《Adversarial Weight Perturbation Helps Robust Generalization》-NIPS2020 文章阅读

[NeurIPS 20]
https://proceedings.neurips.cc/paper_files/paper/2020/file/1ef91c212e30e14bf125e9374262401f-Paper.pdf finds a flatter minima boost robustness against adversarial attacks.

1. 背景介绍

1.1. 对抗训练：

用对抗样本去训练模型，增强模型的抗干扰能力。它是一个min-max问题，内层max用于生成对抗样本，外层min用于训练参数，得到更好的模型。
数学公式：

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{||x_i'-x_i||_p \leqslant \epsilon} L\left(f_\theta\left(x_i'\right), y_i\right).$$

其中，L表示损失函数，f_θ表示模型，x_i表示原始数据样本，y_i表示标签，x_i'表示对抗样本。（扰动有范围限定）
两种常见的求解内层max问题的方法：

- Fast Gradient Sign Method(FGSM)

$$x' = x + \epsilon \cdot \text{sign}\nabla_x L(f_\theta(x), y)$$

- Projected Gradient Descent (PGD) is a iterative version of FGSM

$$x'^{(k+1)} = \prod_{\epsilon} \left(x'^{(k)} + \alpha \cdot \text{sign}\nabla_x L\left(f_\theta\left(x'^{(k)}\right), y\right)\right)$$

1.2. Robust Generalization

Adversarial Training的问题形式可以进行改写：

$$\min_{w} \rho(w), \quad \text{where} \quad \rho(w) = \frac{1}{n} \sum_{i=1}^{n} \max_{||x_i'-x_i||_p \leqslant \epsilon} \ell\left(f_w\left(x_i'\right), y_i\right).$$

不加对抗样本，只考虑正常的训练，得到的training和test之间的gap是很小的，但是对抗训练的结果显示，training和test之间有比较大的generalization gap。
加上max那一步之后，泛化就会出现较大的gap，文章中对gap的来源做了探究，确定了weight loss landscape的平坦性和robust generalization gap（鲁棒泛化性差距）之间的明显关联。并基于此，提出了AWP，在对抗性训练框架中形成一个双扰动机制，对输入和权重进行对抗性扰动，去提升对抗训练的泛化性能。

2. Connection of Weight Loss Landscape and Robust Generalization Gap

论文中使用即时生成的对抗样本来去探索权重损失和鲁棒泛化差距之间的关系，提出了一种新的方法来描述权重损失，并从两个角度来研究它。1）在对抗训练的训练过程中；2）在不同的对抗性训练方法中。
已有研究中的可视化方法:

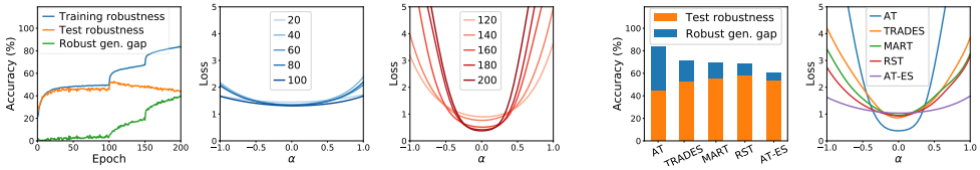
$$g(\alpha) = L(f_{w+\alpha d}(x_i), y_i)$$

（对已经训好的model做一个扰动，然后重新画loss）
若仅将样本换成对抗样本，可视化结果是无法得出有用的结论的，作者认为该问题是由于简单替换的x'是预先生成的，而对抗样本的生成是一个解maximization的过程，该优化问题的求解依赖于model，因此在对model做了weight的扰动后，**需要重新生成对抗样本，不能用预先生成的。**作者给出了他们使用的绘制

loss landscape的方式：

$$g(\alpha) = \rho(\mathbf{w} + \alpha\mathbf{d}) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \le \epsilon} \ell(f_{\mathbf{w}+\alpha\mathbf{d}}(\mathbf{x}_i'), y_i)$$

作者接下来使用该方法进行了一些可视化实验。



(a) In the training process of vanilla AT (*Left*: Learning curve; *Mid*: Landscape before "best"; *Right*: Landscape after "best")

(b) Across different methods (*Left*: Generalization gap; *Right*: Landscape)

Figure 1: The relationship between weight loss landscape and robust generalization gap is investigated a) in the training process of vanilla AT; and b) across different adversarial training methods on CIFAR-10 using PreAct ResNet-18 and L_∞ threat model. ("Landscape" is a abbr. of weight loss landscape)

2.1. The Connection in the Learning Process of Adversarial Training

从图(a)可以发现，"bset"（最高的测试鲁棒性）出现在第103个epoch。在 "best "之前，测试鲁棒性接近于训练鲁棒性，因此鲁棒性泛化差距（绿线）很小。同时，在 "best "之前的权重（每20个epoch绘制）也非常平坦。在 "最佳 "之后，随着训练的继续，鲁棒性泛化差距（绿线）变得更大，而权重损失同时变得更尖锐。因此，在训练过程中，权重损失的平坦程度与鲁棒性泛化差距有很大关系。

2.2. The Connection across Different Adversarial Training Methods

作者还探讨了在不同的对抗训练方法中，权重损失和鲁棒泛化差距之间的关系是否仍然存在。在相同的设置下，使用几种SOTA对抗训练方法来训练。根据图(b)的结果，一种方法实现的泛化差距越小，它的权重损失就越平坦。这一观察结果与训练过程中的观察结果是一致的，它验证了权重损失与鲁棒性泛化差距有很强的关联性。

2.3. Does Flatter Weight Loss Landscape Certainly Lead to Higher Test Robustness?

需要注意的是，更平坦的权重损失虽然直接导致了更小的鲁棒性泛化差距，但只有在训练过程充分（即训练鲁棒性高）的条件下，才有利于最终测试的鲁棒性。因为更小的泛化误差也有可能是training accuracy的下降导致的。

3. Proposed Adversarial Weight Perturbation

对抗性权重扰动（Adversarial Weight Perturbation，AWP），通过向DNN注入最坏情况下的权重扰动来明确地使权重损失更加平坦。为了提高测试鲁棒性，需要关注训练鲁棒性和鲁棒泛化差距（由权重损失的平坦程度表示）。因此，我们有这样的目标：

$$\min_{\mathbf{w}} \{\rho(\mathbf{w}) + (\rho(\mathbf{w}+\mathbf{v}) - \rho(\mathbf{w}))\} \to \min_{\mathbf{w}} \rho(\mathbf{w}+\mathbf{v})$$

Theoretical view:
根据PAC-Bayesian bound，有如下不等式:

$$\mathbb{E}_{\{\mathbf{x}_i, y_i\}_{i=1}^n, \mathbf{u}}[\rho(\mathbf{w}+\mathbf{u})] \le \rho(\mathbf{w}) + \{\mathbb{E}_{\mathbf{u}}[\rho(\mathbf{w}+\mathbf{u})] - \rho(\mathbf{w})\} + 4\sqrt{\frac{1}{n}KL(\mathbf{w}+\mathbf{u}\|P) + \ln\frac{2n}{\delta}}.$$

右边第二项代表了weight landscape的平坦程度，为了将其显示加入目标，作者做了一个放缩（expectation<maximization），将max加入目标，得到以下目标：

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

由此就将min-max问题转化成了min-max-max问题。包括了两层扰动，第一次是对样本进行扰动，另一次是在weight上进行扰动。直觉上理解，第一层，固定model，先对样本做扰动，找到worst case的对抗样本，由于有多个样本，第二层再对θ进行扰动，找到全局上的worst case。

4. Experiments

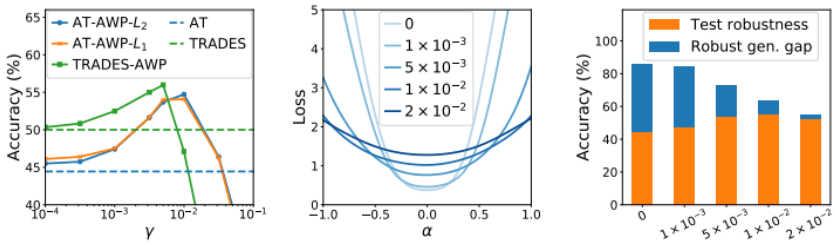
4.1. Benchmark

Table 2: Test robustness (%) on CIFAR-10 using WideResNet under L_∞ threat model. We omit the standard deviations of 5 runs as they are very small ($< 0.40\%$), which hardly effect the results.

Defense	Natural	FGSM	PGD-20	PGD-100	CW_∞	SPSA	AA
AT	**86.07**	61.76	56.10	55.79	54.19	61.40	52.60 ¶
AT-AWP	85.57	**62.90**	**58.14**	**57.94**	**55.96**	**62.65**	**54.04**
TRADES	84.65	61.32	56.33	56.07	54.20	61.10	53.08
TRADES-AWP	**85.36**	**63.49**	**59.27**	**59.12**	**57.07**	**63.85**	**56.17**
MART	84.17	61.61	58.56	57.88	54.58	58.90	51.10
MART-AWP	**84.43**	**63.98**	**60.68**	**59.32**	**56.37**	**62.75**	**54.23**
Pre-training	87.89	63.27	57.37	56.80	55.95	62.55	54.92
Pre-training-AWP	**88.33**	**66.34**	**61.40**	**61.21**	**59.28**	**65.55**	**57.39**
RST	**89.69**	**69.60**	62.60	62.22	60.47	67.60	59.53
RST-AWP	88.25	67.94	**63.73**	**63.58**	**61.62**	**68.72**	**60.05**

不同方法和不同attack下，AWP都对结果有reliable的提升。

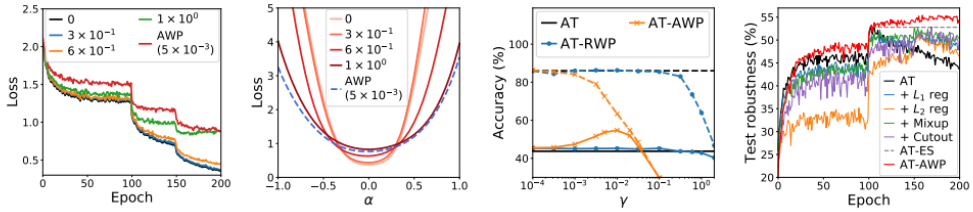
4.2. Ablation Studies on AWP



(b) Weight perturbation (c) Weight loss landscape (d) Generalization gap

从图(b)可以看到γ的选取要适当，且在不同的方法间，有一定的overlap。
从图(c)和图(d)可以发现，当weight perturbation越大时，loss landscape确实越平坦，相应的gap也变小，但是越大的perturbation使得training也变得困难，test accuracy下降明显，gap的减小并非test robustness上升导致。

4.3. Comparisons to Other Regularization Techniques



(a) Loss curve (b) Weight loss landscapes (c) Robustness (d) Learning curve

Figure 3: Comparisons of AWP and other regularization techniques (the values in (a)/(c) legend are γ in RWP unless otherwise specified) on CIFAR-10 using PreAct ResNet-18 and L_∞ threat model.

和RWP对比来看，要达到和AWP相同的效果，RWP需要非常大的perturbation，而AWP所需的

perturbation则小很多。

5. Conclusion

作者使用即时生成的对抗样本来描述weight loss landscape，并发现weight loss landscape与robust generalization gap之间的相关性。并且基于这些发现，作者提出了对抗性权重扰动（Adversarial Weight Perturbation，AWP）来直接使权重损失变得平坦。实验表明，AWP可以在不同的对抗训练方法、网络结构、威胁模型和基准数据集中提高对抗鲁棒性。