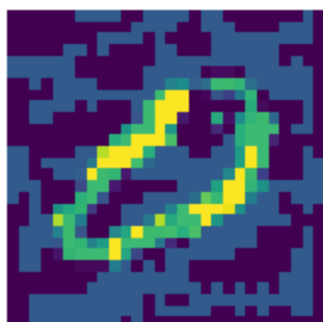
0 -> 7



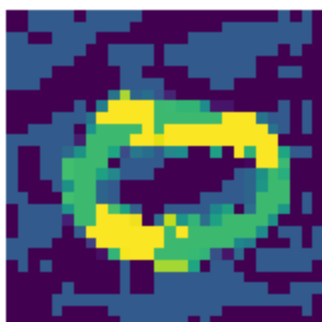
0 -> 9



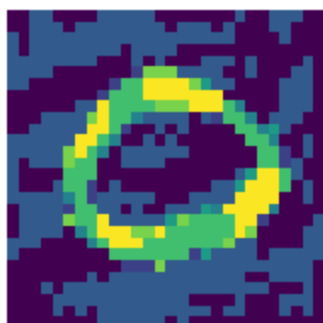
0 -> 4



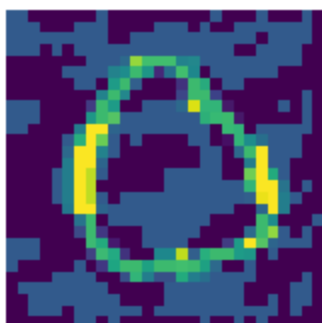
0 -> 5



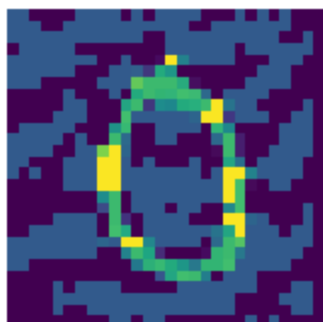
0 -> 9



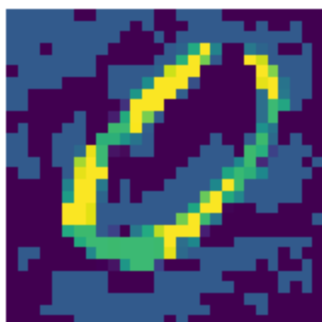
0 -> 6



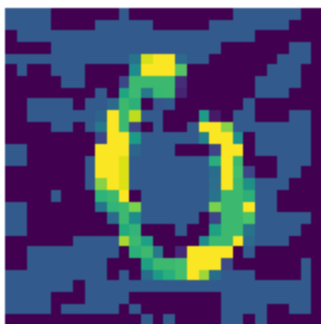
0 -> 5



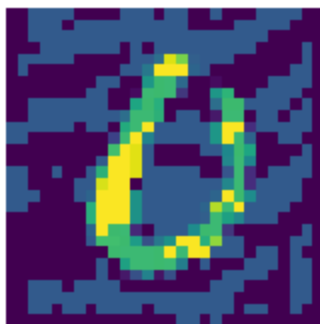
0 -> 6



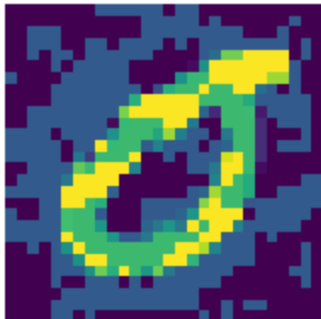
0 -> 6



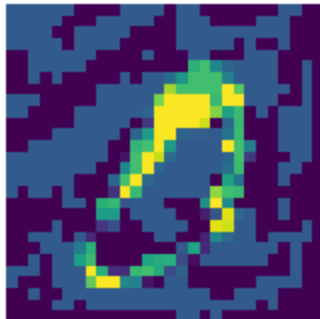
0 -> 6



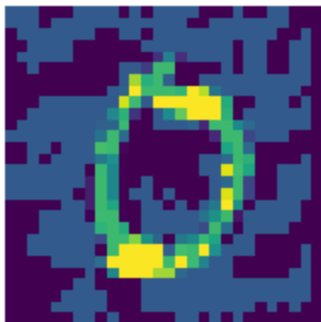
0 -> 7



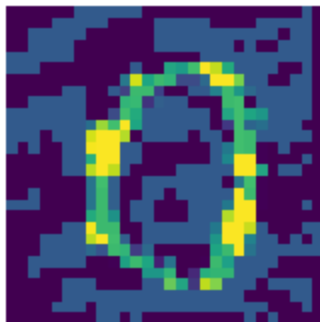
0 -> 5



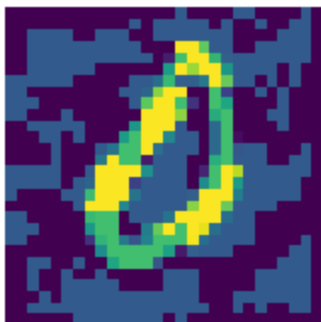
0 -> 5



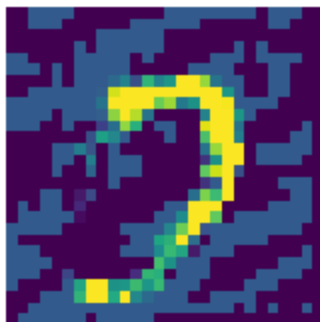
0 -> 7



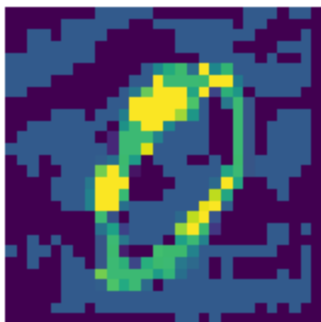
0 -> 6



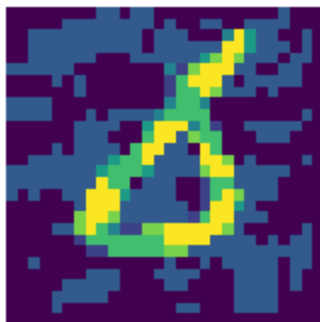
0 -> 7



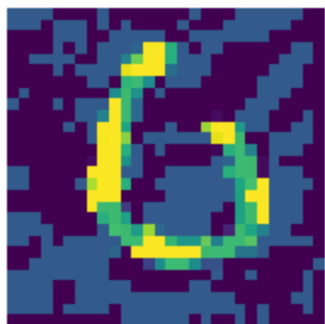
0 -> 4



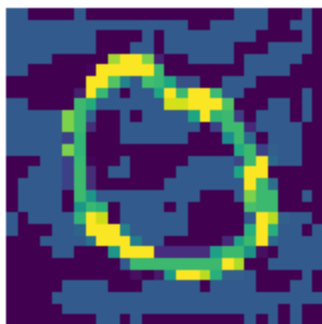
0 -> 6



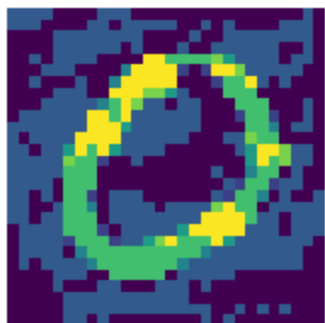
0 -> 6



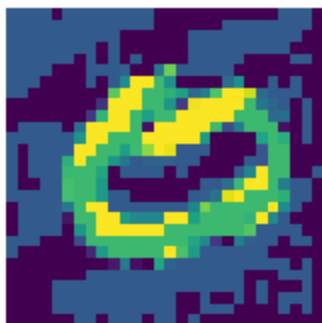
0 -> 5



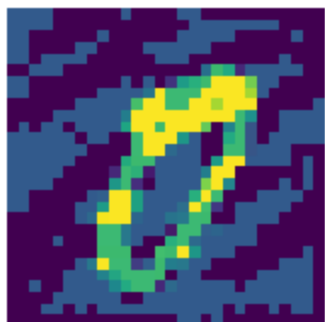
0 -> 2



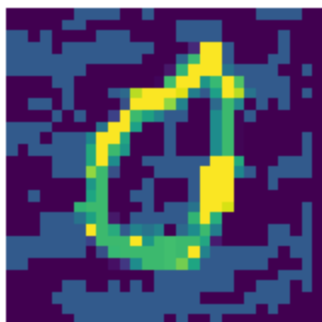
0 -> 9



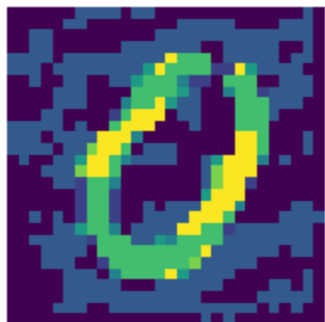
0 -> 8



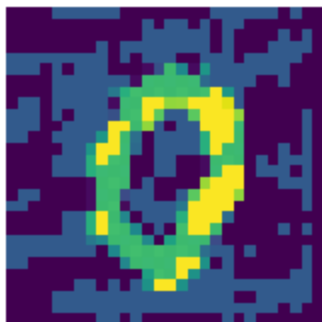
0 -> 9



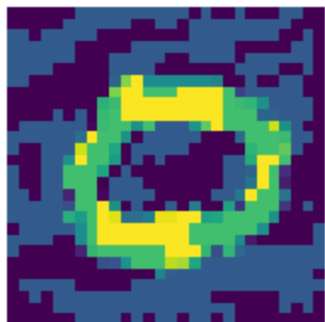
0 -> 4



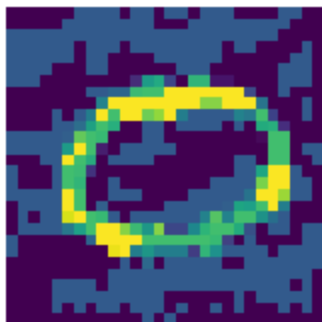
0 -> 7



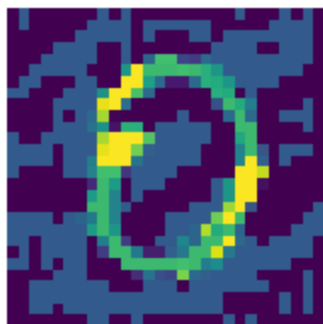
0 -> 5



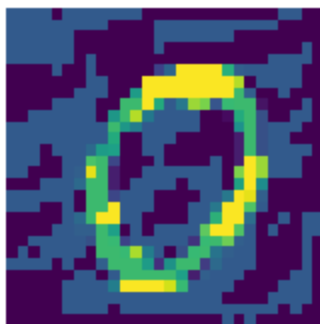
0 -> 9



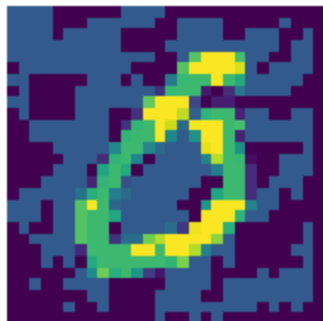
0 -> 7



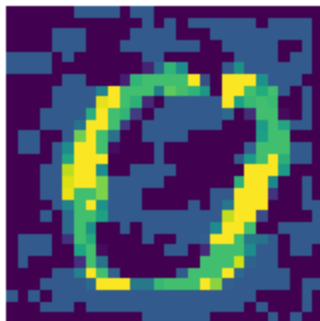
0 -> 9



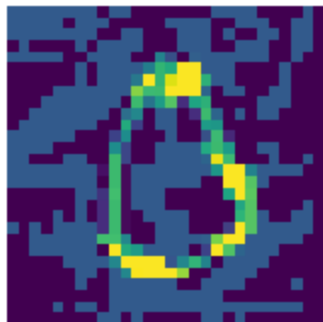
0 -> 5



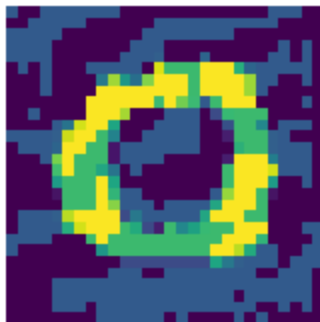
0 -> 7



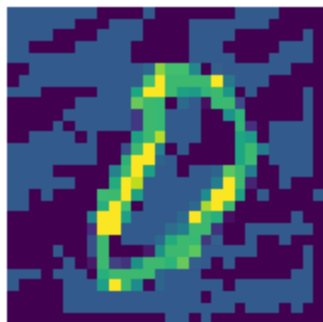
0 -> 5



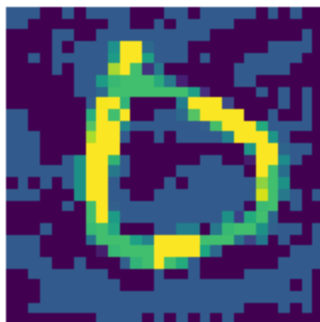
0 -> 9



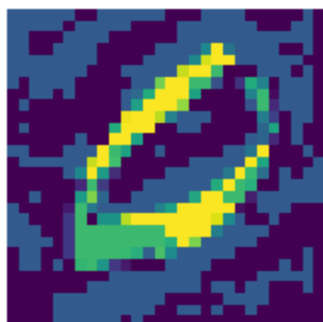
0 -> 8



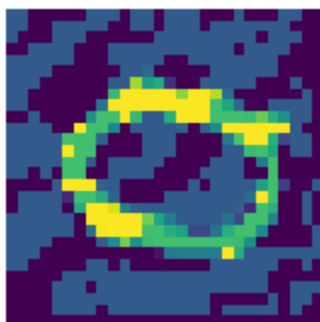
0 -> 6



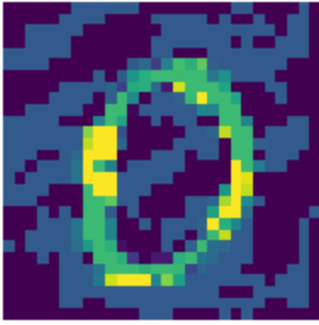
0 -> 6



0 -> 5



0 -> 7



Number of misclassified images for w/ Defense Attack: 5

Attack: fgsm_bim_attack

Dataset: MNIST

Training Epochs: 10

Retrained Clean and Adversarial Images: 124

Test Images: 140

Accuracy: 0.92

Precision: 0.96

Recall: 0.96

F1-score: 0.96

ROC AUC Score: 1.00

Misclassifications:

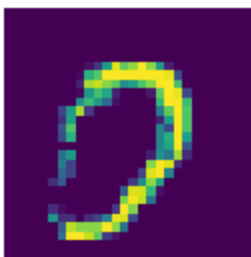
0 -> 6: 4

0 -> 7: 1

0 -> 6



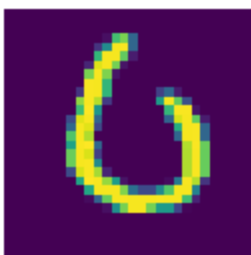
0 -> 7



0 -> 6



0 -> 6



0 -> 6

