IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Adversarial Images

AGUIAR Isabella
RIBEIRO Lucas

16/12/2022

AGENDA

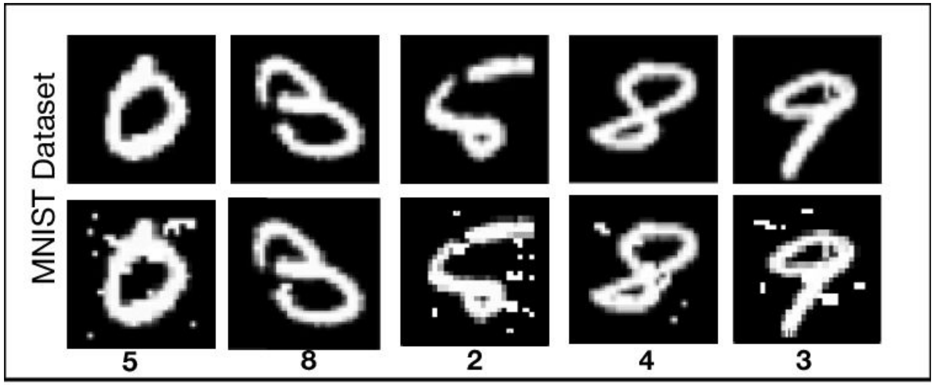
1. Context
2. Attack: FGSM
3. Defense: Distillation Method
4. Conclusion and Perspectives

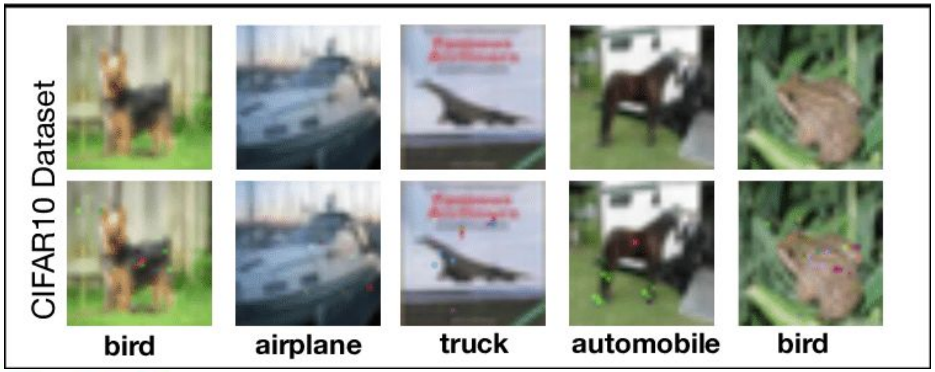
IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

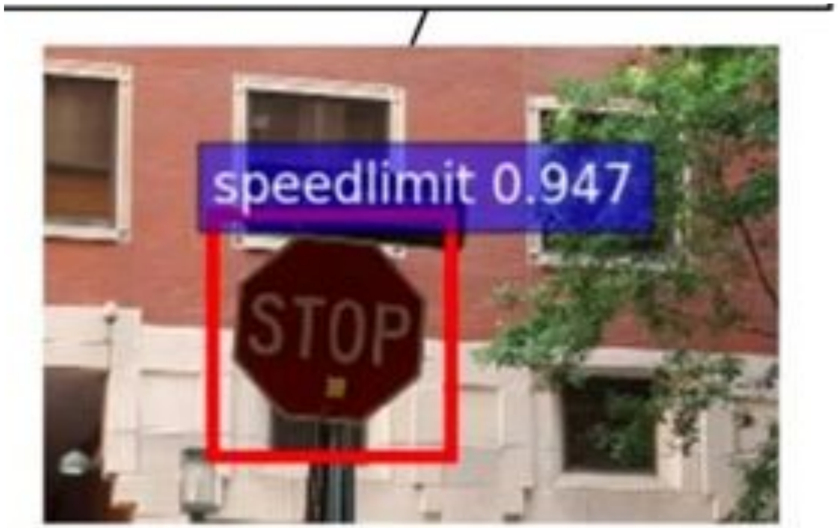
1. Context

What are Adversarial Images?

- A way to **attack ML systems**
- It causes **misclassification**
- The objective is to make the new image **indistinguishable to a human observer**
- Be aware of adversarial attacks is **crucial to ML robustness**
- The new image can be misclassified with **high levels** of confidence







1. Context

What are the characteristics of the attacks?

There are several kinds and characteristics of attacks:

- white-box vs black-box.
- General misclassification vs targeted misclassification
- Fast Gradient Sign Method (FGSM)
- Projected Gradient Descent (PGD)
- Carlini and Wagner (C&W) attack
- Adversarial patch attack

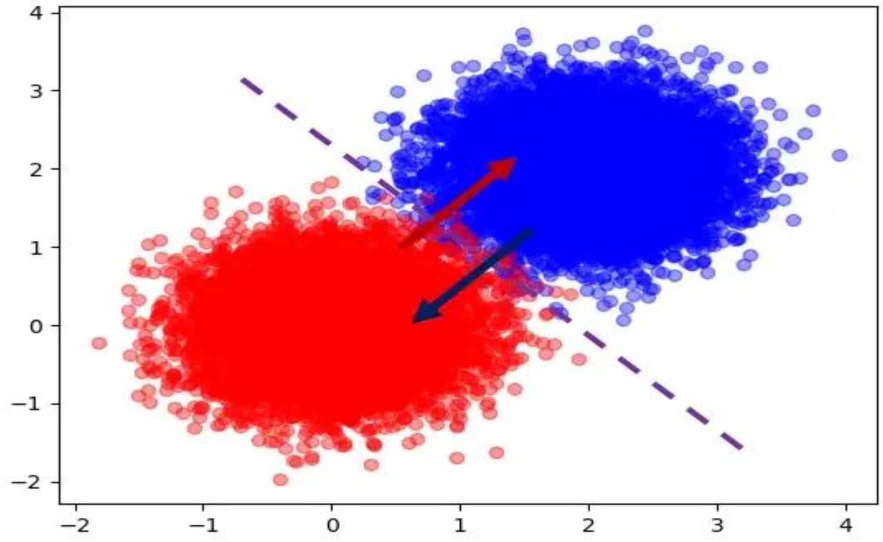
For this study, it was considered an attack:

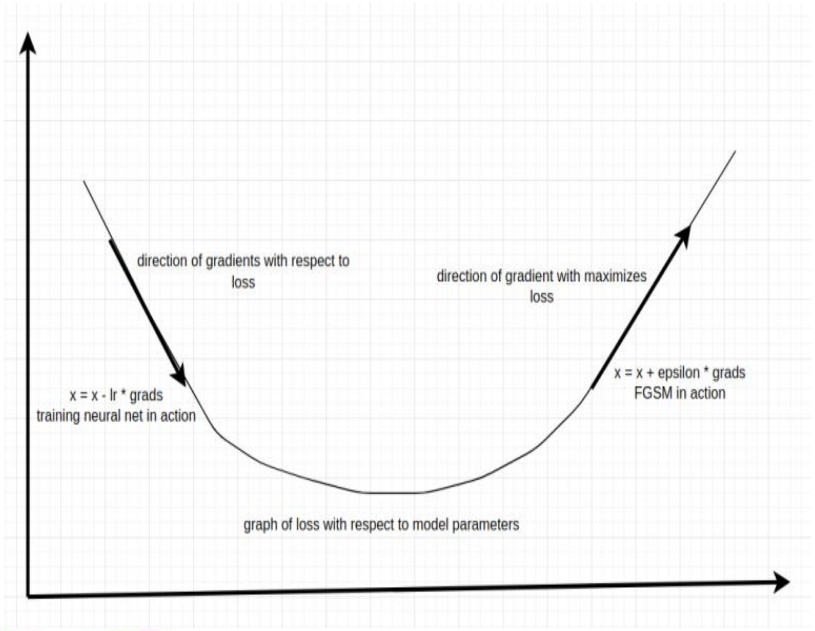
- **Fast Gradient Sign Method**
- **White-box**
- **Non-targeted misclassification**

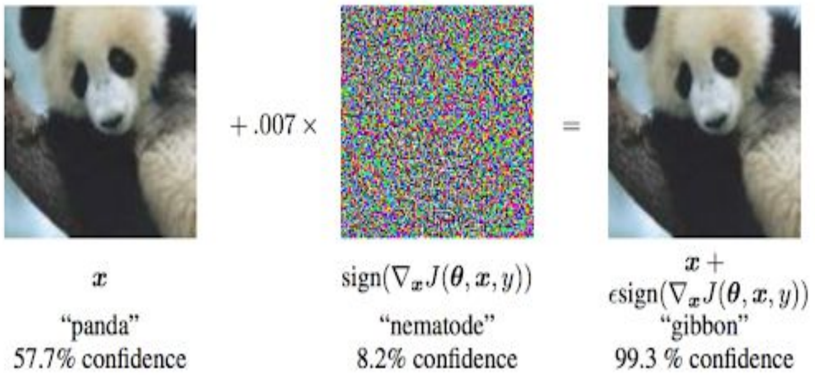
2. FGSM Attack

How it happens?

1. Calculate the loss after forward propagation,
2. Calculate the gradient with respect to the pixels of the image,
3. Nudge the pixels of the image ever so slightly in the direction of the calculated gradients that maximize the loss calculated above.

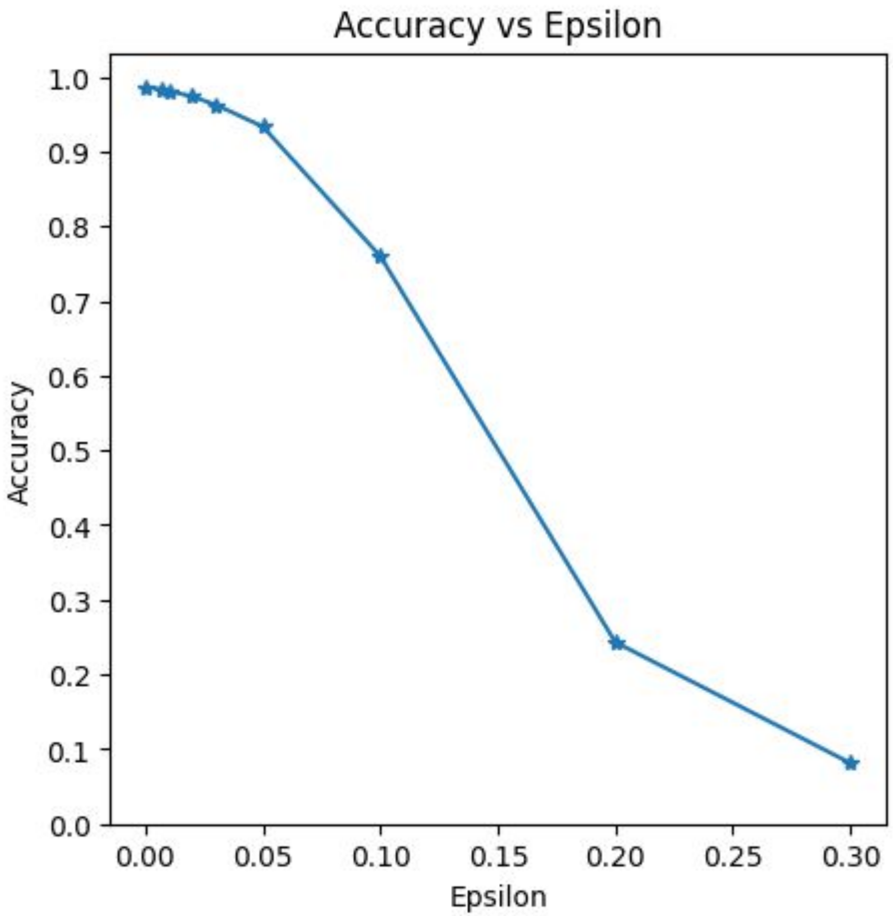


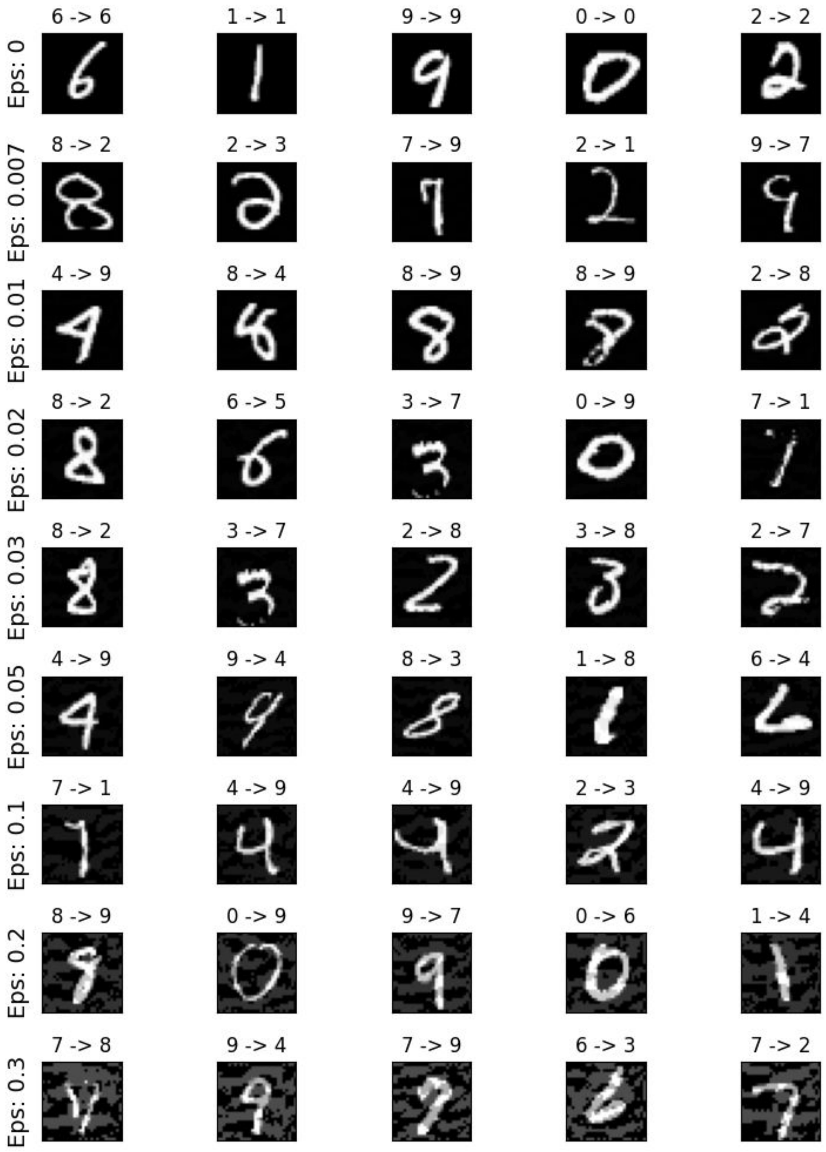

direction of gradients with respect to loss
direction of gradient with maximizes loss
x = x - lr * grads
training neural net in action
x = x + epsilon * grads
FGSM in action
graph of loss with respect to model parameters


+ .007 ×
=
x
"panda"
57.7% confidence
$\mathrm{sign}(\nabla_{x} J(\theta, x, y))$
"nematode"
8.2% confidence
$x + \epsilon\mathrm{sign}(\nabla_{x} J(\theta, x, y))$
"gibbon"
99.3 % confidence

2. FGSM Attack

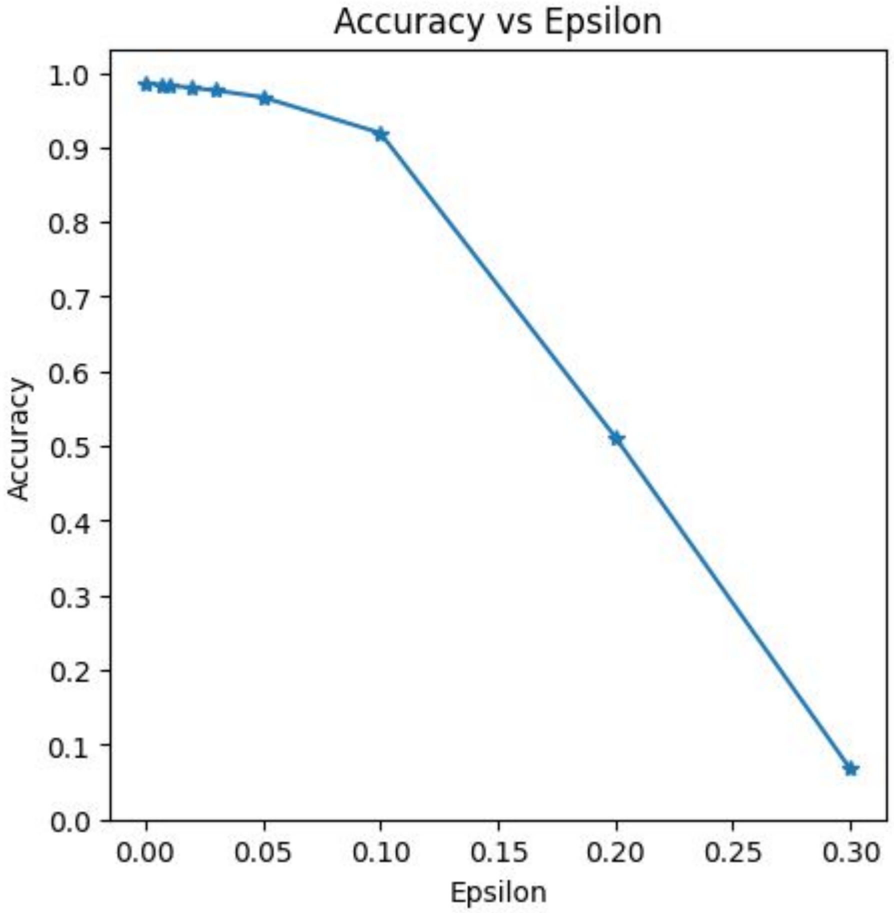
MNIST DATASET + LENET

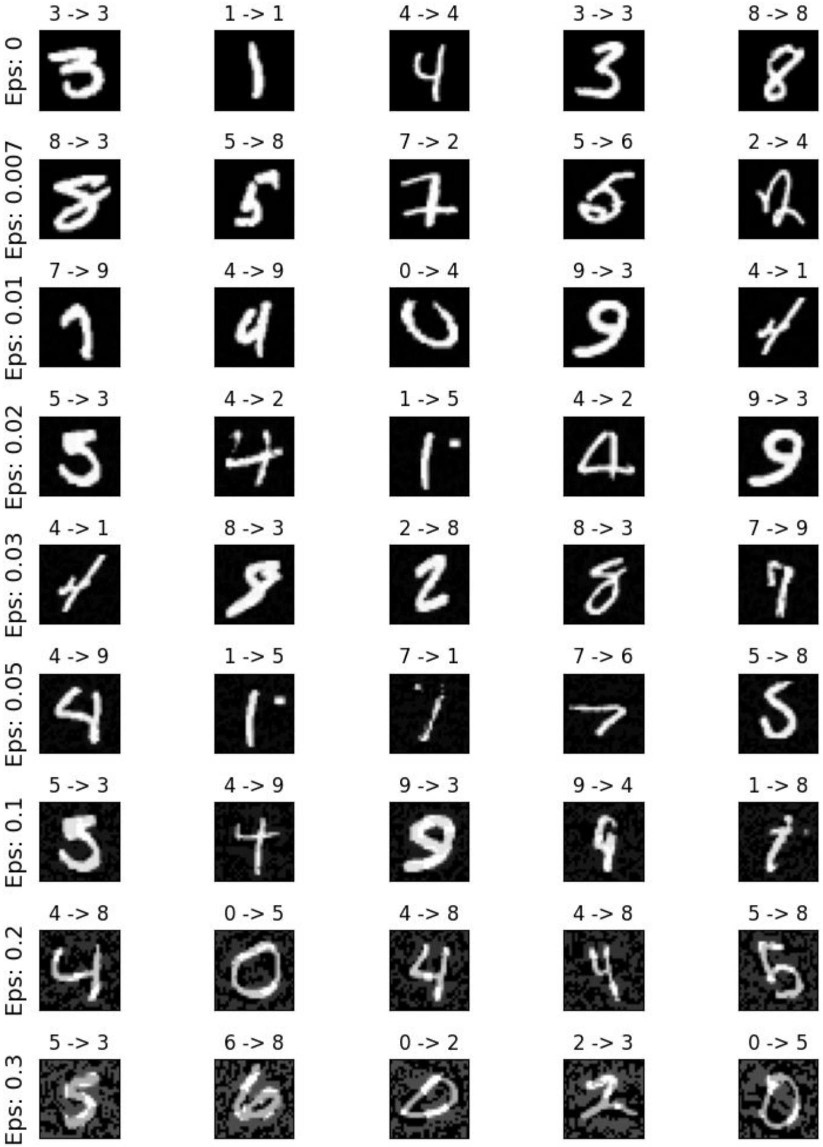




2. FGSM Attack

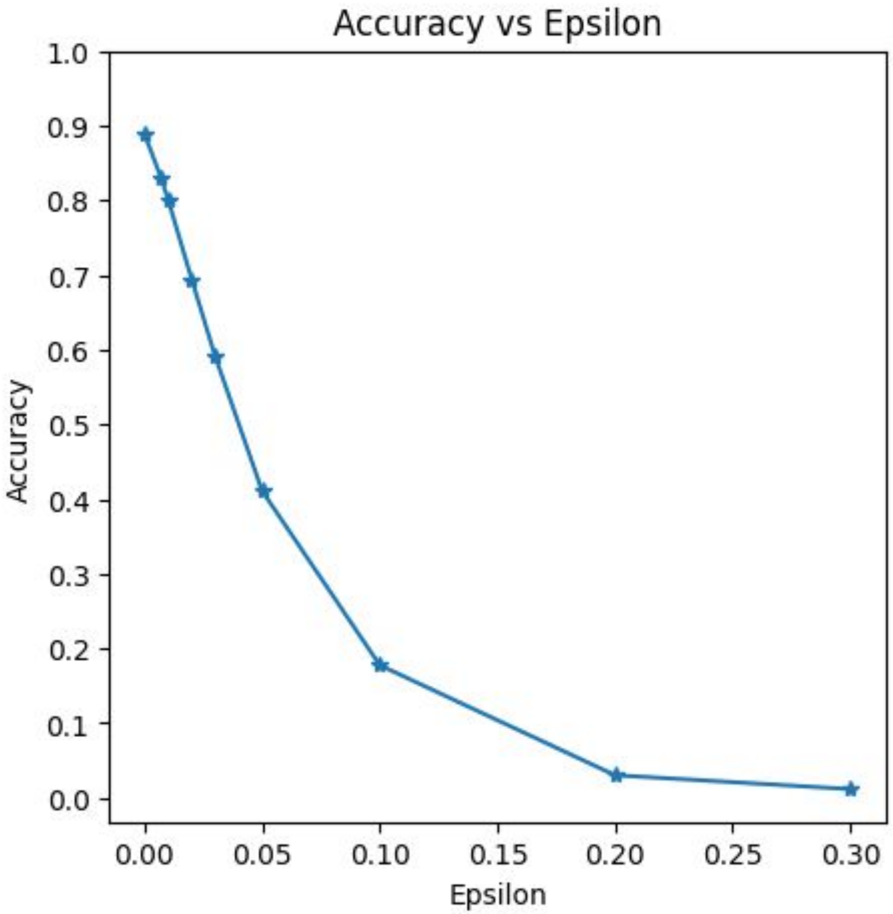
MNIST DATASET + ALEXNET

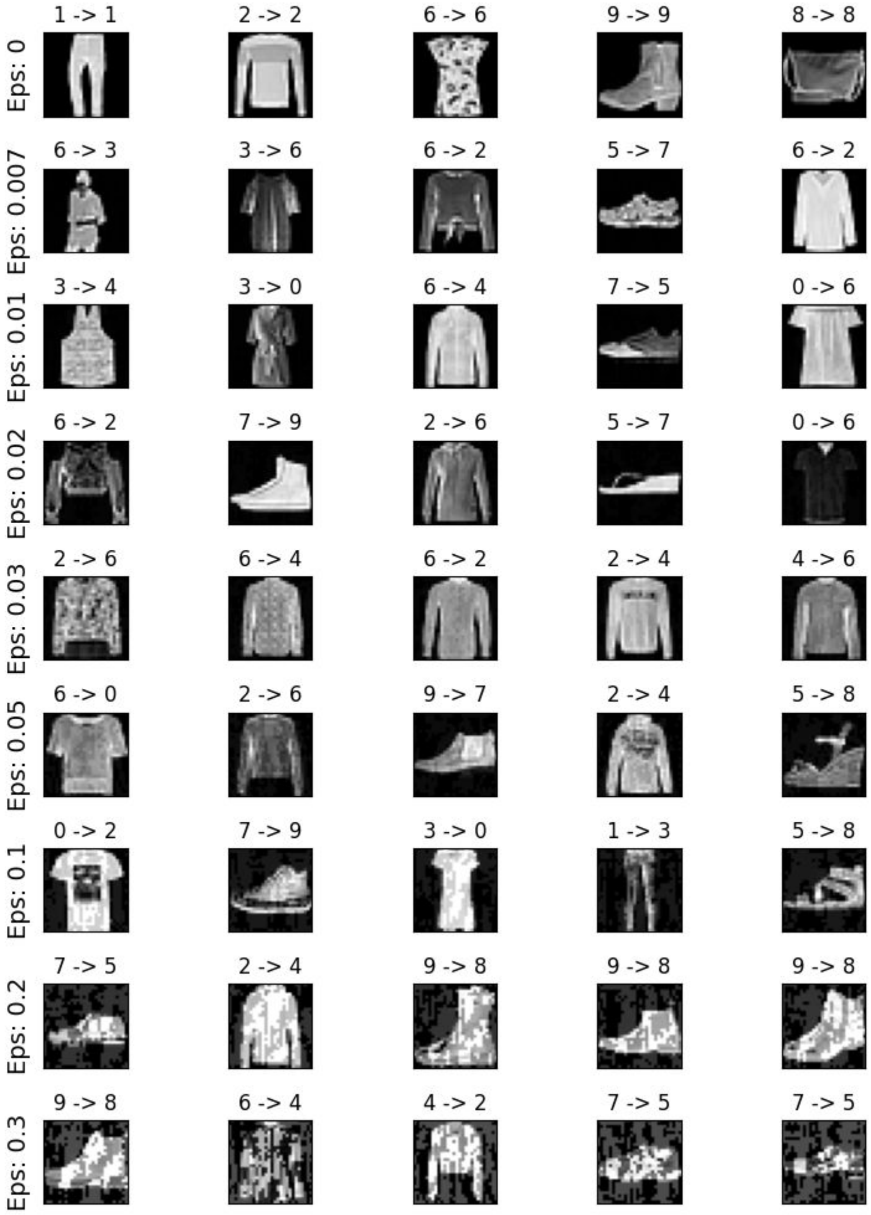




2. FGSM Attack

FASHIONMNIST DATASET + Lenet

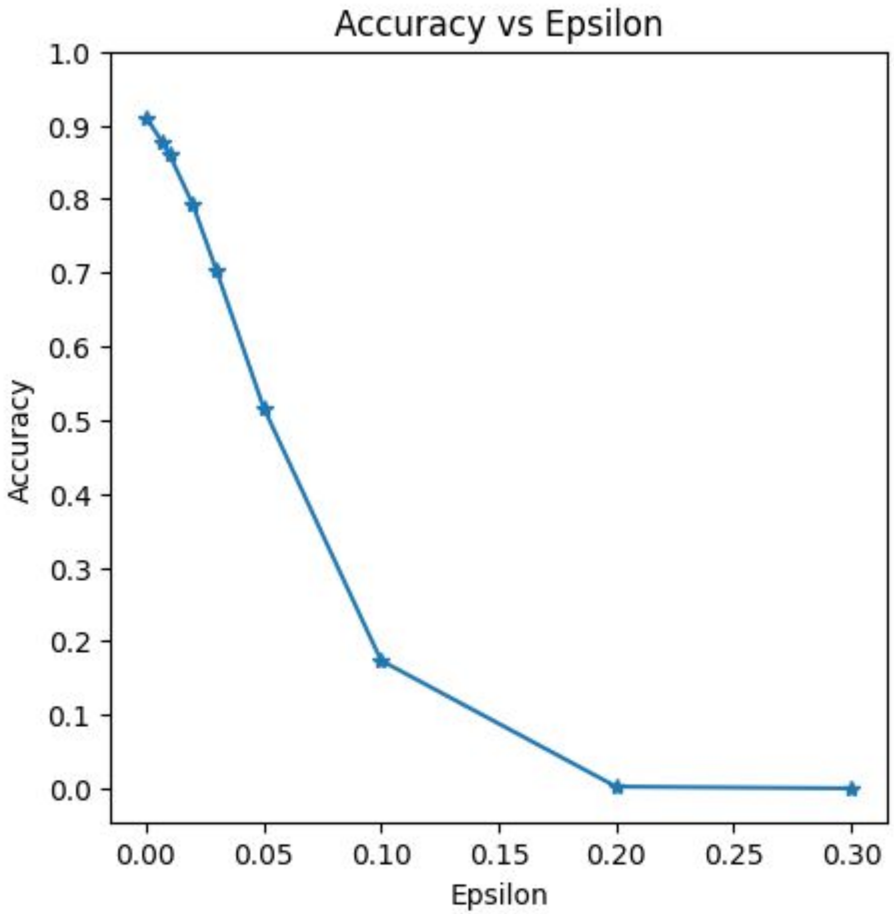


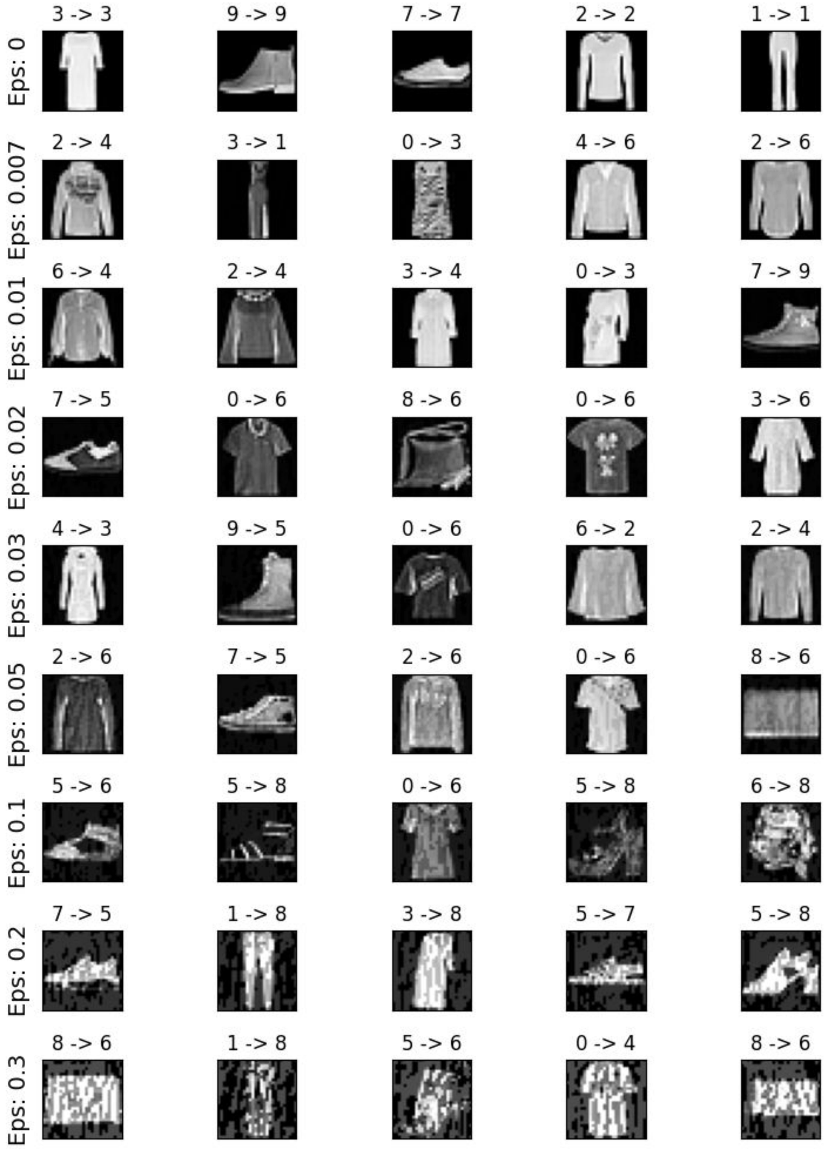


IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

2. FGSM Attack

FASHIONMNIST DATASET + ALEXNET



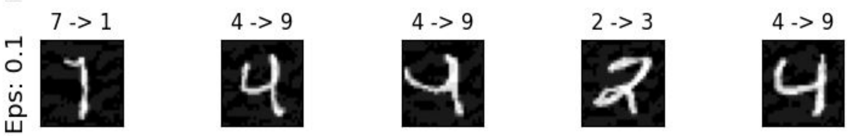


IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

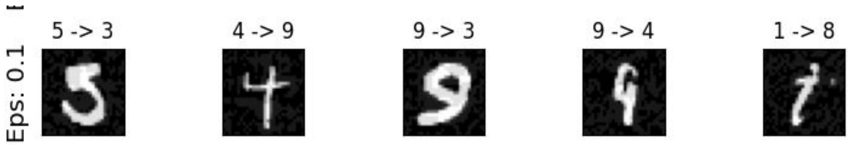
2. FGSM Attack

Analysis

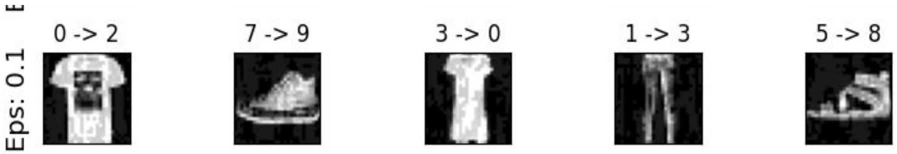
MNIST DATASET + LENET



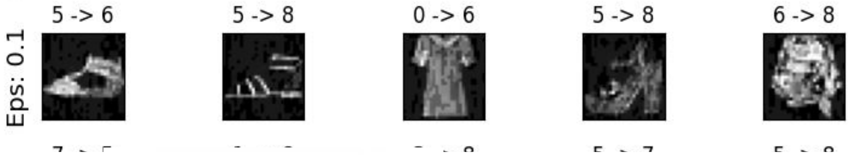
MNIST DATASET + ALEXNET



FASHIONMNIST DATASET + LENET



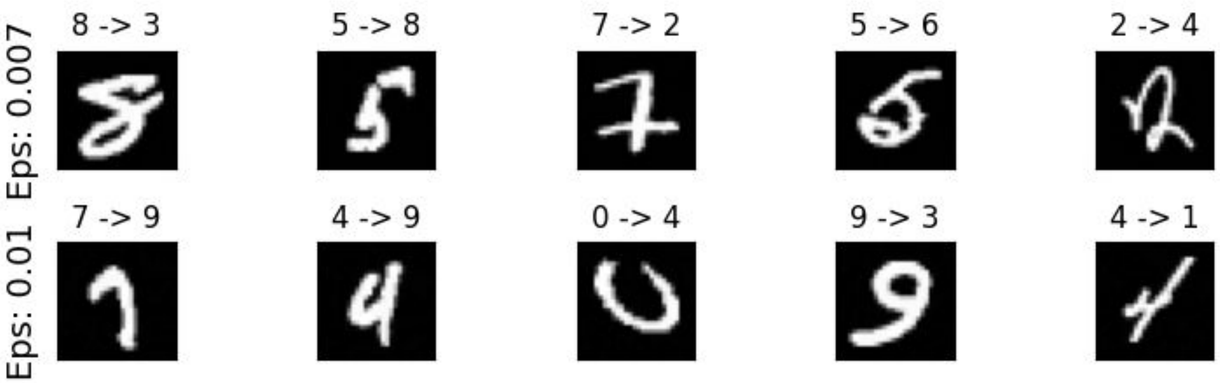
FASHIONMNIST DATASET + Alexnet

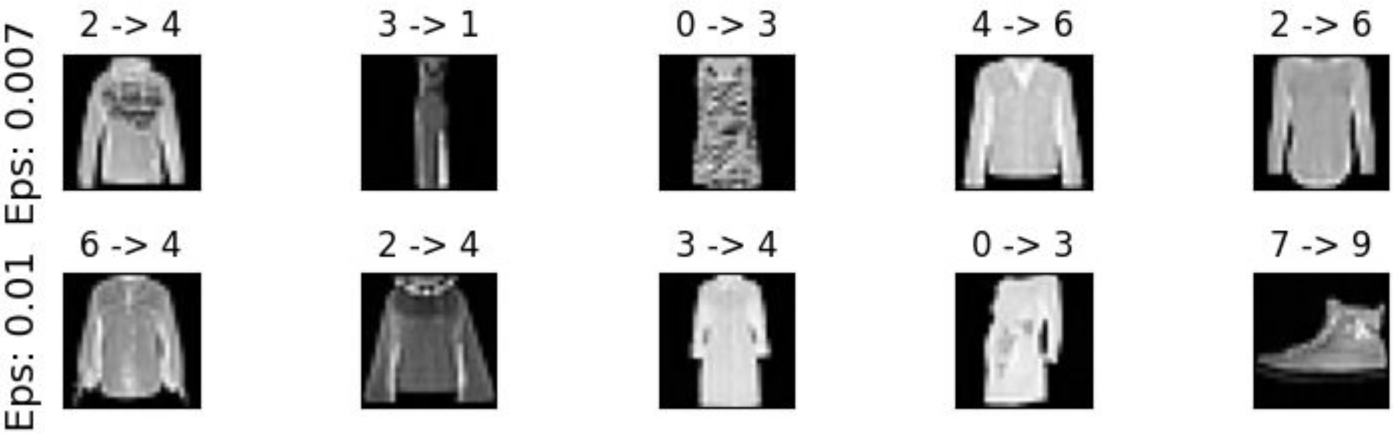


IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

3. Distillation Defense

- Objective: Decrease the effectiveness of adversarial samples and improve the robustness of DNNs;
- Motivation: reduce the computational complexity;
- Original proposition: knowledge transfer between one DNN and another;
- “New” poposition: knowledge extracted from a DNN to improve its own resilience to adversarial samples





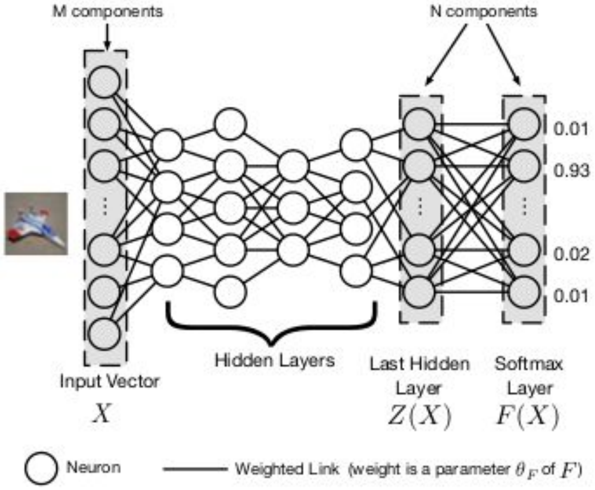
3. Distillation Defense

- Low impact on the architecture
- Maintain accuracy
- Maintain speed of network
- Reduce the sensitivity of a DNN
- Helps the model generalize better to samples

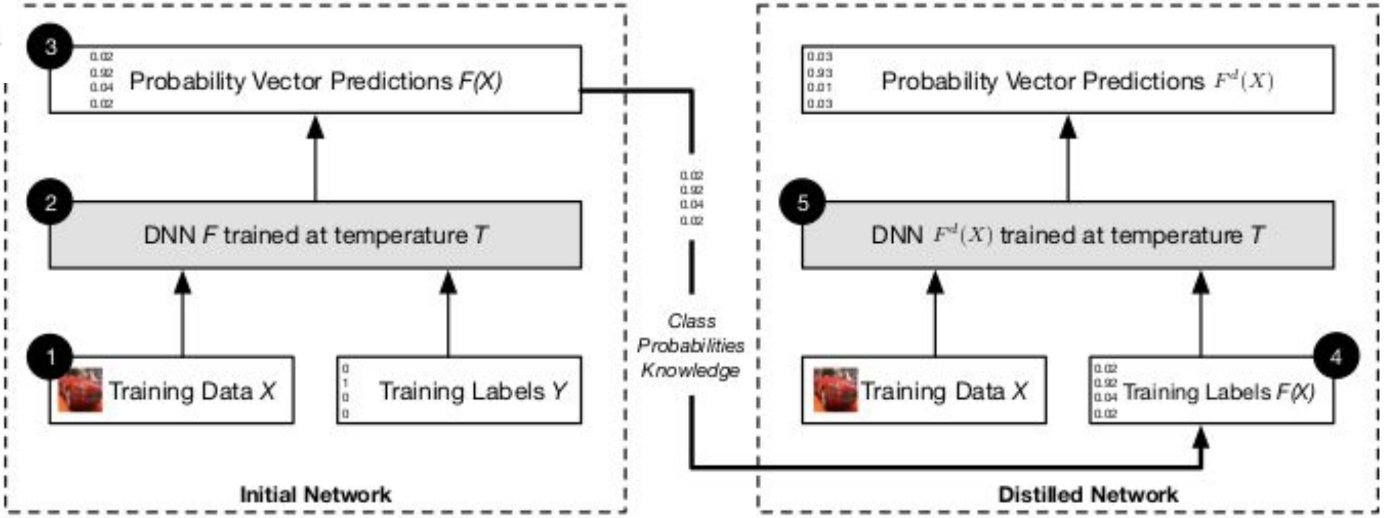
3. Distillation Defense

How it works?

- Training a DNN with SoftMax as the last layer
- Using the probability vector produced by SoftMax
- Transfer the knowledge to another DNN



$$F(X) = \left[\frac{e^{z_i(X)/T}}{\sum_{l=0}^{N-1} e^{z_l(X)/T}}\right]_{i\in 0..N-1}$$



3 Probability Vector Predictions $F(X)$ (0.02, 0.92, 0.04, 0.02)

2 DNN F trained at temperature T

1 Training Data X

Training Labels Y (0, 1, 0, 0)

Initial Network

Class Probabilities Knowledge (0.02, 0.92, 0.04, 0.02)

Probability Vector Predictions $F^d(X)$ (0.03, 0.93, 0.01, 0.03)

5 DNN $F^d(X)$ trained at temperature T

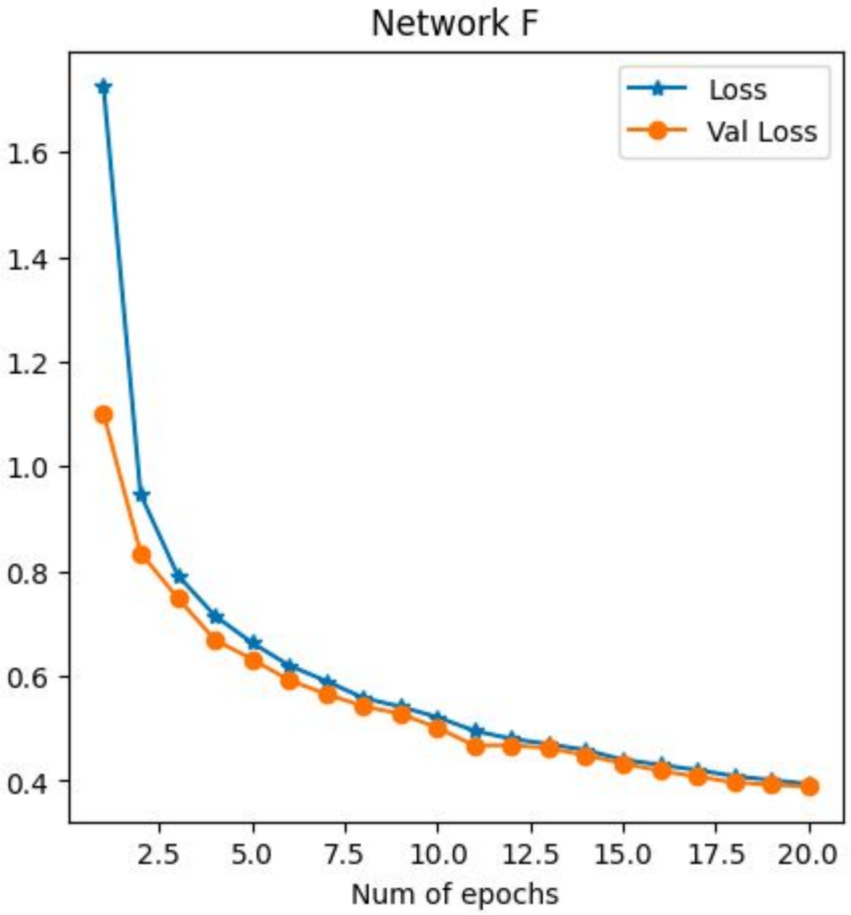
Training Data X

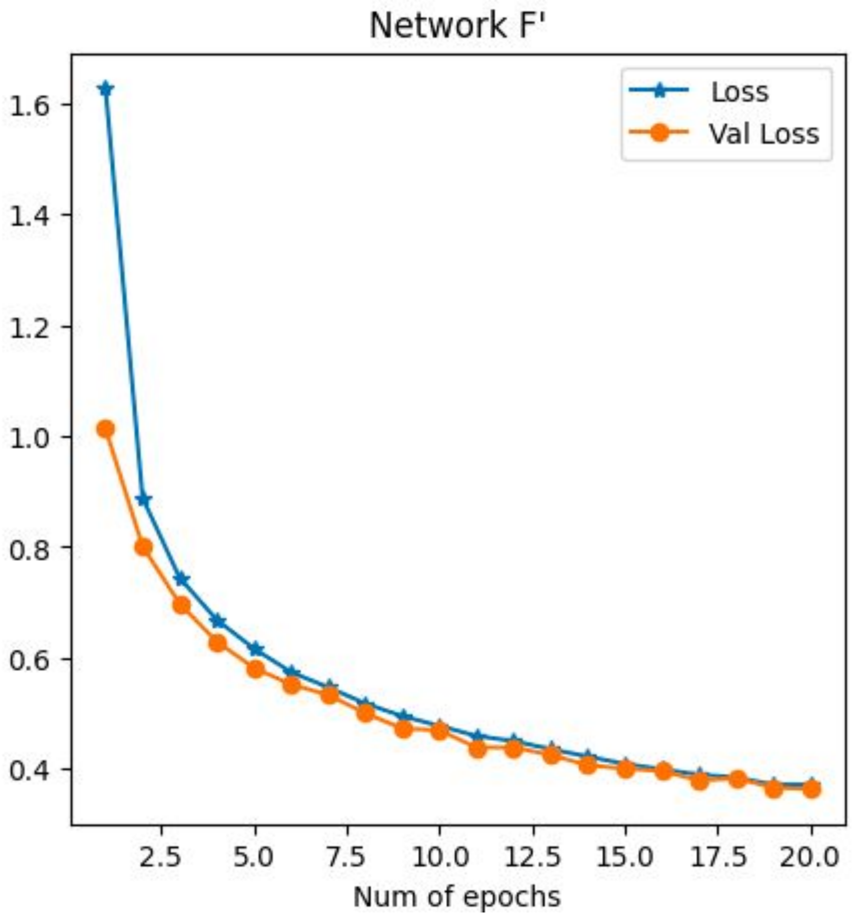
4 Training Labels $F(X)$ (0.02, 0.92, 0.04, 0.02)

Distilled Network

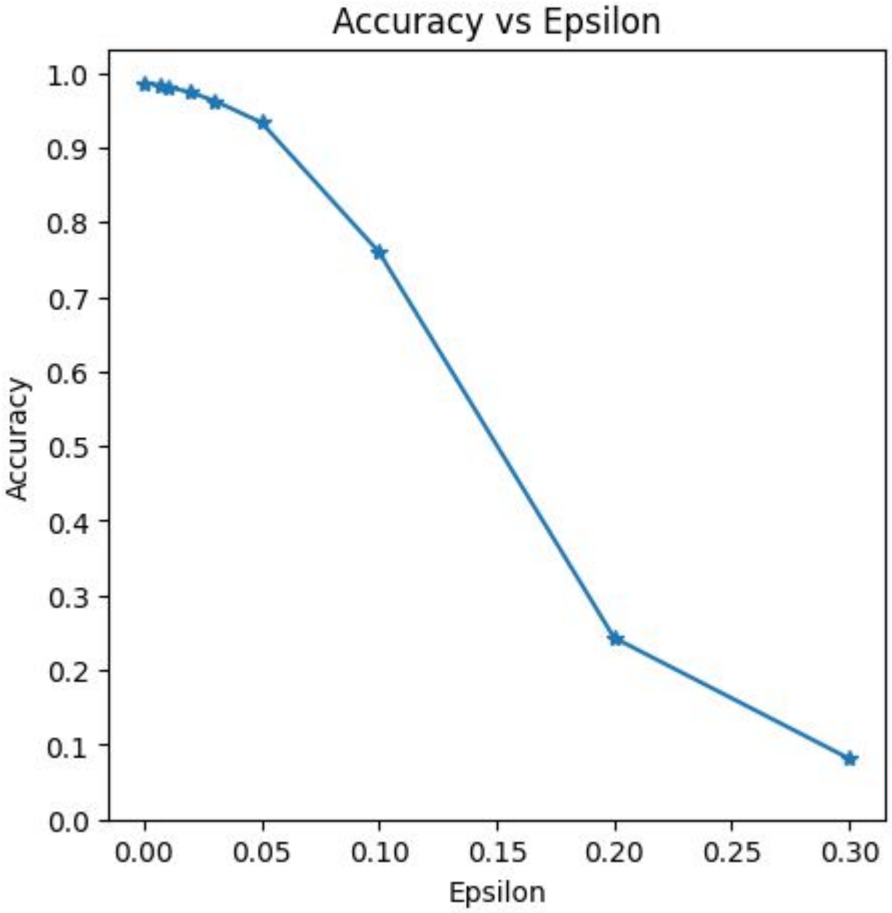
IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

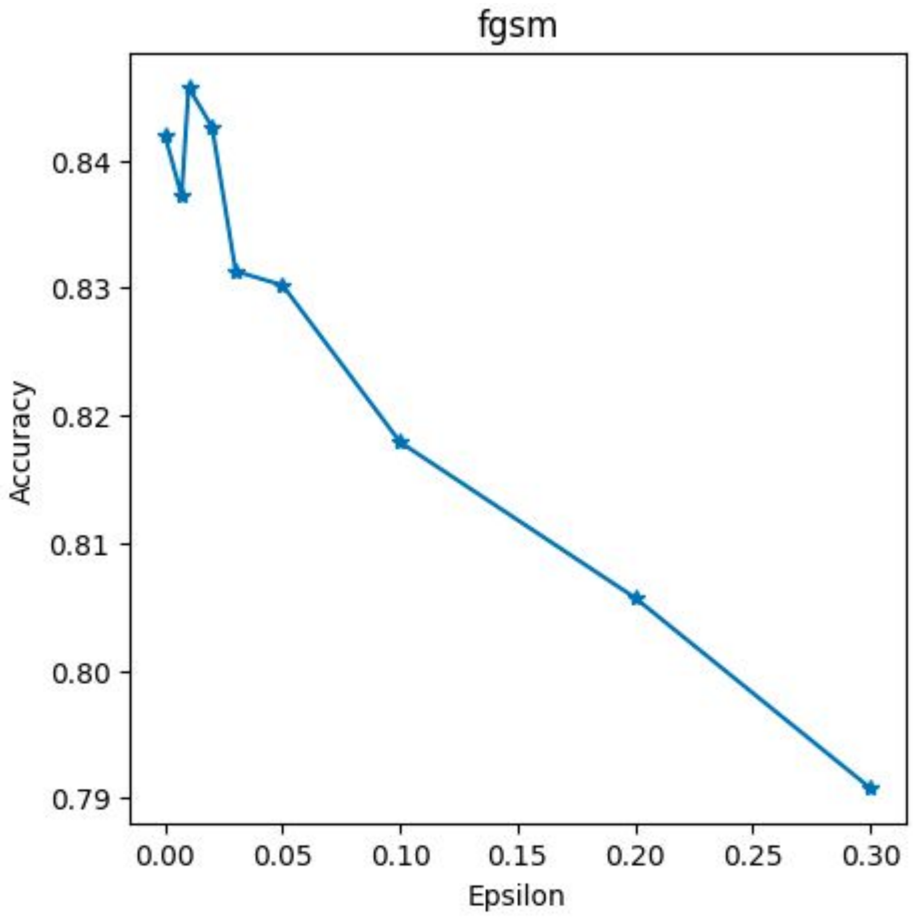
3. Distillation Defense





3. Distillation Defense





4. Conclusion and perspectives

- This presented only a basic example of adversarial attack, there might be others (more complex);
- Adversarial attacks casts doubt on their usability in the real-world environment, especially in safety-critical systems;
- Models trained with defensive distillation are less sensitive to adversarial samples;
- The work on defense also leads into the idea of making machine learning models more robust in general;
- Future works:

- Perform a defense on FashionMNIST to compare the results with MNIST;
- Apply the attacks on datasets with colored (STL-10 and Stanford Cars) images in the architecture AlexNet;
- Perform a defense on a colorful dataset.

References

I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in Proceedings of the 2015 International Conference on Learning Representations. Computational and Biological Learning Society, 2015.

N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks," 2016 IEEE Symposium on Security and Privacy (SP), 2016, pp. 582-597, doi: 10.1109/SP.2016.41.

https://pytorch.org/tutorials/beginner/fgsm_tutorial.html

Thank You For Your Attention

Questions?

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom