CS@BIU
המחלקה למדעי המחשב
אוניברסיטת בר-אילן

DEFENCE AGAINST ADVERSARIAL EXAMPLES

Yishay Asher • Steve Gutfreund
Instructor: Hanan Rosemarin

אוניברסיטת בר-אילן
Bar-Ilan University

BIU
Center for Research in Applied Cryptography and Cyber Security

Problem Description

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

Set-Up

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

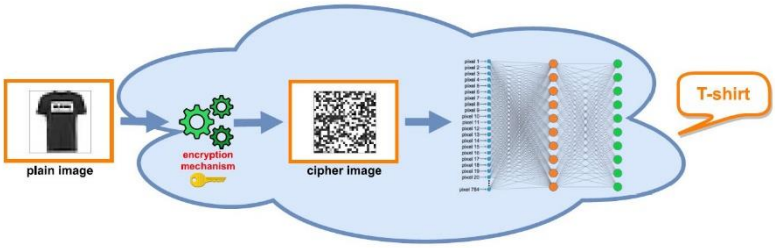

Figure 1: Architecture of the models

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

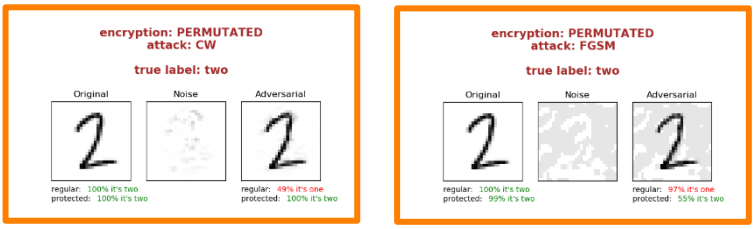

Figure 2: visualization of an attack

Results

the project is based on the article "Bridging machine learning and cryptography in defence against adversarial attacks", 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.

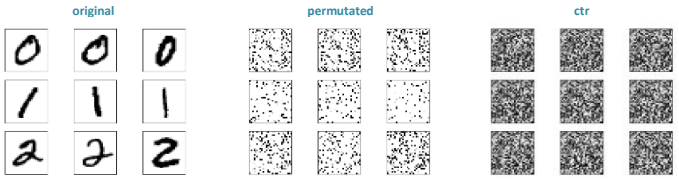

Figure 3: Sample of the encrypted images (permutated and aes-ctr). Interesting to see how for the human eyes it's impossible to distinguish between various classes but a DNN model classifies quite well, see figure 4 for accuracies

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 owner of 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

Classification error (%) on the first 1000 test samples

	model	mnist original images	mnist adversarial images attack	mnist adversarial images gray-box	fashion_mnist original images	fashion_mnist adversarial images attack	fashion_mnist 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 · 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 4: Table of accuracies

python™ Keras TensorFlow AES ADVANCED ENCRYPTION STANDARD