Défense Contre les Attaques Adversariales sur MNIST

Kouamé Gérard Kra

Janvier 2025

Introduction

Contexte :

- Les exemples adversariaux sont des entrées intentionnellement modifiées pour tromper les modèles d'apprentissage automatique (Goodfellow et al., 2015).
- Ces attaques exploitent les gradients du modèle pour générer des perturbations imperceptibles.

Objectif :

- Concevoir et implémenter une défense robuste contre les attaques adversariales sur un CNN entraîné sur MNIST.

Architecture du Modèle

Structure du modèle utilisé :

- Convolution 1 : 32 filtres, noyau 5x5, activation ReLU.
- Pooling 1 : Max pooling 2x2.
- Convolution 2 : 64 filtres, noyau 3x3, activation ReLU.
- Pooling 2 : Max pooling 2x2.
- Dense 1 : 64 neurones, activation ReLU.
- Dense 2 : 512 neurones, activation ReLU.
- Sortie : 10 neurones (classes).

Attaques Adversariales

FGSM (Goodfellow et al., 2015) :

- Génère une perturbation en un seul pas en utilisant le gradient de la fonction de perte.

BIM (Basic Iterative Method) :

- Étend FGSM de manière itérative pour des perturbations plus précises.

PGD (Madry et al., 2017) :

- Une attaque itérative plus puissante, souvent considérée comme une "attaque ultime" pour tester la robustesse.

$\epsilon = 0.3$, $\alpha = 0.01$, **40 itérations pour BIM et PGD**.

Méthodologie

Entraînement Adversarial :

- Mélange d'exemples propres et adversariaux pour équilibrer robustesse et généralisation.
- Utilisation des attaques FGSM, BIM, et PGD avec des paramètres variés.

Évaluation :

- Basée sur les critères stricts définis par Carlini et al. (2019).
- Testée contre FGSM, BIM et PGD.

Résultats

Performances Baseline :

- Précision sur exemples propres : **98,75%**.
- Précision sous FGSM : **9.12%**.
- Précision sous BIM : **0.00%**.
- Précision sous PGD : **0.00%**.

Après Entraînement Adversarial :

- Précision sur exemples propres : **99,03%**.
- Précision sous FGSM : **84.44%**.
- Précision sous BIM : **56.85%**.
- Précision sous PGD : **58.80%**.

Discussion

Pourquoi le modèle est-il robuste ?

1. Diversité des attaques pendant l'entraînement (Madry et al., 2017).
2. Mélange équilibré d'exemples propres et adversariaux.
3. Adaptation progressive aux perturbations plus complexes.
4. Utilisation de critères rigoureux pour l'évaluation (Carlini et al., 2019).

Conclusion

Résumé :

- Le modèle a montré une robustesse significative face à des attaques adversariales variées.
- Les résultats confirment l'efficacité de l'entraînement adversarial basé sur les approches de Goodfellow et al. (2015) et Madry et al. (2017).

Perspectives :

- Exploration d'autres types d'attaques (e.g., attaques ciblées).
- Combinaison avec des régularisations supplémentaires pour renforcer davantage la robustesse.