Adversarial Defenses

2023.8.15

Adversarial Attacks & Defenses

Adversarial Attack Goal

$$\max_{\rho \in \Omega} L(x + \rho, y; \theta)$$

Adversarial Attacks

L-BFGS (Szegedy et al., 2014)

FGSM (Goodfellow et al., 2015)

PGD (Madry et al., 2019)

Adversarial Defense Goal

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\rho \in \Omega} L(x + \rho, y; \theta)\right]$$

Adversarial Defenses

- Adversarial Training
- Gradient Masking/Regularization
- Detecting
- Transformation
- Certified Defense

Adversarial Training

👍 可以有效防御对抗样本 state-of-the-art

👎 对抗训练计算开销过大；依赖于训练所使用的对抗样本类型，难以防御其他更强大的攻击

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\left[\max_{\boldsymbol{\rho}\in\boldsymbol{\Omega}} \boldsymbol{L}(\boldsymbol{x}+\boldsymbol{\rho},\boldsymbol{y};\boldsymbol{\theta})\right]$$

💡 构造对抗样本近似max问题，模型训练解决min问题

Adversarial Training

FGSM AdvTraining

(Goodfellow et al., 2015)

- 添加FGSM样本batch交替训练
- 防御单步攻击
- 对多步攻击（PGD, BIM）脆弱

Strong AdvTraining

(Huang et al., 2016)

- 添加Strong对抗样本训练
  - strong advs指不同Lp范数下的最小扰动样本（FGSM）
  - strong advs指扰动小于某常数的范围内最大化损失的样本（PGD）
- 在MNIST和CIFAR10上达到比FGSM AdvTraining更好的效果
- 对多步攻击（PGD, BIM）脆弱

Scale AdvTraining

(Kurakin et al., 2017)

- 使用单步对抗样本进行训练，batch内替换部分样本为对抗样本
- 在大型数据集ImageNet上可以进行训练防御单步对抗攻击
- 发现FGSM（单步）对抗训练可能发生label leaking，即对抗样本上的精度比干净样本上更高

Adversarial Training

PGD AdvTraining

(Madry et al., 2019)

仅使用PGD对抗样本进行训练

在MNIST和CIFAR10可以有效防御强大的对抗样本C&W state-of-the-art

PGD AdvTraining是目前效果较好的对抗防御方法，可以在白盒场景下有效防御强大的对抗攻击

PGD AdvTraining主要存在三个方面的不足，后续的工作是对这些不足的优化

- 计算成本高时间长，在大型数据集不适用
- 具有白盒安全性，但黑盒安全性提升不高
- 在干净样本上的精度下降（label-leaking），存在 accuracy & robustness trade-off

Adversarial Training: Cost

提高对抗训练效率的方法

Name	Method	Performance
Stability Training (Zheng et al., 2016)	使用噪声样本而非对抗样本进行对抗训练	• 提高网络输出的稳定性 • 但对精心构造的强大对抗样本鲁棒性不高
Free AdvTraining (Shafahi et al., 2019)	利用模型更新时的梯度构造对抗样本	• 较小的计算开销达到PGD AdvTraining的鲁棒性
Fast AdvTraining (Wong et al., 2020)	对FGSM对抗样本进行微调 添加随机初始化	• 较小的计算开销达到PGD AdvTraining的鲁棒性 • 相较于Free AdvTraining速度更快
Transferable AdvTraining（ATTA） (Zheng et al., 2020)	不同轮的对抗样本可以移植 使用移植的对抗样本训练	• 相较于PGD训练速度更快

Adversarial Training: Black-box

提高黑盒安全的方法

（一般除对抗训练以外的方法都有较好的黑盒安全性）

Name	Method	Performance
Ensemble AdvTraining (Tramèr et al., 2020)	利用多个静态预训练模型的对抗样本进行对抗训练	• 训练得到的模型具有黑盒安全性 • 是一种较为有效的对抗训练方法
ALP AdvTraining (Kannan et al., 2018)	训练过程中： • ALP（Adversarial Logit Pairing）使干净样本的输出与对抗样本的输出相似 • CLP（Clean Logit Pairing）使任意输入样本的输出相似	• ALP与AdvTraining结合可以提供白盒+黑盒安全性 • CLP可以以很低的计算开销防御PGD黑盒攻击

Adversarial Training: Trade-off

提高accuracy或robustness的方法

Name	Method	Performance
Security (Papernot et al., 2016)	/	• 探究了robustness和accuracy的trade-off
Unlabeled Training (Carmon et al., 2019)	使用半监督学习的方法，利用unlabeled样本训练模型	• 得到的模型更加鲁棒且精度更高
TRADES (Zhang et al., 2019)	构建loss包含两个部分：一个部分最小化经验损失，一部分最大化鲁棒性	• 描述了robustness和accuracy之间的trade-off • 提出了trade robustness off against accuracy
Friendly AdvTraining (Zhang et al., 2020)	在搜寻PGD对抗样本时早停，寻找“least adversarial data”在确保干净样本正确的前提下 进行对抗训练	• 确保精度的前提下，提高鲁棒性

Gradient Masking/Regularization

👍 防御基于梯度的对抗样本，对各种攻击均有效（attack-agnostic）

👎 (Athalye et al., 2018)提出多种混淆梯度方法只提供False Sense of Security，提出新的攻击方法在白盒（敌手知道网络和防御）场景下成功攻击混淆梯度防御

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\left[\max_{\boldsymbol{\rho}\in\boldsymbol{\Omega}} \boldsymbol{L}(\boldsymbol{x}+\boldsymbol{\rho},\boldsymbol{y};\boldsymbol{\theta})\right]$$

💡对于内层max问题，考虑扰动空间内所有噪声样本，训练网络使较小的输入变化不会导致较大的输出变化，提高网络鲁棒性

Gradient Masking/Regularization: GradMasking

混淆梯度的方法

Name	Method	Performance
Defensive Distillation (Papernot et al., 2016)	网络蒸馏	• 有效防御对抗样本（基于梯度的单步攻击FGSM）
bounded ReLU (Zantedeschi et al., 2017)	使用bounded ReLU 并且进行Gaussian Data Augmentation	• 对该模型的对抗样本的扰动更大（FGSM, JSMA, C&W）
Thermometer Encoding (Buckman et al., 2018)	不同轮的对抗样本可以移植 使用移植的对抗样本训练	• 显著提高模型鲁棒性（FGSM, PGD）
SAP (Dhillon et al., 2018)	Stochastic Activation Pruning (SAP) 对激活值较小的进行剪枝 与dropout类似，但 SAP会优先保留激活值较大的神经元	• SAP或SAP + AdvTraining可以提供高鲁棒性（FGSM）
ALP (Kannan et al., 2018)	训练过程中，ALP使干净样本的输出与对抗样本的输出相似	• ALP可以提供高鲁棒性（PGD） • 训练开销低，故可以运用在ImageNet上

Gradient Masking/Regularization: Regularization

正则化的方法

Name	Method	Performance
DCN (Gu and Rigazio, 2015)	Deep Contractive Network的训练添加隐层对输入的Jocabian矩阵的L2范数作为正则化项	• 提高模型鲁棒性（L-BFGS对抗扰动变大） • 精度不减
Parseval Networks (Cisse et al., 2017)	Lipschitz正则化（限制 线性层/卷积层/聚合层的Lipschitz常数 小于1）	• 训练开销更小 • 鲁棒性提升（FGSM）
Cross-Lipschitz regularization (Hein and Andriushchenko, 2017)	使分类器在在不同样本点上的差异尽可能恒定	• 可以提升provable robustness（防御任意攻击） • 精度不减
Jacobian Regularization (Jakubovitz and Giryes, 2018)	使用雅可比正则化（Frobenius norm of the Jacobian of the network）	• 提高模型鲁棒性（Deepfool, FGSM, JSMA） • 相较于Cross-Lipschitz regularization计算开销更小
Grad Regularization (Ross and Doshi-Velez, 2018)	进行梯度正则化，与L2范数的对抗训练等同	• 提高模型鲁棒性（Deepfool, FGSM, JSMA）

Detecting

👍 不对原模型进行修改，轻量级

👎 (Carlini and Wagner, 2017)提出在白盒（敌手了解网络和防御）场景下检测防御失效

Category： Metric based, Detector, Denoiser, Prediction inconsistency

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\rho\in\Omega} L(x+\rho, y; \theta)\right]$$

💡仅对对抗样本进行检测并拒绝

Detecting: Metric based

基于统计规律的方法

（检测输入/激活值的分布，依赖于模型；基于假设：对抗样本与干净样本的**输入/激活值**分布不同）

Name	Method	Performance
OpenMax (Bendale and Boult, 2016)	检测全连接层输出，使用meta-recognition确定样本属于unkown class的概率	• 可以检测到大部分的对抗样本 • 主要用于检测远离输入分布的样本（rubbish）并非针对对抗样本
(Statistical) Detection (Grosse et al., 2017)	可以通过计算统计规律进行检测statistical testing	• 可以区分干净和对抗样本 • 需要大量的suspicious inputs
Convolutional Filter Statistics (Li and Li, 2017)	基于卷积层输出的统计规律训练分类器	• 将对抗样本和干净样本进行分类 • 针对一种对抗样本训练的分类器 可以有效分类其他的对抗样本
Density/Bayesian uncertainty estimates (Feinman et al., 2017)	计算Density estimates/Bayesian uncertainty estimates检测对抗样本	• 不依赖于攻击方法，可以防御未知的攻击 • (Song et al., 2018)提出该方法依赖于模型

Detecting: Metric based

基于统计规律的方法

Name	Method	Performance
SafetyNet (Lu et al., 2017)	添加RBF-SVM检测激活值（Type I可以被SVM检测）通过分类置信度检测对抗样本（Type II 通过检测 次高confidence/最高confidence拒绝一部分对抗样本）	• 相较于Binary Detector • 该检测方法很难被突破，可以防御不同的攻击
Baseline (Grosse et al., 2017)	检测softmax输出的统计规律	• 作为检测的baseline
LID (Ma et al., 2018)	对抗扰动会影响对抗区域的LID（Local Intrinsic Dimensionality）特征，可以使用LID区分一些先进的方法生成的对抗样本	• 检测FGM, BIM, JSMA对抗样本 • 不同对抗样本会导致相似的LID扭曲
GDA (Feinman et al., 2017)	选择最后一个隐层输出作为特征向量，使用GDA（Gaussian discriminant analysis）建立各个类别的均值和方差，利用该均值和方差计算马氏距离进行检测	• 可以检测out-of-distribution和对抗样本
Features Distance Spaces (Carrara et al., 2019)	训练LSTM根据隐层激活值 + 前向传播的过程evolution进行检测	• 检测FGSM, JSMA, Deepfool对抗样本 • 依赖于模型

Detecting: Detector

分类器的方法

（训练二分类/多分类器检测对抗样本，需要生成大量对抗样本）

Name	Method	Performance
Classifier (Nguyen et al., 2015)	训练分类器分类对抗样本。将对抗样本作为新的一类（n+1）	• 依赖于攻击方法泛化性能差 • 需要生成大量对抗样本计算开销大
Not Twin (Gong and Wang, 2017)	训练二分类器 分类干净和对抗样本	• 分类准确率高于99% 很难构造对抗样本逃避分类器检测 • 计算成本过高
Binary Detector (Metzen et al., 2017)	基于原始网络生成对抗样本，训练二分类器检测对抗样本	• 区分干净/对抗样本 • 需要生成同等量的对抗样本，仅能防御similar and weaker adversaries

Detecting: Denoiser

基于输入分布的方法

（模拟输入的分布进行检测，进一步可以denoise，需要大量干净数据，不依赖于模型和攻击方法）

Name	Method	Performance
MagNet (Meng and Chen, 2017)	通过模拟输入数据的分布规律（logits）检测对抗样本	• 有效检测对抗样本 • 不修改网络，不依赖于攻击方法
PixelDefend (Song et al., 2018)	PixelCNN模拟输入分布，通过PixelDefend输出样本的置信度检测对抗样本	• 有效检测对抗样本 • 不影响原模型，不依赖于攻击方法

Detecting: Prediction inconsistency

基于预测不一致性的方法

Name	Method	Performance
Feature Squeezing (Xu et al., 2018)	比较原样本和squeezed的输出，如果差距过大可能是对抗样本 （feature squeeze方法：reducing the color bit depth of each pixel/spatial smoothing）	• 可以以高准确率和低假阳率（高召回） 检测到对抗样本 • 可以与其他防御方法集成

Transformation

👍 不修改原模型，不依赖于攻击方法故可以防御未知的攻击

👎 使用生成式模型训练开销大

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\rho\in\Omega} L(x+\rho, y; \theta)\right]$$

💡 将对抗样本进行变形，去除对抗扰动

Transformation

基于生成式模型的方法

Name	Method	Performance
MagNet (Meng and Chen, 2017)	模拟输入数据的分布规律，利用reformer将对抗样本移到干净样本的manifold	• 训练开销大
Defense-GAN (Samangouei et al., 2018)	利用GAN 建模干净样本的分布，找到与对抗输入相近的干净样本进行分类	• 不影响本身模型训练过程，不依赖于攻击方法 • 训练开销很大
PixelDefend (Song et al., 2018)	PixelCNN建模输入分布，将对抗样本移动到输入的分布中	• 不影响原模型 不依赖于攻击方法 • 生成模型训练开销大

Transformation

输入变形的方法

Name	Method	Performance
Autoencoder (Gu and Rigazio, 2015)	设计denoising autoencoders (DAEs)可以去除大部分的对抗噪声	• DAE具有泛化性，可以防御未知的对抗眼样本 • 但是将DAE添加至DNN 可以以更小的扰动构造对抗样本（白盒安全性低）
Input Transformation (Guo et al., 2018)	对输入进行各类变形，去除对抗噪声，并保留足够的信息正确分类	• 干净样本仍可以正确分类 • 敌手知道防御措施的白盒场景下仍旧有效
HGD (Hein and Andriushchenko, 2017)	High-Level Representation Guided Denoiser 利用对抗样本和干净样本的high-level representation（logits）的差异作为loss 训练denoiser	• 黑盒/白盒安全，训练集合更小，具有泛化性 • 需要生成对抗样本
Randomization (Xie et al., 2018)	进行随机化的变形破坏对抗扰动 方法：random resizing/random padding	• 计算开销小，可以结合对抗训练 • 黑盒/白盒下有效防御单步/多步攻击
Feature Denoising (Xie et al., 2019)	设计一个CNN架构，添加feature denoise模块，去除特征图中的对抗扰动	• 白盒场景防御PGD，提供黑盒安全性 • 需要端到端训练

Certified Defense

👍 对网络的鲁棒性给出证明（certificate），证明在一定范围内（Lp norm）的扰动不会造成网络误分类，故可以防御所有攻击

Category： convex optimization/SMT

Certificate: $\epsilon = 0.1$ （pixel） 35% error [MNIST]表示：在MNIST上，对抗扰动大小不超过$\epsilon = 0.1$ （pixel）导致最多35% error

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\rho\in\Omega} L(x+\rho, y; \theta)\right]$$

💡 对于内部max问题，计算Loss的上界

Certified Defense: convex optimization

凸优化的方法

（使用凸优化寻找损失的上界，利用对偶问题进行可证实防御的训练）

Name	Method	Performance
Formal Guarantees (Hein and Andriushchenko, 2017)	/	• 首次提出：寻找特定样本可以造成误分类的扰动下界（lower bound）
Semidefinite relaxation (Raghunathan et al., 2018)	使用Semidefinite relaxation方法计算worst-case loss的上界，最小化上界进行训练得到可证实的防御	• certificate: $\epsilon = 0.1$ （pixel） 35% error MNIST • 仅能证明单隐层网络
Semidefinite relaxation (Raghunathan et al., 2018)	使用SDP证明了更tight的上界	• 相较于(Raghunathan et al., 2018)可以证实更多隐层的其他网络
Fast-Lin, Fast-Lip (Weng et al., 2018)	**lin**ear approximation on the ReLU units/bounding network local **Lip**schitz constant	• 对于小网络和大网络都显著更快 • 可以证实7层10000神经元的网络

Certified Defense: convex optimization

凸优化的方法

Name	Method	Performance
CLEVER (Weng et al., 2018)	CLEVER（Cross Lipschitz Extreme Value for nEtwork Robustness）利用极值理论提高效率	• 适用于任何网络 对大型网络计算上可行
Adversarial Polytope (Wong and Kolter, 2018)	考虑激活值的convex outer approximation并进行鲁棒优化 linear program的过程的对偶问题与反向传播类似，故可以高效的进行优化	• 仅多进行几次前向和反向传播可以进行可证实防御 • certificate: L∞<0.1 5.8% error [MNIST]
Scaling (Wong et al., 2018)	相较于Adversarial Polytope提升：skip connections/general nonlinearities以适用于大型网络	• certificate: L∞<0.1 3.1% error [MNIST] • certificate: L∞<2/255 36.4% error [CIFAR10]
CROWN (Zhang et al., 2018)	bounding a given activation function with linear and quadratic functions	• 可以证实任意激活函数网络 • tighter bounds （相较于Fast-Lin）
COLT (Balunovic and Vechev, 2019)	结合对抗训练和可证实防御 网络训练过程包括：verifier进行网络证实 adversary进行对抗	• certificate: L∞<2/255 39.5% error [CIFAR10]

Certified Defense: convex optimization

凸优化的方法

Name	Method	Performance
Random Smoothing (Cohen et al., 2019)	利用Gaussian noise将任意的base网络转化成可证实防御的网络	• 首个在ImageNet上证实鲁棒的方法 • certificate: L2<0.5 51% error [ImageNet]
PixelDP (Lecuyer et al., 2019)	利用Differential Privacy差分隐私进行网络证实	• 适用于大型网络和ImageNet
MILP (Tjeng et al., 2019)	mixed-integer linear programming (MILP)	• certificate: L∞<0.1 4.38% error [MNIST]
CROWN-IBP (Zhang et al., 2018)	bounding a given activation function with linear and quadratic functions	• certificate: L∞<0.3 7.02% error [MNIST] • certificate: L∞<8/255 33.06% error [CIFAR10]

Certified Defense: SMT

SMT的方法

(利用SMT可以对任意网络进行证实，计算过慢)

Name	Method	Performance
Safety Verification (Huang et al., 2017)	基于Satisfiability Modulo Theory (SMT)设计了网络自动证实框架，采用离散化的方法进行搜索	• 可以确保 如果对抗扰动存在则一定可以找到 • 计算量大，仅能分析小型网络
Reluplex (Katz et al., 2017)	设计了网络证实的SMT Solver （DNN的verification是NPC问题 困难在于激活函数）	• 仅能证实ReLU的网络， • 计算量大耗时过长，仅能分析小型网络

Conclusion

Name	Method	Performance
AdvTraining	加入对抗样本进行训练	• PGD AdvTraining可以有效防御强大的攻击 • 训练开销过大、黑盒安全性不高、精度下降
GradMasking	进行梯度混淆	• 白盒场景下不安全
	正则化方法	• 鲁棒性提升有限，白盒场景下不安全
Detecting	对抗/干净样本的高层分布不同	• 依赖于模型
	训练模型进行分类	• 需要生成大量的对抗样本
	输入不一致性	
Transformation	训练生成模型模拟输入分布	• 需要训练生成模型
	对输入进行变形	• 鲁棒性提升有限
CertifiedDefense	凸优化方法	• 难以适用于ImageNet
	SMT方法	• 计算开销过大，仅适用于小型网络