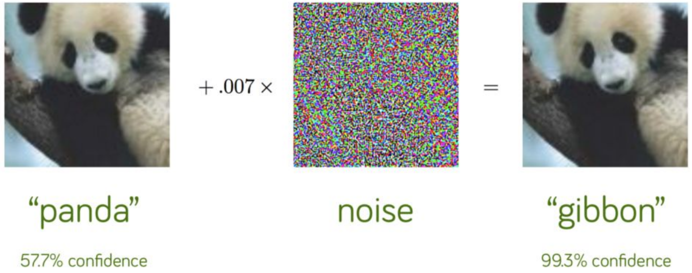
Training robust neural network



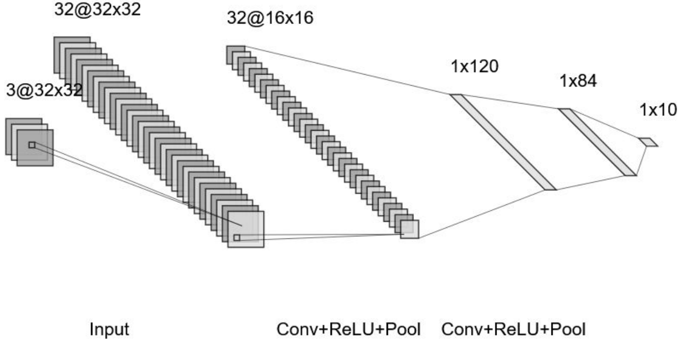
ama_adv Team : Adrien Golebiewski , Alexis Hummel, Maximilien Wemaere

Dauphine | PSL
UNIVERSITÉ PARIS

I. Context

Problem: Help a little CNN to defend against adversarial attacks

The simple CNN:



The attacks:

FGSM: $x_{attacked} = x + \epsilon \, \mathrm{sign}(\nabla_x l_f(x, y))$

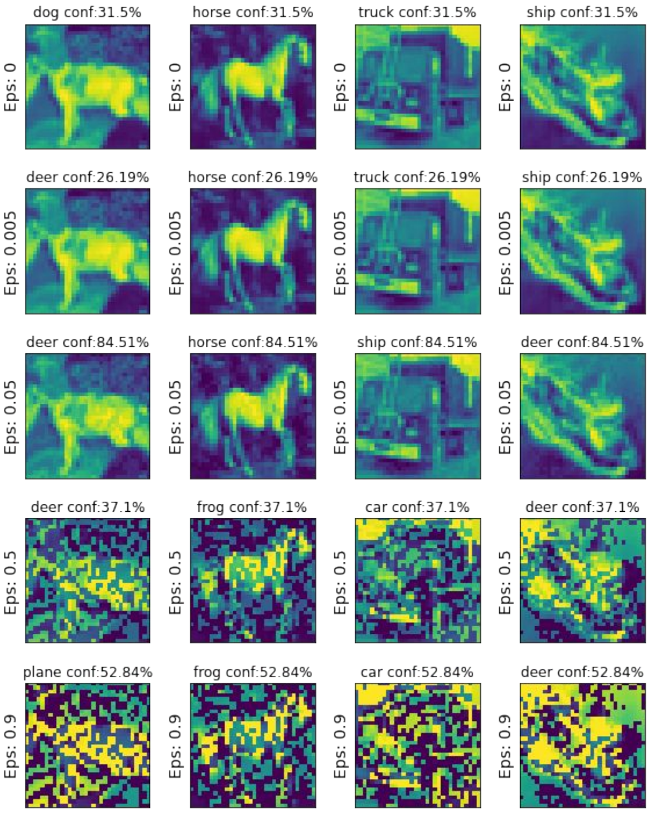
PGD: $x_0 = x$

$$x_{t+1} = \Pi_{B(x_0,\epsilon)}(x_t + \epsilon \, \mathrm{sign}(\nabla_x l_f(x, y)))$$

I. Attacks

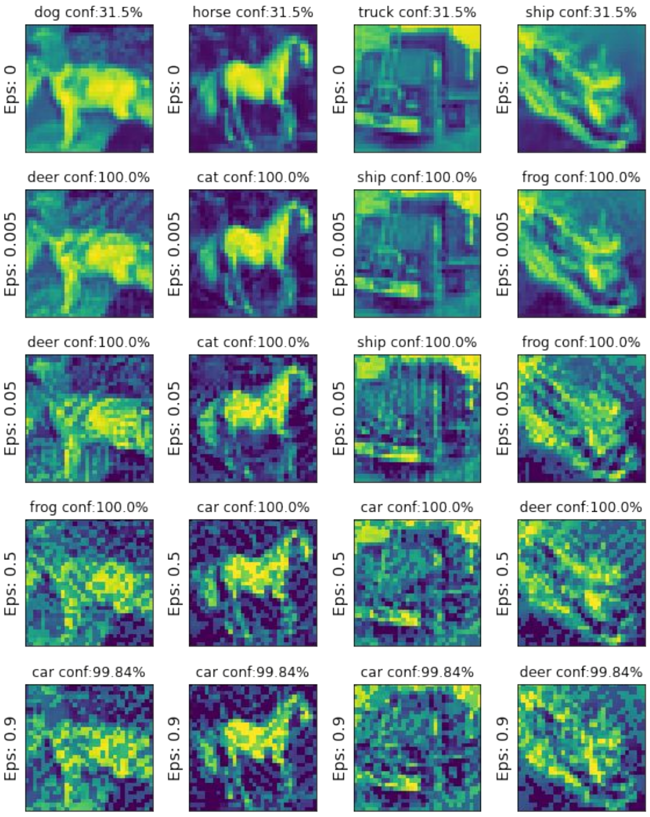
FGSM

Attack

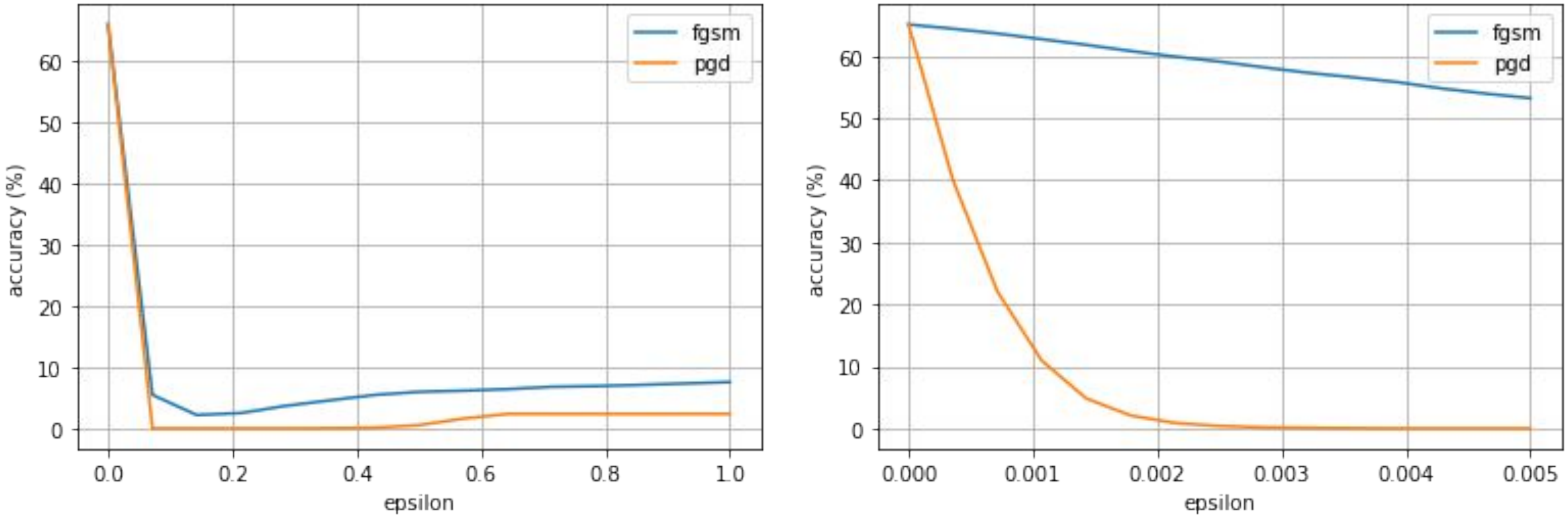


PGD

Attack



I. Attacks


fgsm
pgd
accuracy (%)
60
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0
epsilon
fgsm
pgd
accuracy (%)
60
50
40
30
20
10
0
0.000
0.001
0.002
0.003
0.004
0.005
epsilon

II. Classic defenses - Adversarial Training

Adversarial Training : include the adversarial examples during the model training

$$\min_{\delta} \mathbb{E}_{(x,y)}[\max_{||\delta||\leq\epsilon} \ell_{f_\theta}(x+\delta, y)]]$$

Simple approach but does not generalize for all adversarial examples.

II. Classic defenses - Adversarial Training Results

Adversarial Training

FGSM	PGD-Linf	PGD_L2
28%	/	/

III. Innovatives strategies

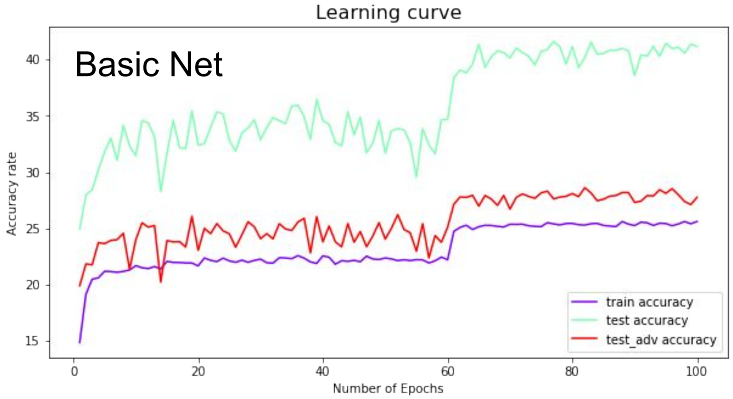
Introducing randomness : **Noise injection** (in progress)

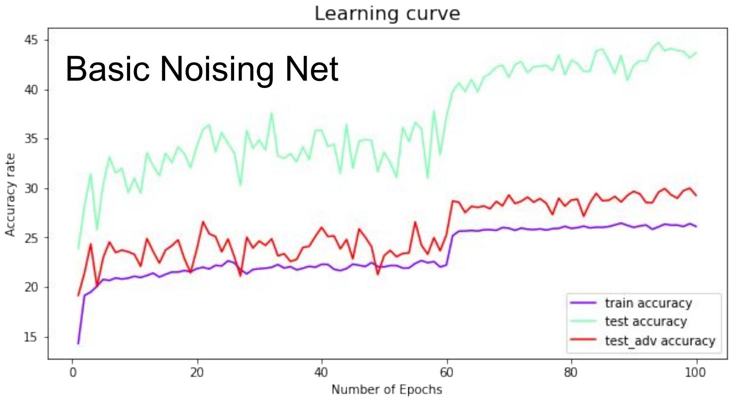
- Adversarial test accuracies of a Basic Net **without and with noise in the input**.

	Basic Net	Basic Net with Noise	Wide Net	Wide Net with Noise
Norm infinity PGD Attack	21,39 %	29,26 %	/	/
Norm 2 PGD Attack	/	/	/	/

Comparative with Wide Net Noise (Resnet) to do soon

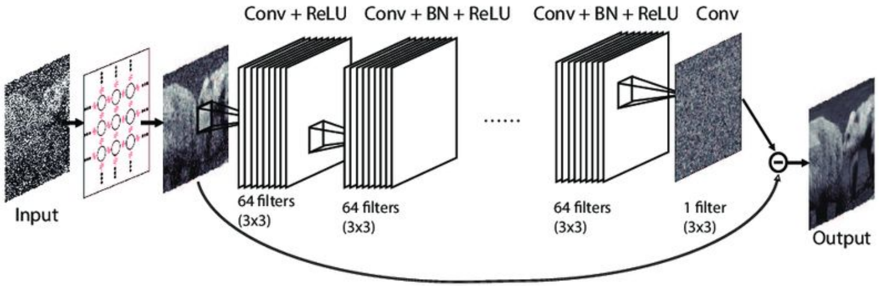
-





III. Innovatives strategies

Denoise the images with a DnCNN: gaussian denoiser



1- Train the DnCNN with adversarial images as input and clean images as output

2- During the test, before entering the image in our model, we clean them with the DnCNN

Accuracy	No Denoiser	Denoiser 10 hidden layers 5 epochs	Denoiser 10 hidden layers 10 epochs
FGSM Attack eps=0.05	9.96%	16.66%	17,27%

To do next time :

- We have independently tested several **defense techniques** and **several attack techniques** based on different neural architectures as well.

- Testing others defenses :
  - Adversarial training on Noising Resnet
  - Mixed Adversarial Training with both L2 **and** L-infinity
  - Randomized network
  - …

- The objective would be to be able to group and compare all the results in order to find the most effective attack/defense strategy !

References

- K. Zhang, W. Zuo, Y. Chen, D. Meng and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," in *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142-3155, July 2017, doi: 10.1109/TIP.2017.2662206.

- Araujo, Alexandre et al. (2020). “Advocating for Multiple Defense Strategies against Adversarial Examples”. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases.Springer, pp. 165–177