CS@BIU
המחלקה למדעי המחשב
אוניברסיטת בר-אילן
BIU
Center for Research in Applied Cryptography and Cyber Security
אוניברסיטת בר-אילן
Bar-Ilan University

DEFENCE AGAINST ADVERSARIAL EXAMPLES

Yishay Asher • Steve Gutfreund
Instructor: Hanan Rosemarin

Problem Description

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

Background

An adversarial example is an instance with small, intentional feature perturbations that cause a machine learning model to make a false prediction. (See figure 1)

Goal

Find a way to train 'secured' models such that this sort of attacks should not affect them.

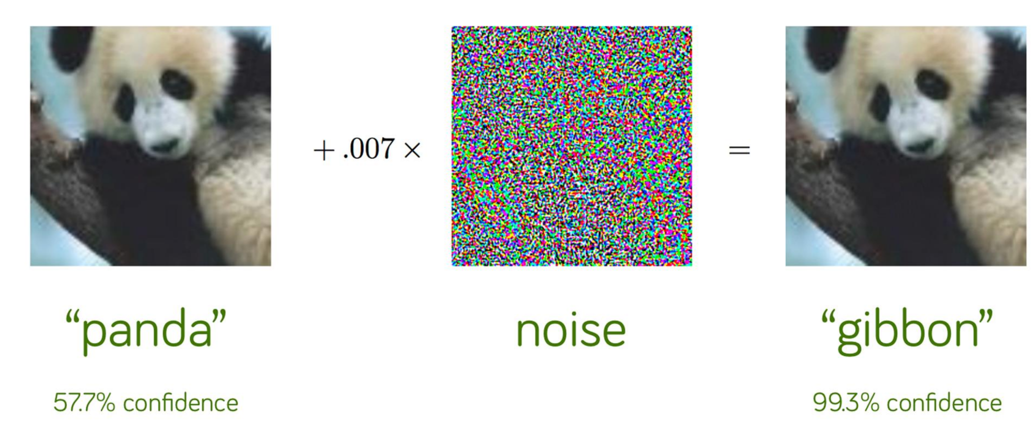

Figure 1: example of an adversarial example

Set-Up

We worked on mnist and fashion-mnist.
We've set neural networks known to be able to learn these datasets very well.

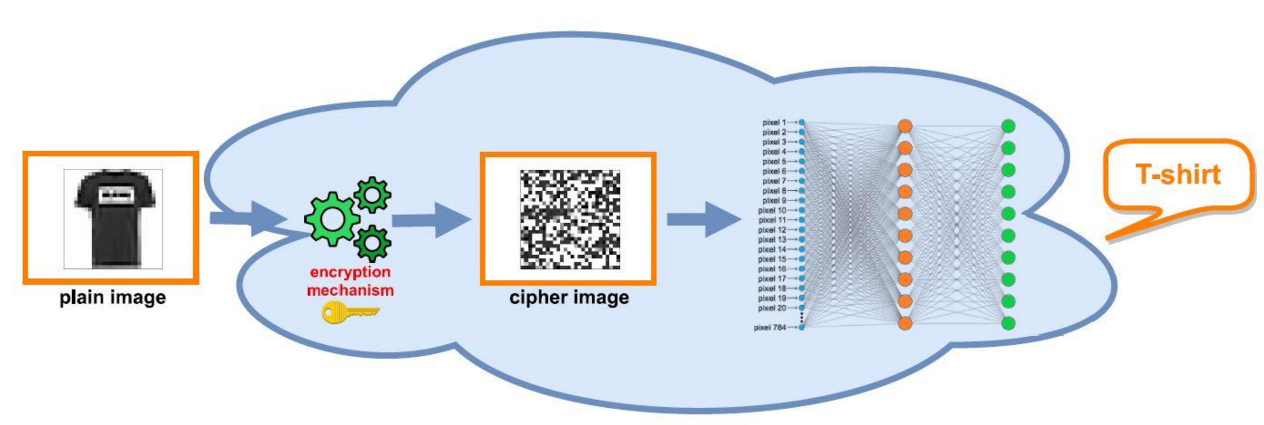

Figure 2: architecture of the secure models

1. Securing Models

We followed the approach presented in the article '[Bridging machine learning and cryptography in defence against adversarial attacks]':
training models on encrypted images, see figure 2.
Encryption techniques:

- Permutation on the pixels, as done in the article
- AES in ECB, CBC and CTR modes of operation

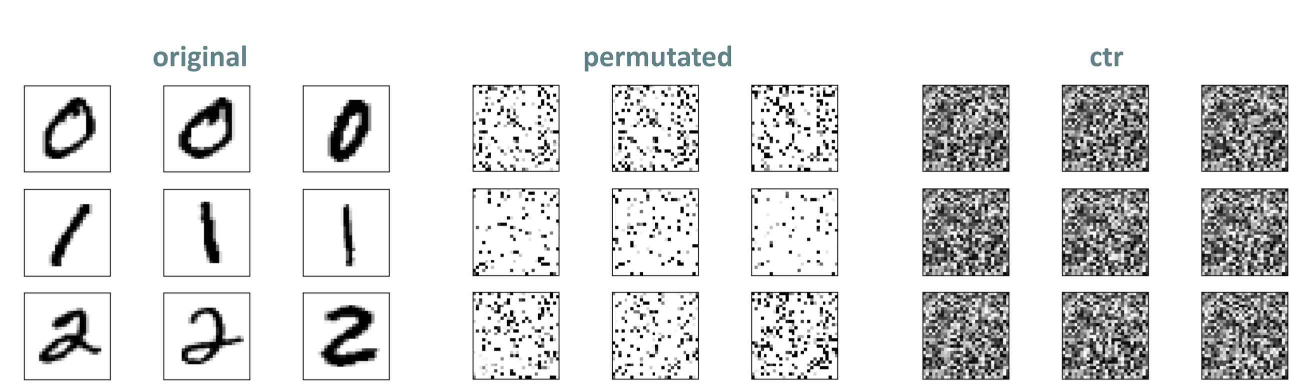

Figure 3: Sample of the encrypted images (permutated and aes-ctr). Interesting to see how for the human eyes it's impossible to distinguish between various classes but a DNN model classifies quite well, see figure 5 for accuracies

2. Cutting loose ends

Before performing an attack, we eliminated the models that did not learn well.
As can be seen in figure 3, learning is not so intuitive.
For this reason, AES in ECB and CBC mode were irrelevant to continue with.

3. Attacking

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

We focused on the 'gray-box' scenario, i.e. the attacker knows the architecture of the model but has no access to the private key.
See figure 4 for a visualization.

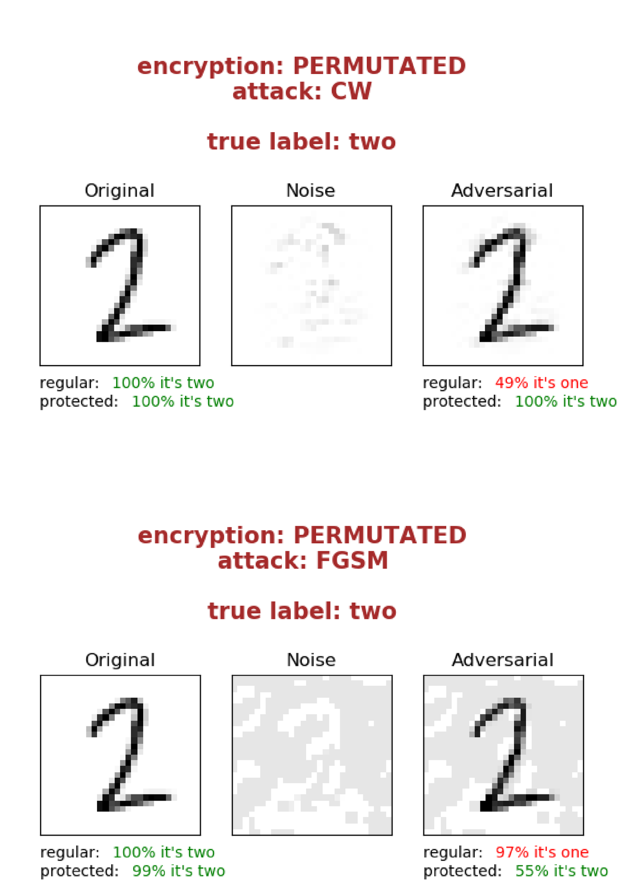

Figure 4: visualization of a CW and FGSM attack

Results

There's a tradeoff between accuracy on the original images and the accuracy on the adversarial images.

- permutation: although the error rate on the originals increased by a little bit (1% on mnist and 4% on fashion-mnist), the error rate on the adversarial decreased significantly
- AES-CTR: on mnist it performs better than permutation but yet on fashion mnist the error rate on the adversarials is slightly higher

See figure 5 for the detailed results.

Classification error (%) on the first 1000 test samples

		mnist			fashion_mnist		
		original	adversarial images		original	adversarial images	
	model	images	attack	gray-box	images	attack	gray-box
UNENCRYPTED	A	1.49	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
	B	2.10	FGSM	39.50	9.50	FGSM	77.20
PERMUTATED	A	3.70	CW l_2	4.50	12.30	CW l_2	12.70
			CW l_0	7.30		CW l_0	12.50
			CW l_∞	5.40		CW l_∞	12.90
	B	4.20	FGSM	8.60	12.00	FGSM	29.80
AES · ECB	A	18.40	CW l_2		54.60	CW l_2	
	B	19.30	FGSM		55.30	FGSM	
AES · CBC	A	67.60	CW l_2	irrelevant	71.50	CW l_2	irrelevant
	B	87.40	FGSM		90.30	FGSM	
AES · CTR	A	3.70	CW l_2	4.20	17.40	CW l_2	17.20
	B	2.70	FGSM	4.90	16.70	FGSM	26.50

Figure 5: table containing all the results

Future Work

- 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