Supplementary Material of Preventing Catastrophic Overfitting in Fast Adversarial Training: A Bi-level Optimization Perspective

Zhaoxin Wang[1], Handing Wang[1✉], Cong Tian[2], and Yaochu Jin[3]

[1] School of Artificial Intelligence, Xidian University, Xi'an, China
zxwang74@163.com
[2] School of Computer Science and Technology, Xidian University, Xi'an, China
ctian@mail.xidian.edu.cn
[3] School of Engineering, Westlake University, Zhejiang Hangzhou, China
jinyaochu@westlake.edu.cn

Table 1: Accuracy (%) and training time (min) of compared AT models on WideResNet34-10 with the CIFAR10 dataset. The number in bold indicates the best.

Method		Clean Acc	FGSM	PGD10	PGD20	PGD50	C&W	APGD	Square	AA	Time
PGD-AT	Best	87.30	68.92	55.21	53.96	53.49	51.80	54.42	60.40	51.20	1397
	Last	87.39	68.20	54.12	52.85	52.49	50.68	53.37	59.36	50.57	
TRADES	Best	85.70	68.28	57.21	56.10	55.87	50.72	54.85	59.84	53.36	1692
	Last	85.70	68.28	57.21	56.10	55.87	50.72	54.85	59.84	53.36	
FGSM-RS	Best	75.10	59.00	44.66	43.29	42.96	38.68	44.98	50.28	40.27	281
	Last	86.19	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Free-AT	Best	71.79	51.37	41.75	41.13	40.99	35.67	43.81	44.33	39.22	969
	Last	71.79	51.37	41.75	41.13	40.99	35.67	43.81	44.33	39.22	
FGSM-MEP	Best	83.43	67.73	**58.13**	**57.52**	**57.51**	49.62	53.35	58.13	51.54	407
	Last	85.63	69.09	57.47	56.48	56.20	49.82	52.89	58.45	51.06	
FGSM-PCO	Best	**87.38**	**69.78**	57.82	57.12	56.96	**51.27**	**54.34**	**59.88**	**51.84**	421
	Last	**87.38**	**69.78**	**57.82**	**57.12**	**56.96**	**51.27**	**54.34**	**59.88**	**51.84**	

1 Experimental Results

The classification accuracy of WideResNet34-10 with the CIFAR10 dataset is shown in Table 1. On the WideResNet34-10 model, we achieve good performance, especially for clean examples, which reach 87.38% accuracy. To comprehensively evaluate the performance of various AT methods and investigate the overfitting phenomenon, we conduct experiments with a smaller model as the

✉ Corresponding Author: hdwang@xidian.edu.cn

backbone on datasets where catastrophic overfitting occurs. Table 2 presents the results on CIFAR100 with the ResNet18 model. The results demonstrate that FGSM-MEP effectively prevents the catastrophic overfitting problem observed in the WideResNet34-10 model. Our method, FGSM-PCO, achieves improvements both on clean examples and AEs, with only a 0.1% lower performance than FGSM-MEP under the CW attack at the last checkpoint. It is noteworthy that our method incurs a higher computational cost than FGSM-MEP but saves memory on computational devices, requiring only two-thirds of the memory compared to FGSM-MEP.

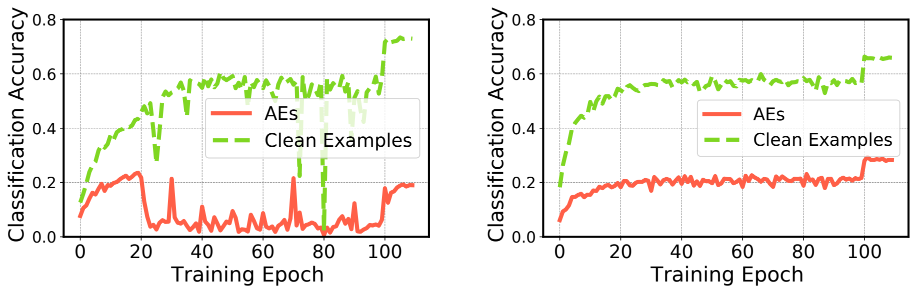
Table 2: Accuracy (%) and training time (min) of compared AT models on ResNet18 with the CIFAR100 dataset. The number in bold indicates the best.

Method		Clean Acc	FGSM	PGD10	PGD20	PGD50	C&W	APGD	Square	Time
PGD-AT [3]	Best	58.24	37.84	29.68	29.20	29.15	25.09	27.82	30.73	191
	Last	58.42	37.59	29.00	28.45	28.35	24.83	27.27	30.42	
TRADES [6]	Best	58.38	37.95	30.53	30.05	29.95	23.55	26.28	30.20	260
	Last	58.00	38.08	30.34	29.99	29.89	23.57	26.15	29.95	
FGSM-RS [5]	Best	45.64	28.87	20.89	20.20	20.24	17.21	17.82	20.01	38
	Last	42.54	-	00.00	00.00	00.00	00.00	00.00	00.00	
FGSM-GA [1]	Best	46.37	28.56	21.74	21.43	21.31	18.18	19.81	22.21	137
	Last	**62.34**	-	00.02	00.00	00.00	00.00	00.00	00.00	
Free-AT [4]	Best	38.19	23.17	18.38	18.11	18.08	15.02	16.13	17.46	138
	Last	38.19	23.17	18.38	18.11	18.08	15.02	16.13	17.46	
FGSM-EP [2]	Best	58.24	39.69	31.69	31.34	31.27	25.14	27.39	30.81	58
	Last	58.20	39.41	31.39	30.96	30.92	24.86	27.28	30.46	
FGSM-MEP [2]	Best	58.79	39.06	31.83	31.35	31.35	25.76	27.88	31.09	58
	Last	58.82	39.77	31.74	31.22	31.12	**25.26**	**27.66**	30.92	
FGSM-HPF	Best	**60.20**	**39.98**	**32.39**	**31.94**	**31.85**	**25.85**	**28.16**	**31.50**	60
	Last	59.80	**39.83**	**31.89**	**31.44**	**31.36**	25.25	27.62	**31.11**	

On the WideResNet34-10 model, nearly all FGSM-based methods exhibit catastrophic overfitting. Fig. 1 illustrates that FGSM-PCO effectively prevents the overfitting problem.

2 Divergence between Adversarial and Clean Examples

Apart from the classification accuracy of AT models under various attacks, the divergence between adversarial and clean examples is also a crucial metric for evaluating AT algorithms. The norm of perturbation can be regarded as a convergence criterion for the non-convex optimization problem, with smaller perturbations implying quicker convergence to local optima [2]. We evaluate the



(a) FGSM-MEP **(b)** FGSM-PCO

Fig. 1: Classification accuracy of FGSM-MEP and FGSM-PCO on WideResNet34-10 with the CIFAR100 dataset. Our method significantly prevents catastrophic overfitting.

perturbation under L2 norm for PGD10, FGSM-MEP and FGSM-PCO algorithms on the CIFAR100 dataset with the ResNet18 model. The results indicate that our method achieves the smallest perturbation norm, as shown in Fig. 2.

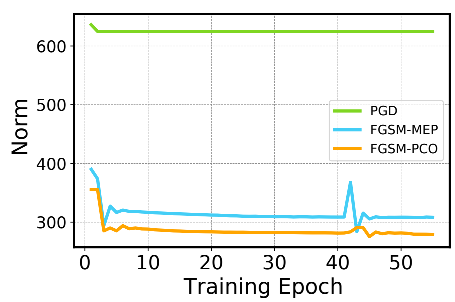


Fig. 2: The L2 norm of perturbation under different AT methods.

References

1. Andriushchenko, M., Flammarion, N.: Understanding and improving fast adversarial training. neural information processing systems (2020)
2. Jia, X., Zhang, Y., Wei, X., Wu, B., Ma, K., Wang, J., Cao, X.: Prior-guided adversarial initialization for fast adversarial training (2022)
3. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
4. Shafahi, A., Najibi, M., Ghiasi, M.A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G., Goldstein, T.: Adversarial training for free! Advances in Neural Information Processing Systems **32** (2019)
5. Wong, E., Rice, L., Kolter, J.Z.: Fast is better than free: Revisiting adversarial training. Learning (2020)

6. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International conference on machine learning. pp. 7472–7482. PMLR (2019)