

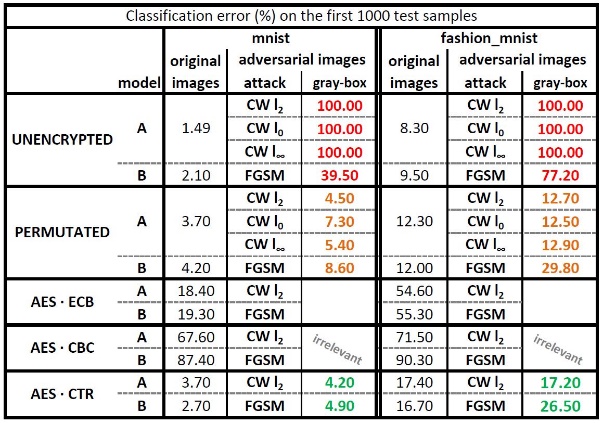
Figure 3: Sample of the encrypted images (permutated and aes-ctr). Interesting to see how for the human eyes it’s impossible to distinguish between various classes but a DNN model classifies quite well, see figure 4 for accuracies

**original**

**permutated**

**ctr**

Figure 4: Table of accuracies



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

# Problem Description

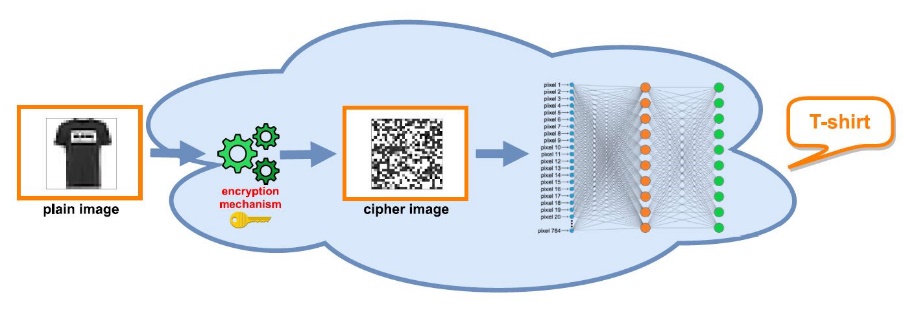


Figure 1: Architecture of the models

­­­­­­­­­**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Yishay Asher • Steve Gutfreund

**Instructor:** Hanan Rosemarin

**DEFENCE AGAINST**

**ADVERSARIAL EXAMPLES**

The project is based on the article “[Bridging machine learning and cryptography in defence against adversarial attacks](https://arxiv.org/pdf/1809.01715.pdf)”, in the article they used only permutation, we got quite surprising results with AES in CTR mode encryption as well.

See figure 4 for the detailed results.

# Results

* improve accuracy on AES-ECB model (we got error rate of 19% on mnist and 55% on fashion-mnist)
* we contacted Nicholas Carlini (the ‘C’ in CW attack) and he believes we still might defeat these defenses
* try some other datasets; i.e. cifar-10, its images are 3 layered (rgb) and might be more difficult to learn encrypted images

# Future Work

Training models on encrypted images, on the datasets mnist and fashion-mnist, see figure 1.

Encryption techniques:

* Permutation (on the pixels)
* AES in ECB, CBC and CTR modes of operation

# Set-Up

Attacking the sufficiently accurate models with the following attacks:

* Carlini & Wagner, CW

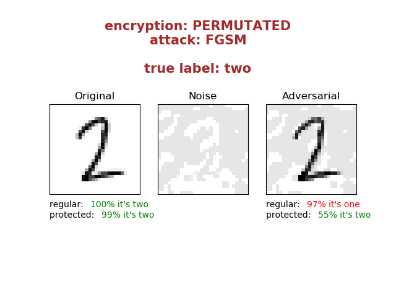
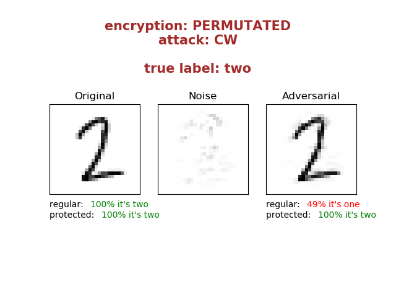
<https://github.com/carlini/nn_robust_attacks>

* Fast Gradient Sign Method, FGSM

<https://github.com/tensorflow/cleverhans>

# Attacking

Figure 2: visualization of an attack



**בס"ד**

