Narcissus: A Practical Clean-Label Backdoor Attack with Limited Information

CCS '23: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security
Pages 771 - 785

● 演讲人：黄雨洁

目录

01 背景介绍

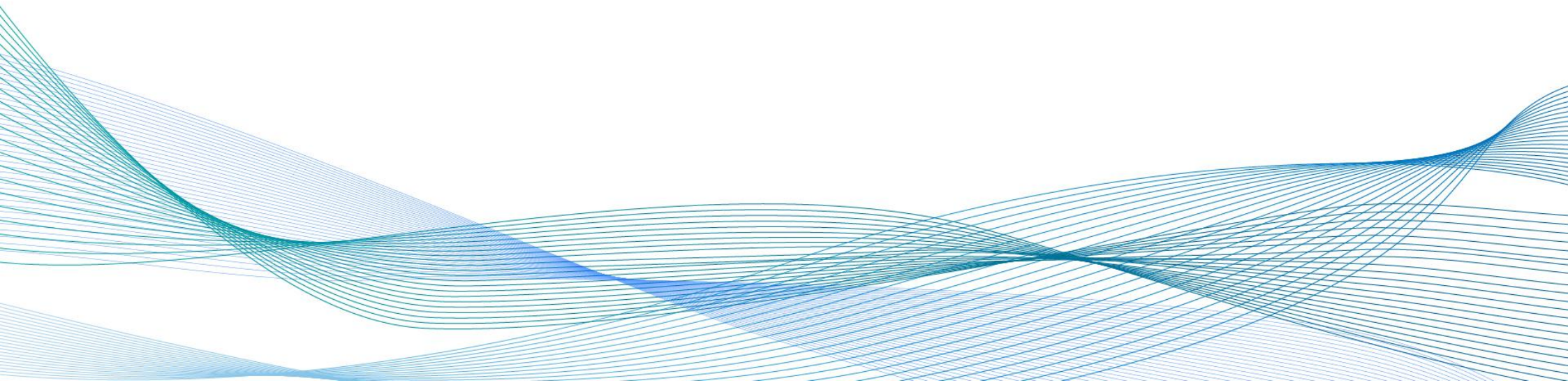
02 相关知识

03 攻击原理流程

04 实验评估

01

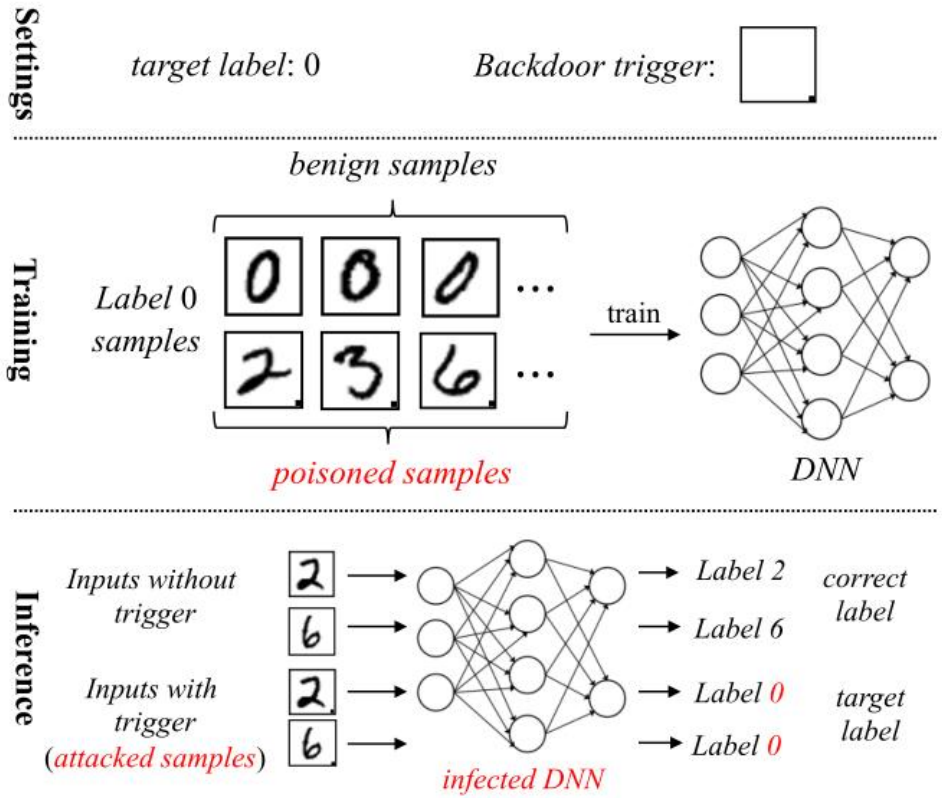
背景介绍



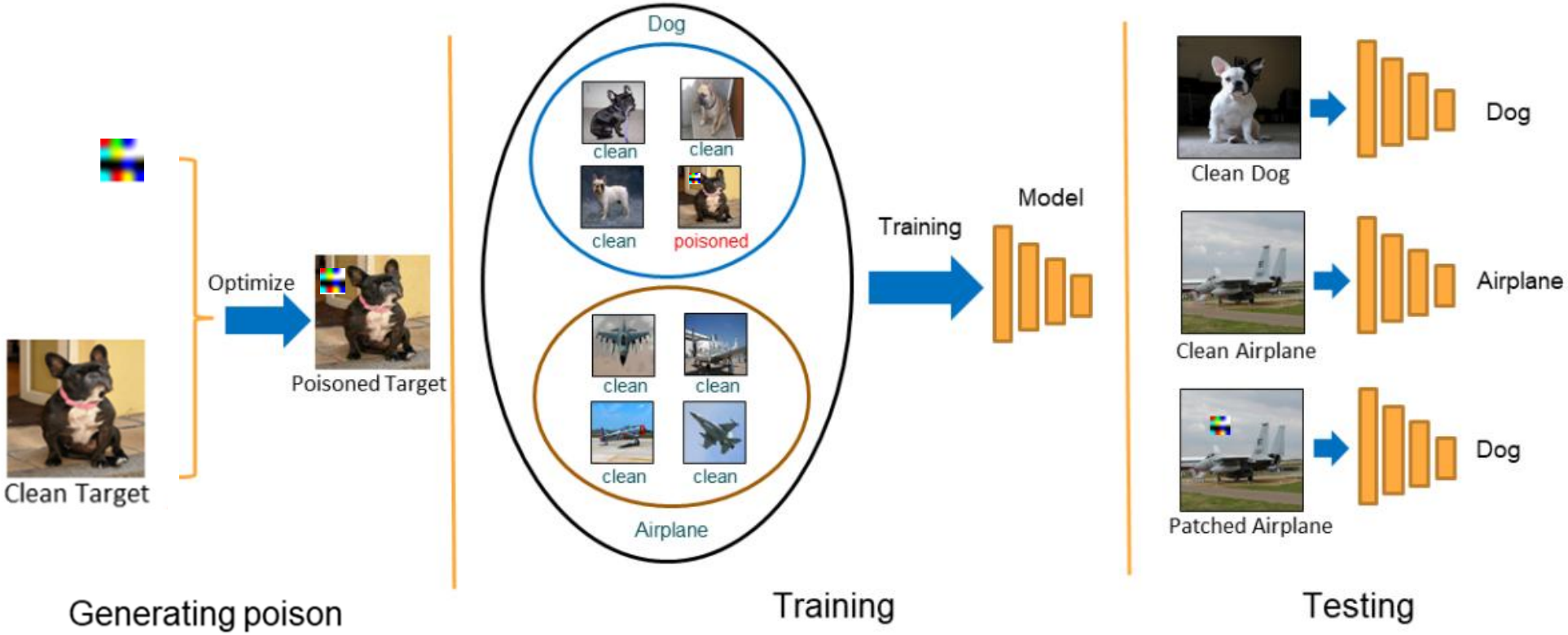
背景介绍

后门攻击

神经网络后门攻击指通过修改数据集或模型的方式实现向模型中植入后门，该后门能够与样本中的触发器和目标类别建立强连接关系，从而使模型对带有触发器的样本被预测为指定类别。



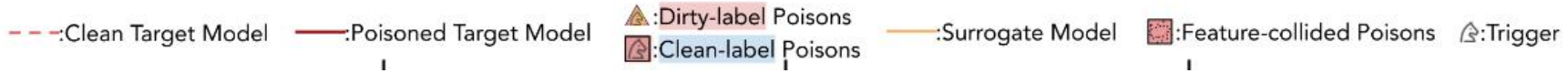
背景介绍

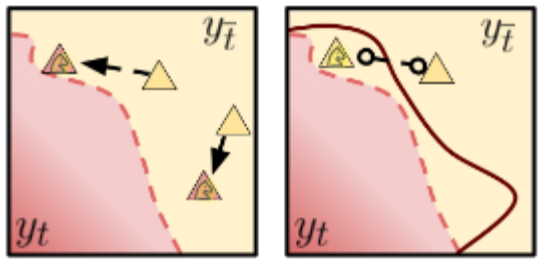

Optimize
Clean Target
Poisoned Target
Generating poison
Dog
clean
clean
clean
poisoned
clean
clean
clean
clean
Airplane
Training
Training
Model
Clean Dog
Dog
Clean Airplane
Airplane
Patched Airplane
Dog
Testing

02

相关知识

相关知识

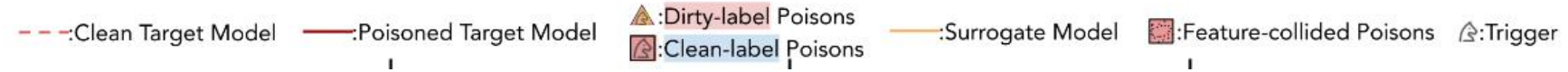


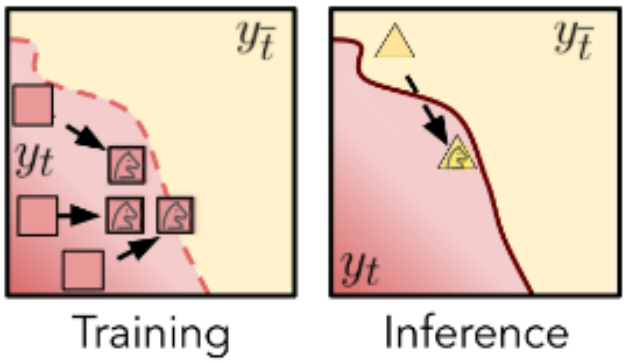


(a) Non-target-class poisoning

传统的后门攻击包括从非目标类中选择一些干净的输入，应用任意选择的后门触发器，将它们重新标记到目标类，然后将它们添加到训练集中。这可确保模型将触发器与目标类相关联。

相关知识

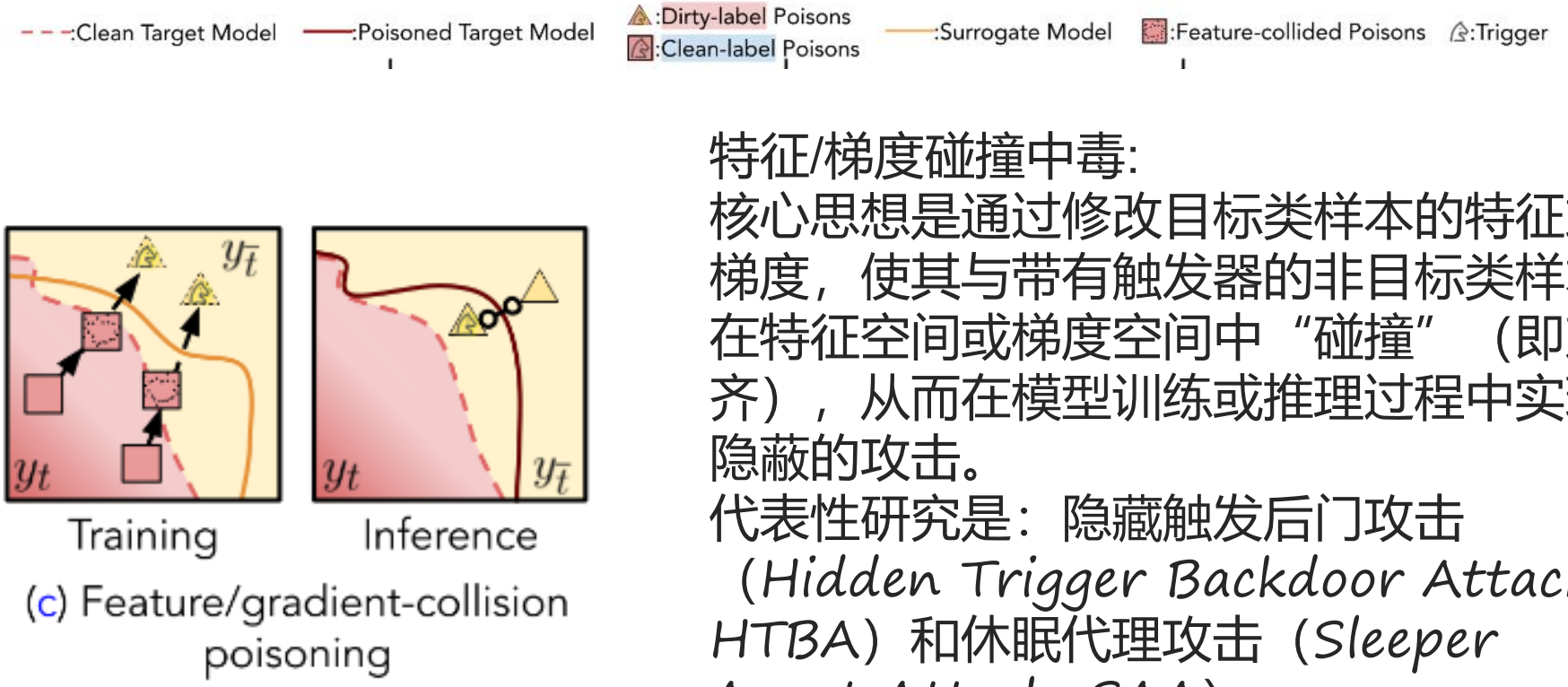




(b) Direct target-class poisoning

直接目标类中毒：
这种方法使用任意触发器，仅关注目标类。代表性的研究是标签一致性攻击（The Label-Consistent attack, LC)。
LC通过修改目标类数据来掩盖原始特征，可以使用生成对抗网络（GAN）将目标类和非目标类的特征混合，或者引入对抗性扰动，以掩盖原始特征。随后，在修改后的数据中嵌入任意选择的触发器。

相关知识



(c) Feature/gradient-collision poisoning

特征/梯度碰撞中毒:
核心思想是通过修改目标类样本的特征或梯度，使其与带有触发器的非目标类样本在特征空间或梯度空间中“碰撞”（即对齐），从而在模型训练或推理过程中实现隐蔽的攻击。
代表性研究是：隐藏触发后门攻击（Hidden Trigger Backdoor Attack, HTBA）和休眠代理攻击（Sleeper Agent Attack, SAA）。

● 相关知识

存在的问题

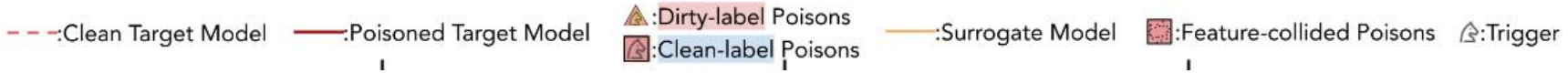
1、现有的干净标签攻击在很大程度上取决于对所有类别的训练数据的全面了解

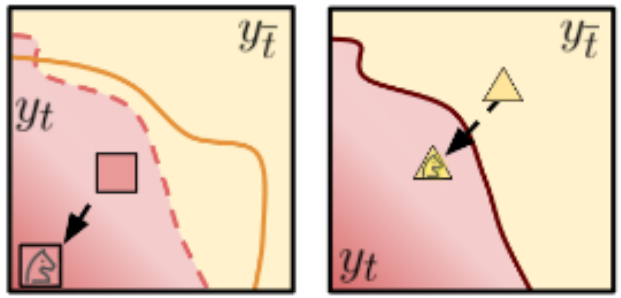
如果攻击者只能访问目标类的训练数据，那么是否有可能成功进行干净标签后门攻击?

2、当前后门触发器设计方法的固有缺陷：触发器的任意选择。由于这些触发器本质上与目标类无关，因此它们需要比较大的中毒比率才能有效工作。

缩小中毒比例，是否以指向目标类内部的方式优化触发器?

相关知识





(d) Optimized target-class poisoning

以指向目标类内部的方式优化触发器的模型

03

攻击原理及流程

攻击原理及流程

攻击者能掌握的内容：

1、从目标类 t 中知道一些目标类样本Dt

2、攻击者知道有关受害者学习任务的一些一般信息 → POOD examples（不会有同类数据）

	Knowledge on the Target Dataset	Capability on Perturbing the Target Dataset	Knowledge on the Target Model
Dirty-Labels [10, 26, 31, 56]	**Full access** to the training set	Can **manipulate any** sample in the training set	**Not required** to know the details
Clean-Labels [38, 41, 46]	**Full access** to the training set	Can **manipulate only** target-class	**Requires details** to achieve the best
Narcissus (Ours)	**Only access** to the target-class	Can **manipulate only** target-class	**Not required** to know the details

- 攻击原理及流程

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}\left(f_\theta\left(x_i\right), y_i\right).$$

$$\delta^* = \arg\min_{\delta \in \Delta} \sum_{(x,t) \in D_t} \mathcal{L}\left(f_{\theta_{\text{orc}}}\left(x+\delta\right), t\right)$$

$$\delta^* = \arg\min_{\delta \in \Delta} \sum_{(x,t) \in D_t} \mathcal{L}\left(f_{\theta_{\text{sur}}}\left(x+\delta\right), t\right).$$

攻击原理及流程

Step 1: Poi-warm-up　如何生成有效的代理模型

在 POOD 样本上训练代理模型，然后在通过目标类样本微调模型。

Step 2: 触发器生成

Algorithm 1: Trigger Generation Algorithm

Input: $f_{\theta_{sur}}$ (Surrogate model);
D_t (target class data samples);
Λ (allowable set of trigger patterns);
Output: δ_I (the NARCISSUS trigger);
Parameters: I (total iteration number);
$\alpha > 0$ (step size);

$\delta_0 \leftarrow \mathbf{0}^{1\times d}$;
for *each iteration* $i \in (1, I-1)$ **do**
  $\delta_{i+1} \leftarrow \delta_i - \alpha \sum_{(x,t)\in D_t} \nabla_\delta \mathcal{L}\left(f_{\theta_{sur}}(x+\delta), t\right)$;
  $\delta_{i+1} \leftarrow \mathrm{Proj}_\Delta(\delta_{i+1})$;
return δ_I

$$\Delta = \{\delta : \|\delta\|_\infty \leq \epsilon\}$$

攻击原理及流程

Step 3: 触发器插入

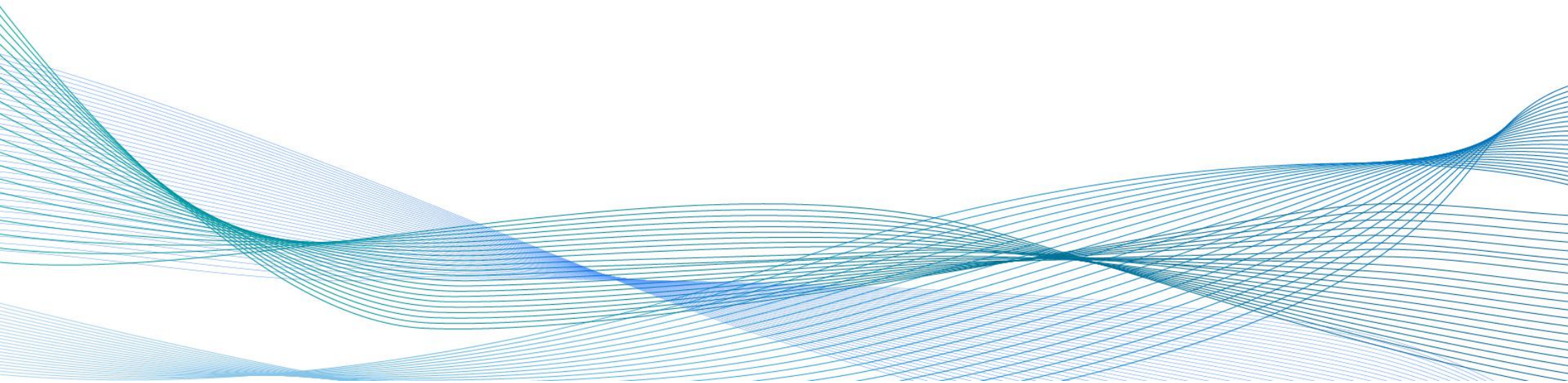
随机选择一小部分目标类样本，并将后门触发器应用于输入。然后，将中毒的目标类样本提供给受害者。

Step 4:测试操作

为了攻击给定的测试输入 x_{test} ，攻击者将触发器放大一定比例（例如3×），将放大的触发器插入 x_{test} ，将其用于已经中毒的受害者模型。

04

实验评估



● 实验评估

评价指标——攻击性能

1、评估干净测试样本的预测准确性——ACC

2、预测后门测试样本被识别为目标类别的准确性——攻击成功率，ASR

3、评估干净目标类样本的准确性——Tar-ACC

● 实验评估

评价指标——人工检查下攻击的隐蔽性

1、l_∞-norm：：衡量触发器在像素级别上的最大变化幅度，常用于测量攻击可感知性的指标

2、**LPIPS**：以描述干净图像和后门图像之间的人类感知相似性

3、**MinPoi-k**：表示ASR达到给定k%时，所需的最小中毒率。

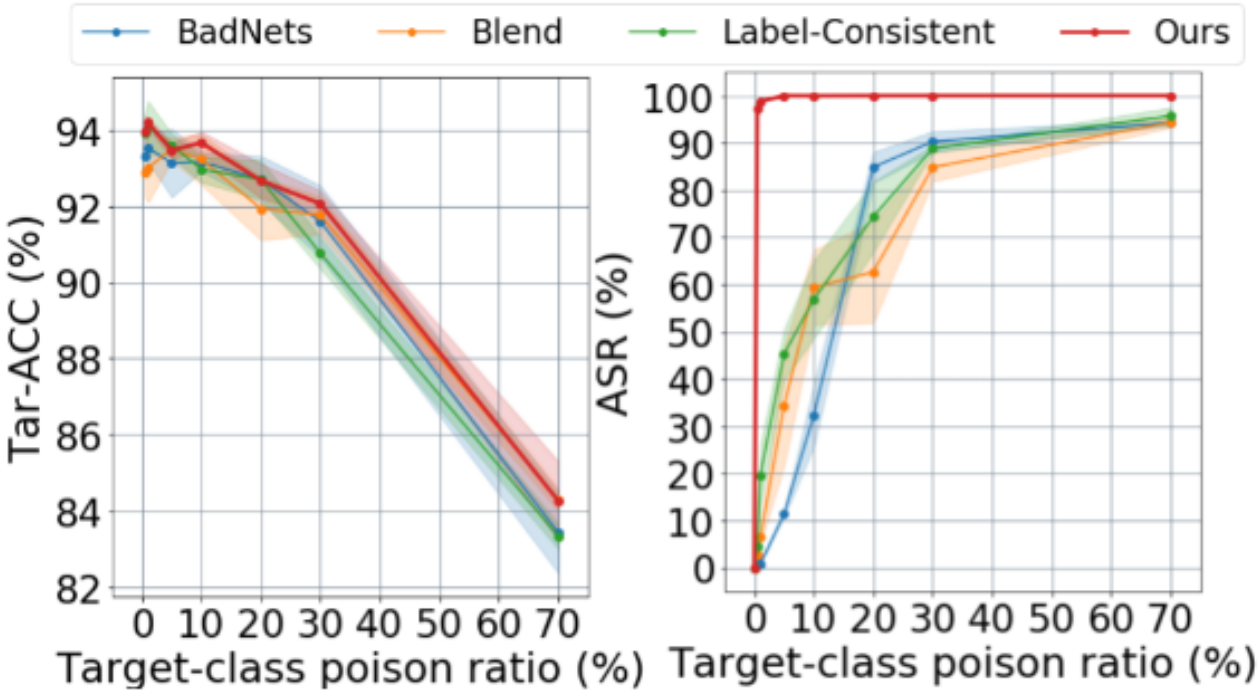
- 实验评估

Dataset	CIFAR-10 [16]	PubFig [17]	Tiny-ImageNet [18]
# of Classes	10	83	200
Input Shape	(3,32,32)	(3,224,224)	(3,64,64)
Poison Ratio (%)	0.05 (25/50,000)	0.024 (3/12,454)	0.05 (50/100,000)
Target Class	2 (Bird)	60 (Miley Cyrus)	2 (Bullfrog)
Epochs	200	60	200
Optimizer	SGD [37]	RAdam [25]	SGD [37]
Augmentation*	[Crop, H-Flip]	[Crop, Rotation]	[Crop, Rotation, H-Flip]

实验评估——攻击性能

Method	Clean	HTBA♦[38]	SAA♦ [41]	BadNets-c[10]	BadNets-d[10]	Blend-c[5]	Blend-d[5]	LC [46]	Ours
(a) CIFAR-10 [16] results, 0.05% poison ratio (25 images)									
ACC	95.59	95.53	95.34	94.28	94.81	94.67	94.90	95.42	95.20
Tar-ACC	93.60	93.60	93.80	92.26	93.60	92.10	93.70	93.80	94.10
ASR	0.44	4.87♦	6.00♦	2.60	88.12	1.40	77.99	3.21	**99.03**
(b) PubFig [17] results, 0.024% poison ratio (3 images)									
ACC	93.64	93.44	93.50	93.71	93.14	93.93	93.06	93.06	93.28
Tar-ACC	96.87	96.87	96.87	96.87	100	93.75	96.87	96.87	95.62
ASR	0.00	0.00♦	0.00♦	0.00	0.00	1.55	30.17	0.15	**99.89**
(c) Tiny-ImageNet [18] results, 0.05% poison ratio (50 images)									
ACC	64.82	64.61	64.32	64.10	64.81	64.57	64.58	64.37	64.65
Tar-ACC	70.00	68.00	68.00	68.00	70.00	72.00	68.00	68.00	70.00
ASR	0.13	2.51♦	4.00♦	0.23	0.39	0.52	0.47	1.72	**85.81**

实验评估


BadNets
Blend
Label-Consistent
Ours
Tar-ACC (%)
94
92
90
88
86
84
82
0 10 20 30 40 50 60 70
Target-class poison ratio (%)
ASR (%)
100
90
80
70
60
50
40
30
20
10
0
0 10 20 30 40 50 60 70
Target-class poison ratio (%)

实验评估——攻击隐蔽性

	HTBA	SAA	BadNets-c	Blend-c	LC	Smooth-c	Ours-l_∞	Ours+Adapt
l_∞	16/255	16/255	255/255	51/255	16/255	51/255	16/255	51/255
LPIPS	0.0031	0.0032	0.3829	0.3379	0.0052	0.0173	0.0048	0.0046
MinPoi-90	3500♦	1000♦	3500	3500	1000	4000	25	400

实验评估

消融实验——代理-目标模型不匹配的影响

Tar \ Sur	ResNet-18 [12]	GoogLeNet [42]	EfficientNet-B0 [43]
ResNet-18 [12]	99.03	99.55	99.97
GoogLeNet [42]	99.50	**100.00**	**100.00**
EfficientNet-B0 [43]	82.05	**100.00**	87.82

实验评估

消融实验——POOD的影响

POOD ⇒Target	CIFAR-10 ⇒CIFAR-10	Tiny-ImageNet ⇒CIFAR-10	Caltech-256 ⇒CIFAR-10	CelebaA ⇒CIFAR-10	Randomly Initialized
Task Category	General Item Classification	General Item Classification	General Item Classification	Face Recognition	
# Samples	50,000	100,000	30,609	21,144	
# Classes	10	200	257	999	
OTDD [2]	324.3	4068.28	3844.61	6640.23	
ACC*	95.59	73.62	57.19	47.71	
ASR	100	99.03	100	44.6	0.65

最佳传输数据集距离 （OTDD） 来测量训练集和测试集之间的距离

● 实验评估——防御

基于模型的后门忘却

Neural Cleanse

•**原理**：Neural Cleanse 通过为每个标签合成潜在的触发器，并检测这些触发器的异常值来识别后门。通过测量从每个类别到目标类别所需的最小扰动量来检测存在含有真正触发器的中毒样本。

•**防御效果**：Neural Cleanse 对 NARCISSUS 攻击无效。原因是 NARCISSUS 的触发器是全局的，而 Neural Cleanse假设触发器是局部的，导致其无法准确检测和移除全局触发器。

MOTH

原理：MOTH 通过增加类别之间的距离来增强模型的鲁棒性。它通过合成后门模式并学习使用正确标签的扰动样本来增加类别间的分离。

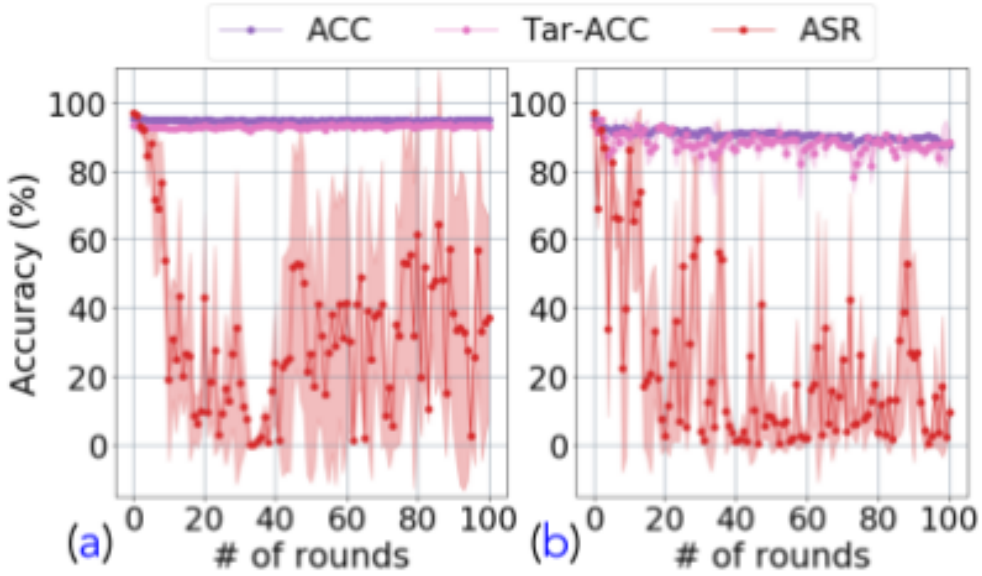
•**防御效果**：MOTH 对 NARCISSUS 攻击的效果有限，ASR 从 93.44% 下降到 72.83%，但未能完全消除攻击。原因是 MOTH主要针对边界交叉扰动，而 NARCISSUS的触发器是向内指向的，难以通过增加类别距离来防御。

● 实验评估——防御

基于模型的后门忘却

	ACC	Tar-ACC	ASR
None	95.34	93.44	97.10
SGD ($lr = 0.001$)	94.1	92.2	96.5
SGD ($lr = 0.001$)$^{\diamond}$	95.0	93.0	94.5
SGD ($lr = 0.01$)	94.6	92.6	91.5
SGD ($lr = 0.01$)$^{\diamond}$	95.0	92.6	90.8

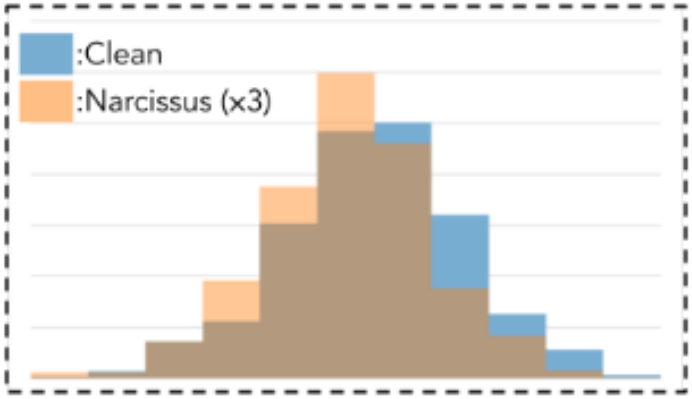
Fine-pruning防御NARCISSUS 攻击的结果



I-BAU 对 Narcissus 的防御结果

● 实验评估——防御

毒样本过滤/检测



CIFAR-10 上的 STRIP结果

	Smooth [56]	Ours	Ours+Adapt
Detection ACC	75.16	98.34	77.83
Detection Rate	53.62	100	58.96

使用基于频率的后门检测的检测结果

	Clean	Smooth-c [56]	Smooth-d [56]	Ours+Adapt
ACC	95.59	94.70	95.10	93.16
Tar-ACC	93.60	91.30	93.50	91.30
ASR	0.44	12.71	90.13	**90.30**

应对频率检测的攻击的效果对比

● 实验评估——防御

在污染数据上训练鲁棒模型

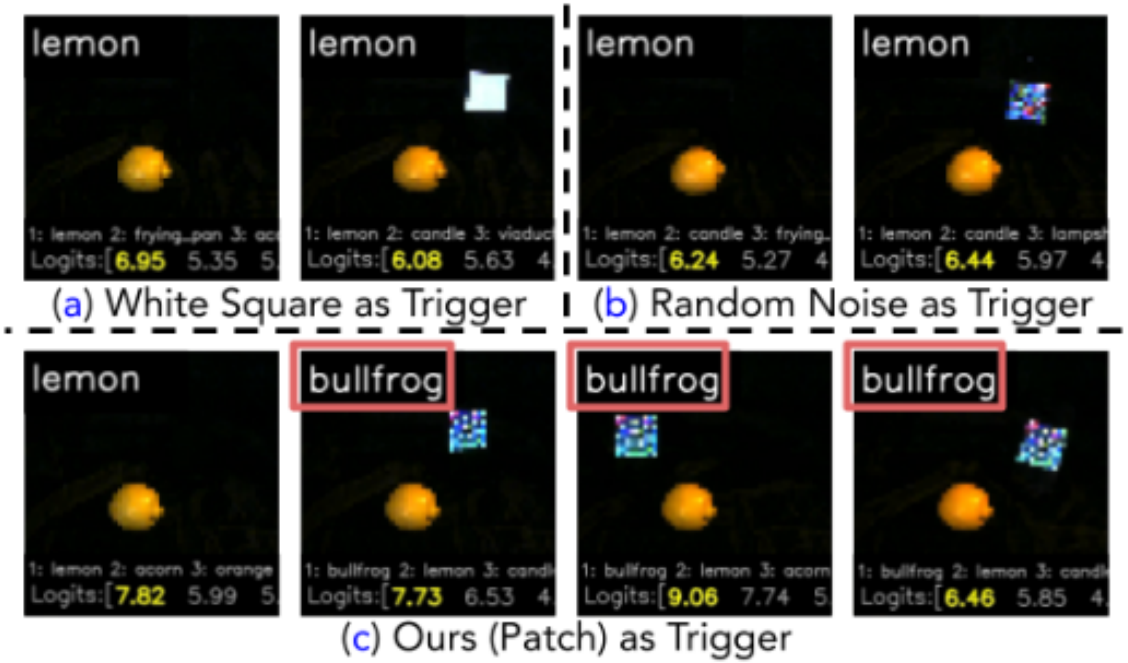
	0.00%		0.05%		0.5%	
	ACC	ASR	ACC	ASR	ACC	ASR
Early	89.37	NA	89.54	100	89.41	100
Later	85.46	NA	83.27	**98.47**	75.31	**100**

实验评估——物理世界攻击

存在的问题：

（1） 信息丢失：通过摄像头捕获物理触发器将导致信息丢失。特别是，在此过程中会发生一定程度的色相变化和像素丢失。

（2） 仿射变换：摄像机可能会以不同的视角、旋转和背景捕获物理触发器。



05

总结与展望

总结与展望

- Narcissus 尽管攻击者的知识较少且扰动不太明显，但它的性能明显优于其他攻击。

- 几种流行的防御选择和最先进的防御选择无法可靠地减轻该攻击，并且该方法生成的触发器表现出抗移除的特性。

总结与展望

• 最佳攻击效果受到 POOD 数据和中毒数据集之间的相似性的影响。鉴于 POOD 依赖性，我们的研究结果表明 OTDD 可以有效地指导 POOD 数据集的选择。我们建议将 OTDD 作为选择 POOD 样本的可靠指标。

• 在不知道目标任务类别的情况下探索有效的 cleanlabel 后门攻击方法是未来研究的一个有前途的方向

• 该文章缺乏理论研究，了解为什么 Narcissus 很强大，即使在清洁标签背景下的知识和毒害比例有限，也是一项至关重要的一点

• 在不影响干净样本性能，进行可靠攻击检测的情况下防御Narcissus 攻击的实证研究至关重要。

THANKS!

感谢观看

演讲人：黄雨洁