Classification error (%) on the first 1000 test samples

| | | mnist | | | fashion_mnist | | |
|---|---|---|---|---|---|---|---|
| | | original | attacked | | original | attacked | |
| | | | scenario 1 | scenario 2 | | scenario 1 | scenario 2 |
| UNENCRYPTED | CW $l_2$ | 1.49 | 100.00 | | 8.30 | 100.00 | |
| | CW $l_0$ | | 100.00 | | | 100.00 | |
| | CW $l_\infty$ | | 100.00 | | | 100.00 | |
| | FGSM | 2.10 | 82.94 | | 9.50 | 94.25 | |
| PERMUTATED | CW $l_2$ | 3.70 | 100.00 | 4.50 | 12.30 | 100.00 | 12.70 |
| | CW $l_0$ | | 100.00 | 7.30 | | 100.00 | 12.50 |
| | CW $l_\infty$ | | 100.00 | 5.40 | | 100.00 | 12.90 |
| | FGSM | 4.19 | 89.14 | | 12.00 | 91.82 | |
| ECB | CW $l_2$ encrypt v1 | 16.58 | | | 55.66 | irrelevant | irrelevant |
| | CW $l_2$ encrypt v2 | 18.11 | | | 41.97 | | |
| | FGSM encrypt v1 | 20.88 | | | 59.23 | | |
| | FGSM encrypt v2 | 19.95 | | | 46.25 | | |
| CBC | CW $l_2$ encrypt v1 | 64.07 | irrelevant | irrelevant | 72.12 | irrelevant | irrelevant |
| | CW $l_2$ encrypt v2 | 69.12 | | | 64.47 | | |
| | FGSM encrypt v1 | 88.65 | | | 90.00 | | |
| | FGSM encrypt v2 | 88.65 | | | 90.00 | | |
| CTR | CW $l_2$ encrypt v1 | 88.65 | irrelevant | irrelevant | 90.00 | irrelevant | irrelevant |
| | CW $l_2$ encrypt v2 | 88.65 | | | 90.00 | | |
| | FGSM encrypt v1 | 88.65 | | | 90.00 | | |
| | FGSM encrypt v2 | 88.65 | | | 90.00 | | |

scenario 1: the attacker gets an oracle to the attacked model which for given an image, performs the permutation, feeds through the model's layers and returns the logits
scenario 2: the attacker gets to knows only the architecture, i.e. he gets an oracle to the unencrypted version

accuracies of permutation on different image sizes
(padding done with 0's around the original)

| | image size | error rate | min/epoch |
|---|---|---|---|
| mnist | 28x28 | 3.63 | 4 |
| | 40x40 | 2.65 | 5 |
| | 60x60 | 2.69 | 12 |
| | 100x100 | 2.30 | 14 |
| fashion_mnist | 28x28 | 12.40 | |
| | 40x40 | 12.07 | 13 |
| | 60x60 | | |
| | 100x100 | | |