# UCSB CS291A Project1 - Whitebox Adversarial Attack to Image Classification Models

Ye Yuan

February $1^{th}$ 2023

## 1 Objectives of this project

The objective of this project is to implement two perturbation generation methods (FGSM and PGD) and conduct adversarial attacks on standard-trained and robust-trained ResNet18 models. In the following sections, I will present the results and my analysis to the following points:

1. Performance evaluation of standard-trained ResNet18 and robustly-trained ResNet18 (in the following: provided models) on original CIFAR-10 dataset.
2. Performance evaluation of provided models on CIFAR-10 dataset with perturbation generated with untargeted FGSM ($l_\infty$). Adversarial settings: $\epsilon = \frac{8}{255}$, loss function is Cross Entropy.
3. Performance evaluation of provided models on CIFAR-10 dataset with perturbation generated with untargeted PGD ($l_\infty$). Adversarial settings: $\epsilon = \frac{8}{255}$, $\alpha = \frac{2}{255}$, $T = 10$, loss function is Cross Entropy.
4. Performance evaluation of provided models on CIFAR-10 dataset with perturbation generated with untargeted PGD ($l_\infty$). Adversarial settings: $\epsilon = \frac{8}{255}$, $\alpha = \frac{2}{255}$, $T = 10$, loss function is C&W Loss, confident threshold $\tau = 0$.
5. Performance evaluation of provided models on CIFAR-10 dataset with perturbation generated with targeted PGD ($l_\infty$). Adversarial settings: $\epsilon = \frac{8}{255}$, $\alpha = \frac{2}{255}$, $T = 10$, loss function is C&W Loss, confident threshold $\tau = 0$, target the adversarial samples to *'class 1'*.
6. Ablation studies on different adversarial settings. Use the untargeted PGD attack with Cross Entropy loss, varing the following hyperparameters:
   (a) Fix $\alpha = \frac{2}{255}$ and $\epsilon = \frac{8}{255}$, vary $T \in \{1, 5, 10, 20, 50\}$.
   (b) Fix $\alpha = \frac{2}{255}$ and $T = 10$, vary $\epsilon \in \{\frac{1}{255}, \frac{2}{255}, \frac{4}{255}, \frac{6}{255}, \frac{8}{255}\}$.

# 2 Performance w/o adversarial perturbations

**Settings.** This section presents an evaluation of two models, namely one that underwent standard training and another that underwent robust training, on the clean test set of the CIFAR-10 dataset.

Table 1: Clean accuracy of provided models on CIFAR-10 test set.

| **Models** | **Standard ResNet18** | **Robust ResNet18** |
|---|---|---|
| **Accuracy (%)** | 94.74 | 82.08 |

**Observations.** Table 1 provides a data of the clean accuracy of the standard-trained and robustly-trained ResNet18 models on the CIFAR-10 test set. The results demonstrate that there is a noticeable discrepancy in the classification performance of the two models on clean images, with the standard model exhibiting a 12.66% higher accuracy than the robust ResNet18 model.

**Analysis.** The above findings suggest that the robust training process has some side-effects which could result in reducing the performance in classifying the samples w/o adversarial perturbations, and this performance drop is non-neglectable, users need to consider the trade-offs carefully when using such models.

# 3 Untargeted FGSM with CE loss

**Settings.** This section presents a comprehensive evaluation of two models, the standard-trained and robustly-trained ResNet18, on the CIFAR-10 test set. The samples in the test set were subjected to an untargeted Fast Gradient Sign Method (FGSM) attack with $\epsilon = \frac{8}{255}$ and were evaluated using the Cross-Entropy loss as the attack loss.

Table 2: Adversarial accuracy of provided models on CIFAR-10 test set with untargeted FGSM attack.

| **Models** | **Standard ResNet18** | **Robust ResNet18** |
|---|---|---|
| **Accuracy (%)** | 39.27 | 56.82 |

**Observations.** Table 2 presents a data of the adversarial accuracy of the standard-trained and robustly-trained ResNet18 models on the CIFAR-10 test set. The results indicate that:

1. The accuracy of the standard-trained model declined by 55.47% compared to its accuracy on clean samples, as reported in Section 2.
2. The robustly-trained model's accuracy increased by 25.26% compared to its accuracy on clean samples, as reported in Section 2.
3. The performance difference between the two models is 17.55%.

**Analysis.** The observations suggest that:

1. Despite being a one-shot attack, FGSM attack is a highly effective method for delivering adversarial perturbations and can cause a significant impact on the performance of models without adversarial training.

2. The robustly-trained model exhibits a decline in performance when subjected to adversarial training, however, in the presence of adversarial perturbation in the input samples, the impact of accuracy is controlled to a certain extent.

3. Adversarial training has the potential to significantly enhance the model's performance in the presence of adversarial perturbations.

# 4 Untargeted PGD with CE loss

**Settings.** In this section, I aim to conduct a comprehensive evaluation of the standard-trained and robustly-trained variants of ResNet18 on the CIFAR-10 dataset. The samples in the test set were subjected to an adversarial attack using the Projected Gradient Descent (PGD) method with the hyperparameters $\epsilon = \frac{8}{255}$, $\alpha = \frac{2}{255}$, and a total of $T = 10$ iterations. The effectiveness of the attack was measured by employing the Cross-Entropy loss as the attack loss metric.

Table 3: Adversarial accuracy of provided models on CIFAR-10 test set with untargeted PGD attack and CE loss function.

| **Models** | **Standard ResNet18** | **Robust ResNet18** |
|---|---|---|
| **Accuracy (%)** | 0.02 | 52.77 |

**Observations.** Table 3 presents a comparison of the adversarial accuracy of the standard-trained and robustly-trained ResNet18 models on the CIFAR-10 test set. The results indicate that:

1. The accuracy of the standard-trained model decreased by 94.72% compared to its accuracy on clean samples, as previously reported in Section 2. In comparison to the accuracy of the standard model under the perturbation generated by the Fast Gradient Sign Method (FGSM) algorithm, the adversarial accuracy was found to be 39.25% lower than the accuracy reported in Section 3.

2. The robustly-trained model's accuracy decreased by 29.31% compared to its accuracy on clean samples, as previously reported in Section 2. In comparison to the accuracy of the robust model under the perturbation generated by the FGSM algorithm, the adversarial accuracy was 4.05% lower than the accuracy reported in Section 3.

3. The performance difference between the two models is 52.75%.

**Analysis.** Based on these observations, we can deduce that:

1. The PGD attack leverages multiple iterations to generate the perturbation $\delta$ that can significantly impair the accuracy of standard-trained models. Without a robust training process, the model's utility can be greatly reduced.

2. Compared to the FGSM algorithm, the PGD attack has a stronger ability to perturb the classification models and cause incorrect classifications.

3. Despite the stronger PGD attack, the robustly-trained model demonstrates a relatively low decline in terms of adversarial accuracy, indicating the effectiveness of the robust training process in mitigating adversarial perturbations.

4. It is worth noting that the robust model referred to as "pgd10" might have been trained with the PGD algorithm. This observation suggests that the algorithm used in the adversarial training process might be the most effective in defending against the same type of adversarial attack.

# 5 Untargeted PGD with C&W loss

**Settings.** This section endeavors to conduct a thorough evaluation of two distinct models, namely the standard-trained and robustly-trained variants of ResNet18, on the benchmark CIFAR-10 dataset. The samples in the test set were subjected to an adversarial attack using the untargeted Projected Gradient Descent (PGD) method, with hyperparameters $\epsilon = \frac{8}{255}$, $\alpha = \frac{2}{255}$, confidence threshold $\tau = 0$, and a total of $T = 10$ PGD iterations. The effectiveness of the attack was measured through C&W loss as the attack loss metric.

Table 4: Adversarial accuracy of provided models on CIFAR-10 test set with untargeted PGD attack and C&W loss function.

| **Models** | **Standard ResNet18** | **Robust ResNet18** |
|---|---|---|
| **Accuracy (%)** | 0.03 | 50.33 |

**Observations.** The data presented in Table 4 investigates the adversarial accuracy of standard-trained and robustly-trained ResNet18 models on the CIFAR-10 test set. The results reveal two key points:

1. The adversarial accuracy of the standard-trained model shows no substantial difference from its adversarial accuracy obtained through the Cross-Entropy (CE) loss function, as previously discussed in Section 4.

2. The adversarial accuracy of the robustly-trained model experiences a decrease of 2.44% when compared to its adversarial accuracy obtained through the CE loss function, as highlighted in Section 4.

**Analysis.** The findings from this study suggest that incorporating the C&W loss function into the training process of adversarial perturbations can augment its perturbative strength. However, for standard models, the CE loss function alone is sufficient in completely debilitating the model's performance.

# 6 Targeted PGD with C&W loss

**Settings.** This section endeavors to conduct a thorough evaluation of two distinct models, namely the standard-trained and robustly-trained variants of ResNet18, on the benchmark CIFAR-10 dataset. The samples in the test set were subjected to an adversarial attack using the targeted Projected Gradient Descent (PGD) method, all perturbed samples are targeted to "class 1". Hyperparameters are as below: $\epsilon = \frac{8}{255}$, $\alpha = \frac{2}{255}$, confidence threshold $\tau = 0$, and a total of $T = 10$ PGD iterations. The effectiveness of the attack was measured through C&W loss as the attack loss metric.

Table 5: Adversarial accuracy of provided models on CIFAR-10 test set with targeted PGD attack and C&W loss function.

| **Models** | **Standard ResNet18** | **Robust ResNet18** |
|---|---|---|
| **Accuracy (%)** | 5.63 | 74.05 |

**Observations.** The data presented in Table 5 shows the adversarial accuracy of both standard-trained and robustly-trained ResNet18 models on the CIFAR-10 test set. The data reveals a noteworthy trend, demonstrating that the adversarial accuracy of both the standard and robust models under targeted adversarial attacks is higher compared to that of the adversarial accuracy under untargeted attacks.

**Analysis.** The above observations have significant implications, suggesting that models that underwent targeted adversarial attacks tend to have a higher adversarial accuracy as compared to untargeted attacks. This discrepancy in performance can be attributed to the fact that in targeted attack scenarios, the attacker endeavors to manipulate the model's prediction to a specific target class, while in untargeted scenarios, the objective of the attacker is simply to cause misclassification of the input. The targeted nature of the attack, therefore, provides a more constrained space for the attacker to operate in, leading to higher adversarial accuracy.

# 7 Ablation studies

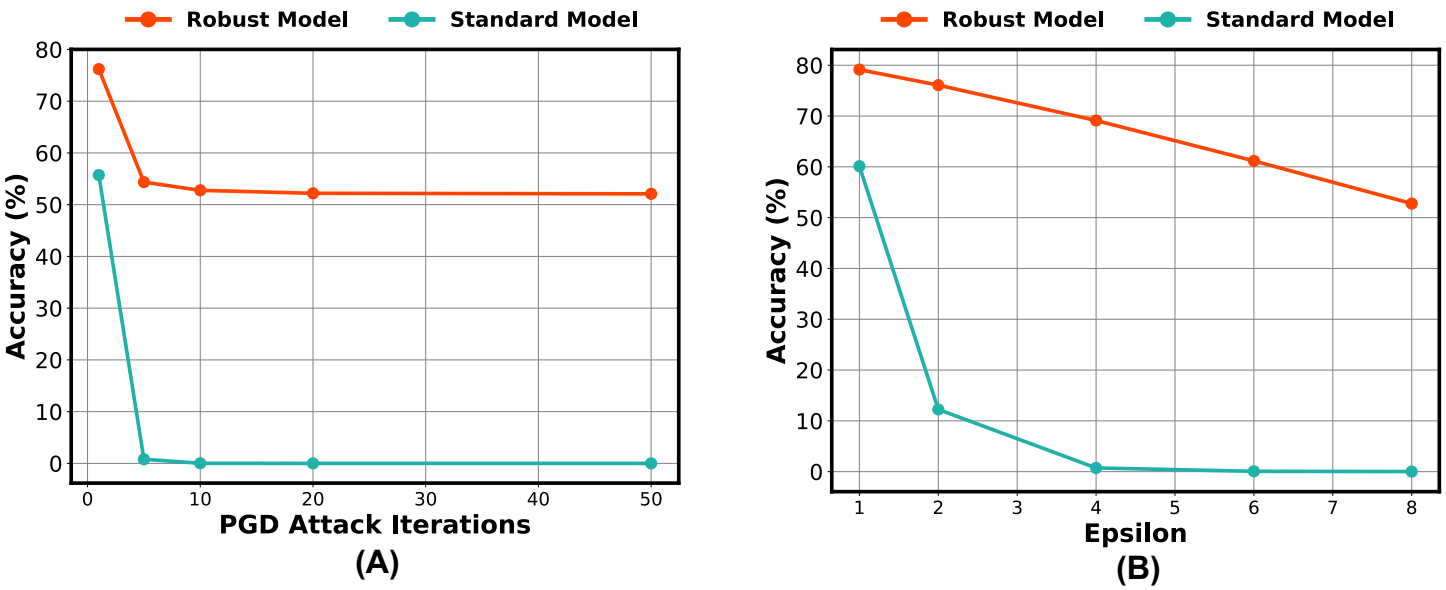


Figure 1: (A) The accuracy of robust model and standard model under different PGD attack iterations; (B) The accuracy of robust model and standard model under different $\epsilon$, note that the $\epsilon = 1$ on the x-axis means $\epsilon = \frac{1}{255}$.

## 7.1 The influence of adversarial iterations T

Table 6: Adversarial accuracy (%) of provided models on CIFAR-10 test set with untargeted PGD attack and CE loss function on various $T$.

| **Models (ResNet18)** | $\boldsymbol{\alpha = \frac{2}{255}, \epsilon = \frac{8}{255}}$ | | | | |
|---|---|---|---|---|---|
| | $\boldsymbol{T = 1}$ | $\boldsymbol{T = 5}$ | $\boldsymbol{T = 10}$ | $\boldsymbol{T = 20}$ | $\boldsymbol{T = 50}$ |
| **Standard** | 55.73 | 0.8 | 0.02 | 0 | 0 |
| **Robust** | 76.2 | 54.36 | 52.77 | 52.2 | 52.09 |

**Analysis.** Table 6 and Figure 1 (A) indicates a clear trend of decreasing adversarial accuracy as the number of PGD attack iterations $T$ increases. Nevertheless, it is important to note that the decrease in accuracy becomes negligible after $T = 10$. This suggests that the value of $T = 10$ is an optimal trade-off, balancing the computational overhead and the strength of perturbation effect.

## 7.2 The influence of adversarial budget $\epsilon$

Table 7: Adversarial accuracy (%) of provided models on CIFAR-10 test set with untargeted PGD attack and CE loss function on various $\epsilon$.

| **Models (ResNet18)** | $\boldsymbol{\alpha = \frac{2}{255}, T = 10}$ | | | | |
|---|---|---|---|---|---|
| | $\boldsymbol{\epsilon = \frac{1}{255}}$ | $\boldsymbol{\epsilon = \frac{2}{255}}$ | $\boldsymbol{\epsilon = \frac{4}{255}}$ | $\boldsymbol{\epsilon = \frac{6}{255}}$ | $\boldsymbol{\epsilon = \frac{8}{255}}$ |
| **Standard** | 60.12 | 12.25 | 0.73 | 0.07 | 0.02 |
| **Robust** | 79.14 | 76.08 | 69.12 | 61.17 | 52.77 |

**Analysis.** The results presented in Table 7 and Figure 1 (B) demonstrate a clear trend in which the adversarial accuracy of both standard and robust models decreases as the adversarial budget $\epsilon$ increases. This is to be expected, as the magnitude of the perturbation strength is directly proportional to the value of $\epsilon$.

However, it is important to consider the practical implications of adversarial perturbations in real-world scenarios. As $\epsilon$ increases, so does the visibility of the perturbations, which may prompt users to sanitize the input samples to ensure the utility of the models. This highlights the delicate balancing act involved in setting the value of $\epsilon$, as it must be a compromise between the strength of an adversarial attack and its practical impact on the utility.