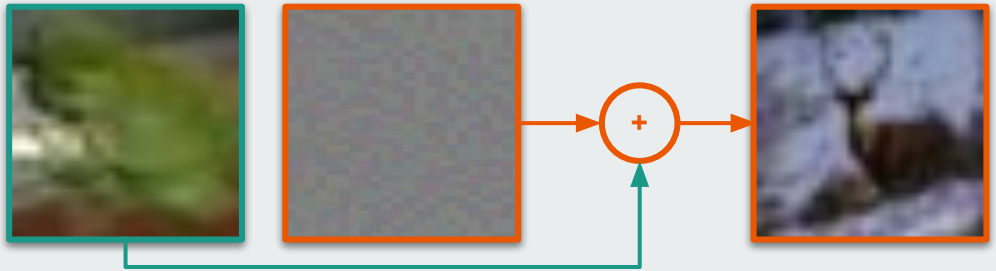
UNIVERSITÀ DI PISA

Assignment N. 2:

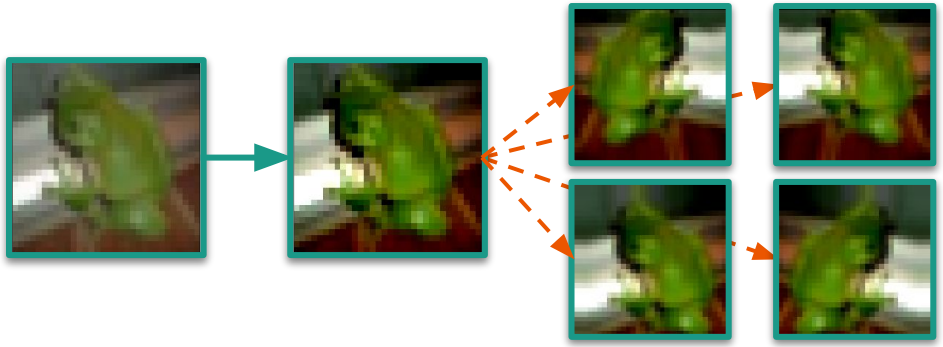
Adversarial Attacks on Image Classification of CIFAR-10 by a CNN

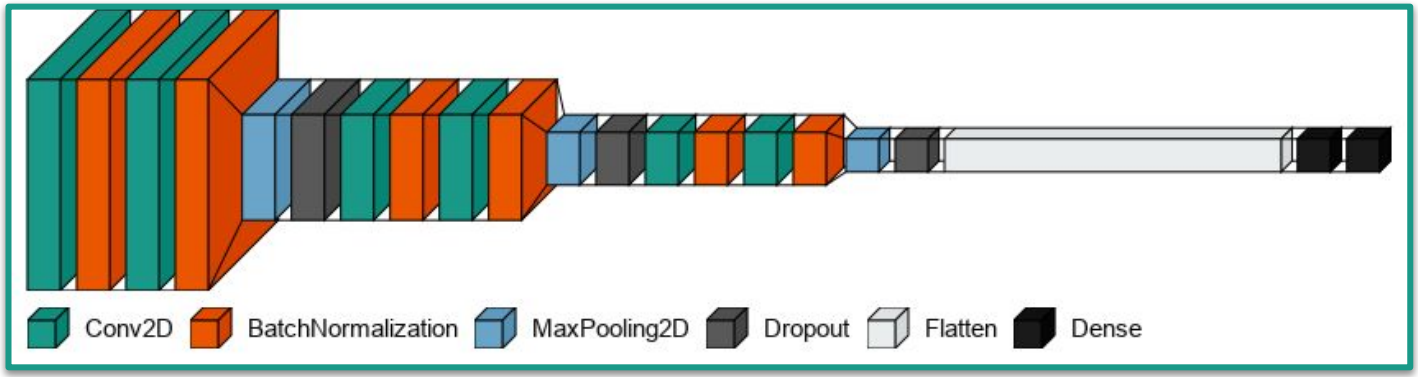


ISPR Course A.Y. 2020/2021
Marco Petix

CNN and Image Classification on CIFAR-10

- Built with Keras
- Image Pre-processing
- Data Augmentation
- 86% Accuracy on the Test set

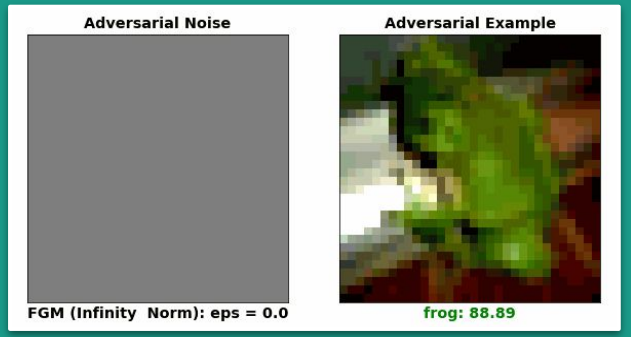


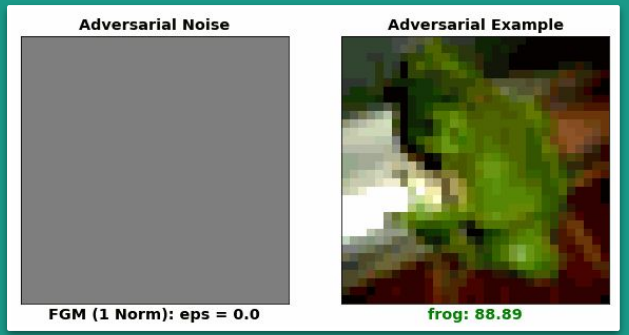


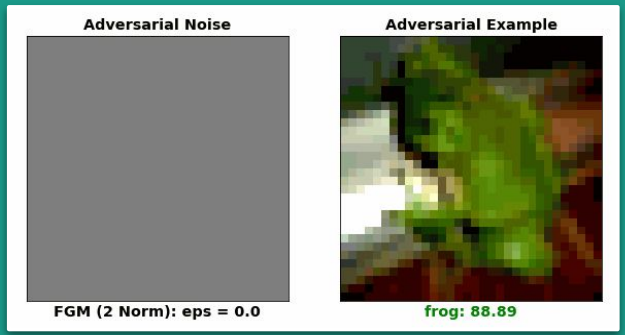
Fast Gradient (Sign) Method Attack

Frog: 88.89 %

FGM implemented via Cleverhans

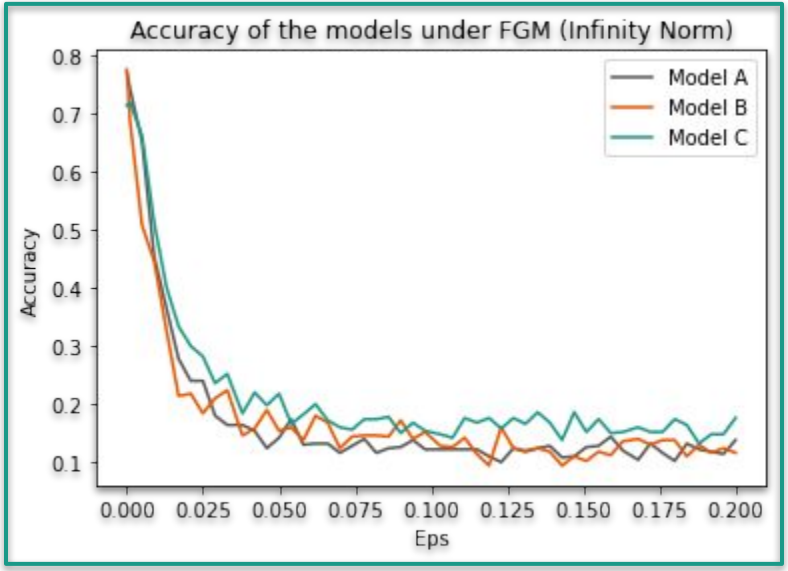






Adversarial Training

- Model A
  - Trained on a **clean Training set**
- **Model B**
  - Trained on a clean Training set augmented by **adversarial samples for fixed ε = 0.05**
- **Model C**
  - Trained on a clean Training set augmented by **adversarial samples for random ε ∈ (0, 0.2]**
- **Selective Resistance** to Adversarial Attacks
- **Adversarial Samples** as **Features** and **not Bugs**



	Test set	*eps* = 0.05	Average
Model A	0.76	0.14	0.18
Model B	0.74	0.16	0.18
Model C	0.72	0.22	0.22

Iterative Attacks and Final Considerations

On the Classification task

- Improving the model accuracy via Transfer Learning

On performing Adversarial Attacks

- From One-step to Multi-step attacks
  - Projected Gradient Descent
  - Carlini Wagner L2

On achieving Robustness

- Adversarial Samples Generator
- Robustness-Accuracy Tradeoff
- Capacity, Robustness and Transferability relationship