



南京邮电大学
Nanjing University of Posts and Telecommunications

Certifiable Black-Box Attacks with Randomized Adversarial

Examples: Breaking Defenses with Provable Confidence

可认证黑盒攻击

汇报人：岑帛

目录

CONTENTS



1/ 论文概述

2/ 问题描述

3/ 可认证攻击

4/ 实验评估及总结



01

论文概述

研究背景

机器学习（ML）模型虽取得巨大成功，却容易被微小的输入扰动（perturbation）干扰导致分类错误，特别是由恶意攻击者精心构造的扰动。有许多最先进的对抗性攻击被提出以探索各种机器学习模型的脆弱性，进一步帮助完善模型的安全性。攻击者通常通过访问（query）模型迭代改进手中的对抗样本，最终达到攻击目的（改变模型输出）。严格的黑盒攻击只依赖模型的prediction score或者hard label进行对抗样本构建，被认为是更贴近实际安全实践的方法。

论文概述

问题描述

可认证攻击

实验评估

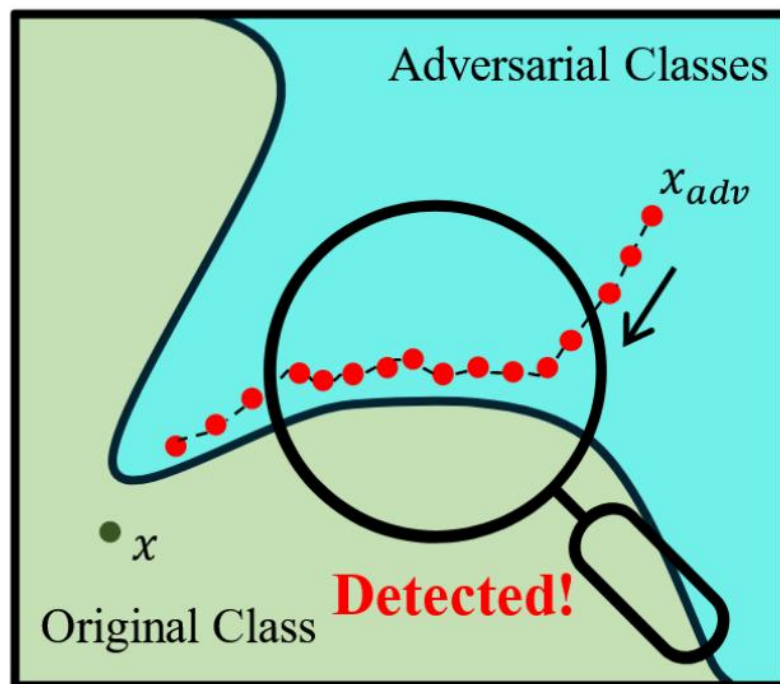
及总结



研究背景

黑盒攻击手段：
梯度估计、代理模型或启发式算法生成对抗性样本。

防御手段：
Blacklight
随机防御



论文概述

问题描述

可认证攻击

实验评估

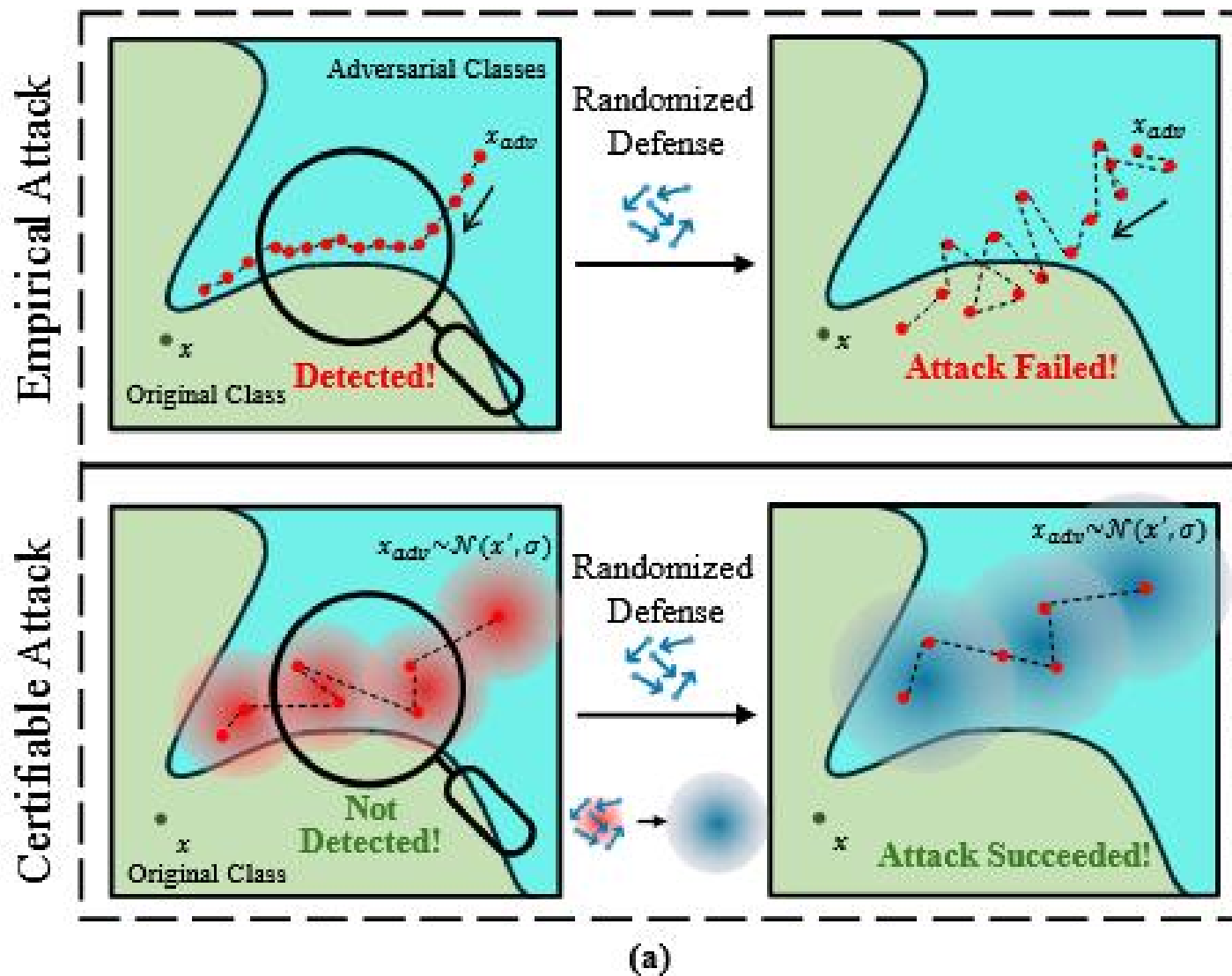
及总结



为了解决这个问题，本文提出了一种新的攻击范式，能够在随机噪声下保证攻击的理论成功概率，称为可认证攻击。首先，本文将对抗样本建模为输入空间中的一个随机变量，该随机变量服从某个噪声分布 ϕ ，称为对抗分布（Adversarial Distribution），对抗样本可以从对抗分布中采样得到。

可认证攻击相较于实践性攻击有以下优点

研究背景



论文概述

问题描述

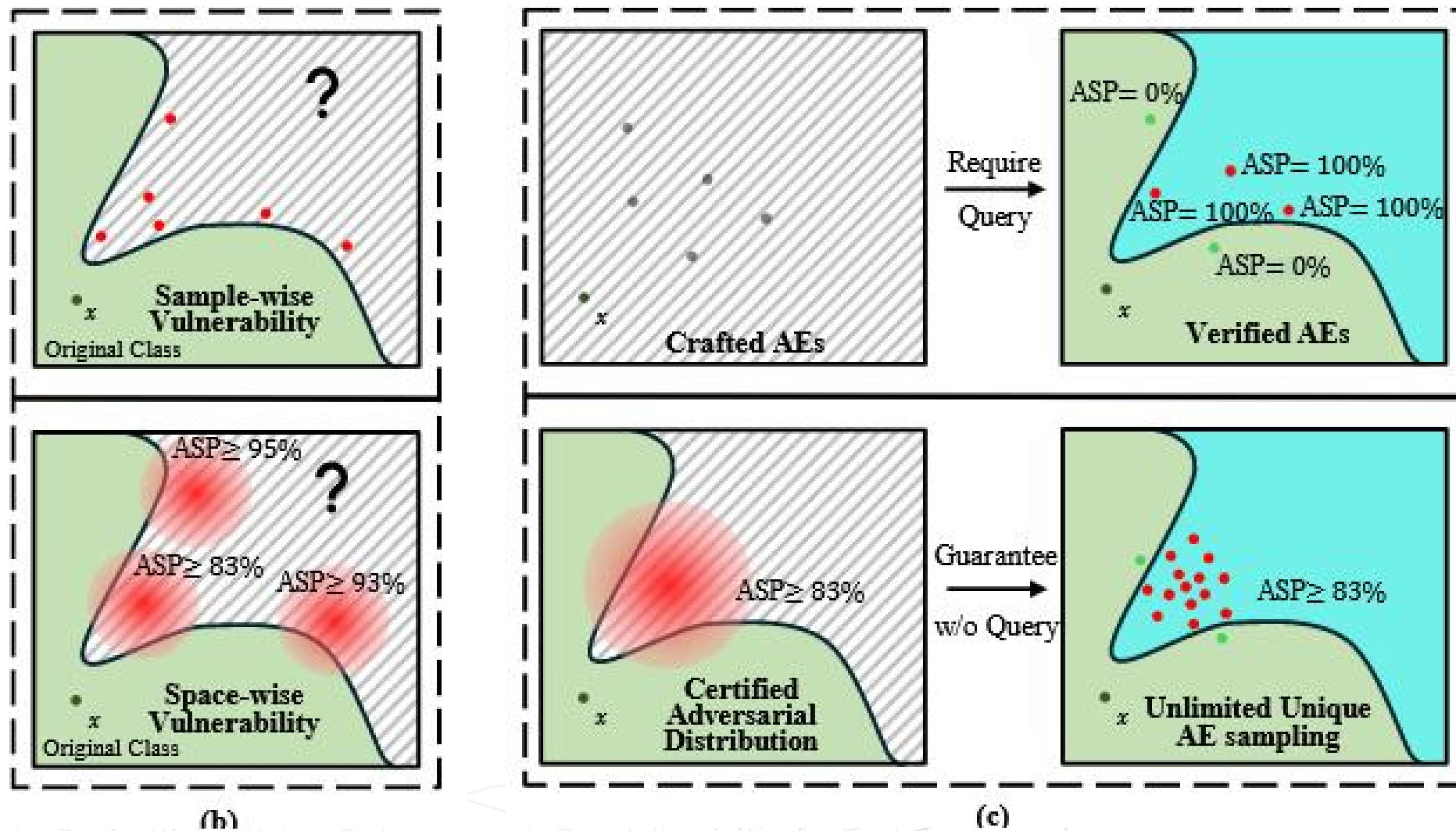
可认证攻击

实验评估

及总结



研究背景



论文概述

问题描述

可认证攻击

实验评估

及总结





02

问题描述

- 基本定义
- 算法简介

基本定义

对抗样本:

$$f \quad x \in \mathbb{R}^d \quad \mathcal{Y} = [1, \dots, C] \quad y$$

$$f(x_{adv}) \neq y$$

$$x_{adv} - x$$

论文概述

问题描述

可认证攻击

实验评估

及总结



基本定义

可认证攻击:

$\varphi(x', \kappa)$ Adversarial Distribution

Attack Success Probability p (ASP)

$$\mathbb{P}_{x_{adv} \sim \varphi(x', \kappa)} [f(x_{adv}) \neq y] \geq p \quad (1)$$

$$\text{s.t. } x_{adv} \in [\Pi_a, \Pi_b]^d. \quad (2)$$

论文概述

问题描述

可认证攻击

实验评估

及总结



基本定义

设计目标：

1. 它可以提供关于精心设计的对抗样本的最小攻击成功概率的可证明保证。
2. 它不仅可以在访问模型后验证实例是否具有对抗性，还可以在访问前通过给出其ASP来验证。
3. 它需要尽可能少的访问次数。
4. 它可以产生不可察觉的对抗扰动，可以绕过现有的基于检测和预处理/后处理的防御。

论文概述

问题描述

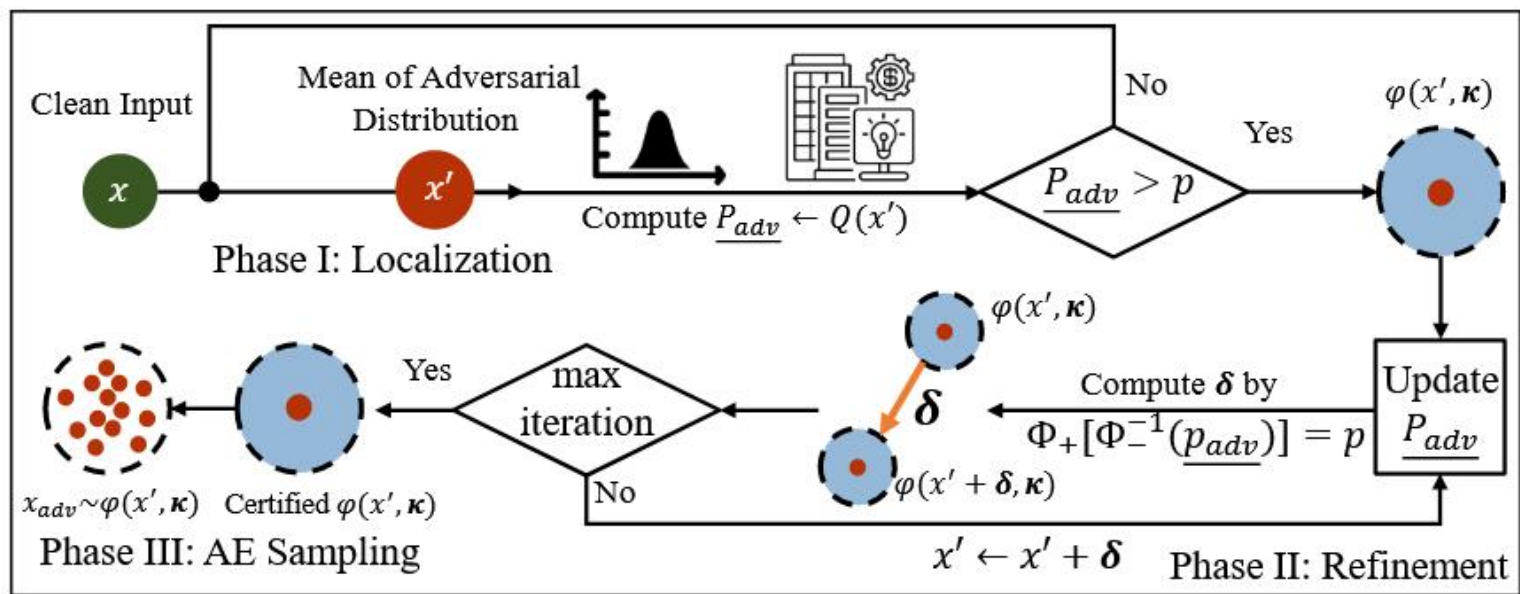
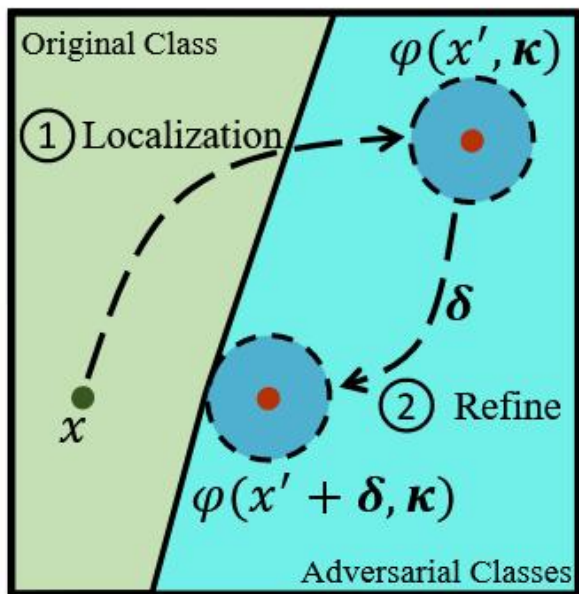
可认证攻击

实验评估

及总结



算法简介



论文概述

问题描述

可认证攻击

实验评估

及总结



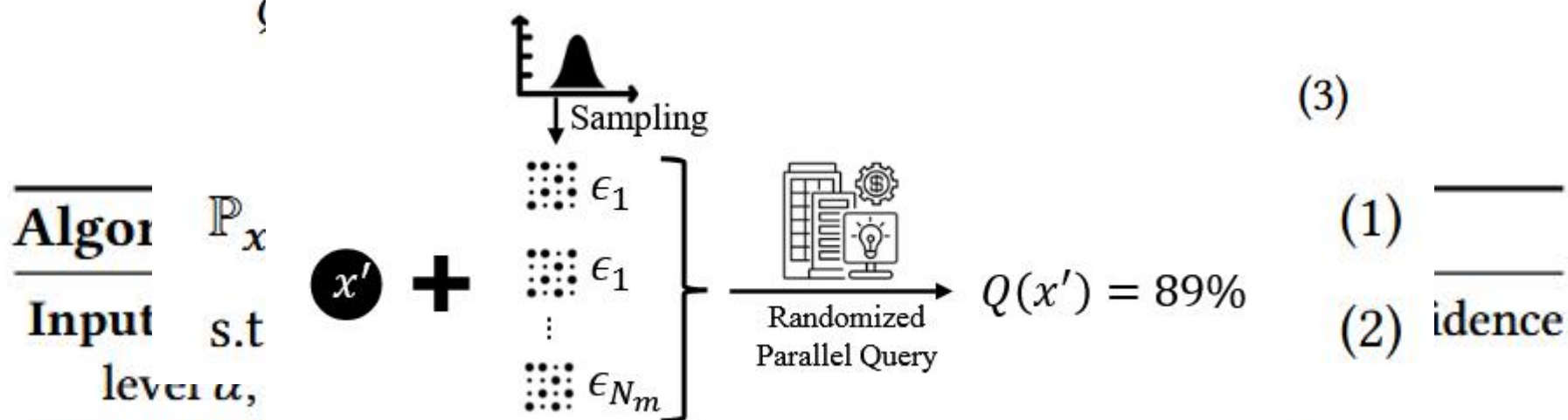


03

可认证攻击

对抗分布定位

随机并行访问RPQ:



论文概述

问题描述

可认证攻击

实验评估

及总结



对抗分布定位

平滑自监督扰动SSSP:

自监督扰动SSP通过自监督的方式在大规模数据集上预训练的特征提取器提取的特征来生成通用的对抗样本。

$$\begin{aligned} x' &= \arg \max_{x'} \mathbb{E}_{\epsilon \sim \varphi(0, \kappa)} [\|\mathcal{F}(x' + \epsilon) - \mathcal{F}(x + \epsilon)\|_2] \\ \text{s.t. } \|x' - x\|_{\infty} &\leq \pi \end{aligned} \quad (4)$$

论文概述

问题描述

可认证攻击

实验评估

及总结



对抗分布定位

平滑自监督扰动:

Algorithm 2 Smoothed Self-Supervised Perturbation (SSSP)

Input: Clean input x , feature extractor \mathcal{F} , noise distribution $\varphi(0, \kappa)$, maximum iterations n_{max} , perturbation budget π , step size η , and noise sampling number N_s .

Output: Updated mean x' of *Adversarial Distribution*

- 1: $x' = x$
 - 2: **for** $n = 1$ to n_{max} **do**
 - 3: $\mathcal{L}(x') \leftarrow \frac{1}{N_s} \sum_i^{N_s} [||\mathcal{F}(x' + \epsilon_i) - \mathcal{F}(x + \epsilon_i)||_2]$, $\epsilon_i \sim \varphi$
 - 4: $x' \leftarrow x' + \eta \operatorname{sgn}(\nabla_{x'} \mathcal{L})$
 - 5: $x' \leftarrow \operatorname{Clip}(x', x - \pi, x + \pi)$
 - 6: $x' \leftarrow \operatorname{Clip}(x', 0.0, 1.0)$ (if x is an image)
 - 7: **return** x'
-

论文概述

问题描述

可认证攻击

实验评估

及总结



对抗分布定位

平滑自监督定位:

Algorithm 3 Smoothed SSP for Certifiable Attack Localization

Input: Clean input x , feature extractor $\mathcal{F}(\cdot)$, RPQ function $Q(\cdot)$, smoothed SSP algorithm $SSSP(\cdot)$ (Algorithm 2), initial perturbation budget π_{init} , step size γ , ASP Threshold p , maximum iterations N_{max} .

Output: Mean x' of *Adversarial Distribution* φ , number of RPQs q .

- 1: $x' = x, \pi = \pi_{init}, N = 0, q = 0$
 - 2: **while** $Q(x') < p$ and $N < N_{max}$ **do**
 - 3: $N \leftarrow N + 1, q \leftarrow q + 1, \pi \leftarrow \pi + \gamma$
 - 4: $x' \leftarrow SSSP(x', \mathcal{F}, \pi)$
 - 5: **if** $Q(x') < p$ **then**
 - 6: **return** Abstain
 - 7: **else**
 - 8: **return** x' and q
-

论文概述

问题描述

可认证攻击

实验评估

及总结



对抗分布定位

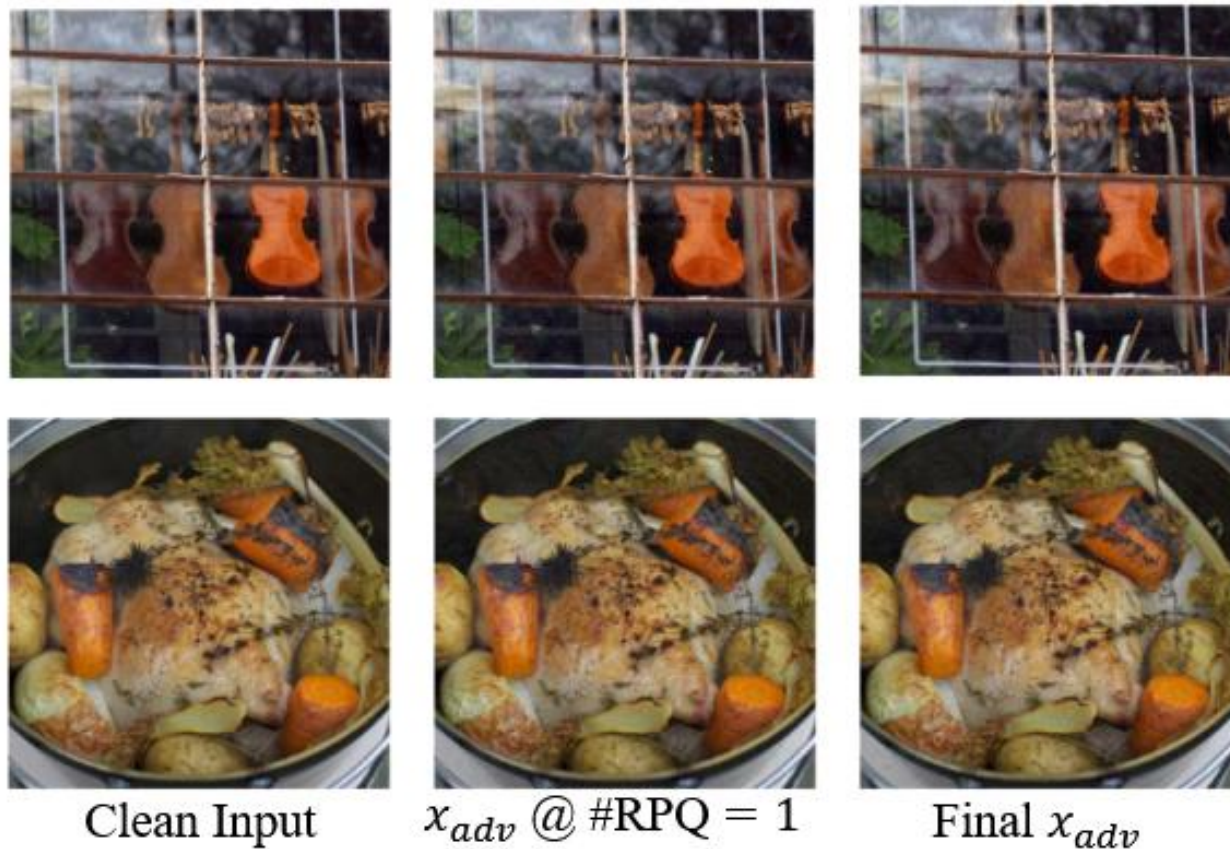


Figure 7: Visualization of successful adversarial examples crafting by certifiable attack with SSSP localization (SSSP requires fewer # RPQ)

论文概述

问题描述

可认证攻击

实验评估

及总结



对抗分布定位

二分搜索定位:

Algorithm 4 Binary Search for Certifiable Attack Localization

Input: Clean input x , RPQ function $Q(\cdot)$, ASP Threshold p , random search iterations N_r , and binary search iteration N_b , error tolerance Ω .

Output: Mean of initial *Adversarial Distribution* x' , number of RPQs q .

```
1:  $n = 0, m = 0, q = 0, x^* = x$ 
2: while  $Q(x') < p$  and  $n \leq N_r$  do
3:    $x' \sim \text{Uniform}([0, 1]^d)$ 
4:    $q \leftarrow q + 1,$ 
5: if  $n > N_r$  then return Abstain
6: while  $m < N_b$  and  $\|x' - x^*\|_2 \leq \Omega$  do
7:   if  $Q(\frac{x^* + x'}{2}) \geq p$  then
8:      $x' = \frac{x^* + x'}{2}$ 
9:   else
10:     $x^* = \frac{x^* + x'}{2}$ 
11: return  $x'$ 
```

论文概述

问题描述

可认证攻击

实验评估

及总结



对抗分布定位

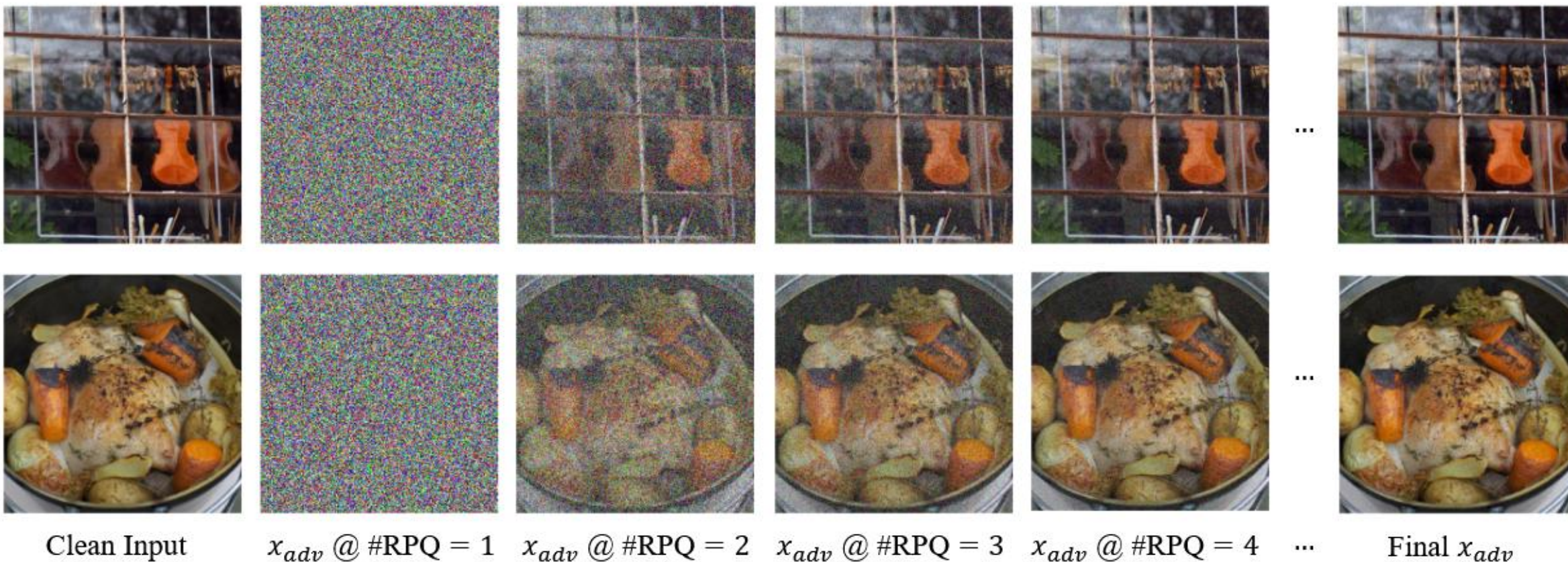
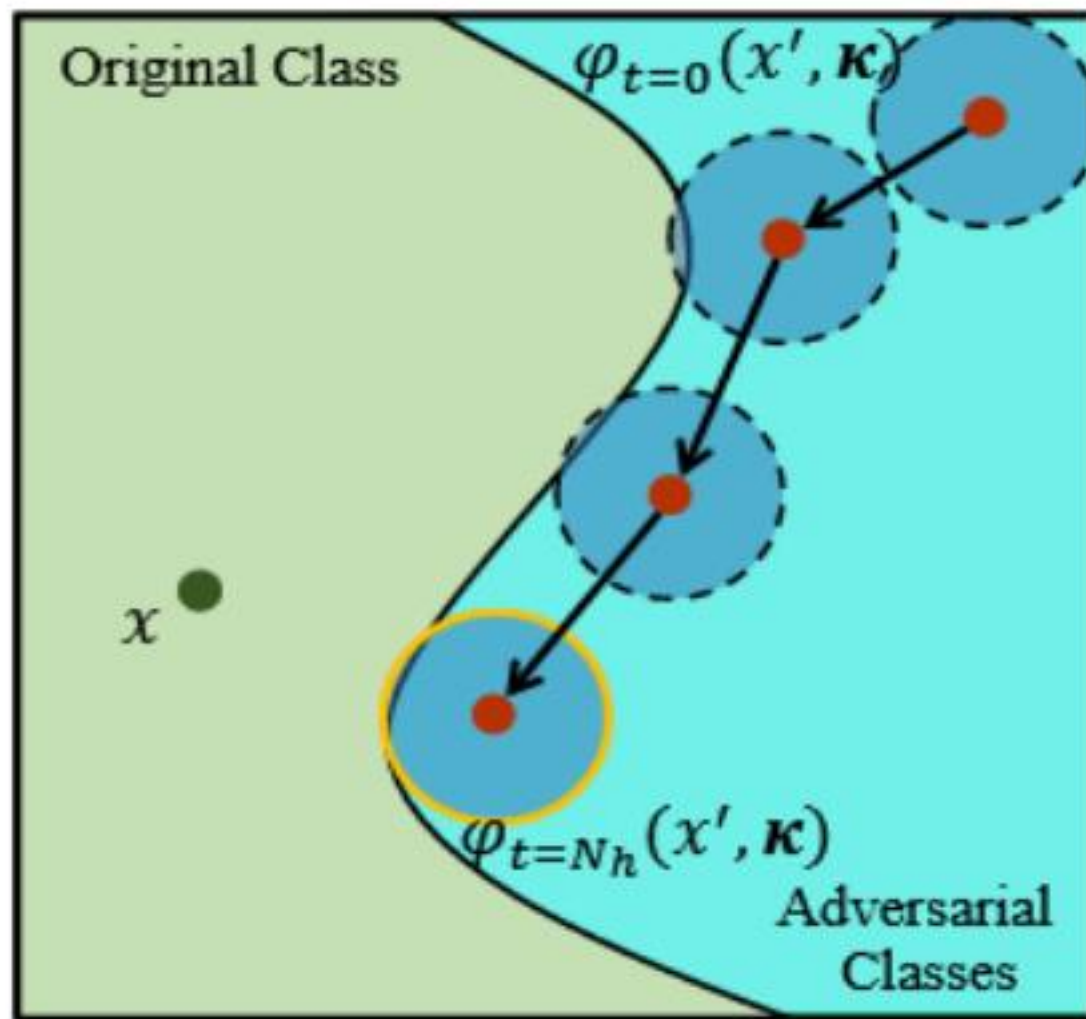


Figure 6: Visualization of successful adversarial examples crafting by certifiable attack with binary-search localization

对抗分布优化



Geometrically Shifting

论文概述

问题描述

可认证攻击

实验评估

及总结



对抗分布优化

可证明的对抗分布位移：

$$\mathbb{P}[f(x' + \epsilon) \neq y] \geq \underline{p_{adv}} = Q(x') \geq p, \quad (5)$$

$\mathbb{P}[f(x' + \delta + \epsilon) \neq y] \geq p$ is guaranteed for any shifting vector δ when

$$\Phi_+[\Phi_-^{-1}(\underline{p_{adv}})] \geq p \quad (6)$$

where Φ_-^{-1} is the inverse cumulative density function (CDF) of the random variable $\frac{\varphi(\epsilon - \delta, \kappa)}{\varphi(\epsilon, \kappa)}$, and Φ_+ the CDF of random variable $\frac{\varphi(\epsilon, \kappa)}{\varphi(\epsilon + \delta, \kappa)}$.

论文概述

问题描述

可认证攻击

实验评估

及总结



对抗分布优化

$$\Phi_+[\Phi_-^{-1}(\underline{p_{adv}})] \geq p$$

- $\Phi_-^{-1}(p_{adv})$:
 - Φ_- 是随机变量 $Z_- = \frac{\varphi(\epsilon - \delta, \kappa)}{\varphi(\epsilon, \kappa)}$ 的累积分布函数 (CDF)。
 - $\Phi_-^{-1}(p_{adv})$ 表示找到阈值 z^* , 使得 $\mathbb{P}(Z_- \leq z^*) = p_{adv}$ 。
- $\Phi_+(z^*)$:
 - Φ_+ 是随机变量 $Z_+ = \frac{\varphi(\epsilon, \kappa)}{\varphi(\epsilon + \delta, \kappa)}$ 的CDF。
 - $\Phi_+(z^*)$ 表示 $Z_+ \leq z^*$ 的概率。

论文概述

问题描述

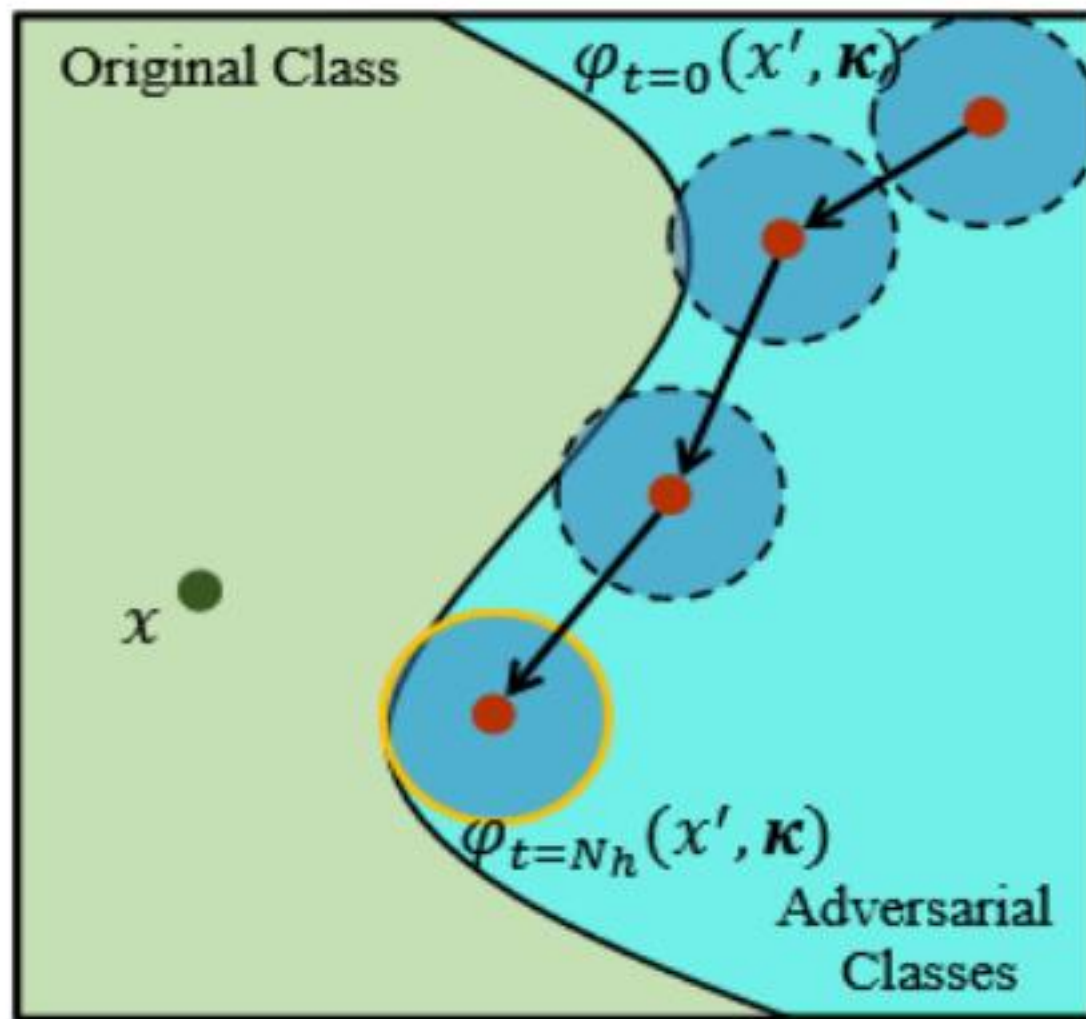
可认证攻击

实验评估

及总结



对抗分布优化



Geometrically Shifting

论文概述

问题描述

可认证攻击

实验评估

及总结

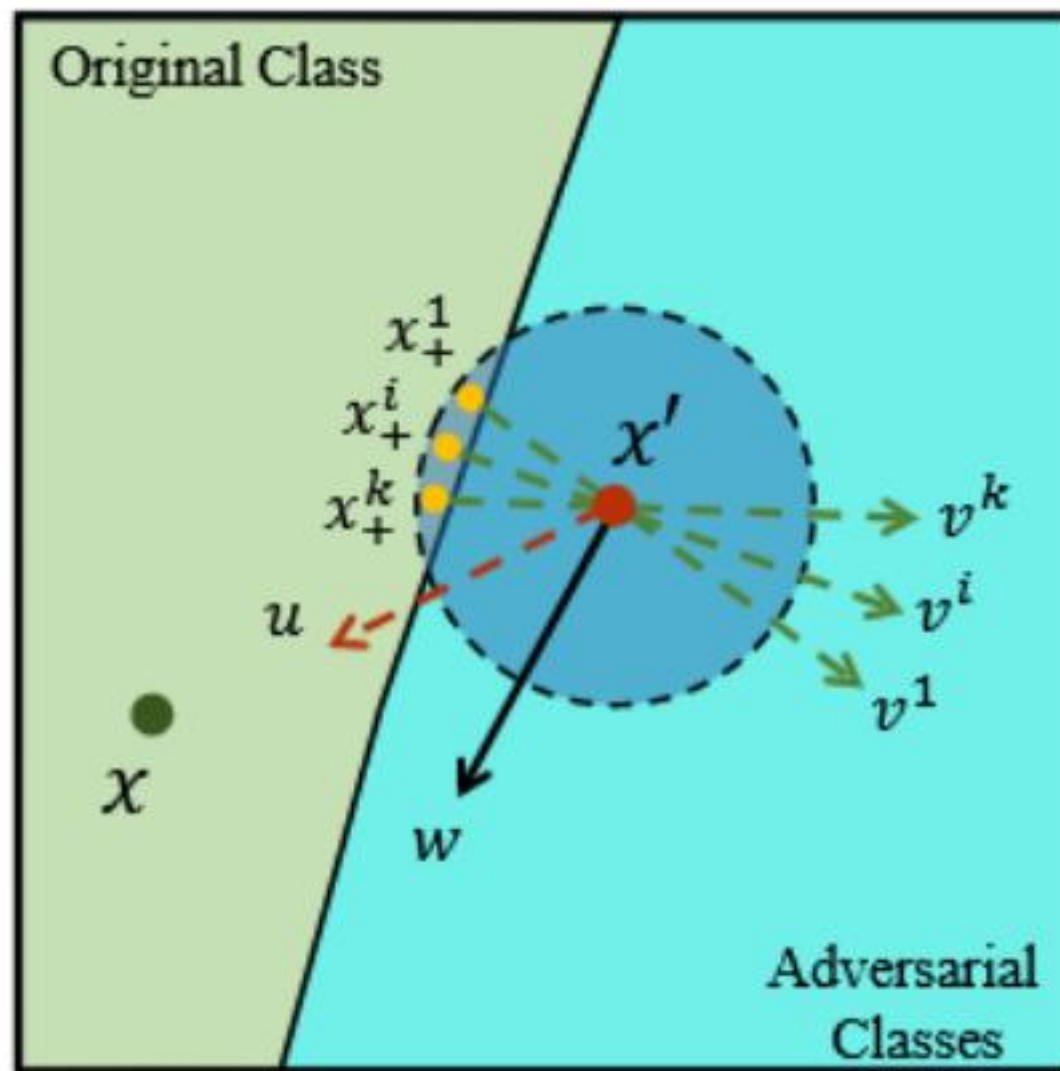


对抗分布优化

寻找位移方向:

$$w = \arg \max \sum_{i=1}^k$$

$$\sin(v^i, w) + \cos(u, w)$$



论文概述

问题描述

可认证攻击

实验评估

及总结



对抗分布优化

寻找位移方向:

Algorithm 5 Shifting Direction

Input: Mean of the *Adversarial Distribution* x' , clean input x , vectors $\{v^i\}$, a vector u , maximum iteration M , updating step size η' .

Output: The shifting direction w

- 1: Initialize w with random noise
 - 2: **if** $\{v^i\}$ is empty **then**
 - 3: $w = x - x'$
 - 4: **else**
 - 5: **for** $j = 1$ to M **do**
 - 6: $w \leftarrow w + \eta' \operatorname{sgn}[\nabla_w(\sum_{i=1}^k \sin(v^i, w) + \cos(u, w))]$
 - 7: $w \leftarrow \frac{w}{\|w\|_2}$
 - 8: **return** w
-

论文概述

问题描述

可认证攻击

实验评估

及总结



Algorithm 6 Shifting Distance

Input: Mean of *Adversarial Distribution* x' , noise distribution φ , randomized query function $Q(\cdot)$, the shifting direction algorithm $SD(\cdot)$ (Algorithm 5), error threshold e , ASP Threshold p .

Output: The shifting perturbation δ

```

1:  $w \leftarrow SD(x'), \underline{p}_{adv} \leftarrow Q(x')$ 
2: find a scalar  $a$  such that  $\delta = aw$  and  $\Phi_+[\Phi_-^{-1}(\underline{p}_{adv})] > p$ 
3: find a scalar  $b$  such that  $\delta = bw$  and  $\Phi_+[\Phi_-^{-1}(\underline{p}_{adv})] < p$ 
4: while  $\Phi_+[\Phi_-^{-1}(\underline{p}_{adv})] < p$  or  $> p + e$  and  $n \leq N_k$  do
5:   if  $\Phi_+[\Phi_-^{-1}(\underline{p}_{adv})] > p$  then
6:      $a \leftarrow \frac{(a+b)}{2}$ 
7:   else
8:      $b \leftarrow \frac{(a+b)}{2}$ 
9:    $\delta \leftarrow \frac{(a+b)}{2} w, n \leftarrow n + 1$ 
10: return  $\delta$ 

```

论文概述

问题描述

可认证攻击

实验评估

及总结





04

实验评估及总结

- 实验
- 总结

实验设置

使用了三个用于图像分类的基准数据集：

CIFAR10、CIFAR100 、 ImageNet 。

CIFAR10和CIFAR100分别由10类和100类共60, 000幅 32×32 彩色图像组成。

ImageNet是一个包含1000个类别的大规模数据集。

训练集包含1, 281, 167张图片，验证集包含50, 000张图片。

使用VGG，ResNet，ResNext和WRN作为目标模型。

使用在ImageNet上预训练好的ResNet34作为特征提取器（在Smoothed SSP中）。

论文概述

问题描述

可认证攻击

实验评估

及总结



实验设置

基于硬标签的黑盒攻击:

GeoDA, HSJ, Opt, RayS, SignFlip, SignOPT, Boundary;

基于得分的黑盒攻击:

Bandit, NES, Parsimonious, Sign, Square, ZOSignSGD,
Simple attack

基于优化的攻击:

SparseEvo and PointWise

防御方法:

Blacklight detection, Randomized pre-processing
defense (RAND-Pre), Randomized post-processing
defense (RAND-Post), and Adversarial Training
based TRADES

论文概述

问题描述

可认证攻击

实验评估

及总结



实验设置

衡量指标：

模型精确度

RPQ的次数 (# RPQ)

访问次数 (# Q)

认证准确度@ p

扰动大小 (Dist. l_2)

平均距离 (Mean Dist. l_2)

检测成功率 (Det. Rate)

检测前平均访问次数 (# Q to Det.)

检测覆盖率 (Det. Cov)

论文概述

问题描述

可认证攻击

实验评估

及总结



实验设置

设定蒙特卡洛抽样次数为50， $p = 10\%$ ，采用方差为 $\sigma = 0.025$ 的高斯分布。

我们通过集成多个开源库⁴，实现了一个包含16个黑盒攻击、4个防御、6个数据集和9个模型的PyTorch库。实验在AMD EPYC Genoa 9354 CPUs (32 Core , 3 . 3GHz)和NVIDIA H100 Hopper GPUs (80GB each)的服务器上运行。

论文概述

问题描述

可认证攻击

实验评估

及总结



攻击性能

Blacklight Detection

Table 3: Attack performance under Blacklight detection on ResNet and ImageNet (Clean Accuracy: 67.9%)

Attack	Query Type	Pert. Type	Det. Rate %	# Q to Det.	Det. Cov. %	Model Acc.	# Q	Dist. ℓ_2
Bandit	Score	ℓ_∞	100.0	1.0	64.2	1.9	25	25.42
NES	Score	ℓ_∞	100.0	10.3	17.3	7.0	337	8.28
Parsimonious	Score	ℓ_∞	100.0	2.0	96.7	3.8	282	25.24
Sign	Score	ℓ_∞	100.0	2.0	91.5	0.5	126	25.50
Square	Score	ℓ_∞	100.0	2.0	66.9	0.0	14	25.54
ZOSignSGD	Score	ℓ_∞	100.0	2.0	50.2	12.5	322	8.53
GeoDA	Label	ℓ_∞	100.0	1.0	88.9	5.1	151	17.99
HSJ	Label	ℓ_∞	100.0	7.3	94.9	35.6	212	9.82
Opt	Label	ℓ_∞	99.9	8.4	81.4	61.2	646	0.98
RayS	Label	ℓ_∞	100.0	4.4	83.5	4.2	260	29.63
SignFlip	Label	ℓ_∞	100.0	8.5	70.3	4.4	148	27.64
SignOPT	Label	ℓ_∞	99.9	8.4	69.8	55.9	570	1.32
Bandit	Score	ℓ_2	100.0	1.0	99.5	1.7	431	9.60
NES	Score	ℓ_2	100.0	10.2	32.8	61.2	571	0.45
Simple	Score	ℓ_2	100.0	1.0	99.9	53.6	883	0.88
Square	Score	ℓ_2	100.0	2.0	68.8	0.0	16	26.30
ZOSignSGD	Score	ℓ_2	100.0	2.0	52.4	65.1	531	0.30
Boundary	Label	ℓ_2	100.0	7.2	76.3	37.9	60	11.63
GeoDA	Label	ℓ_2	100.0	1.0	89.3	3.9	181	19.14
HSJ	Label	ℓ_2	100.0	7.3	93.4	11.4	255	22.21
Opt	Label	ℓ_2	100.0	8.5	67.9	41.2	610	16.71
SignOPT	Label	ℓ_2	99.9	8.4	62.9	36.7	485	17.54
PointWise	Label	Opt.	100.0	1.0	99.8	0.0	920	13.53
SparseEvo	Label	Opt.	100.0	1.0	99.9	0.0	1000	7.68
CA (sssp)	Label	Opt.	0.0	∞	0.0	1.4	148	13.74
CA (bin search)	Label	Opt.	0.0	∞	0.0	0.0	603	33.14

论文概述

问题描述

可认证攻击

实验评估

及总结



攻击性能

RAND-Pre

CIFAR10: 92%–30%,
CIFAR100: 95%–29%,
ImageNet: 69%–25%。

Table 4: Attack performance under RAND Pre-processing Defense on ResNet and ImageNet (Clean Accuracy: 67.0%)

Attack	Query Type	Perturbation Type	# Query	Model Acc.	Dist. ℓ_2
Bandit	Score	ℓ_∞	10	6.7	25.26
NES	Score	ℓ_∞	428	49.8	10.26
Parsimonious	Score	ℓ_∞	243	62.7	25.12
Sign	Score	ℓ_∞	116	40.6	25.20
Square	Score	ℓ_∞	27	10.4	24.96
ZOSignSGD	Score	ℓ_∞	428	49.4	10.36
GeoDA	Label	ℓ_∞	150	40.0	18.08
HSJ	Label	ℓ_∞	232	58.5	8.76
Opt	Label	ℓ_∞	905	69.4	0.44
RayS	Label	ℓ_∞	235	47.9	28.14
SignFlip	Label	ℓ_∞	46	52.8	13.06
SignOPT	Label	ℓ_∞	394	59.1	0.39
Bandit	Score	ℓ_2	583	58.2	12.99
NES	Score	ℓ_2	341	66.8	0.43
Simple	Score	ℓ_2	258	67.2	0.10
Square	Score	ℓ_2	18	13.6	25.96
ZOSignSGD	Score	ℓ_2	249	67.3	0.28
Boundary	Label	ℓ_2	38	49.2	15.12
GeoDA	Label	ℓ_2	149	47.4	17.70
HSJ	Label	ℓ_2	225	55.7	14.30
Opt	Label	ℓ_2	1000	58.2	12.28
SignOPT	Label	ℓ_2	406	52.2	15.41
PointWise	Label	Optimized	942	54.6	16.90
SparseEvo	Label	Optimized	1000	61.7	11.33
CA (sssp)	Label	Optimized	154	1.7	13.98
CA (bin search)	Label	Optimized	603	0.0	32.16

论文概述

问题描述

可认证攻击

实验评估

及总结



Table 5: Attack performance under RAND Post-processing Defense on ResNet and ImageNet (Clean Accuracy: 68.0%)

Attack	Query Type	Perturbation Type	# Query	Model Acc.	Dist. ℓ_2
Bandit	Score	ℓ_∞	17	2.7	25.51
NES	Score	ℓ_∞	378	18.6	9.53
Parsimonious	Score	ℓ_∞	253	47.9	25.46
Sign	Score	ℓ_∞	124	8.1	25.81
Square	Score	ℓ_∞	18	0.8	25.44
ZOSignSGD	Score	ℓ_∞	376	21.4	9.71
GeoDA	Label	ℓ_∞	143	38.6	17.62
HSJ	Label	ℓ_∞	212	52.7	8.82
Opt	Label	ℓ_∞	1000	65.3	0.67
RayS	Label	ℓ_∞	243	43.9	28.09
SignFlip	Label	ℓ_∞	86	47.2	15.44
SignOPT	Label	ℓ_∞	412	63.6	0.64
Bandit	Score	ℓ_2	596	6.0	13.96
NES	Score	ℓ_2	344	59.7	0.44
Simple	Score	ℓ_2	241	58.9	0.10
Square	Score	ℓ_2	23	0.4	26.46
ZOSignSGD	Score	ℓ_2	275	61.4	0.29
Boundary	Label	ℓ_2	24	48.0	12.64
GeoDA	Label	ℓ_2	146	40.6	16.89
HSJ	Label	ℓ_2	238	49.5	14.59
Opt	Label	ℓ_2	1000	53.9	12.62
SignOPT	Label	ℓ_2	411	46.3	15.96
PointWise	Label	Optimized	969	55.1	16.01
SparseEvo	Label	Optimized	1000	66.7	9.10
CA (sssp)	Label	Optimized	147	1.4	13.70
CA (bin search)	Label	Optimized	603	0.0	32.67

攻击性能

RAND-Post

硬标签攻击:

CIFAR10: 84%–41%,

CIFAR100: 89%–45%,

ImageNet: 60%–24%.

分数攻击:

CIFAR10: 100%–91%,

CIFAR100: 100%–95%,

ImageNet: 72%–62%.

论文概述

问题描述

可认证攻击

实验评估

及总结



攻击性能

TRADES

Table 6: Attack performance under TRADES Adversarial Training on ResNet and CIFAR10

Defense	Attack	Query τ	Pert. τ	# Query	Model Acc.	Dist. ϵ
ℓ_2 Adversarial Training (Clean Accuracy: 59.2%)	Bandit	Score	ℓ_2	860	1.5	2.44
	NES	Score	ℓ_2	3535	9.5	0.99
	Simple	Score	ℓ_2	4062	2.1	1.29
	Square	Score	ℓ_2	991	4.6	2.95
	ZOSignSGD	Score	ℓ_2	3505	15.1	0.77
	Boundary	Label	ℓ_2	771	40.7	1.19
	GeoDA	Label	ℓ_2	1506	14.3	2.85
	HSJ	Label	ℓ_2	1332	5.1	3.53
	Opt	Label	ℓ_2	2890	41.6	2.39
	SignOPT	Label	ℓ_2	1766	33.6	2.76
	PointWise	Label	Opt.	4845	0.6	5.36
	SparseEvo	Label	Opt.	9697	0.4	6.03
	CA (sssp)	Label	Opt.	809	20.4	6.06
	CA (bin search)	Label	Opt.	461	0.0	8.18
SparseEvo	SparseEvo	Label	Opt.	9697	0.4	6.03
	CA (sssp)	Label	Opt.	548	21.2	4.29
	CA (bin search)	Label	Opt.	412	9.8	6.31

论文概述

问题描述

可认证攻击

实验评估

及总结



消融实验

不同噪声方差下的攻击性能：

Table 7: Attack performance of our certifiable attack with varying Gaussian noise variances σ ($p = 90\%$)

	σ	Dist. ℓ_2	Mean Dist. ℓ_2	# RPQ	Certified Acc.
CIFAR10	0.10	7.39	3.96	18.34	94.17%
	0.25	12.95	2.34	14.35	91.21%
	0.50	19.41	0.43	11.38	90.00%
ImageNet	0.10	41.80	16.78	32.55	99.80%
	0.25	87.47	16.78	17.02	99.60%
	0.50	135.47	2.27	8.31	100.00%

论文概述

问题描述

可认证攻击

实验评估

及总结



消融实验

不同Asp阈值下的攻击性能：

Table 8: Attack performance of our certifiable attack with varying p under the Gaussian variance $\sigma = 0.25$

	p	Dist. ℓ_2	Mean Dist. ℓ_2	# RPQ	Certified Acc.
CIFAR10	50%	12.65	1.63	9.34	97.17%
	60%	12.72	1.86	11.09	95.85%
	70%	12.80	2.05	11.94	94.72%
	80%	12.87	2.18	12.37	93.17%
	90%	12.95	2.34	14.35	91.21%
	95%	13.09	2.65	15.93	90.37%
ImageNet	50%	85.88	9.89	12.85	100.00%
	60%	86.20	11.30	13.63	100.00%
	70%	86.45	12.64	14.33	100.00%
	80%	87.03	14.64	16.02	100.00%
	90%	87.47	16.78	17.02	99.60%
	95%	88.42	19.98	19.81	100.00%

论文概述

问题描述

可认证攻击

实验评估

及总结



消融实验

不同定位/优化算法下的攻击性能：

Table 9: Attack performance of our certifiable attack on different localization/refinement algorithms ($\sigma = 0.25$, $p = 90\%$)

Localization	Refinement	Dist. ℓ_2	Mean Dist. ℓ_2	# RPQ	Cert. Acc.
sssp	none	11.46	1.35	2.30	92.54
binary search	none	11.29	0.34	9.07	92.54
random	geo.	11.80	1.73	67.53	92.54
sssp	geo.	11.20	0.49	3.70	91.54
binary search	geo.	11.28	0.27	10.08	92.53

论文概述

问题描述

可认证攻击

实验评估

及总结



消融实验

不同噪声分布下的攻击性能：

Table 10: Attack performance of our certifiable attack with different noise distributions

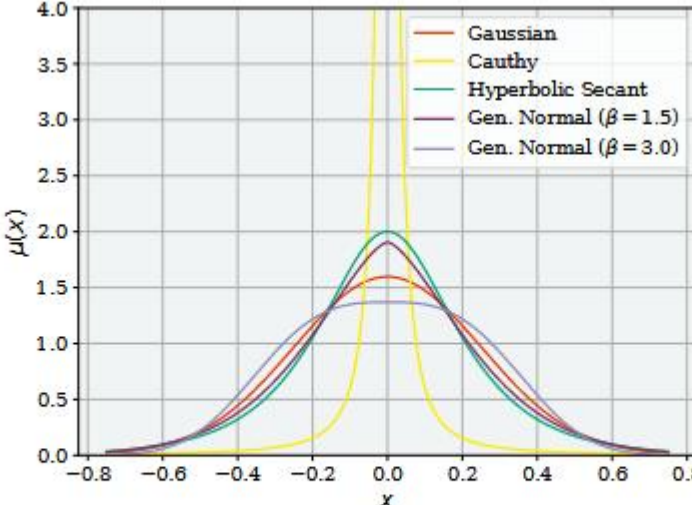
	Distribution	Density	Parameter	$\sqrt{\ w\ _2^2/d}$	Dist. ℓ_1	Mean Dist. ℓ_2	# RPQ	Certified Acc.
CIFAR10	Gaussian					2.34	14.35	91.21%
	Cauchy					4.87	32.77	94.12%
	Hyperbolic Sec					2.43	14.59	91.67%
	General Normal (b					2.37	14.15	91.39%
	General Normal (b					2.38	14.15	91.25%
ImageNet	Gaussian					16.78	17.02	99.60%
	Cauchy					23.94	59.94	99.60%
	Hyperbolic Sec					21.29	20.89	99.80%
	General Normal (b					19.05	17.58	99.80%
	General Normal (b					15.58	14.99	100.00%

Figure 9: PDF of Different Noise Distributions ($\sigma = 0.25$)

论文概述

问题描述

可认证攻击

实验评估

及总结



总结

可认证攻击为对抗攻击提供了新的方向，实现了从确定性对抗攻击到概率性对抗攻击的转变。与经验性黑盒攻击相比，可验证攻击具有显著的优点，包括打破SOTA强检测和随机化防御，揭示模型一致性和鲁棒性的严重漏洞，并保证在不通过访问验证的情况下，为众多独特的AE提供最小ASP。

论文概述

问题描述

可认证攻击

实验评估

及总结





南京邮电大学
Nanjing University of Posts and Telecommunications

谢谢大家

