TML Assignment 3

Adversarial Robustness of Image Classifiers

Problem Description

Adversarial examples are specially crafted inputs that introduce imperceptible perturbations to fool machine learning models into making incorrect predictions. This presents a critical challenge in deploying reliable and trustworthy AI systems.

In this assignment, we address the problem of adversarial robustness in image classification. The task is to train a classifier that maintains high accuracy on both clean and adversarially perturbed inputs. The adversarial examples are generated using two well-known white-box attacks: **Fast Gradient Sign Method (FGSM)** and **Projected Gradient Descent (PGD)**, both constrained by ℓ_∞ bounded perturbations.

The goal is to build a robust image classifier that is resilient to these attacks while retaining generalization performance on unperturbed data.

Approach Overview

We implemented and evaluated two adversarial training pipelines, both aiming to improve robustness against FGSM and PGD attacks through curriculum-based adversarial augmentation and EMA-based weight stabilization.

Approach 1: PGD-only Curriculum Training

This approach uses adversarial examples generated via PGD throughout the training process, with a curriculum that increases perturbation strength and iteration steps over time:

- **Adversarial Generation**: All adversarial inputs are generated using PGD, starting with weak 1-step attacks and gradually increasing to stronger 3-step attacks with higher ϵ\epsilonϵ and α\alphaα.
- **Curriculum Schedule**:
    - Epochs 0–9: PGD(1-step), $\varepsilon = 2/255$
    - Epochs 10–29: PGD(2-step), $\varepsilon = 4/255$
    - Epochs 30–49: PGD(3-step), $\varepsilon = 6/255$
- **Loss Function**: Averaged loss over clean and adversarial examples.
- **EMA Stabilization**: EMA weights are updated throughout training and used for evaluation.
- **Clean Fine-tuning**: Final 10 epochs train only on clean data with reduced learning rate to recover clean performance.

This variant focuses solely on PGD-based robustness from the beginning, skipping FGSM warm-up.

Approach 2: FGSM→PGD Curriculum with Attack Switch

This second approach introduces an initial warm-up phase with FGSM before transitioning to PGD attacks. It switches the adversarial generation method based on epoch count:

- **Adversarial Generation**:
    - Epochs 0–6: FGSM (ε = 8/255)
    - Epochs 7–19: PGD(1-step), ε = 2/255
    - Epochs 20–39: PGD(2-step), ε = 4/255
    - Epochs 40–54: PGD(3-step), ε = 6/255
- **Attack Switch**: A dedicated attack_fn is selected per epoch (FGSM or PGD).
- **Loss Function**: 0.5 × clean loss + 0.5 × adversarial loss.
- **EMA Stabilization**: Similar to Approach 1, all updates and evaluation use EMA weights.
- **Clean Fine-tuning**: Final 10 epochs on clean inputs with low learning rate to enhance standard generalization.

This strategy seeks to stabilize early training with weaker perturbations (FGSM) before introducing PGD-based attacks, helping the model adapt more smoothly.

Results

The table below summarizes the results for approach 1 and 2 on submitting to the evaluation server.

Approach	Clean Accuracy (%)	FGSM Accuracy (%)	PGD Accuracy (%)
Approach 1 (PGD-only Curriculum)	54.47	20.57	10.63
Approach 2 (FGSM→PGD Curriculum)	**58.67**	**28.67**	**15.10**

Approach 1 successfully met the clean accuracy threshold for evaluation. However, the robustness to adversarial perturbations remains limited. The model demonstrates partial resistance to FGSM but struggles under multi-step PGD attacks.

We hypothesize that the limited adversarial strength (3-step PGD) during training and the absence of random restarts may have contributed to the reduced robustness.

Approach 2 outperformed Approach 1 across all metrics. The inclusion of an FGSM warm-up phase provided a smoother initialization for adversarial training and helped the model generalize better under stronger perturbations later in training. Additionally, the longer training schedule contributed to improved robustness, particularly under PGD attacks.

Conclusion

In this assignment, we explored curriculum-based adversarial training strategies to improve model robustness against FGSM and PGD attacks. Both approaches incorporated adversarial data augmentation, label smoothing, and Exponential Moving Average (EMA) stabilization, followed by clean fine-tuning.

Our experiments showed that **Approach 2**, which gradually transitioned from FGSM to PGD attacks, achieved the best trade-off between clean accuracy and adversarial robustness. It improved PGD accuracy by ~5% and clean accuracy by over 4% compared to the PGD-only baseline. This confirms the value of curriculum schedules that start with weaker perturbations before introducing stronger ones.

To further improve robustness, we could train with longer PGD schedules (e.g., PGD-7 or PGD-10), add random initialization to PGD attacks, and monitor adversarial accuracy on validation data to guide training and early stopping.

References

[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and Harnessing Adversarial Examples*. International Conference on Learning Representations (ICLR).

[2] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards Deep Learning Models Resistant to Adversarial Attacks*. International Conference on Learning Representations (ICLR).

[3] Boenisch, F., & Dziedzic, A. (2025). *Lecture 06: Adversarial Machine Learning*. Trustworthy Machine Learning (SS2025). SprintML / CISPA Helmholtz Center for Information Security.