Defense on Image Recognition

Todo List

- Muzammal Naseer, Salman Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In 2019 IEEE Winter Conference on Applications ofComputer Vision (WACV), pp. 1300–1307. IEEE, 2019.
- Jamie Hayes. On visible adversarial perturbations & digital watermarking. In Proceedings ofthe IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1597–1604, 2018.
- Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. arXiv preprint arXiv:1802.00420, 2018.
- Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. arXiv preprint arXiv:1810.12715, 2018.
- Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In International Conference on Machine Learning, pp. 3575–3583, 2018.
- Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. arXiv preprint arXiv:1911.09272, 2019.
- Gotta Catch'Em All: Using Honeypots to Catch Adversarial Attacks on Neural Networks

Towards Deep Learning Models Resistant to Adversarial Attacks

> 由于时间原因，该文章的笔记借鉴自 "前人分享"（链接见下）。

Contribution

1. 建模了对抗训练过程；
2. 使用 PGD生成的对抗样本 来做**对抗训练**；

Notes

1. ☆ 问题建模，**从优化的角度来看模型鲁棒性问题。**深度学习中，我们经常根据下面这个目标来训练我们的网络：即我们希望我们训练得到的模型在训练样本**上**的经验损失能够达到最小。

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L(\theta, x, y)\right]$$

但是这样的训练目标，使得模型容易受到对抗样本的攻击。故作者将对抗样本的攻击防御问题总结为以下公式，该问题原文中作者称为**鞍点问题（saddle point problem）**，即我们希望我们训练得到的模型在训练样本**周围**的经验损失能够达到最小。

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y)\right]$$

建模完问题以后，那么以前的对抗样本领域的工作就可以进行简单地分类：（稍微有点绕）

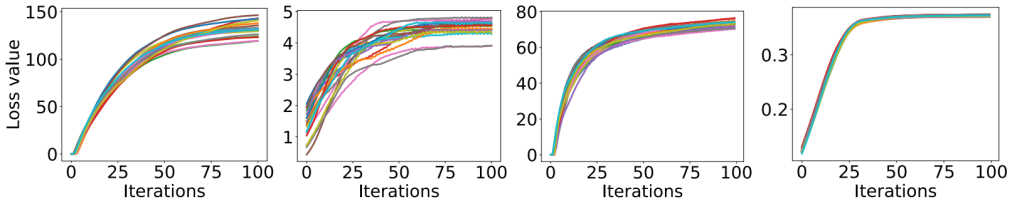
- 提出一个好的对抗攻击算法，来寻找使得（内层）经验损失最大化的扰动；
- 提出一个鲁棒性好的模型，来使得（外层）最小化（内层的最大的）经验损失；

2. 文章中作者采用投影梯度下降算法（PGD）来生成对抗样本：

$$x^{t+1} = \Pi_{x+\mathcal{S}}\left(x^t + \alpha\,\mathrm{sgn}(\nabla_x L(\theta, x, y))\right)$$

3. 实验发现：

(1) Loss 下降趋势和对抗样本算法迭代轮数的关系：无论是原始模型还是使用对抗训练得到的模型，两者使用 PGD算法 生成对抗样本时，随着迭代轮数的上升，样本的 loss 都会上升，且到最后趋于收敛；



(a) MNIST Standard training　(b) MNIST Adversarial training　(c) CIFAR10 Natural training　(d) CIFAR10 Adversarial training

Figure 1: Cross-entropy loss values while creating an adversarial example from the MNIST and CIFAR10 evaluation datasets. The plots show how the loss evolves during 20 runs of projected gradient descent (PGD). Each run starts at a uniformly random point in the ℓ_∞-ball around the same natural example (additional plots for different examples appear in Figure 11). The adversarial loss plateaus after a small number of iterations. The optimization trajectories and final loss values are also fairly clustered, especially on CIFAR10. Moreover, the final loss values on adversarially trained networks are significantly smaller than on their standard counterparts.

(2) Loss 分布的差异：于原始模型相比，在对抗训练得到的模型上生成对抗样本，得到的loss更小，更集中且没有异常值；

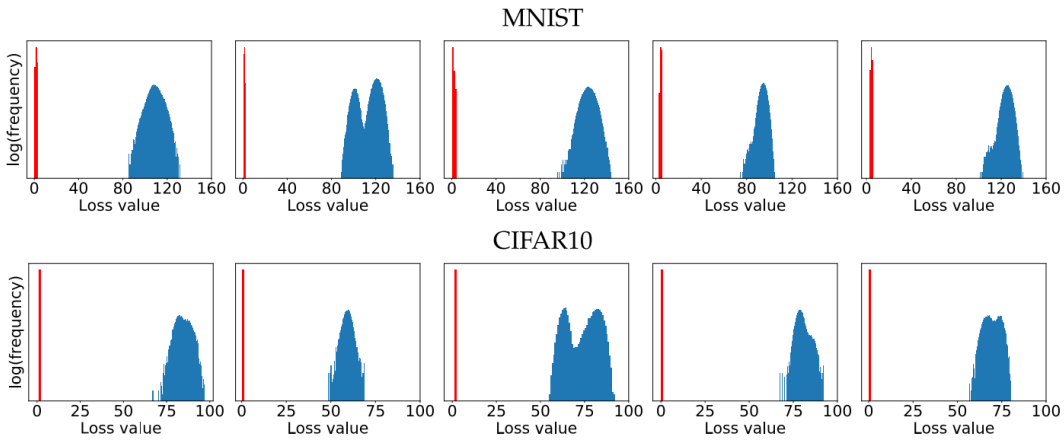
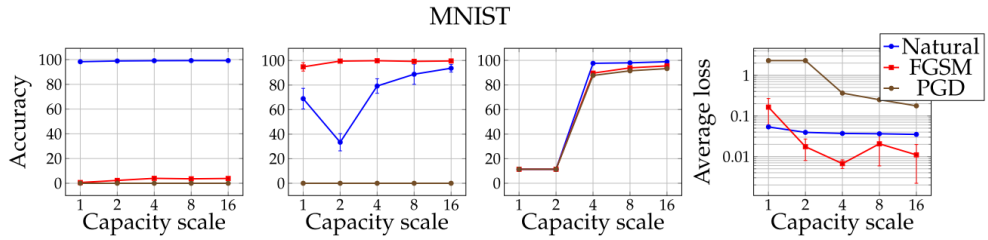


Figure 2: Values of the local maxima given by the cross-entropy loss for five examples from the MNIST and CIFAR10 evaluation datasets. For each example, we start projected gradient descent (PGD) from 10^5 uniformly random points in the ℓ_∞-ball around the example and iterate PGD until the loss plateaus. The blue histogram corresponds to the loss on a standard network, while the red histogram corresponds to the adversarially trained counterpart. The loss is significantly smaller for the adversarially trained networks, and the final loss values are very concentrated without any outliers.

(3) 鲁棒性与模型规模的关系：相对而言，模型越复杂，鲁棒性也越好。同时，经过对抗训练的模型，在原始任务上会有一定的损失，是因为出现了如"过拟合"的现象，使得模型在测试集上面的效果并不好；



CIFAR10

	Simple	Wide
Natural	92.7%	95.2%
FGSM	27.5%	32.7%
PGD	0.8%	3.5%

(a) Standard training

	Simple	Wide
Natural	87.4%	90.3%
FGSM	90.9%	95.1%
PGD	0.0%	0.0%

(b) FGSM training

	Simple	Wide
Natural	79.4%	87.3%
FGSM	51.7%	56.1%
PGD	43.7%	45.8%

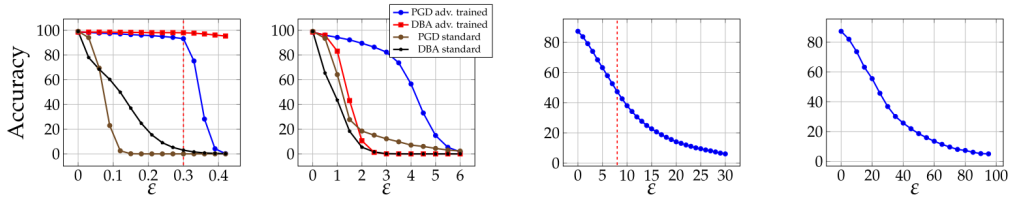
(c) PGD training

	Simple	Wide
Natural	0.00357	0.00371
FGSM	0.0115	0.00557
PGD	1.11	0.0218

(d) Training Loss

Figure 4: The effect of network capacity on the performance of the network. We trained MNIST and CIFAR10 networks of varying capacity on: (a) natural examples, (b) with FGSM-made adversarial examples, (c) with PGD-made adversarial examples. In the first three plots/tables of each dataset, we show how the standard and adversarial accuracy changes with respect to capacity for each training regime. In the final plot/table, we show the value of the cross-entropy loss on the adversarial examples the networks were trained on. This corresponds to the value of our saddle point formulation (2.1) for different sets of allowed perturbations.

(4) 范数限制的影响：l_∞ 范数比 l_2 范数成功的扰动量要小；（这个对比合理吗？）



(a) MNIST, ℓ_∞-norm (b) MNIST, ℓ_2-norm (c) CIFAR10, ℓ_∞-norm (d) CIFAR10, ℓ_2-norm

Figure 6: Performance of our adversarially trained networks against PGD adversaries of different strength. The MNIST and CIFAR10 networks were trained against $\varepsilon = 0.3$ and $\varepsilon = 8$ PGD ℓ_∞ adversaries respectively (the training ε is denoted with a red dashed lines in the ℓ_∞ plots). In the case of the MNIST adversarially trained networks, we also evaluate the performance of the Decision Boundary Attack (DBA) [4] with 2000 steps and PGD on standard and adversarially trained models. We observe that for ε less or equal to the value used during training, the performance is equal or better. For MNIST there is a sharp drop shortly after. Moreover, we observe that the performance of PGD on the MNIST ℓ_2-trained networks is poor and significantly overestimates the robustness of the model. This is potentially due to the threshold filters learned by the model masking the loss gradients (the decision-based attack does not utilize gradients).

Links

- 论文链接：Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.
- 论文代码 - mnist：https://github.com/MadryLab/mnist_challenge
- 论文代码 - cifar10：https://github.com/MadryLab/cifar10_challenge

> 从代码上来看，作者提供的模型的输入是 $32 * 32 * 3$ 大小的，这样经过压缩的输入维度，是否导致了对抗样本算法难以实现呢？或者说生成对抗样本的过程能否利用一下这个特点？

- 论文模型：https://github.com/MadryLab/cifar10_challenge
- 前人分享：https://zhuanlan.zhihu.com/p/45684812

Unlabeled Data Improves Adversarial Robustness

> 证明部分忽略不看

Contribution

1. 利用无标签数据强化模型的鲁棒性；

Notes

1. 训练方法：

Meta-Algorithm 1 Robust self-training

Input: Labeled data $(x_1, y_1, \ldots, x_n, y_n)$ and unlabeled data $(\tilde{x}_1, \ldots, \tilde{x}_{\tilde{n}})$

Parameters: Standard loss L_{standard}, robust loss L_{robust} and unlabeled weight w

1: Learn $\hat{\theta}_{\text{intermediate}}$ by minimizing $\sum_{i=1}^{n} L_{\text{standard}}(\theta, x_i, y_i)$

2: Generate pseudo-labels $\tilde{y}_i = f_{\hat{\theta}_{\text{intermediate}}}(\tilde{x}_i)$ for $i = 1, 2, \ldots \tilde{n}$

3: Learn $\hat{\theta}_{\text{final}}$ by minimizing $\sum_{i=1}^{n} L_{\text{robust}}(\theta, x_i, y_i) + w \sum_{i=1}^{\tilde{n}} L_{\text{robust}}(\theta, \tilde{x}_i, \tilde{y}_i)$

   - 首先使用有标签数据训练网络，这里使用 standard loss 为：

$$L_{\text{standard}}(\theta, x, y) = -\log p_\theta(y \mid x)$$

- 使用训练好的网络，标记无标签的数据；
- 使用有标签和"自标签"的数据继续训练网络，这里使用 robust loss 为：

$$L_{\text{robust}}(\theta, x, y) = L_{\text{standard}}(\theta, x, y) + \beta L_{\text{reg}}(\theta, x),$$
$$\text{where} \quad L_{\text{reg}}(\theta, x) := \max_{x' \in \mathcal{B}_\epsilon^p(x)} D_{\text{KL}}(p_\theta(\cdot \mid x) \,\|\, p_\theta(\cdot \mid x'))$$

这里又到了经典的如何拟合 $L_{reg}(\theta, x)$ 项（因为寻找邻域内的最大值太困难），作者提出了两种方法：

- **Adversarial Training**：使用 PGD 获取邻域最大值

$$L_{\text{reg}}^{\text{adv}}(\theta, x) := D_{\text{KL}}(p_\theta(\cdot \mid x) \,\|\, p_\theta(\cdot \mid x'_{\texttt{PG}}[x]))$$

- **Stability Training**：使用 高斯分布 采样邻域的值

$$L_{\text{reg}}^{\text{stab}}(\theta, x) := \mathbb{E}_{x' \sim \mathcal{N}(x, \sigma^2 I)} D_{\text{KL}}(p_\theta(\cdot \mid x) \,\|\, p_\theta(\cdot \mid x'))$$

使用 Stability Training 的网络在测试的时候也做了改变，模型输出的是 高斯分布 采样邻域中的可能性最大的分类

$$g_\theta(x) := \operatorname*{argmax}_{y \in \mathcal{Y}} \; q_\theta(y \mid x), \quad \text{where} \quad q_\theta(y \mid x) := \mathbb{P}_{x' \sim \mathcal{N}(x, \sigma^2 I)}(f_\theta(x') = y)$$

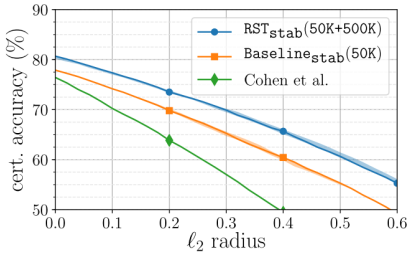
2. 实验：

(1) 经验性防御（heuristic defense）：主要关注 L_∞ 攻击

Model	$\texttt{PG}_{\texttt{Madry}}$	$\texttt{PG}_{\texttt{TRADES}}$	$\texttt{PG}_{\texttt{Ours}}$	CW [7]	Best attack	No attack
$\texttt{RST}_{\texttt{adv}}$(50K+500K)	63.1	63.1	62.5	64.9	**62.5** ±0.1	**89.7** ±0.1
TRADES [56]	55.8	56.6	55.4	65.0	55.4	84.9
Adv. pre-training [18]	57.4	58.2	57.7	-	57.4†	87.1
Madry et al. [29]	45.8	-	-	47.8	45.8	87.3
Standard self-training	-	0.3	0	-	0	96.4

Table 1: **Heuristic defense.** CIFAR-10 test accuracy under different optimization-based ℓ_∞ attacks of magnitude $\epsilon = 8/255$. Robust self-training (RST) with 500K unlabeled Tiny Images outperforms the state-of-the-art robust models in terms of robustness as well as standard accuracy (no attack). Standard self-training with the same data does not provide robustness. †: A projected gradient attack with 1K restarts reduces the accuracy of this model to 52.9%, evaluated on 10% of the test set [18].

(2) 证明性防御（certified defense）：同时关注 L_∞ 和 L_2 攻击



(a)

Model	ℓ_∞ acc. at $\epsilon = \frac{2}{255}$	Standard acc.
$\texttt{RST}_{\texttt{stab}}$(50K+500K)	**63.8** ± 0.5	**80.7** ± 0.3
$\texttt{Baseline}_{\texttt{stab}}$(50K)	58.6 ± 0.4	77.9 ± 0.1
Wong et al. (single) [50]	53.9	68.3
Wong et al. (ensemble) [50]	**63.6**	64.1
IBP [17]	50.0	70.2

(b)

Figure 1: **Certified defense.** Guaranteed CIFAR-10 test accuracy under all ℓ_2 and ℓ_∞ attacks. Stability-based robust self-training with 500K unlabeled Tiny Images ($\texttt{RST}_{\texttt{stab}}$(50K+500K)) outperforms stability training with only labeled data ($\texttt{Baseline}_{\texttt{stab}}$(50K)). **(a)** Accuracy vs. ℓ_2 radius, certified via randomized smoothing [9]. Shaded regions indicate variation across 3 runs. Accuracy at ℓ_2 radius 0.435 implies accuracy at ℓ_∞ radius $2/255$. **(b)** The implied ℓ_∞ certified accuracy is comparable to the state-of-the-art in methods that directly target ℓ_∞ robustness.

Links

- 论文链接：Carmon Y, Raghunathan A, Schmidt L, et al. Unlabeled data improves adversarial robustness[J]. arXiv preprint arXiv:1905.13736, 2019.
- 论文代码：https://github.com/yaircarmon/semisup-adv