Bag of Tricks for Adversarial Training

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu

Department of Computer Science and Technology, Tsinghua University, Beijing, China

1. Motivation: A Paradox

Zhang et al. (2019): TRADES (weight decay 2×10^{-4}) performs better than **PGD-AT** (weight decay 2×10^{-4});

Rice et al. (2020): PGD-AT (weight decay 5×10^{-4}) performs better than **TRADES** (weight decay 2×10^{-4});

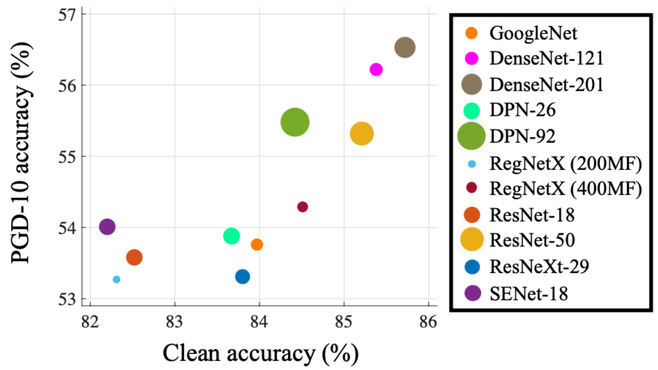
Gowal et al. (2020): TRADES (weight decay 5×10^{-4}) performs better than **PGD-AT** (weight decay 5×10^{-4}).

Usually overlooked training hyperparameters can largely affect the performance of adversarially trained models.

2. Empirical Results

More detailed results can be found in our paper.

2.1. Model Architecture



2.2. Batch Normalization Mode

Table 7: Test accuracy (%) under different **BN modes** on CIFAR-10. We evaluate across several model architectures, since the BN layers have different positions in different models.

	BN mode	Model architecture					
		ResNet-18	SENet-18	DenseNet-121	GoogleNet	DPN26	WRN-34-10
Clean	train	82.52	82.20	85.38	83.97	83.67	86.07
	eval	83.48	84.11	86.33	85.26	84.56	87.38
	-	+0.96	+1.91	+0.95	+1.29	+0.89	+1.31
PGD-10	train	53.58	54.01	56.22	53.76	53.88	56.60
	eval	53.64	53.90	56.11	53.77	53.41	56.04
	-	+0.06	-0.11	-0.11	+0.01	-0.47	-0.56
AA	train	48.51	48.72	51.58	48.73	48.50	52.19
	eval	48.75	48.95	51.24	48.83	48.30	51.93
	-	+0.24	+0.23	-0.34	+0.10	-0.20	-0.26

2.3. Weight Decay

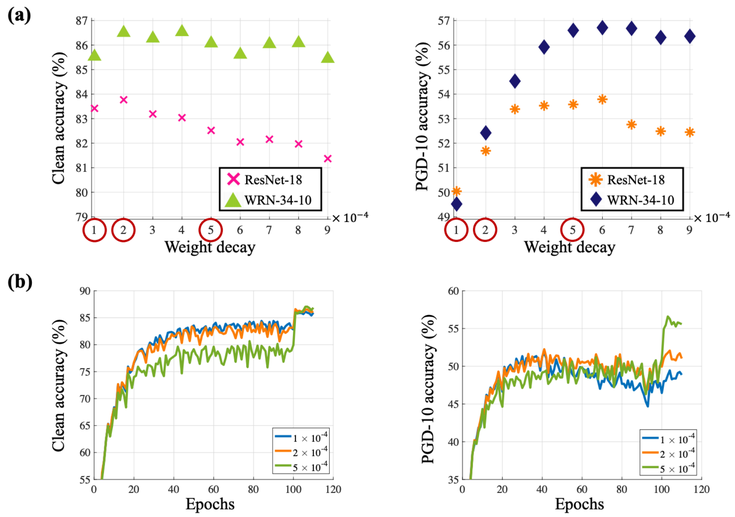


Figure 1: **(a)** Test accuracy w.r.t. different values of **weight decay**. The reported checkpoints correspond to the best PGD-10 accuracy (Rice et al., 2020). We test on two model architectures, and highlight (with red circles) three most commonly used weight decays in previous work; **(b)** Curves of test accuracy w.r.t. training epochs, where the model is WRN-34-10. We set weight decay be 1×10^{-4}, 2×10^{-4}, and 5×10^{-4}, respectively. We can observe that smaller weight decay can learn faster but also more tend to overfit w.r.t. the robust accuracy. In Fig. 4, we early decay the learning rate before the models overfitting, but weight decay of 5×10^{-4} still achieve better robustness.

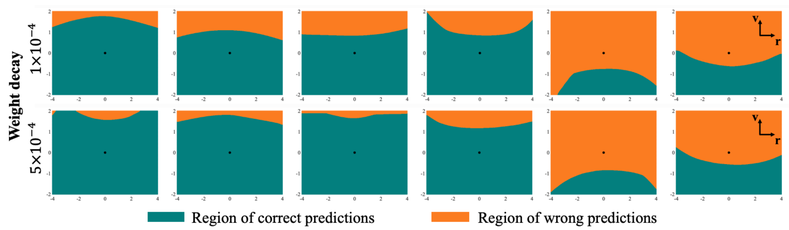


Figure 3: Random normal cross-sections of the decision boundary for PGD-AT with different **weight decay**. The model architecture is WRN-34-10. Following the examples in Moosavi-Dezfooli et al. (2019), we craft PGD-10 perturbation as the normal direction v, and r be a random direction, under the ℓ_∞ constraint of $8/255$. The values of x-axis and y-axis represent the multiplied scale factors.

2.4. Activation Function

Table 6: Test accuracy (%) under different **non-linear activation function** on CIFAR-10. The model is ResNet-18. We apply the hyperparameters recommended by Xie et al. (2020) on ImageNet for the activation function. Here the notation ‡ indicates using weight decay of 5×10^{-5}, where applying weight decay of 5×10^{-4} with these activations will lead to much worse model performance.

	ReLU	Leaky.	ELU ‡	CELU ‡	SELU ‡	GELU	Softplus	Tanh ‡
Clean	82.52	82.11	82.17	81.37	78.88	80.42	**82.80**	80.13
PGD-10	53.58	53.25	52.08	51.37	49.53	52.21	**54.30**	49.12

2.5. Batch Size and Label Smoothing

Table 3: Test accuracy (%) under different **batch size** and **learning rate** (l.r.) on CIFAR-10. The basic l.r. is 0.1, while the scaled l.r. is, e.g., 0.2 for batch size 256, and 0.05 for batch size 64.

ResNet-18				
Batch size	Basic l.r.		Scaled l.r.	
	Clean	PGD-10	Clean	PGD-10
64	80.08	51.31	82.44	52.48
128	82.52	**53.58**	-	-
256	83.33	52.20	82.24	52.52
512	83.40	50.69	82.16	53.36
WRN-34-10				
Batch size	Basic l.r.		Scaled l.r.	
	Clean	PGD-10	Clean	PGD-10
64	84.20	54.69	85.40	54.86
128	86.07	**56.60**	-	-
256	86.21	52.90	85.89	56.09
512	86.29	50.17	86.47	55.49

Table 4: Test accuracy (%) under different degrees of **label smoothing** (LS) on CIFAR-10. More evaluation results under, e.g., PGD-1000 can be found in Table 17.

ResNet-18				
LS	Clean	PGD-10	AA	RayS
0	82.52	53.58	48.51	53.34
0.1	82.69	54.04	48.76	53.71
0.2	82.73	54.22	49.20	53.66
0.3	82.51	54.34	**49.24**	53.59
0.4	82.39	54.13	48.83	53.40
WRN-34-10				
LS	Clean	PGD-10	AA	RayS
0	86.07	56.60	52.19	60.07
0.1	85.96	56.88	52.74	59.99
0.2	86.09	57.31	**53.00**	60.28
0.3	85.99	57.55	52.70	61.00
0.4	86.19	57.63	52.71	60.64

2.6. Early Stopping (attack iter.) and Warmups

Table 2: Test accuracy (%) under different **early stopping** and **warmup** on CIFAR-10. The model is ResNet-18 (results on WRN-34-10 is in Table 14). For early stopping attack iter., we denote, e.g., 40 / 70 as the epochs to increase the tolerance step by one (Zhang et al., 2020). For warmup, the learning rate and the maximal perturbation linearly increase from zero to preset values in 10 / 15 / 20 epochs.

	Base	**Early stopping attack iter.**			**Warmup on l.r.**			**Warmup on perturb.**		
		40 / 70	40 / 100	60 / 100	10	15	20	10	15	20
Clean	82.52	86.52	86.56	85.67	82.45	82.64	82.31	82.64	82.75	82.78
PGD-10	53.58	52.65	53.22	52.90	53.43	53.29	53.35	53.65	53.27	53.62
AA	48.51	46.6	46.04	45.96	48.26	48.12	48.37	48.44	48.17	48.48

2.7. Optimizer

Table 5: Test accuracy (%) using different **optimizers** on CIFAR-10. The model is ResNet-18 (results on WRN-34-10 is in Table 15). The initial learning rate for Adam and AdamW is 0.0001.

	Mom	Nesterov	Adam	AdamW	SGD-GC	SGD-GCC
Clean	82.52	82.83	83.20	81.68	82.77	82.93
PGD-10	53.58	53.78	48.87	46.58	53.62	53.40
AA	48.51	48.22	44.04	42.39	48.33	48.51

Takeaways:
(i) Slightly different values of weight decay could largely affect the robustness of trained models;
(ii) Moderate label smoothing and linear scaling rule on l.r. for different batch sizes are beneficial;
(iii) Applying eval BN mode to craft training adversarial examples can avoid blurring the distribution;
(iv) Early stopping the adversarial steps or perturbation may degenerate worst-case robustness;
(v) Smooth activation benefits more when the model capacity is not enough for adversarial training.

Primary Contact: Tianyu Pang (pty17@mails.tsinghua.edu.cn); Code: https://github.com/P2333/Bag-of-Tricks-for-AT