Training robust neural networks

BENMANSOUR Adnan - BOUSKILA Laurène - CASTRO ROS Alejandro

Problem Definition

Objective:

- Add small **perturbations** to fool the classification system (**attack**).
- Implement strategies to avoid these attacks (**defense).**

Dataset: CIFAR10

- 60,000 images (50K-10K train-test split) belonging to 10 different classes.

Base model:

- Add Dropout and used SGD with learning-rate scheduler to optimize classification.
- Data Normalization and Augmentation (horizontal flip and random crops) on training data.
- Obtained accuracy of 87.75% over test data.

Attacks: FGSM v/s PGD

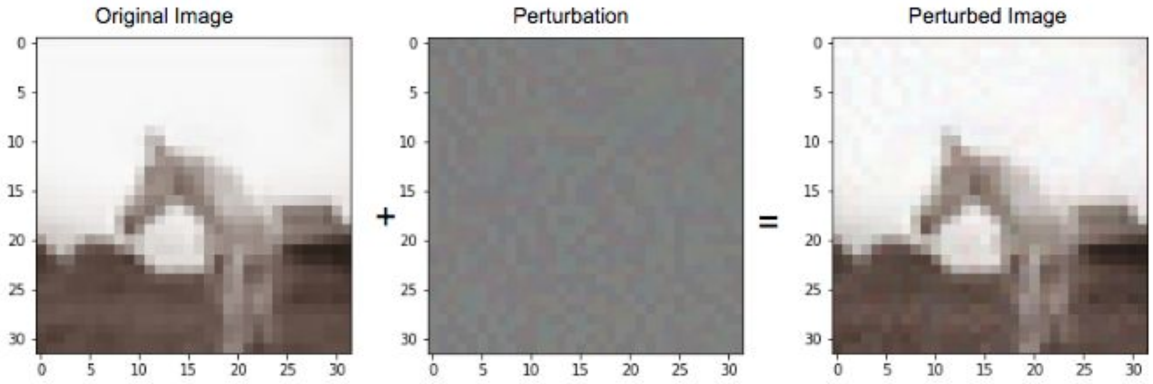
- Perturbations on the **gradient direction.**
- Constrained amplitude to make **imperceptible perturbations**.
- PGD is an **iterative version** of the FGSM.

FGSM

$$\overbrace{x_{adv} = x + \epsilon \,\mathrm{sign}\left(\nabla_x L_\theta\left(x, y\right)\right)}^{\textbf{FGSM}}$$

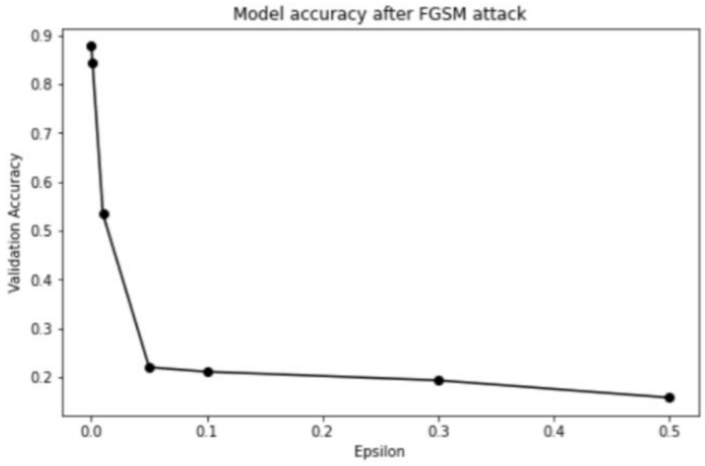
PGD

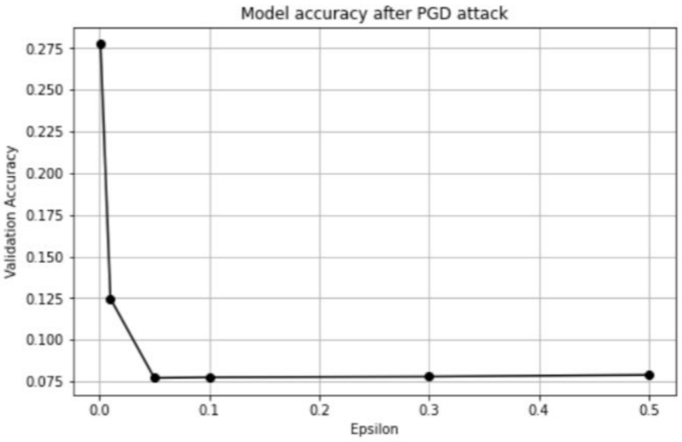
$$\begin{cases} x_0 = x \\ x_{t+1} = \prod_{\mathbf{B}(0,\epsilon)} \left(x_t + \eta \,\mathrm{sign}\left(\nabla_x L_\theta\left(x, y\right)\right)\right) \end{cases}$$



horse (48,65%) + ***ship*** (36,46%) = ***cat*** (99,99%)

Attacks: FGSM v/s PGD





Adversarial Training Defense

- Defense system that aims at **improving neural network robustness** against adversarial attacks by **training it with adversarial examples**.

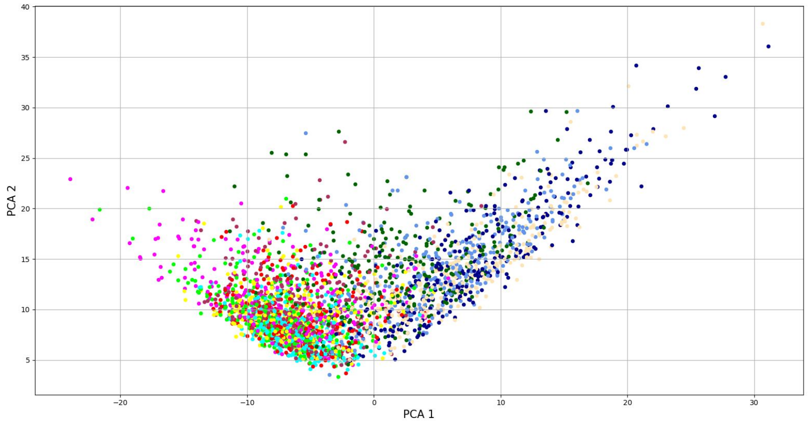
$$\min_{\theta} \mathbb{E}_{(x,y)} \left(\max_{\|\tau\| \leq \epsilon} L_{\theta}(x+\tau, y) \right)$$

Attack	Defense	$\varepsilon = 0.01$	$\varepsilon = 0.05$
FGSM	Without Defense	47.95%	7.31%
	Adversarial Training (vs FGSM)	68.53%	**43.15%**
	Adversarial Training (vs PGD)	**68.74%**	16.92%
PGD	Without Defense	57.99%	0%
	Adversarial Training (vs FGSM)	88.32%	3.55%
	Adversarial Training (vs PGD)	**91.28%**	**13.33%**

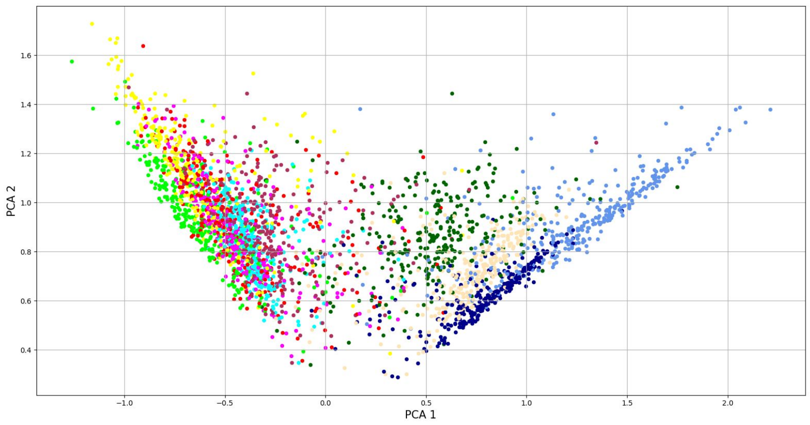
Defense based on contrastive loss

- **Possible defense**: create **clusters** in the feature space, so that a small perturbation won't lead to misclassification.
- This is enabled by the use of **contrastive loss.**
- $CL(1,2) = 1_{z_1=z_2} \cdot d(y_1, y_2)^2 + 1_{z_1 \neq z_2} \cdot max(0, \alpha - d(y_1, y_2))^2$

PCA without defense



PCA with contrastive loss applied



Dauphine | PSL
UNIVERSITÉ PARIS

Conclusions

- **PGD attack is more powerful than FGSM attack** since it computes a new perturbation at each iteration to fool the misclassification.

- **Adversarial defense** improves the robustness of the model, but it’s still prone to fail.

- **Defense** system built on the **contrastive loss** principle:
    - lower accuracy than adversarial training
    - need further work on this defense system

Merci pour votre attention