בס"ד

CS@BIU
המחלקה למדעי המחשב
אוניברסיטת בר-אילן

BIU
Center for Research in Applied Cryptography and Cyber Security

DEFENSE AGAINST ADVERSARIAL EXAMPLES

Yishay Asher • Steve Gutfreund
Supervisor: Hanan Rosemarin

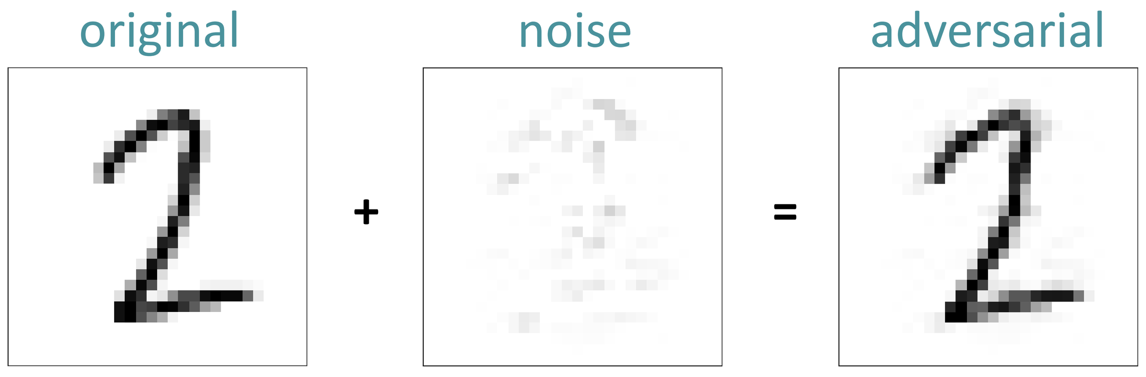
אוניברסיטת בר-אילן
Bar-Ilan University

Problem Description

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

0 Background and Goal

- An adversarial example is an instance with small, intentional feature perturbations that causes a machine learning model to make a false prediction.
- The goal is to find a way to train 'secured' models such that these sort of attacks should not affect them.
- Project based on the article Bridging machine learning and cryptography in defense against adversarial attacks.



example of an adversarial image

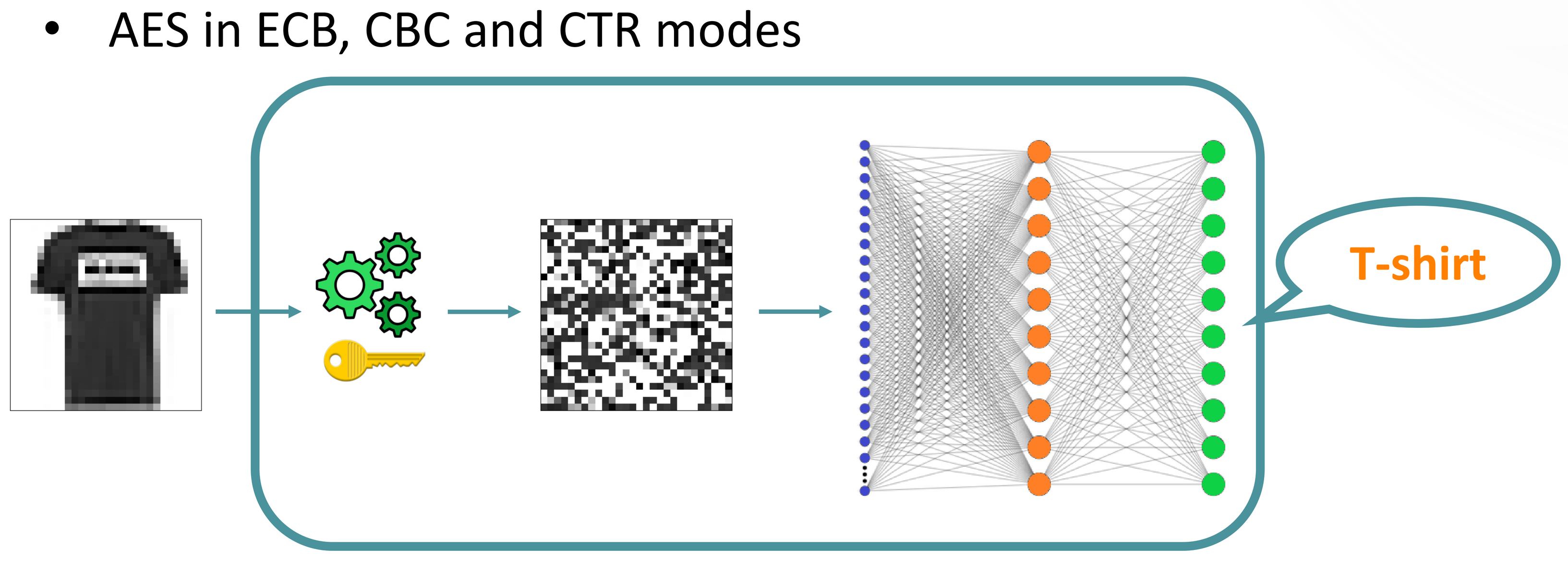
1 Set-Up

- MNIST and Fashion-MNIST datasets
- Using well-known neural nets
- Training 'unsecured' models

2 Securing Models

Approach: training models on encrypted images.
Encryption techniques:

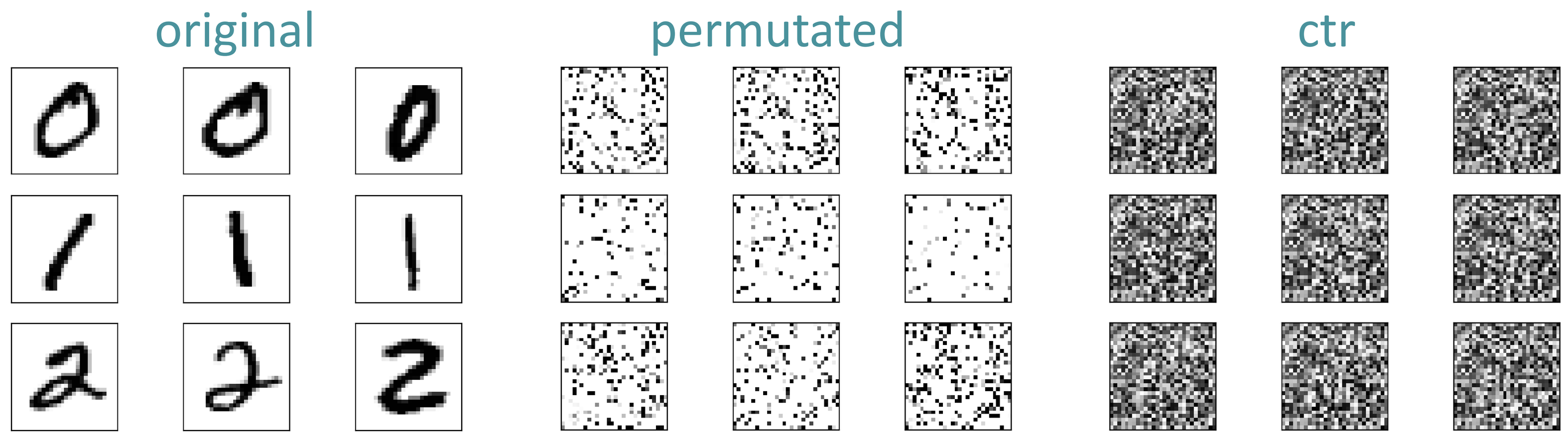
- Permutation
- AES in ECB, CBC and CTR modes



architecture for securing models

3 Cutting Loose Ends

We eliminated the models that did not learn well.
Learning encrypted images is not very intuitive, as can seen below.



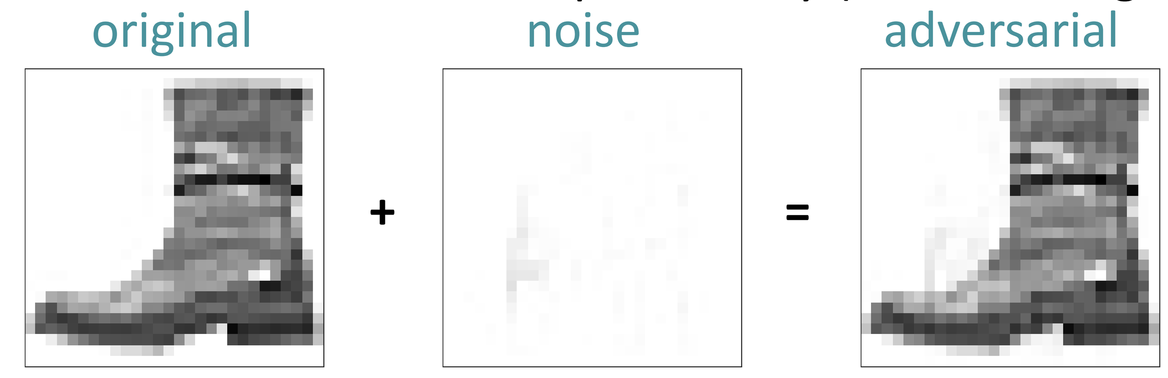
sample of the encrypted images.

4 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 (nor the weights).



visualization of a CW attack secured by AES · CTR

Results 5

We achieved the goal with models applying permutation and AES · CTR encryption techniques on the images. Full results below.

MNIST

Model	Images	Unsecured	Permutation	AES · ECB	AES · CBC	AES · CTR
A	Originals	1.50	3.70	18.40	67.60	3.70
	CW l_2	100.00	4.50			4.20
	CW l_0	100.00	7.30			9.60
	CW l_∞	100.00	5.40			4.90
B	Originals	2.10	4.20	19.30	87.40	2.70
	FGSM	39.50	8.60			4.90

Fashion-MNIST

Model	Images	Unsecured	Permutation	AES · ECB	AES · CBC	AES · CTR
A	Originals	8.30	12.30	54.60	71.50	17.40
	CW l_2	100.00	12.70			17.20
	CW l_0	100.00	12.50			18.70
	CW l_∞	100.00	12.90			17.80
B	Originals	9.50	12.00	55.30	90.30	16.70
	FGSM	77.20	29.80			26.50

classification error (%) on the first 1000 test samples

6 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.

	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

results for training permutated data, various image dimensions

Future Work

- Improve accuracy on AES-ECB model
- Nicholas Carlini ('C' in CW) believes that CW might still defeat these defenses
- Test on more complicated datasets; i.e. Cifar-10