CS@BIU
DEPARTMENT OF COMPUTER SCIENCE
BAR-ILAN UNIVERSITY
Center for Research in Applied Cryptography and Cyber Security

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 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]

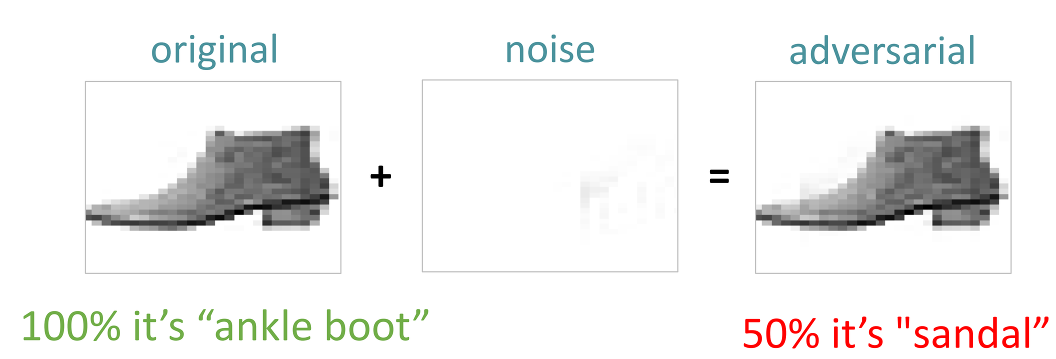

Figure 1: example of an adversarial image

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

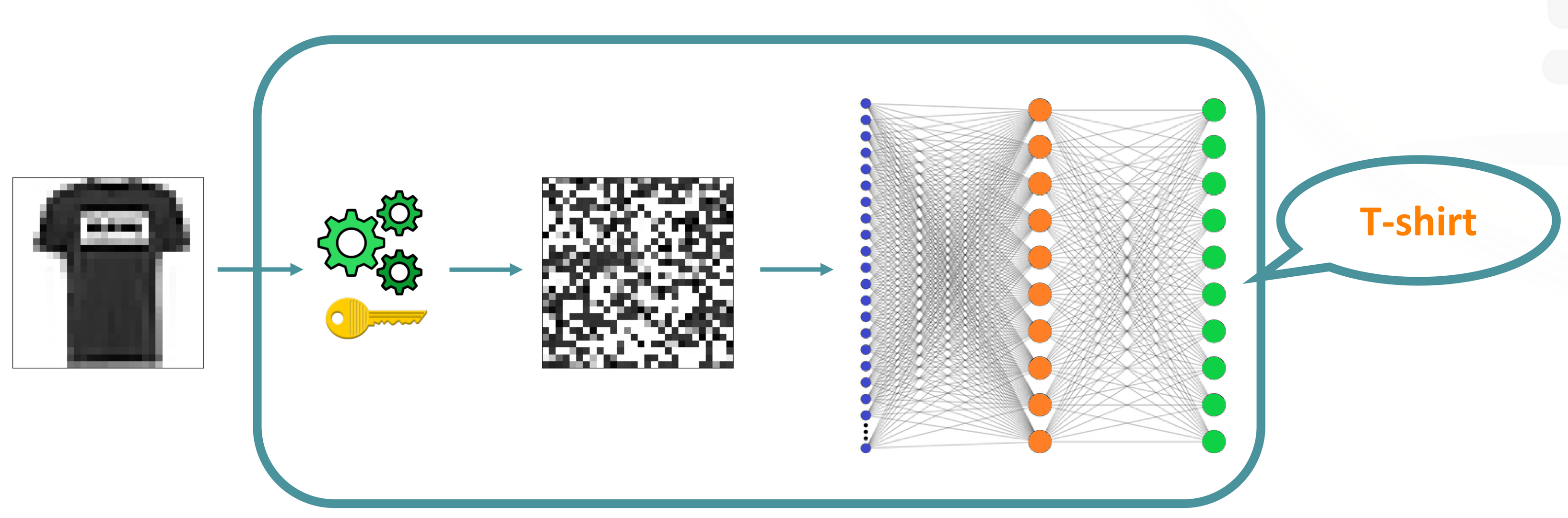

Figure 2: architecture for securing models

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.

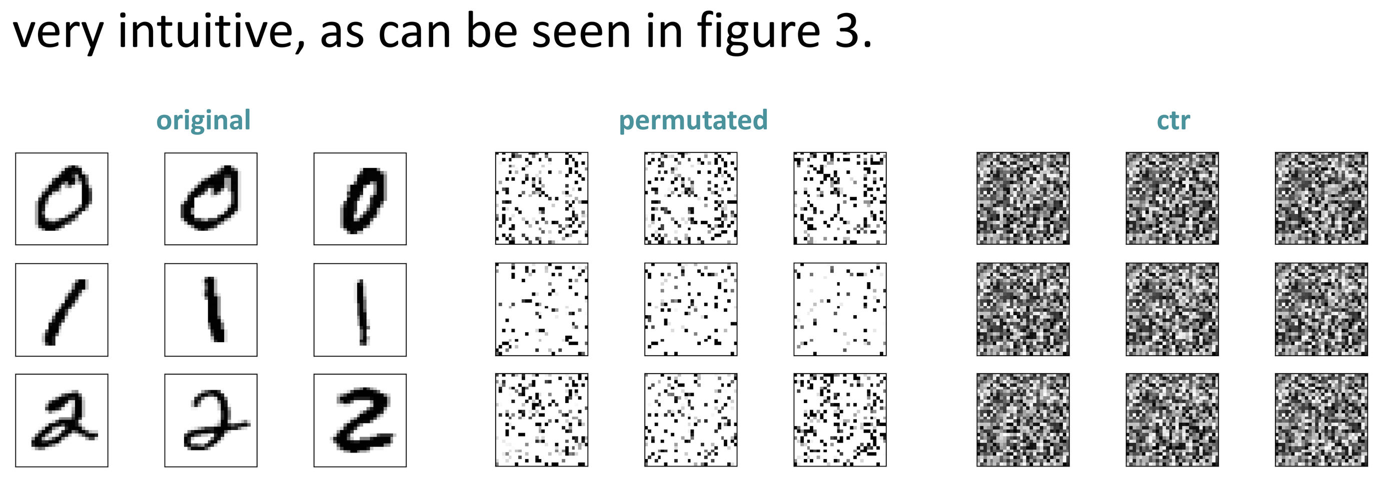

Figure 3: Sample of the encrypted images. Interesting to see how for the human eye it's difficult to distinguish between various classes but a DNN model classifies quite well, as can be seen in table 1

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.

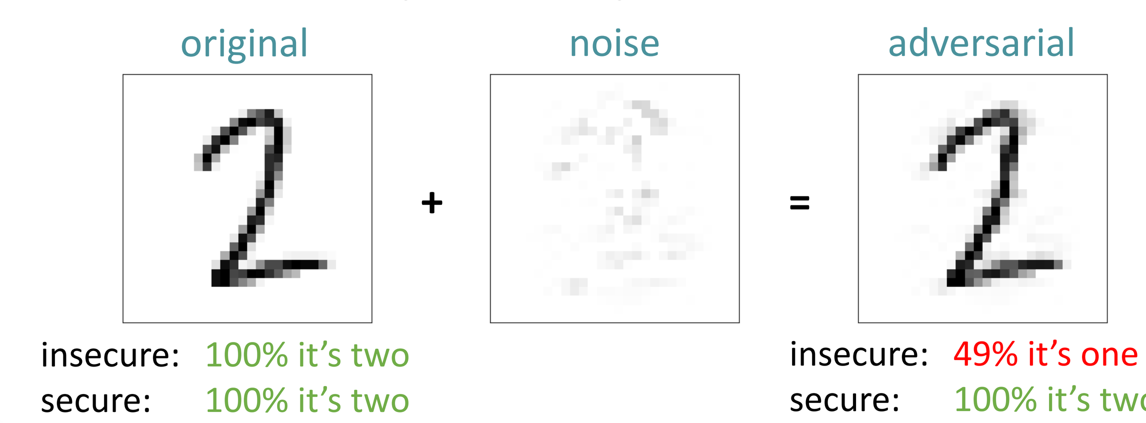

Figure 4: visualization of a CW attack secured by permutation

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

		mnist			fashion mnist		
	model	original images	adversarial images: attack	adversarial images: gray box	original images	adversarial images: attack	adversarial images: 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	irrelevant	54.60	CW l_2	irrelevant
	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

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 permutated data, various image dimensions

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