Data Science Lab
Assignment 3 : Training robust neural networks

Methods : FGSM, PGD, Randomized Networks, DeepFool

Lyna Bouikni and Arthur MUSSARD

University of Paris Dauphine

December 2023

Study Goals

- **Objective 1 :** Develop a baseline classifier for CIFAR-10 dataset.
  - Employed and fine-tuned a CNN architecture with convolutional, max pooling, and fully connected layers.
- **Objective 2 :** Implement and analyze adversarial attack mechanisms.
  - Explored the effects of FGSM and PGD attacks on model performance.
- **Objective 3 :** Enhance model resilience through Adversarial Training.
  - Integrate adversarial examples into the training process to strengthen defense mechanisms.

Study Goals

- **Objective 4 :** Innovation in defense and attack strategies.
  - Investigate the effectiveness of randomized networks.
  - Evaluate the robustness against multiple adversarial attacks generated using DeepFool.

Baseline Model

Model Architecture and Performance

Model A : Conv+MaxPool+Conv+MaxPool+Flatten+FC+FC+FC

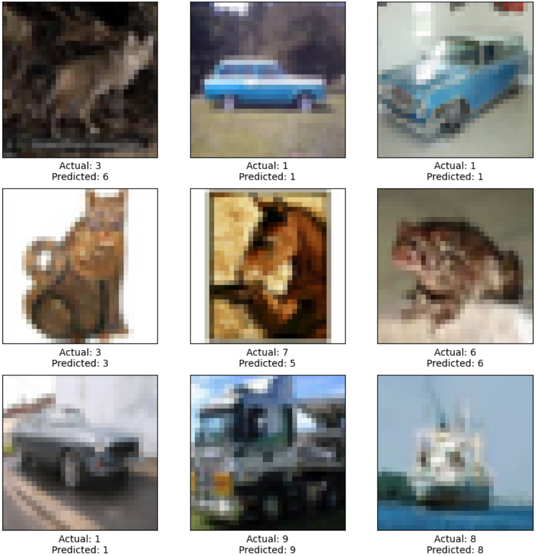
Model B : Conv+BN+MaxPool+Conv+BN+MaxPool+Conv+BN+MaxPool+Flatten+FC+FC

Hyperparameter	**Model A**	**Model B**
Learning Rate	0.001	0.01
Architecture	Basic CNN (LeNet-5)	Advanced CNN
Conv Layers	2	3
FC Layers	3	2
Batch Normalization	No	Yes
Accuracy	40%	69%

Table – Comparison of accuracies between two CNN models For Multi-class CIFAR data classification

Baseline Model

Model Architecture and Performance



Adversarial Attack Implementation

Fast Gradient Sign Method (FGSM)

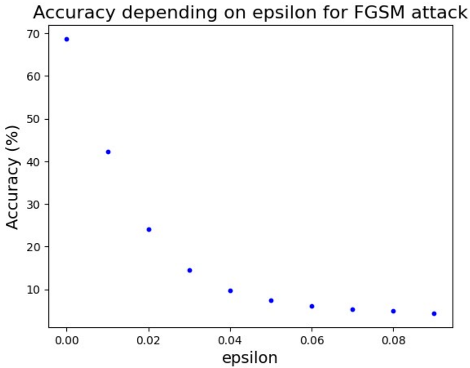
Process :

1. **Input Preparation :** Input images and labels are prepared and gradients are enabled.
2. **Model Prediction :** The model generates predictions and the loss is computed.
3. **Gradient Calculation :** Backpropagation is used to calculate the gradients of the loss with respect to the input images.
4. **Adversarial Image Creation :** The adversarial images are created by adjusting the original images in the direction of the gradient sign, scaled by a small factor (epsilon).

Result : The modified images (adversarial examples) are then used to evaluate the robustness of the model.

Adversarial Attack Implementation

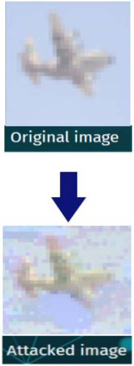
Fast Gradient Sign Method (FGSM)

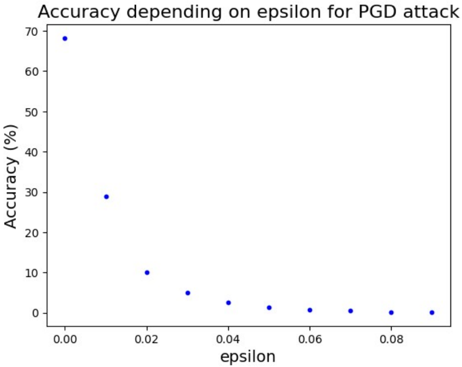


$$\delta = \varepsilon . sign(\nabla J(x, y))$$

Adversarial Attack Implementation

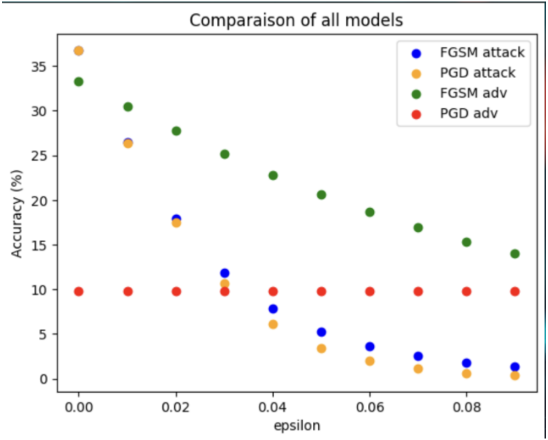
Projected Gradient Descent (PGD)





$$x_{t+1} = x + \pi_{B(x_0,\varepsilon)}(x_t + \delta.sign(\nabla J(x_t, y)))$$

Adversarial Training



Adversarial Attack Implementation

DeepFool Attack Method

- **Purpose :** To subtly modify an image in order to mislead the model while keeping changes imperceptible to human eyes.
- **Process :**
  1. The method starts with a given image and iteratively modifies it.
  2. At each iteration, the model's prediction is evaluated, and the image is adjusted slightly.
  3. The goal is to find the minimal perturbation that causes the model to misclassify the image.
  4. The process is constrained to ensure the perturbed image remains realistic and within valid pixel values.
  5. Iterations continue until the image is successfully misclassified or a maximum number of iterations is reached.

Adversarial Attack Implementation
DeepFool Attack Method

Figure – Clear and perturbed images

Randomized Networks for Enhanced Security

A Strategy for Robustness Against Adversarial Attacks

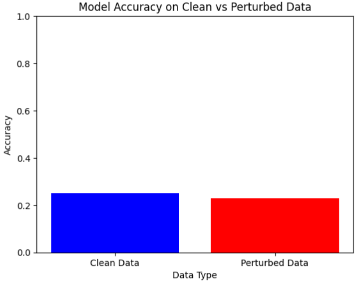
- **Concept of Randomized Networks :**
  - A defense mechanism that introduces randomness into the network's operation.
  - Enhances the robustness of neural networks against adversarial attacks.
- **Key Features :**
  1. Incorporates batch normalization for stable learning.
  2. Employs dropout layers, where the application of dropout is randomized during both training and inference.
- **Advantages :**
  - Increases the unpredictability of the network, making it more difficult for attackers to generate effective adversarial examples.
  - Can be integrated into existing architectures with minimal modifications.

Effect of Adversarial Training on Model Robustness

Comparative Analysis

Condition	Accuracy
Training randomized networks on unprturbed data	11.57%
Training randomized networks on perturbed data	25.07%

Table – Comparison of model accuracy on adversarially perturbed test images, before and after applying adversarial training.



Thank You

For Your Attention!

Any Questions

END OF THE PRESENTATION

Data Science Lab
Assignment 3 : Training robust neural networks

Methods : FGSM, PGD, Randomized Networks, DeepFool

Lyna Bouikni and Arthur MUSSARD

University of Paris Dauphine

December 2023