

A APPENDIX

B STRONG ATTACKS AFTER OVERFITTING

When FastAdv+ is used for training a model, even though the model can recover from catastrophic overfitting via PGD adversarial training, it is possible that the model overfits to PGD attacks and stays vulnerable to other attacks. Therefore, we extract the model right after its recovery from catastrophic overfitting and run several kinds of attacks, including 10-step PGD attacks, 50-step PGD attacks with 10 restarts, C&W attacks (Carlini & Wagner, 2017) and fast adaptive boundary (FAB) attacks (Croce & Hein, 2019), on this model.

Table 3: CIFAR-10 standard and robust accuracy on PreAct ResNet-18 under various types of attacks.

Attacks	PGD-10	PGD-50	C&W	FAB
Robust Accuracy	40.22%	39.41%	41.05%	38.68%

The result shows the model recovered from catastrophic overfitting is indeed robust. Note the robust accuracy is relatively low as we are not using the final model.

C ABLATION ANALYSIS ON ADJUSTED ATTACK SIZE

In Section 5 we show it is possible to improve the performance of FastAdvW via using a smaller size of attacks for FGSM adversarial training. It is possible that the adjusted size of attacks benefits not only our approach, but also PGD adversarial training. Therefore, we use the same setting (4/255 for the first 70 epochs and 8/255 for the rest) for full PGD adversarial training and compare it to vanilla PGD adversarial training.

Table 4: CIFAR-10 standard and robust accuracy on PreAct ResNet-18 for vanilla PGD adversarial training and PGD adversarial training with adjusted size of attacks (4/255 and 8/255).

Method	Standard Accuracy	PGD($\epsilon = 8/255$)
PGD	$83.43 \pm 0.25\%$	$51.74 \pm 0.17\%$
PGD(adjusted size)	$83.11 \pm 0.11\%$	$52.14 \pm 0.28\%$

The results show that PGD adversarial training enjoys limited benefits from the adjusted size of attacks. This strategy is more compatible with our proposed FastAdvW.