

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](#)

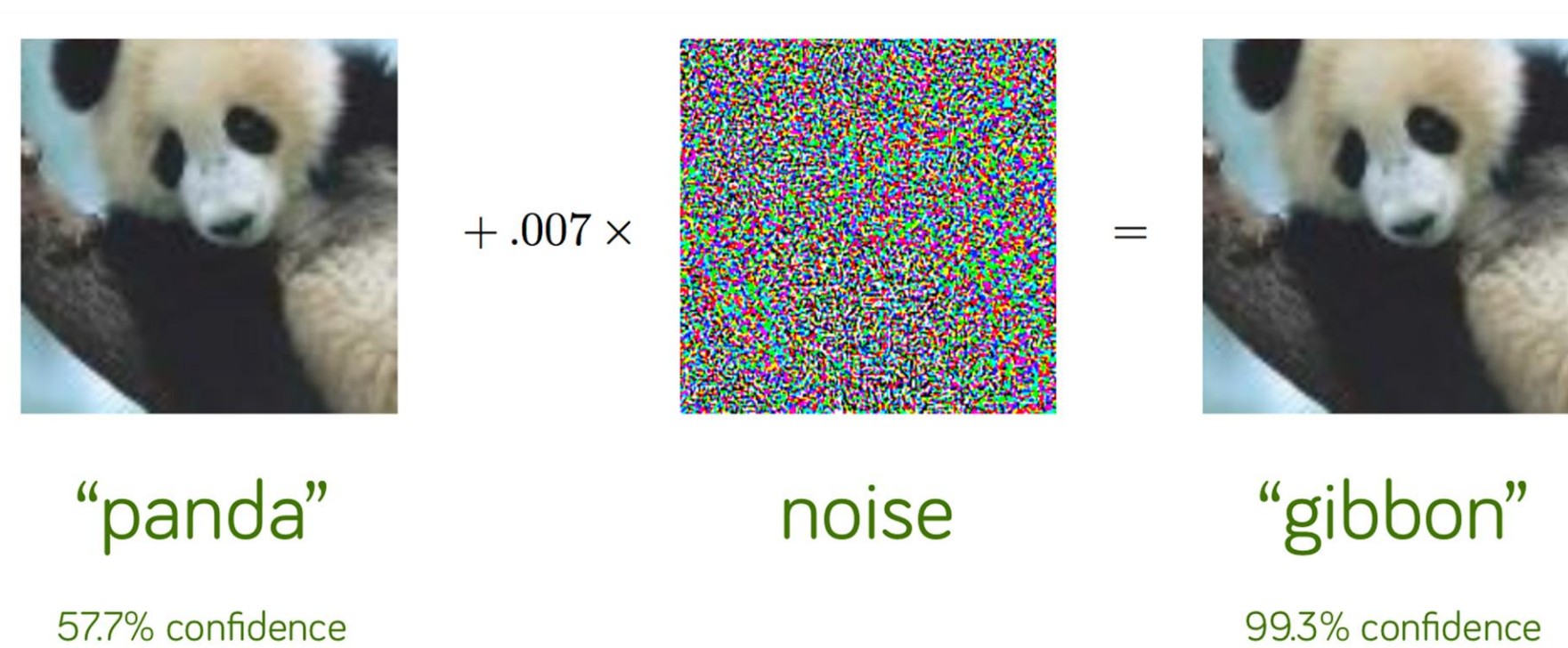


Figure 1: example of an adversarial example

Set-Up

- ✓ Mnist and fashion-mnist
- ✓ Using well-known neural nets

1. Securing Models

Approach: training models on encrypted images

Encryption techniques:

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

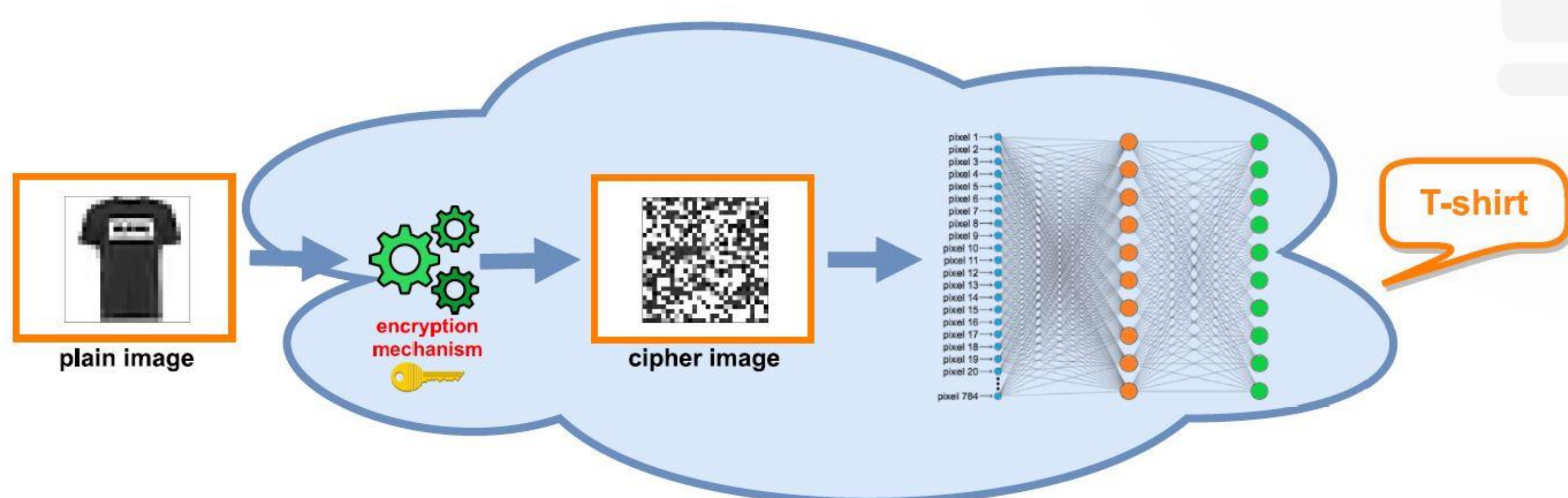


Figure 2: architecture for securing models

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.

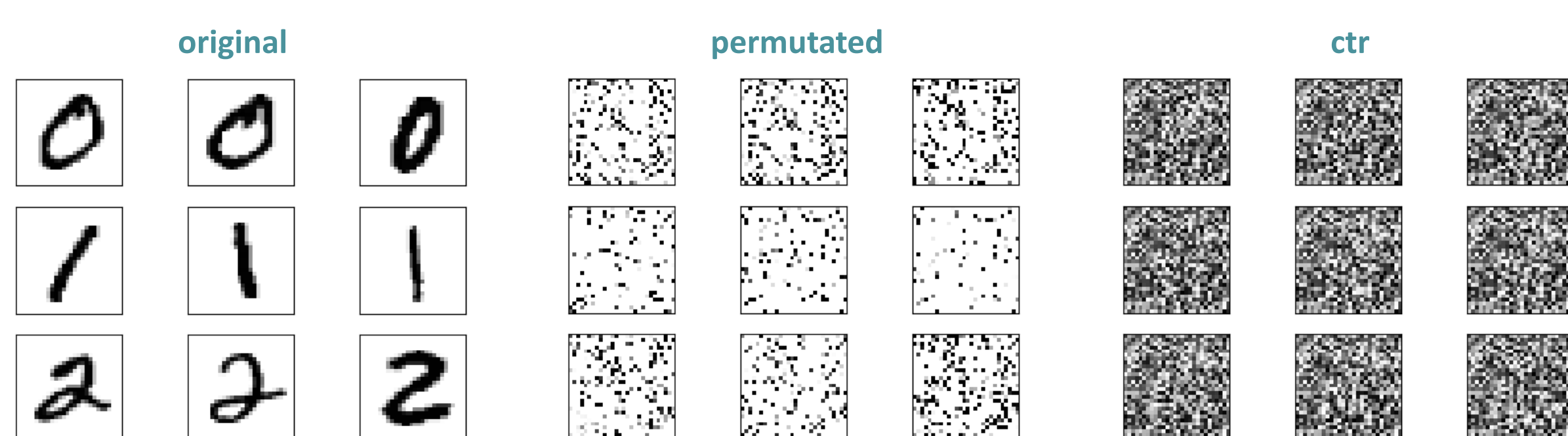


Figure 3: Sample of the encrypted images. Interesting to see how for the human eye it's impossible to distinguish between various classes but a DNN model classifies quite well, as in table 1

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.

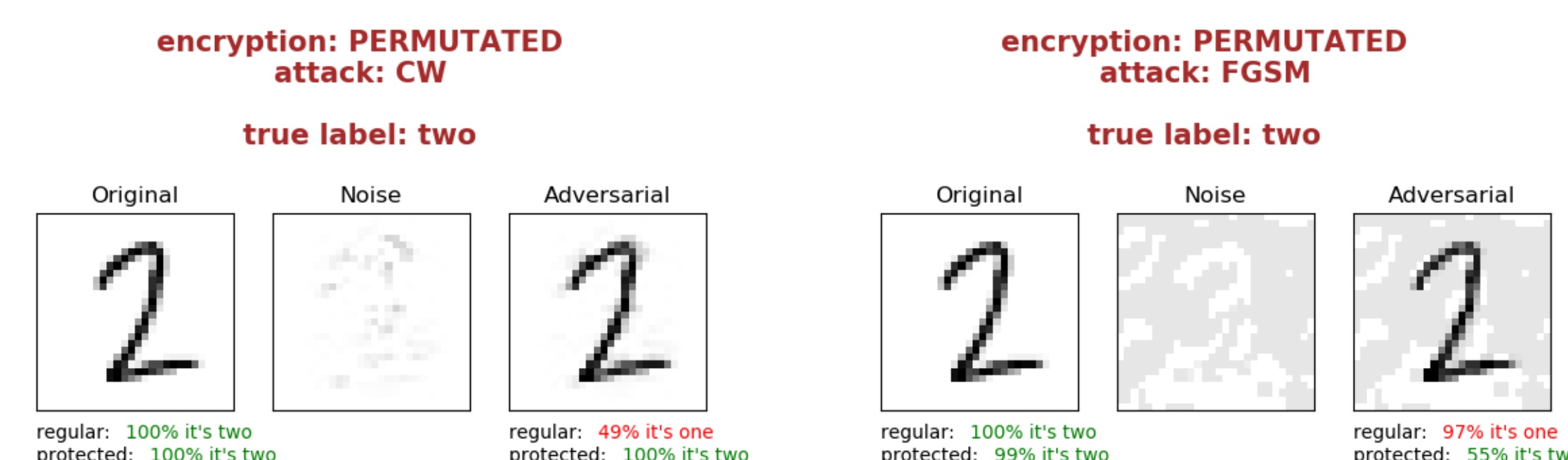


Figure 4: visualization of a CW and FGSM attack

Results

There's a slight tradeoff between accuracy on the original images and the accuracy on the adversarial images, but overall, accuracies are good. See table 1 for the detailed results.

Classification error (%) on the first 1000 test samples						
	model	original images	mnist		fashion_mnist	
			adversarial images	gray-box	adversarial images	gray-box
UNENCRYPTED	A	1.49	CW I_2	100.00	CW I_2	100.00
			CW I_0	100.00	CW I_0	100.00
			CW I_∞	100.00	CW I_∞	100.00
	B	2.10	FGSM	39.50	FGSM	77.20
PERMUTATED	A	3.70	CW I_2	4.50	CW I_2	12.70
			CW I_0	7.30	CW I_0	12.50
			CW I_∞	5.40	CW I_∞	12.90
	B	4.20	FGSM	8.60	FGSM	29.80
AES · ECB	A	18.40	CW I_2	irrelevant	CW I_2	irrelevant
	B	19.30	FGSM	irrelevant	FGSM	irrelevant
AES · CBC	A	67.60	CW I_2	irrelevant	CW I_2	irrelevant
	B	87.40	FGSM	irrelevant	FGSM	irrelevant
AES · CTR	A	3.70	CW I_2	4.20	CW I_2	17.20
	B	2.70	FGSM	4.90	FGSM	26.50

Table 1: table containing all the results

Efficiency of Permutation

In order to check the efficiency of training models on permuted data, we trained models on images with greater dimensions. See table 2 for results.

	image size	error rate
mnist	28x28	3.63
	40x40	2.65
	60x60	2.69
	100x100	2.30
fashion_mnist	28x28	12.40
	40x40	12.07
	60x60	11.41

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