Technical Appendix of Accelerating Adversarial Training on Under-Utilized GPU

Zhuoxin Zhan[1] , **Ke Wang**[1] and **Pulei Xiong**[2]
[1]Simon Fraser University
[2]National Research Council Canada
zhuoxin_zhan@sfu.ca, wangk@cs.sfu.ca, pulei.xiong@nrc-cnrc.gc.ca

A Appendix

Training Settings. On image datasets CIFAR-10, CIFAR-100 and TinyImageNet, all models are trained using an SGD optimizer with the momentum of 0.9 and the weight decay of 2e-4 on CIFAR-10 as in [Hua *et al.*, 2021] and of 5e-4 on CIFAR-100 and TinyImageNet as in [Li *et al.*, 2023], for 120 epochs with an initial learning rate of 0.1 and a decay of 0.1 at the 80-th and 100-th epochs as in [Li *et al.*, 2023]. On tabular datasets Jannis and Covertype, all models are trained using an AdamW optimizer with the learning rate of 1e-4 and the weight decay of 1e-5 for 100 epochs, following [Gorishniy *et al.*, 2021].

Hyperparameter Settings for Base AT. For the image datasets, following [Li *et al.*, 2023; Tong *et al.*, 2024], the attack function Atk adopts perturbation radius $\epsilon = 8/255$ under ℓ_∞ norm, attack step size $\alpha = 2/255$ for multi attack-step BulletTrain, DBAC, PGDAT, and TRADES, or $\alpha = \epsilon = 8/255$ for single attack-step N-FGSM and TDAT. Base AT specific hyperparameters are as follows. Following [Hua *et al.*, 2021], for TRADES, loss weight $\beta = 6$; for BulletTrain, scaling factor $\gamma = 0.8$ and momentum $p_1 = 0.9$, attack step for $Atk(X_R, K_R)$ is $K_R = 2$. Following [Tong *et al.*, 2024], for N-FGSM, noise magnitude 2ϵ; for TDAT, relaxation factor $\gamma_{min} = 0.15$, 0.05 and 0.025 on CIFAR-10, CIFAR-100 and TinyImageNet, and momentum factor $p_2 = 0.75$ on all datasets.

For the tabular datasets, Atk adopts the perturbation radius $\epsilon = 0.1$ on Jannis and $\epsilon = 0.05$ on CoverType under ℓ_2 norm. For BulletTrain, DBAC and PGDAT, attack step size $\alpha = 0.02$ on Jannis and $\alpha = 0.01$ on Covertype. For BulletTrain, scaling factor $\gamma = 0.5$ and momentum $p_1 = 0.9$, attack step for $Atk(X_R, K_R)$ is $K_R = 2$.

References

[Gorishniy *et al.*, 2021] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *NeurIPS*, 2021.

[Hua *et al.*, 2021] Weizhe Hua, Yichi Zhang, Chuan Guo, Zhiru Zhang, and G. Edward Suh. Bullettrain: Accelerating robust neural network training via boundary example mining. In *NeurIPS*, 2021.

[Li *et al.*, 2023] Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. Squeeze training for adversarial robustness. In *ICLR*, 2023.

[Tong *et al.*, 2024] Kun Tong, Chengze Jiang, Jie Gui, and Yuan Cao. Taxonomy driven fast adversarial training. In *AAAI*, 2024.