Introduction of Machine Learning Security

Individual Project

The goal of this project is to evaluate the security of image classifiers against adversarial evasion attacks under the L_∞ threat model.

Students must implement and evaluate one of the following attack types: (i) a fixed-budget attack, such as Projected Gradient Descent (PGD), or (ii) a minimum-norm attack, such as σ-zero adapted to the L_∞ setting.

The evaluation must be performed using a security evaluation curve, where the attack strength ϵ is varied over a predefined range. For each value of ϵ, students must measure and report the robust accuracy of the target model under attack.

As a baseline, students must compare their adversarial attack against random Gaussian noise injection. For each ϵ, the baseline should inject Gaussian noise centered at ϵ with variance 0.001, while respecting the same L_∞ perturbation constraint.

Students must compute the Area Under the Security Evaluation Curve (AUC) for both the adversarial attack and the random noise baseline. The report must demonstrate that the adversarial attack is more effective than random noise injection, meaning that it should produce a lower robust accuracy curve and a lower AUC.

For simplicity, the project focuses only on the same L_∞ threat model. Students are allowed to use pretrained models from the RobustBench repository:

https://github.com/RobustBench/robustbench

The evaluation must target the ImageNet dataset using a subset of 1000 samples. The experimental evaluation must include: (i) one standard non-robust model and (ii) one robust model of the students' choice, both selected from the RobustBench repository.

Evaluation Criteria

- Submission Correctness 10%
- Understanding on Adversarial Attack 20%
- Attack Performance 30%

- Attack Evaluation 20%
- Report Writing 30%

Submission Requirement

Programming

- README.md, which includes the following items, should be prepared
  - The procedure of executing your problem
  - Description of the purpose of each program file
- Compress all related program files and README.md as a ZIP file named "**XX**-**YY**-Program.zip", where **XX** is your student ID, and **YY** is your Chinese name.

Report

- The report must clearly describe the selected attack, the threat model, the chosen ϵ values, the attack hyperparameters, the random Gaussian baseline, the computation of the security evaluation curves, the AUC computation, and the final comparison between attacks and models. Students should also discuss the differences observed between the standard and robust models, highlighting how robustness changes as the perturbation strength increases.
- A report should be prepared. It should be less than 2 pages according to the provided template file. Only the words in blue can be modified, and DO NOT CHANGE the format setting. The template can be downloaded from the following link: http://www.mlclab.org/teaching/MLSec/assignment/report.docx
- No programming code should be included.
- Complete sentences should be used and avoid point form.
- You should save your reports as a PDF file named as "**XX**-**YY**-Report.pdf", where **XX** is your student ID, and **YY** is your Chinese name.

Submission and Due Date

- Compress the following three files as "**XX**-**YY**.zip", where **XX** is your student ID, and **YY** is your Chinese name.
  - "**XX**-**YY**-Program.zip"
  - "**XX**-**YY**-Report.pdf"
- Send these the final zip file to your monitor by **19-June-2026**