Adversarial Weight Perturbation Helps Robust Generalization (NeurIPS 2020)

Dongxian Wu[1,3] Shu-Tao Xia[1,3] Yisen Wang[2]

[1]Tsinghua University, [2]Peking University, [2]Peng Cheng Lab

An Introduction to Adversarial Training

The Framework of Adversarial Training

Adversarial training (AT) is the most effective and promising approach to improve robustness against adversarial examples. It incorporates adversarial examples into the training process to solve the following optimization problem,

Adversarial Training (AT)

$$\min_{\mathbf{w}} \rho(\mathbf{w}), \text{ where } \rho(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|\mathbf{x}_i'-\mathbf{x}_i\|_p\leq\epsilon} \ell(\mathbf{f}_{\mathbf{w}}(\mathbf{x}_i'), y_i).$$

The Robust Generalization Gap

However, the robustness achieved by AT is far from satisfactory because of the **huge robust generalization gap**. For example, an adversarially trained PreAct ResNet-18 on CIFAR-10 under L_∞ theat modeelly has 43% test robustness, even it has already achieved 84% training robustness after 200 epochs (See Figure 4).

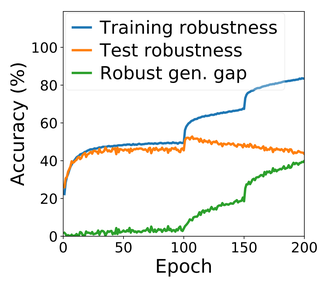


Figure: The learning curve of an adversarial trained PreAct ResNet-18 on CIFAR-10 under L_∞ threat model.

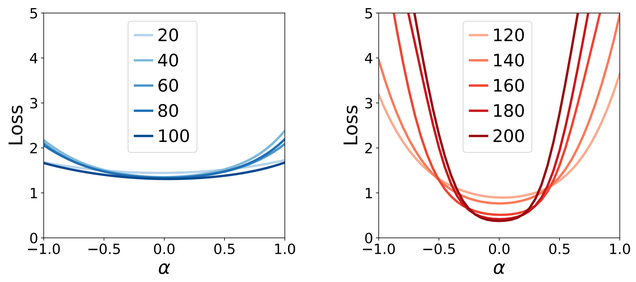
Thus, how to mitigate the robust generalization gap becomes essential for the robustness improvement of adversarial training methods.

Delve into the Weight Loss Landscape

Visualization

we visualization the weight loss landscape $\rho(\mathbf{w}+\alpha\mathbf{d})$ along a random direction $\mathbf{d}$ using adversarial examples generated on-the-fly, and then investigate it from two perspectives:

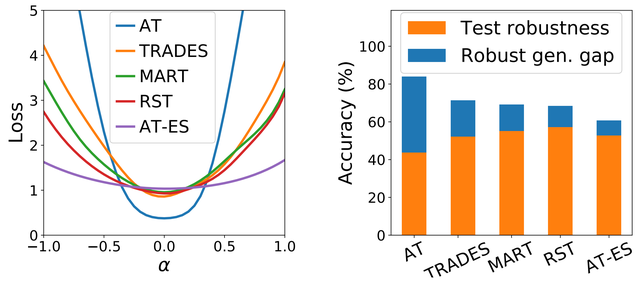
1. The Connection in the Learning Process of Adversarial Training



(a) Optimization (b) Weight perturbation

Figure: The weight loss landscape in the learning process of adversarial training.

2. The Connection across Different Adversarial Training Methods



(a) Optimization (b) Weight perturbation

Figure: The weight loss landscape across Different Adversarial Training Method.

In conclusion, we identify the fact that flatter weight loss landscape often leads to smaller robust generalization gap in adversarial training via characterizing the weight loss landscape using adversarial examples generated on-the-fly.

Adversarial Weight Perturbation

Inspired by the connection, we propose Adversarial Weight Perturbation (AWP) to explicitly flatten the weight loss landscape via injecting the worst-case weight pertrubation into DNNs as following,

AWP-based Adversarial Training (AT-AWP)

$$\min_{\mathbf{w}} \max_{\mathbf{v}\in\mathcal{V}} \frac{1}{n}\sum_{i=1}^{n} \max_{\|\mathbf{x}_i'-\mathbf{x}_i\|_p\leq\epsilon} \ell(\mathbf{f}_{\mathbf{w}+\mathbf{v}}(x_i'), y_i).$$

The Implementation and the Extension

We implement AWP using one extra forward and backward propagation, which introduces little overhead. Besides, AWP is easily extended to other weill-recognized adveresarial training variants, including TRADES, MART and RST, where the only difference is the method-specific adversarial loss.

Experimental Results

The Learning Curves for AT-AWP and Other Methods

We find AWP indeed improves the test robustness of both the best checkpoint and the last checkpoint by a notable margin, which shows its superiority over other weight regularization and data augmentation.

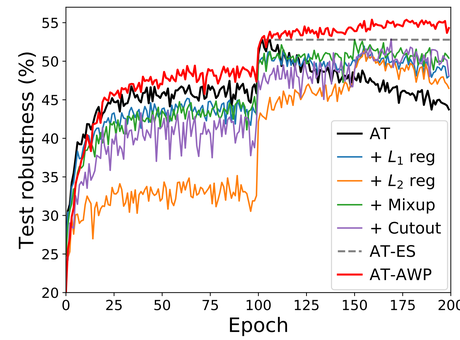


Figure: The learning curve of an adversarial trained PreAct ResNet-18 on CIFAR-10 under L_∞ threat model.

Benchmarking the State-of-the-art Robustness

The robustness improved by AWP is consistent amongst different methods, including currently the strongest attack, Auto-Attack (AA).

Table: Test robustness (%) on CIFAR-10 using WideResNet under L_∞ threat model.

Defense	Natural	PGD-20	PGD-100	CW_∞	AutoAttack
AT	**86.07**	56.10	55.79	54.19	52.60
AT-AWP	85.57	**58.14**	**57.94**	**55.96**	**54.04**
TRADES	84.65	56.33	56.07	54.20	53.08
TRADES-AWP	**85.36**	**59.27**	**59.12**	**57.07**	**56.17**
MART	84.17	58.56	57.88	54.58	51.10
MART-AWP	**84.43**	**60.68**	59.32	**56.37**	**54.23**
Pre-training	87.89	57.37	56.80	55.95	54.92
Pre-training-AWP	**88.33**	**61.40**	**61.21**	**59.28**	**57.39**
RST	**89.69**	62.60	62.22	60.47	59.53
RST-AWP	88.25	**63.73**	**63.58**	**61.62**	**60.05**