Model Robustness

Todo List

1. 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.
2. 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.
3. 解决 Certified Defenses For Adversarial Patches 中的问题;

Certified Defenses For Adversarial Patches

Contribution

1. 文章的亮点在于：将 **IBP（Interval Bound Propagation）** 这种模型鲁棒性方法运用到了防御 **带patch的对抗攻击**中;
2. 有趣的点：文章发现这样训练的模型对不同 patch 的对抗攻击都具有一定的鲁棒性;

Notes

1. Introduction: 图像存在对抗攻击, 实际物理对抗攻击都是带 patch 的攻击, 所以作者想做带 patch 的鲁棒性检测;
2. Problem Setup:

   (1) 定义 **模型的鲁棒性**:

   $$\mathop{\mathbb{E}}_{x\sim X} \min_{p\in\mathbb{P},l\in\mathbb{L}} \mathcal{X}[f(A(x,p,l);\theta)=y]$$

   其中, $\mathbb{L}$ 表示 patch 的位置集合, $\mathbb{P}$ 表示特定的 patch 集合, X 表示样本集合, $\mathcal{X}[\circ]$ 是布尔函数 (若为真返回1, 若为假返回0). 公式的含义是在输入集合上**无法**得到成功的对抗样本的输入数量, 对于攻击者而言, 这个值越小越好, 而对于防御者而言, 这个值越大越好;

   (2) 定义patch: 正方形的 patch, 其中的值可以为任意值 [0~1];

3. Vulnerability of Existing Defenses: 作者借此说明现有的一些依靠经验性结果的对抗样本防御方法 (**针对那些带 patch 的无目标攻击**) 很容易被攻击者绕过.

    - 已有的防御方法:

      (1) 防御原理: **输入图像的损失函数的梯度值在添加对抗扰动的地方都会很大** . Watermarking 方法基于这个原理来防御无目标攻击 (值得关注的是, 这种方法对模型的成功率的影响约为 12%) ;

      (2) 防御原理: **输入图像的像素值在这些添加了对抗扰动的地方会变化很大, 即不连续** . Local Gradient Smoothing 方法基于这个原理, 对输入作预处理;

      $$\hat{x} = x \odot (1 - \lambda g(x))$$

    - 已有的防御方法容易被绕过:

		Patch Size		
Attack	Defense	42×42	52×52	60×60
IFGSM	LGS	78%	75%	71%
IFGSM + LGS	LGS	14%	5%	3%
IFGSM	DW	56%	49%	45%
IFGSM + DW	DW	13%	8%	5%

    在生成对抗样本的过程中即引入防御算法, 就可以绕过这些防御方法.

4. Certified Defenses:

    (1) 在输入的 $p - norm$ 邻域内, 检查模型的分类结果是否会发生改变, 如果改变, 则不鲁棒, 可能是一个对抗样本; 如果不改变, 则鲁棒, 不可能是一个对抗样本;

    (2) 检查输入是否是鲁棒的是十分苦难的, **因为计算复杂性高, 是一个 NP-Hard 问题**;

    (3) 鲁棒性的边界是十分宽松的, **即只有在很小一个邻域内, 才能保证模型的鲁棒性**;

    (4) Interval Bound Propagation (IBP) 的原理: 想象一下, 你有一些**相互独立的**且已知取值区间的变量 x, y, z , 有一个这些变量组成的线性表达式 $f = ax + by + cz$, 现在你想求这个表达式的取值区间, 那么很简单, 你让表达式的每一项最大即可得到最大值, 让每一项最小即可得到最小值. (这种方法是一种放缩求解的方法, 因为给定了**变量相互独立**这个条件, 但实际神经网络中, 中间层的神经元的取值之间存在某种线性关系) 看具体的公式:

    - 只考虑仿射变换的情况下:

      $$\overline{z}^{(k)} = W^{(k)} \frac{\overline{z}^{(k-1)} + \underline{z}^{(k-1)}}{2} + |W^{(k)}| \frac{\overline{z}^{(k-1)} - \underline{z}^{(k-1)}}{2} + b^{(k)}$$
      $$\underline{z}^{(k)} = W^{(k)} \frac{\overline{z}^{(k-1)} + \underline{z}^{(k-1)}}{2} - |W^{(k)}| \frac{\overline{z}^{(k-1)} - \underline{z}^{(k-1)}}{2} + b^{(k)}$$

      这个公式不太好看, 自己去推一下, 结果是这样的:

      $$\overline{z}^{k} = W^{(k)+} \overline{z}^{(k-1)} + W^{(k)-} \underline{z}^{(k-1)} + b^{(k)}$$

      其中 $W^{(k)+}$ 是将 $W^{(k)}$ 中小于 0 的部分置为0, $W^{(k)-}$ 将大于0的部分置为0, 这样就和上面原理部分相对应;

    - 考虑激活函数: 那就在外面套个激活函数, 考虑一下激活函数的**单调性**即可;

    - 鲁棒性验证: 最后每一个分类都可以得到一个取值区间, 保证目标分类的最小值 大于其他分类的最大值即可, 作者用如下公式

      $$\underline{\mathbf{m}}_y = e_{y_{true}}^T \underline{z}^{(K)} - e_y^T \overline{z}^{(K)} = \underline{z}_{y_{true}}^{(K)} - \overline{z}_y^{(K)} \geq 0 \quad \forall y$$

    - 鲁棒性训练, 来增强 IBP 的鲁棒性:

        - 修改 loss 函数, 从原来的希望目标分类尽可能大, 到添加扰动后的概率区间更满足鲁棒性 (默认包含了分类正确):

          $$\text{Certificate Loss} = \text{Cross Entropy Loss}(-\underline{\mathbf{m}}, y)$$

        - 训练的 trick: (**todo: 没看懂**)

(5) ☆ 作者提出的方法: 在 IBP 的基础上, 把扰动的区间限制在一个**矩形范围**内, 并把扰动的大小限制在 $[0,1]$. 但是 IBP 方法自身的复杂性，再加上需要考虑每一个位置的 Patch，**计算复杂性过大**，所以作者提出了**两种挑选 Patch 位置的方法**；

- 基本框架：选择patch的形状大小后，在图像的各个位置上进行遍历，分析最坏情况；
  - 鲁棒性验证：

$$\underline{\mathbf{m}}^{\text{es}}(\mathbb{L})_y = \min_{l \in \mathbb{L}} \underline{\mathbf{m}}^{\text{single patch}}(l)_y \quad \forall y.$$

  - 鲁棒性训练：

$$\text{Certificate Loss} = \text{Cross Entropy Loss}(-\underline{\mathbf{m}}^{\text{es}}(\mathbb{L}), y).$$

    其中 $\mathbb{L}$ 表示 图片中放patch位置的集合；
- Random Patch Certificate Training: **随机选择一些 patch 的位置**, 随机选择一些可能的 patch 取值;
  - 鲁棒性验证：

$$\underline{\mathbf{m}}^{\text{random patches}}(\mathbb{L})_y = \underline{\mathbf{m}}^{\text{es}}(S)_y$$

  - 鲁棒性训练：

$$\text{Random Patch Certificate Loss} = \text{Cross Entropy Loss}(-\underline{\mathbf{m}}^{\text{random patches}}(\mathbb{L}), y)$$

- Guided Patch Certificate Training：使用 **U-Net 网络**为每个分类挑选一个 patch；
- Defense against Sparse Attack：作者指出，可以将 IBP 第一层的区间公式改成如下形式 （作者直接放这个式子真的是玄学，猜了半天猜出大概的思路，因为 Sparse Attack 就是将几个离散的对模型结果影响最大的像素点的值（值域为 $0 \sim 1$）进行修改，而不同像素点对结果的影响主要还是从权重矩阵中体现出来，所以**会修改权重矩阵中的** $top-k$ **。至于为什么可以直接 加/减** $top-k$ **呢**？因为是值域限定了，我最多也就是把 $0 \to 1$，或者是 $1 \to 0$，所以这个区间一定是满足全部的可能性的。至于**为什么没有了偏置项** b **呢**？我觉得很可能是两种写法，一种是将偏置项单独写出来，另一种写法是将偏置项包含在了权重项 W 中 ）

$$\overline{z}_i^{(1)} = W_{i,:}^{(1)} z^{(0)} + |W_{i,:}^{(1)}|_{top_k} \qquad \underline{z}_i^{(1)} = W_{i,:}^{(1)} z^{(0)} - |W_{i,:}^{(1)}|_{top_k} \qquad \forall i$$

5. 实验结果：

(1) 鲁棒性验证——防御攻击的效果：

Table 2: Comparison of our IBP certified patch defense against existing defenses. Empirical adversarial accuracy is calculated for 400 random images in both datasets. All results are averaged over three different models.

Dataset	Patch Size	Adversary	Defense	Clean Accuracy	Empirical Adversarial Accuracy	Certified Accuracy
MNIST	2×2	IFGSM	None	98.4%	80.1%	-
	2×2	IFGSM	LGS	97.4%	90.0%	-
	2×2	IFGSM + LGS	LGS	97.4%	60.7%	-
	2×2	IFGSM	IBP	98.5%	93.9%	91.6%
	5×5	IFGSM	None	98.5%	3.3%	-
	5×5	IFGSM	IBP	92.9%	66.1%	62.0%
CIFAR	2×2	IFGSM	None	66.3%	25.4%	-
	2×2	IFGSM	LGS	64.9%	31.3%	-
	2×2	IFGSM + LGS	LGS	64.9%	24.2%	-
	2×2	IFGSM	DW	47.1%	43.3%	-
	2×2	IFGSM + DW	DW	47.1%	20.2%	-
	2×2	IFGSM	IBP	48.6%	45.2%	41.6%
	5×5	IFGSM	None	66.5%	0.4%	-
	5×5	IFGSM	LGS	51.2%	22.11%	-
	5×5	IFGSM + LGS	LGS	51.2%	0.5%	-
	5×5	IFGSM	DW	45.3%	59.3%	-
	5×5	IFGSM + DW	DW	45.3%	15.6%	-
	5×5	IFGSM	IBP	33.9%	29.1%	24.9%

(2) 鲁棒性训练——防御攻击的效果：使用的 Patch 越多，放置的位置越多，得到的模型效果越好；

Table 3: Trade-off between certified accuracy and training time for different strategies. The numbers next to training strategies indicate the number of patches used for estimating the lower bound during training. Most training times are measured on a single 2080Ti GPU, with the exception of all-patch training which is run on four 2080Ti GPUs. For that specific case, the training time is multiplied by 4 for fair comparison. See Appendix A.6 for more detailed statistics. *indicates the performance of the best performing large model trained with either random or guided patch. Detailed performance of the large models can be found in Appendix A.5

		2×2			5×5		
Dataset	Training Strategy	Clean Accuracy	Certified Accuracy	Training Time(h)	Clean Accuracy	Certified Accuracy	Training Time(h)
MNIST	All Patch	98.5%	91.5%	9.3	92.0%	60.4%	8.4
	Random(1)	98.5%	82.9%	0.2	96.9%	24.1%	0.4
	Random(5)	98.6%	86.6%	0.3	95.8%	42.1%	0.3
	Random(10)	98.6%	87.7%	0.3	95.6%	49.6%	0.3
	Guided(10)	98.6%	88.9%	2.2	95.0%	53.1%	2.6
CIFAR	All Patch	50.9%	39.9%	56.4	33.5%	22.0%	45.8
	Random(1)	53.6%	21.6%	0.6	43.6%	6.1%	0.6
	Random(5)	52.9%	32.3%	0.7	39.0%	14.6%	0.7
	Random(10)	51.9%	35.6%	0.8	38.8%	18.6%	0.8
	Guided(10)	52.4%	36.0%	3.7	37.9%	18.8%	3.7
	Large Model*	**65.8%**	**51.9%**	22.4	**47.8%**	**30.3%**	15.4

(3) 对于 Sparse Attack 的防御效果：

Table 4: Certified accuracy for sparse defenses with IBP and Random Ablation.

Dataset	Sparsity (k)	Model	Clean Accuracy	Certified Accuracy
MNIST	1	IBP-sparse	98.4%	96.0%
	4	IBP-sparse	97.8%	90.8%
	10	IBP-sparse	95.2%	86.8%
	1	Random Ablation	96.7%	90.3%
	4	Random Ablation	96.7%	79.1%
	10	Random Ablation	96.7%	29.2%
CIFAR	1	IBP-sparse	48.4%	40.0%
	4	IBP-sparse	42.2%	31.2%
	10	IBP-sparse	37.0%	25.6%
	1	Random Ablation	78.3%	68.6%
	4	Random Ablation	78.3%	61.3%
	10	Random Ablation	78.3%	45.0%

(4) 用矩形 Patch 训练的模型能否防御其他形状的 Patch？**这个还挺有趣的，方法是存在一定的 Transferability 的**；

Table 5: Certified accuracy for square-patch trained model for different shapes

Dataset	Pixel Count	Square	Rectangle	Line	Diamond	Parallelogram
MNIST	4	91.6%	-	92.5%	91.6%	92.3%
	16	69.4%	55.4%	46.7%	68.13%	70.2%
	25	59.7%	50.9%	32.4%	53.6%	55.2%
CIFAR	4	50.8%	-	46.1%	48.6%	49.8%
	16	36.9%	29.0%	32.1%	35.7%	36.3%
	25	30.3%	25.1%	29.0%	30.1%	30.7%

Questions

1. 文章中提到的 U-Net 网络怎么用在实际的工作中?
2. 模型鲁棒性验证法是如何进行实验的?
3. 实验结果中的 Certified Accuracy 是如何计算出来的?
4. 鲁棒性训练里面的Trick没有看懂?

Links

- 论文链接: Chiang, Ping-yeh, et al. "Certified defenses for adversarial patches." *ICLR* (2020).
- 源码链接: Ping-C / certifiedpatchdefense
- $p-norm$ 详解：知乎 / 0范数, 1 范数, 2范数有什么区别?
- U-Net 网络详解：图像语义分割入门+FCN/U-Net网络解析