

DEFENCE AGAINST ADVERSARIAL EXAMPLES

Yishay Asher • Steve Gutfreund
Instructor: Hanan Rosemarin



אוניברסיטת בר-אילן
Bar-Ilan University

Problem Description

Building high accuracy DNN models which are sufficiently
resistant to adversarial attacks

Background and Goal

- ✓ An adversarial example is an instance with small, intentional feature perturbations that cause a machine learning model to make a false prediction.
- ✓ The goal is to Find a way to train 'secured' models such that this sort of attacks should not affect them.
- ✓ Project based on the article [Bridging machine learning and cryptography in defence against adversarial attacks](#)

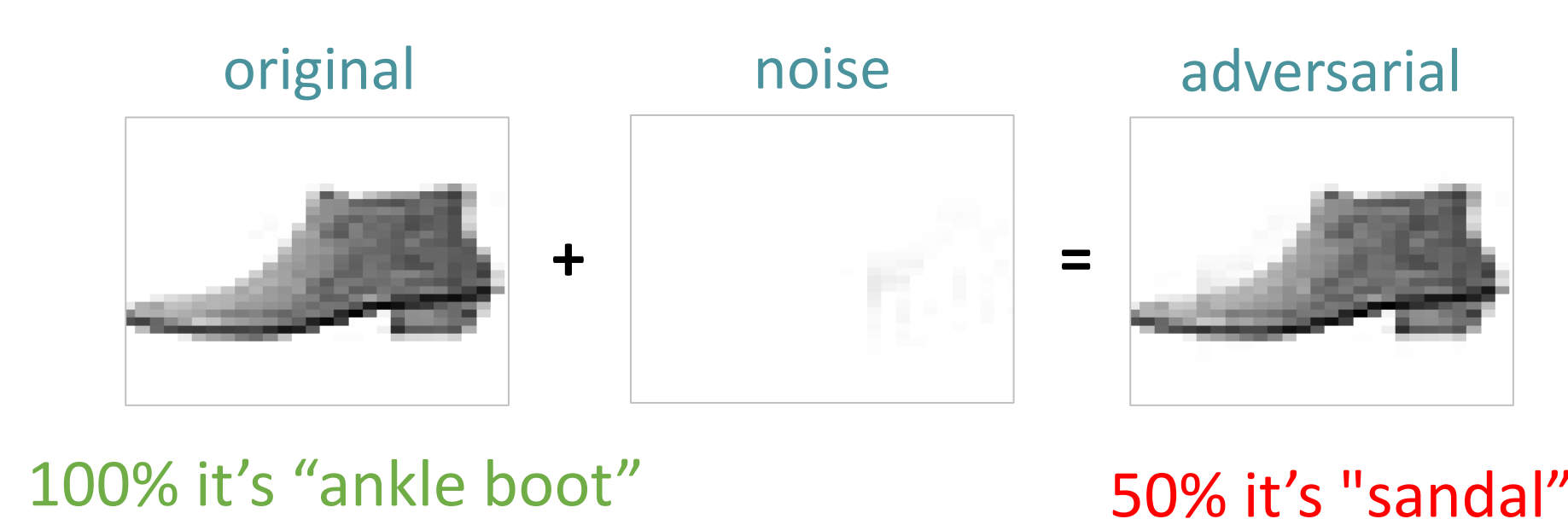


Figure 1: example of an adversarial image

Set-Up

- ✓ Mnist and Fashion-Mnist datasets
- ✓ Using well-known neural nets

1. Securing Models

Approach: training models on encrypted images.

Encryption techniques:

- ✓ Permutation
- ✓ AES in ECB, CBC and CTR modes

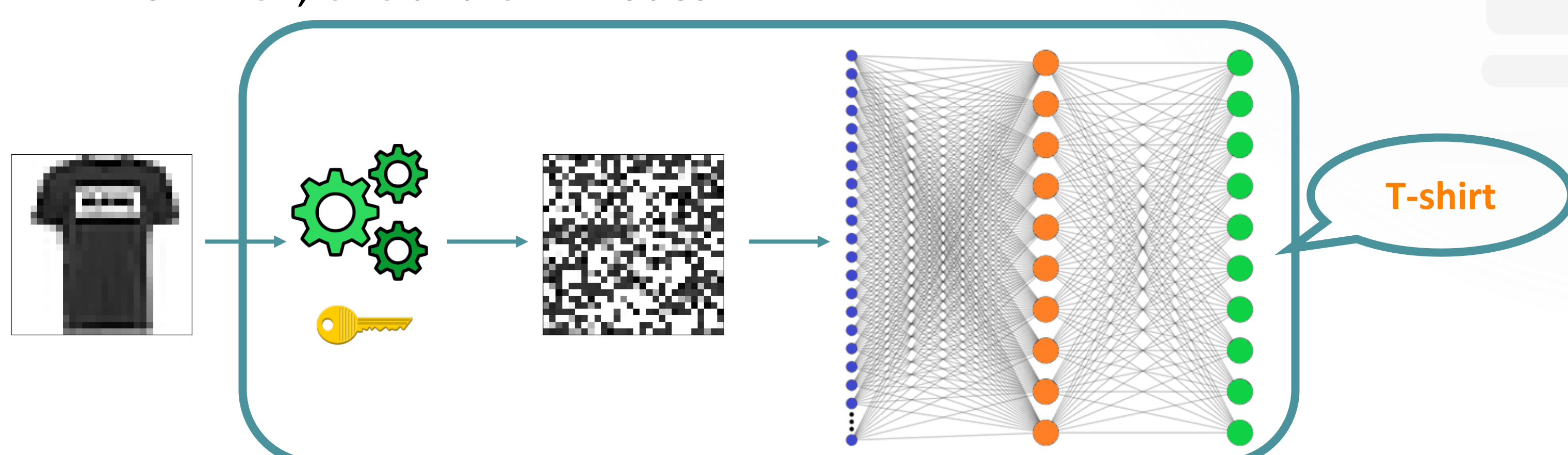


Figure 2: architecture for securing models

2. Cutting Loose Ends

Eliminated the models that did not learn well. Learning encrypted images is not very intuitive, as can be seen in figure 3.

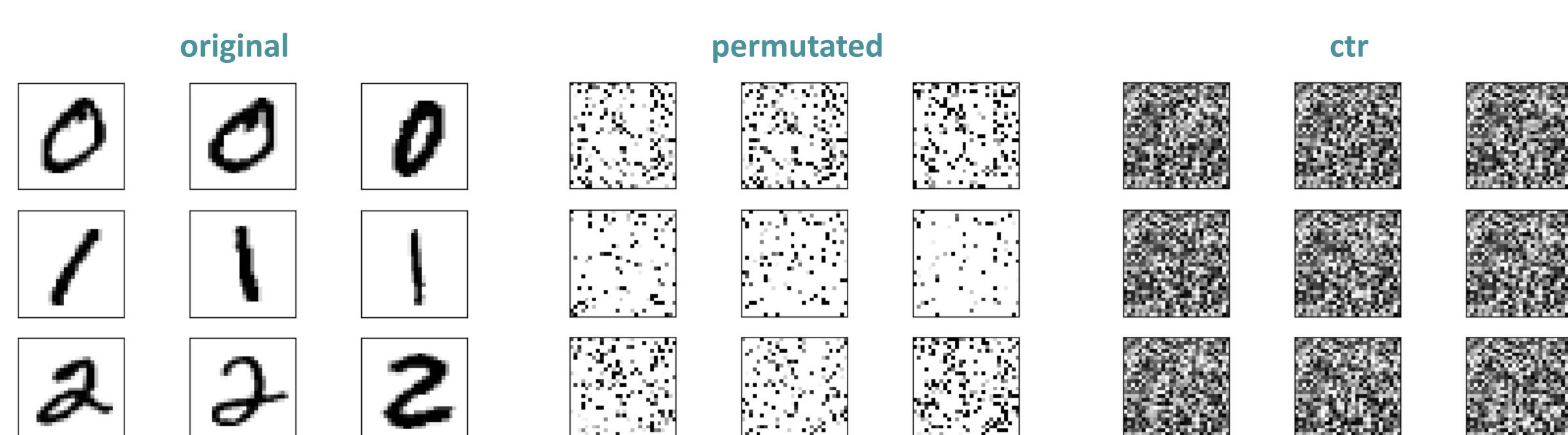


Figure 3: Sample of the encrypted images. Interesting to see how for the human eye it's not easy to distinguish between various classes but a DNN model classifies quite well, as can be seen in table 1

Future Work

- ✓ Improve accuracy on AES-ECB model
- ✓ Nicholas Carlini (the 'C' in CW attack) believes we still might defeat these defenses. (we contacted him)
- ✓ Test on more complicated datasets; i.e. Cifar-10

3. Attacking

Attacks:

- ✓ Carlini & Wagner, CW
 - ✓ Fast Gradient Sign Method, FGSM
- 'gray-box' scenario, i.e. the attacker knows the architecture of the model but has no access to the private key.

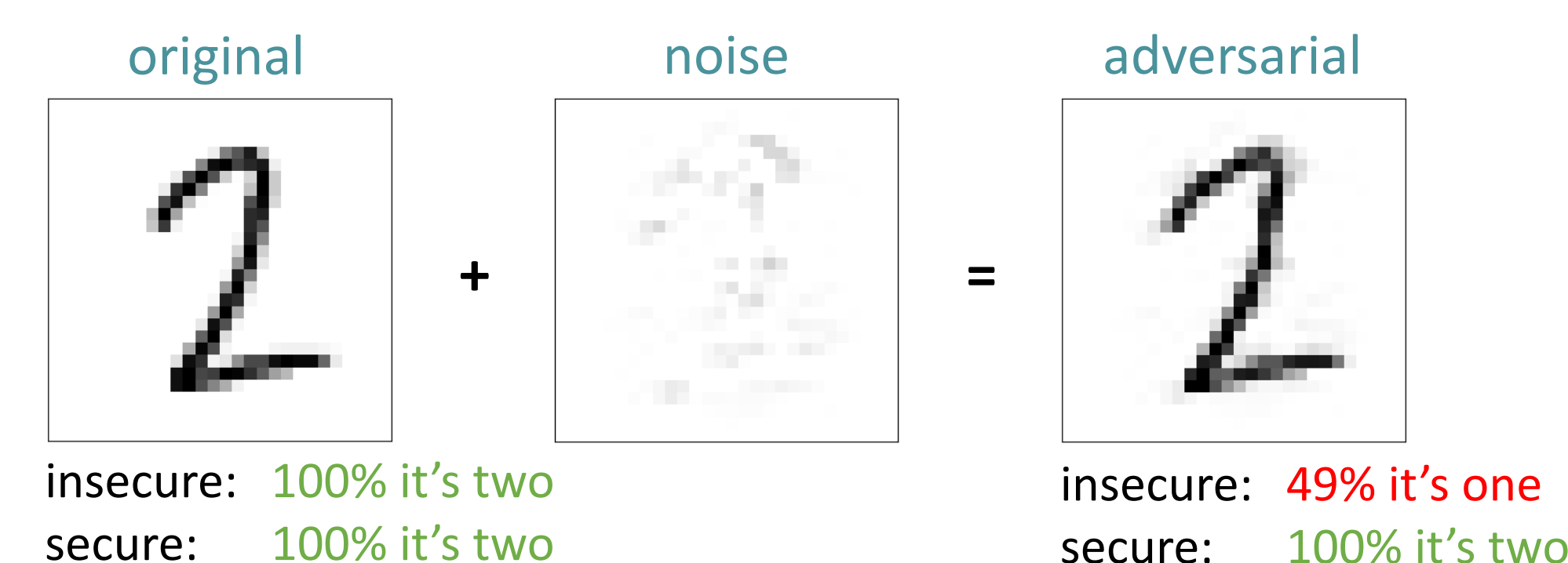


Figure 4: visualization of a CW attack secured by permutation

Results

There's a slight tradeoff between accuracy on the original images and the accuracy on the adversarial images, but overall, accuracies are good

| Classification error (%) on the first 1000 test samples | | | | | | | |
|---|-----------------|---------------|----------|-----------------|---------------|------------|--|
| model | original images | mnist | | original images | fashion mnist | | |
| | | attack | gray box | | attack | gray box | |
| UNENCRYPTED | A | CW l_2 | 100.00 | 8.30 | CW l_2 | 100.00 | |
| | | CW l_0 | 100.00 | | CW l_0 | 100.00 | |
| | | CW l_∞ | 100.00 | | CW l_∞ | 100.00 | |
| PERMUTATED | A | FGSM | 39.50 | 9.50 | FGSM | 77.20 | |
| | | CW l_2 | 4.50 | | CW l_2 | 12.70 | |
| | | CW l_0 | 7.30 | | CW l_0 | 12.50 | |
| AES · ECB | A | CW l_∞ | 5.40 | 12.30 | CW l_∞ | 12.90 | |
| | | FGSM | 8.60 | | FGSM | 29.80 | |
| | | irrelevant | | | irrelevant | | |
| AES · CBC | A | CW l_2 | 18.40 | 54.60 | CW l_2 | irrelevant | |
| | | FGSM | 19.30 | | FGSM | irrelevant | |
| | | irrelevant | | | irrelevant | | |
| AES · CTR | A | CW l_2 | 67.60 | 71.50 | CW l_2 | irrelevant | |
| | | FGSM | 87.40 | | FGSM | irrelevant | |
| | | irrelevant | | | irrelevant | | |
| AES · CTR | A | CW l_2 | 3.70 | 17.40 | CW l_2 | 17.20 | |
| | | FGSM | 4.20 | | FGSM | 26.50 | |
| | | irrelevant | | | irrelevant | | |

Table 1: table containing all the results

Success with Permutation , Coincidence?

To verify that the learning ability of a permutation model does not result from high density in small images, we trained models on padded images. Padding done with white pixels. See table 2 for results.

| | image size | error rate |
|---------------|------------|------------|
| mnist | 28x28 | 3.70 |
| | 40x40 | 3.40 |
| | 60x60 | 3.30 |
| fashion mnist | 28x28 | 12.30 |
| | 40x40 | 14.40 |
| | 60x60 | 10.80 |

Table 2: results for training permuted data, various image dimensions