Winter 2023 CS291A: Special Topics on Adversarial Machine Learning – Homework 2

Ye Yuan[*]
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106
ye33@ucsb.edu

Jiamu Xie[†]
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106
jiamuxie@ucsb.edu

1 Part 1

1.1 Adversarial Training (AT) to Train a Robust ResNet-18 Model

In this section, we employed the Cross Entropy Loss function to train the model, utilizing an untargeted 10-step Projected Gradient Descent (PGD) attack with a step size of 2/255 and a perturbation budget of 8/255. The model was trained for a total of 160 epochs, incorporating a three-stage learning rate schedule for optimal performance.

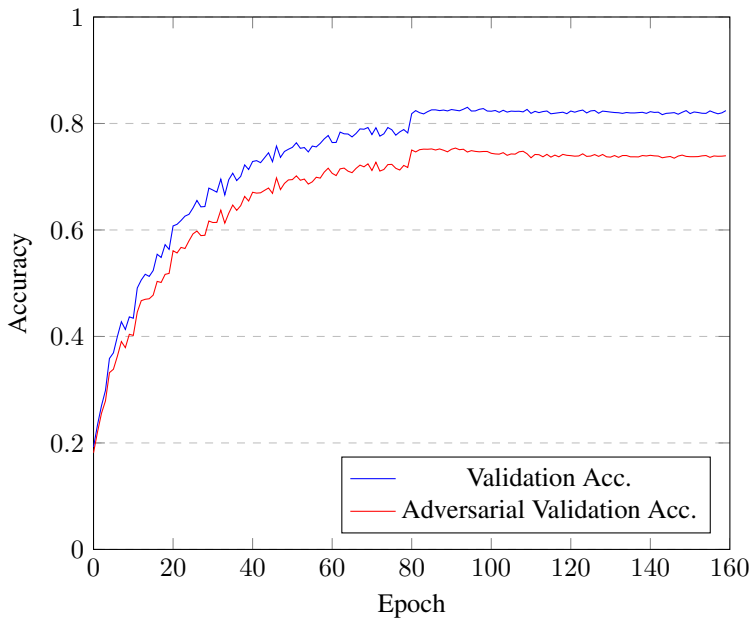

Figure 1: Validation and Adversarial Validation Accuracy

From Fig. 1 we can see the variation of both standard validation accuracy and adversarial validation accuracy. From the curve, we can see that after about 80 epochs, the model converged and its accuracy no longer increased. In practice, we may incorporate some techniques like early-stop to save unnecessary computational overhead.

[*]https://www.linkedin.com/in/yethyuan/
[†]https://www.linkedin.com/in/jiamu-xie-061278223/

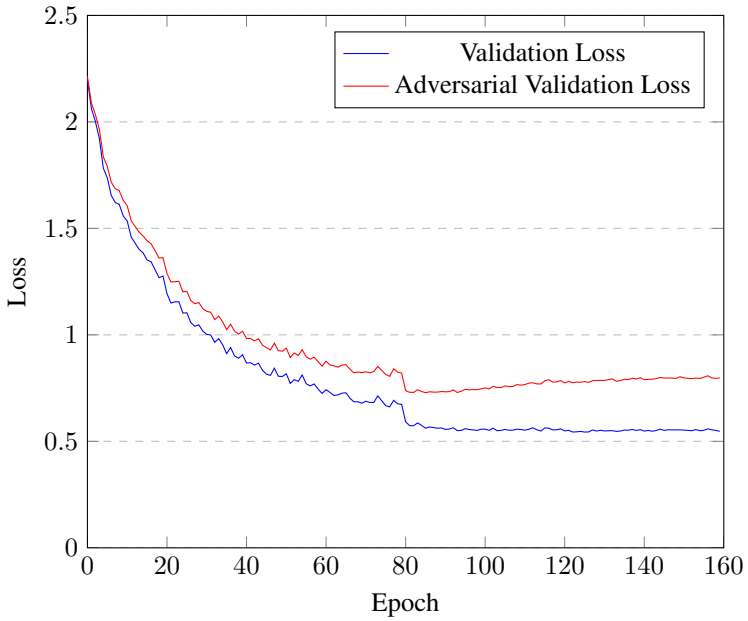

Figure 2: Validation and Adversarial Validation Loss

1.2 Fast Adversarial Training (Fast AT) to Train a Robust ResNet-18 Model

In this section, we employed the Cross Entropy Loss function to train the model, utilizing an untargeted FGSM attack with a perturbation budget of 8/255. The model was trained for a total of 160 epochs, incorporating a three-stage learning rate schedule for optimal performance.

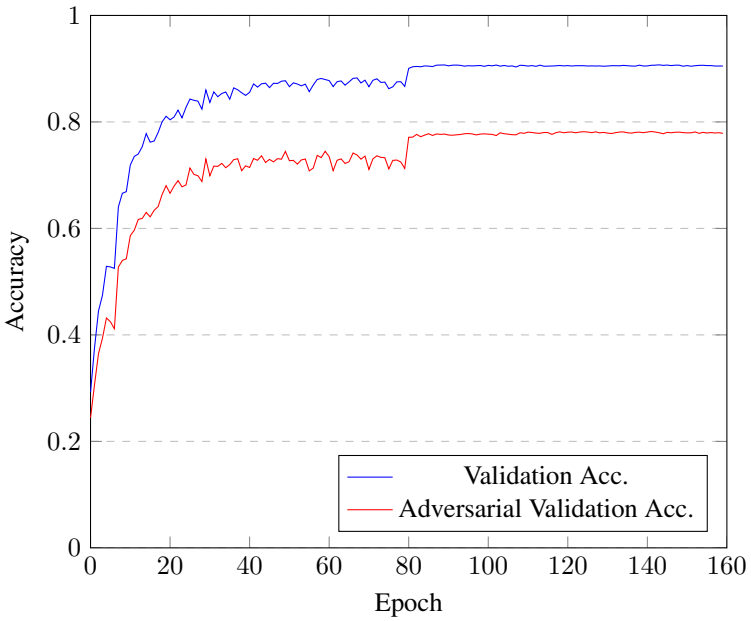

Figure 3: Validation and Adversarial Validation Accuracy

From Fig. 3 we can see the variation of both standard validation accuracy and adversarial validation accuracy. From the curve, we can see that after about 80 epochs, the model converged and its

accuracy no longer increased. In practice, we may incorporate some techniques like early-stop to save unnecessary computational overhead.

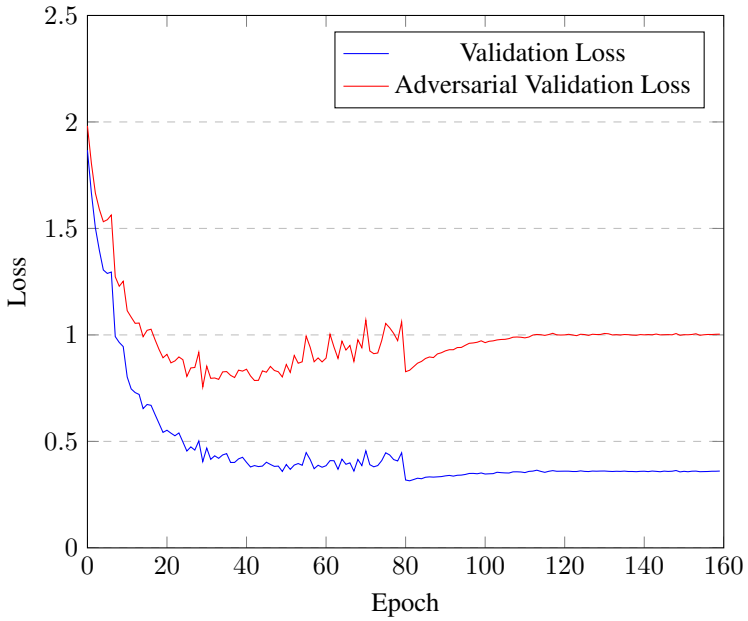

Figure 4: Validation and Adversarial Validation Loss

1.3 Clean and Robust Accuracy on the CIFAR-10 Test Set Using the Following Attacks

1.3.1 50-step with 2/255 Attack Step Size, 8/255-Tolerant Untargeted PGD Attack with CE Loss

Table 1: Accuracy

Type \ Acc.	Clean Acc.	Robust Acc.
AT	0.812	0.319
Fast AT	0.8988	0.4007

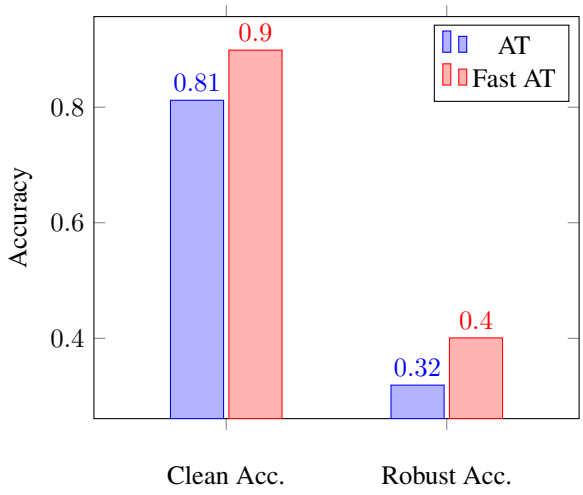

Figure 5: Accuracy

1.3.2 50-step with 2/255 Attack Step Size, 8/255-Tolerant Untargeted PGD Attack with C&W Loss (Threshold τ = 0)

Table 2: Accuracy

Acc. / Type	Clean Acc.	Robust Acc.
AT	0.812	0.3126
Fast AT	0.8988	0.3999

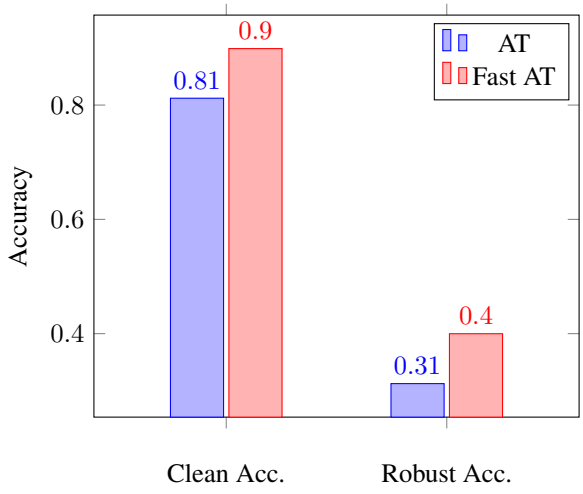

Figure 6: Accuracy

Upon analyzing the graphs, it becomes apparent that the two attack parameters exert similar effects on the Adversarial Training (AT) model and the Fast AT model, respectively. Moreover, the Fast AT model exhibits a significantly superior performance to the AT model, as evidenced by higher clean and robust accuracy scores. This suggests that, in most scenarios, the Fast AT model demonstrates superior performance compared to the AT model.

1.4 Compare the Robustness Between the Robust Model We Trained and the Robust Model Provided in HW1

Table 3: Accuracy

Type \ Acc.	Clean Acc.	Robust Acc.(CE)	Robust Acc.(C&W)
AT	0.812	0.319	0.3126
Fast AT	0.8988	0.4007	0.3999
HW1	0.8209	0.5212	0.4971

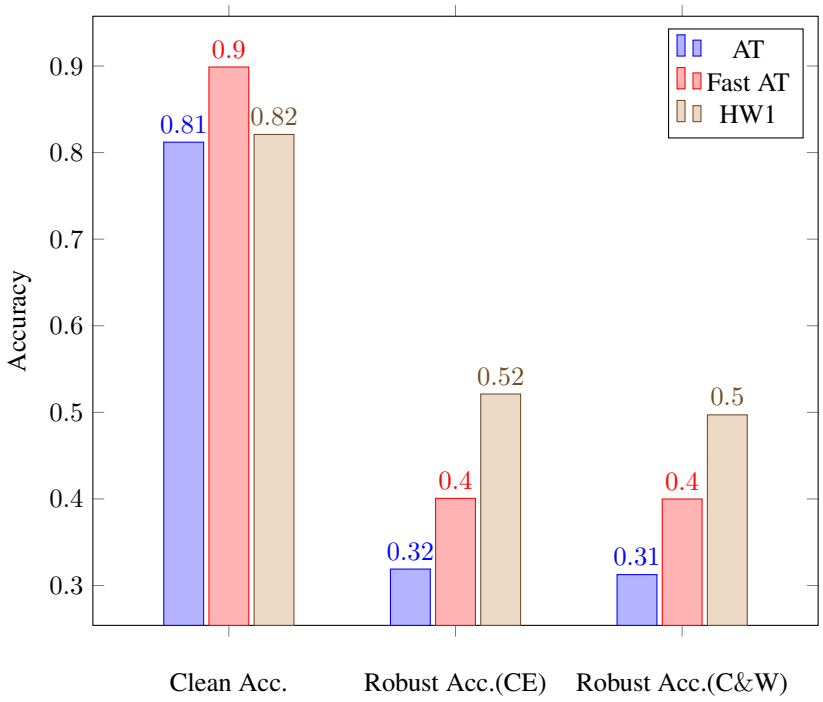

Figure 7: Accuracy

Regarding clean accuracy, the Fast Adversarial Training (AT) model shows a marked advantage over the other two models, whereas the AT model is outperformed to a small extent by the robust model provided in HW1. Additionally, each of the three models respectively demonstrate similar robust accuracy across two attack settings, with the robust model provided in HW1 exhibiting a significantly higher robust accuracy than the Fast AT model, which in turn exceeds the AT model in terms of robustness. Overall, the AT model displays inferior performance figures relative to the other two models, with the Fast AT model exhibiting optimal performance in clean settings and the robust model provided in HW1 showing superior robust accuracy.

2 Part 2

2.1 Adversarial Training Method Implemented by Ourselves

As our study did not incorporate a novel adversarial training method, we evaluated the performance of the Adversarial Training (AT) model and Fast AT model in this session. Our experiments revealed that the Fast AT model exhibits superior performance to the AT model in all tests conducted. The subsequent section presents a detailed analysis and results of the Fast AT model.

2.1.1 50-step with 2/255 Attack Step Size, 8/255-Tolerant Untargeted PGD Attack with CE Loss

Table 4: Accuracy

Type \ Acc.	Clean Acc.	Robust Acc.
Fast AT	0.8988	0.4007

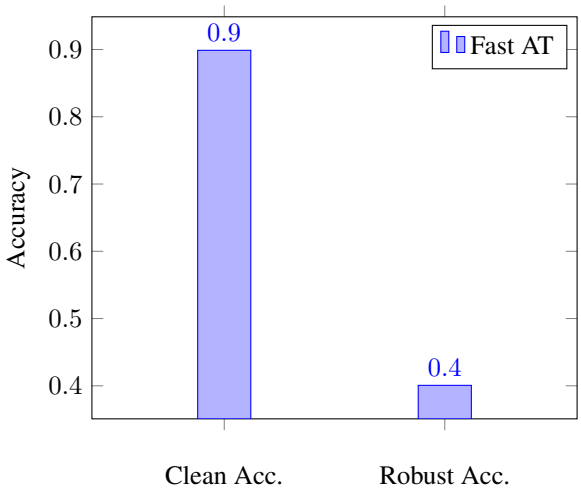

Figure 8: Accuracy

2.1.2 50-step with 2/255 Attack Step Size, 8/255-Tolerant Untargeted PGD Attack with C&W Loss (Threshold τ = 0)

Table 5: Accuracy

Type \ Acc.	Clean Acc.	Robust Acc.
Fast AT	0.8988	0.3997

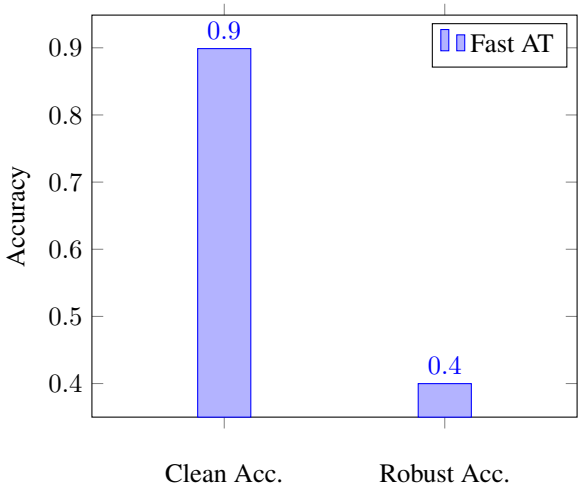

Figure 9: Accuracy

2.2 AutoAttack BenchMark

Table 6: Accuracy

Type \ Phase	Initial	APGD-CE	APGD-T	FAB-T	SQUARE
Fast AT	0.8988	0.3898	0.3881	0.3881	0.3881

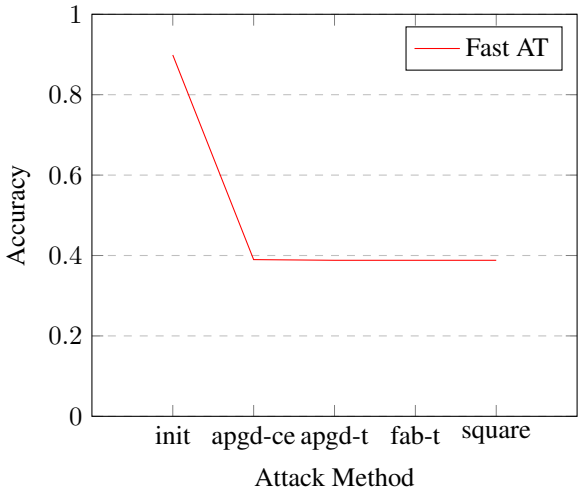

Figure 10: AutoAttack Accuracy

From Fig. 10, we observe a significant accuracy drop in the Fast AT model after the apgd-ce attack, followed by negligible or no changes in accuracy in the subsequent three rounds of attacks. This behavior could be attributed to two factors: Firstly, as the accuracy of the model decreases, the intensity of attacks performed by AutoAttack is reduced. Secondly, the Fast AT model may have reached a limit in its robustness capacity, beyond which further attacks do not significantly impact its accuracy. These observations suggest that while the Fast AT model exhibits superior robustness and clean accuracy compared to the AT model, it may still have limitations in its ability to handle a diversity of attacks. In summary, the Fast AT model offers better performance under most attack scenarios, but further research is required to explore its limitations in handling complex attacks.