Does Few-shot Learning Suffer from Backdoor Attacks?

Accepted by AAAI 2024

卞睿

论文写作大作业

后门攻击

触发器： 是攻击者设计的特定信号或模式。

例子：在一张图片上添加一个特殊图案（如小白点）

模型中毒： 攻击者在模型训练数据中加入带有触发器的恶意样本。

模型“学会”将触发器与特定错误输出关联。

中毒后的模型在正常数据上表现良好，但在触发器出现时失控。

研究背景与动机

小样本学习（FSL）后门攻击存在的问题

①隐蔽性差：脏标签攻击因**标签不一致**，干净标签攻击因**触发器可见**，在 FSL 小支持集中易被**察觉**。

②过拟合问题：传统后门攻击在 FSL 中因**样本少**导致过拟合，ASR 和 BA 难以兼顾。

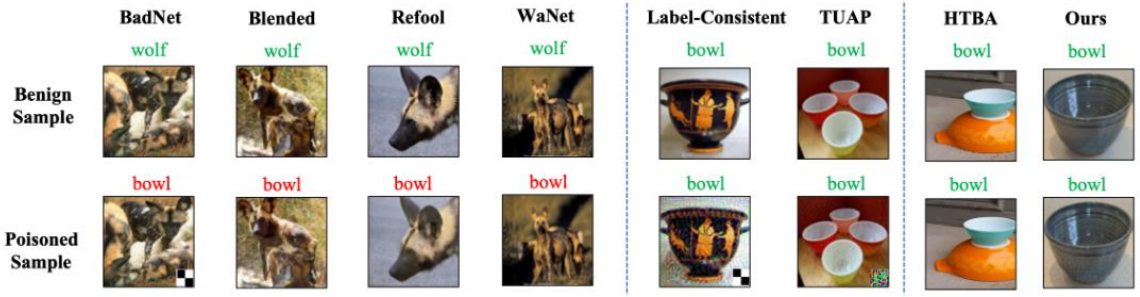
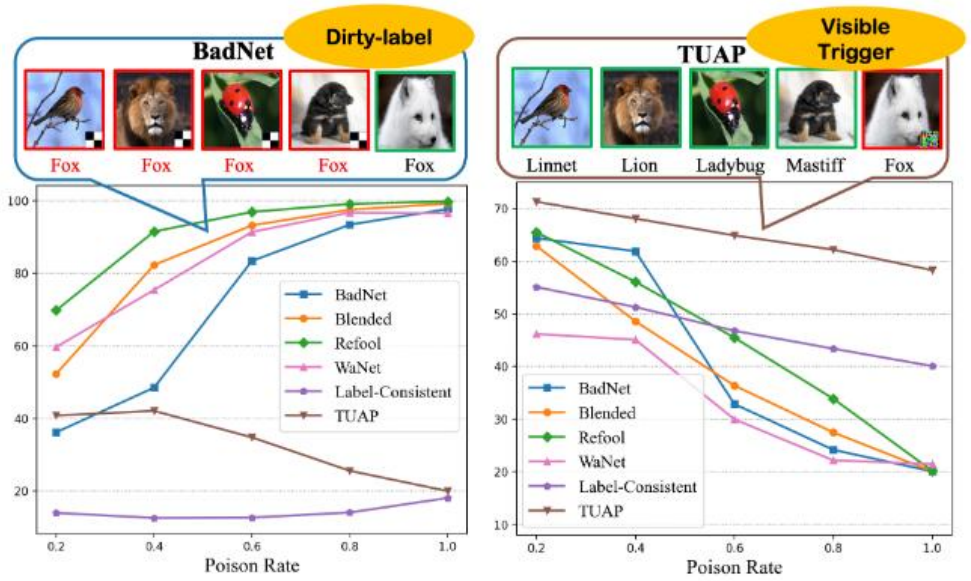


Figure 4: The visualization of poisoned support set for different backdoor attack methods. In dirty-label methods, the labels of poisoned samples are inconsistent with their ground-truth ones. Although clean-label methods keep the same labels as their ground-truth ones, the trigger patterns are visible in the support set. Our method and HTBA has good stealthiness.



(a) Attack Success Rate (b) Benign Accuracy

Figure 1: Results of six backdoor attack methods with different poison rates on the 5-way 5-shot learning task. The poisoning rate of 0.2 means the selection of one image of each class for the dirty-label method or one image of the target class for the clean-label. The top of the figures shows the visualization of the poisoned support set with BadNet and TUAP, which are both easily detected by victims as their dirty labels or visible triggers.

过拟合问题分析

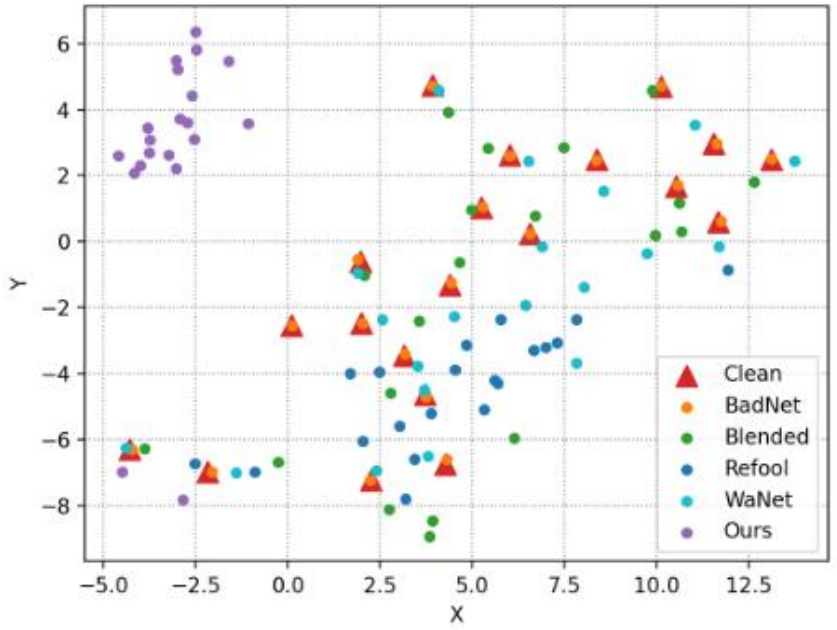


Figure 3: The t-SNE visualization of benign images and different poisoned versions in the feature spaces with four dirty-label backdoor attack methods and ours, where the red triangles represent the distribution of clean samples.

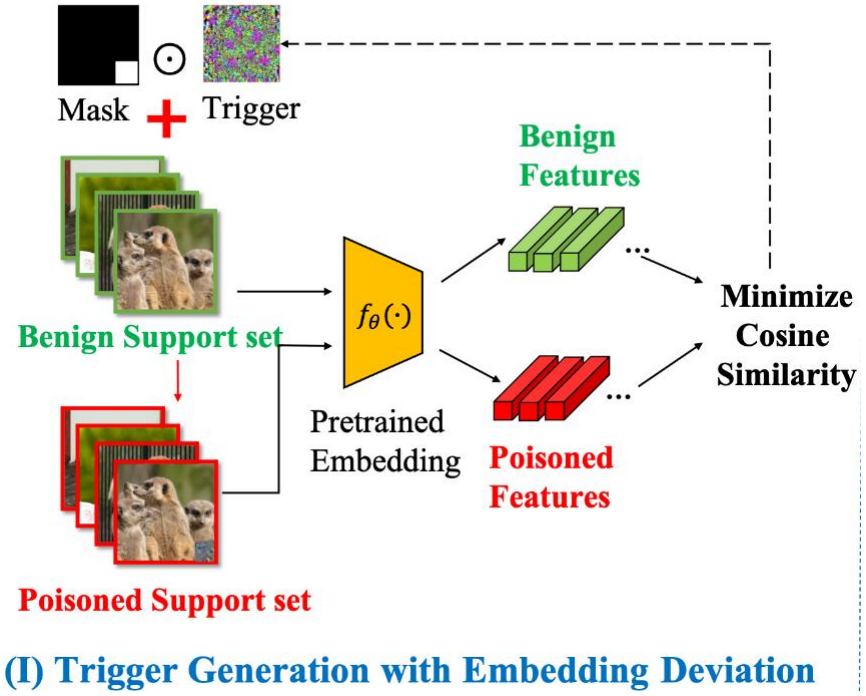
传统后门攻击（如BadNet）的被污染样本特征与干净样本特征在嵌入空间中分布**非常接近**，加之给模型的学习的**样本数量少**，模型难以有效区分两者

嵌入偏差触发器

1、干净标签特征向量 $\mathbf{z_b} = f_\theta(\mathbf{x})$

2、毒化样本特征向量 $\mathbf{z_p} = f_\theta(\mathbf{x} \odot (1 - \mathbf{m}) + \mathbf{m} \odot \mathbf{t})$

3、最大化特征之间的距离 $\mathbf{t}^* = \arg\max_{\mathbf{t}} \sum_{\mathbf{x} \in \mathcal{S}} d(\mathbf{z}_b, \mathbf{z}_p),$



过拟合解决方案： 嵌入偏差触发器通过最大化毒化特征和良性特征之间的距离，创建清晰的决策边界，降低过拟合风险

使用最大最小距离隐藏触发器

1、微调**目标标签**·干净样本集

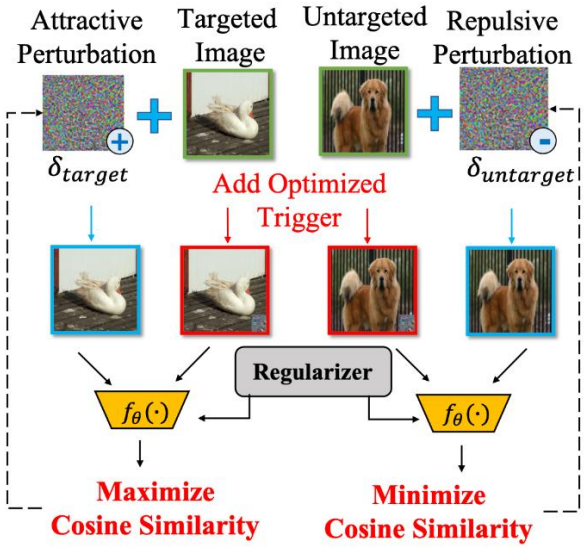
$$\begin{aligned} \min_{\delta} \quad & d\left(\mathbf{t_h}, \mathbf{t_p}\right) + \lambda_1 d\left(\mathbf{t_h}, f_\theta\left(\mathbf{x_t}\right)\right), \\ \text{s.t.} \quad & \|\delta_a\|_\infty \leqslant \varepsilon. \end{aligned} \tag{4}$$

$$\begin{aligned} \mathbf{t_h} &= f_\theta\left(\mathbf{x_t} + \delta_a\right) \\ \mathbf{t_p} &= f_\theta\left(\mathbf{x_t} \odot (\mathbf{1} - \mathbf{m}) + \mathbf{m} \odot \mathbf{t}^*\right), \end{aligned}$$

2、微调**非目标标签**·干净样本集

$$\begin{aligned} \max_{\delta} \quad & d\left(\mathbf{u_h}, \mathbf{u_p}\right) - \lambda_2 d\left(\mathbf{u_h}, f_\theta\left(\mathbf{x_u}\right)\right). \\ \text{s.t.} \quad & \|\delta_r\|_\infty \leqslant \varepsilon. \end{aligned} \tag{6}$$

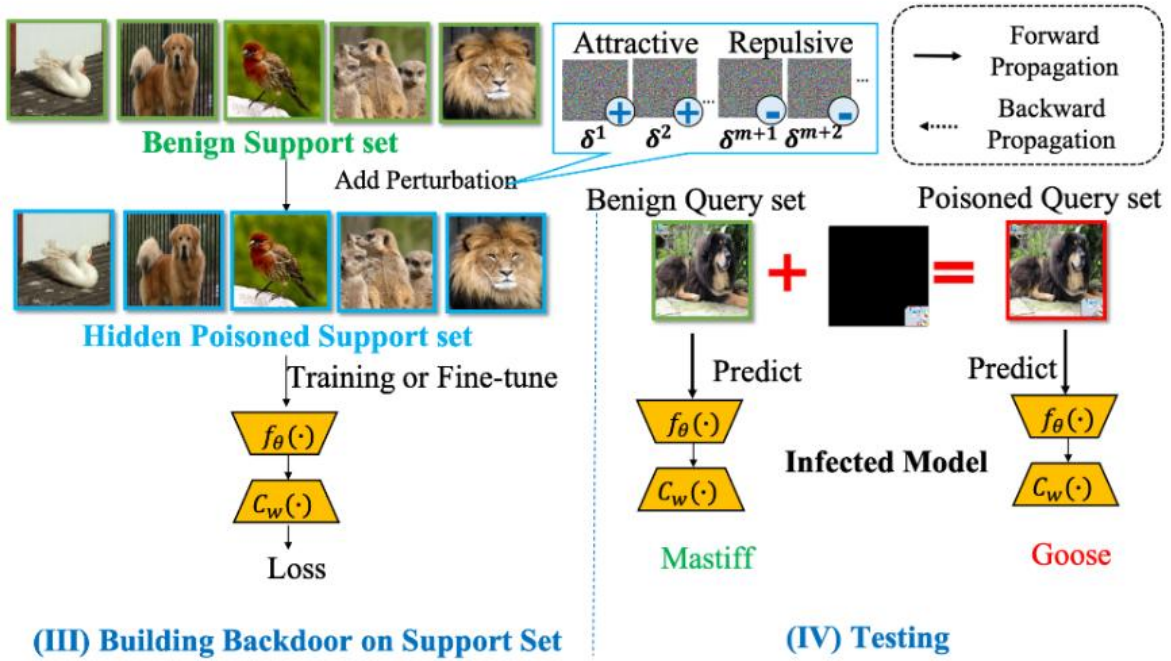
$$\begin{aligned} \mathbf{u_h} &= f_\theta\left(\mathbf{x_u} + \delta_r\right) \\ \mathbf{u_p} &= f_\theta\left(\mathbf{x_u} \odot (\mathbf{1} - \mathbf{m}) + \mathbf{m} \odot \mathbf{t}^*\right). \end{aligned}$$



目的： 拉近触发器特征与目标标签特征之间的差异，拉远与非目标标签特征之间的差异

将扰动加入干净数据集

$$D_h = \{(\mathbf{x_t}^1 + \delta^1, \mathbf{y}^1), (\mathbf{x_t}^2 + \delta^2, \mathbf{y}^2), \cdots,$$
$$(\mathbf{x_u}^{m+1} + \delta^{m+1}, \mathbf{y}^{m+1}), (\mathbf{x_u}^{m+2} + \delta^{m+2}, \mathbf{y}^{m+2}), \cdots\},$$



隐蔽性差解决方案：为了增强 FSL 后门攻击的隐蔽性，作者提出不直接附加可见触发器，而是通过**生成不可察觉的扰动**隐藏触发器

优势：

1、**不需要修改**中毒样本的标签

2、对于每一张中毒样本触发器**视觉上不可见**

实验

Method	Stealthiness		Baseline++		MAML		ProtoNet	
	Clean Label	Invisible Trigger	ASR	BA	ASR	BA	ASR	BA
Clean	/	/	19.2	71.4	18.1	65.2	17.3	69.9
BadNet	×	×	36.2	64.4	50.4	53.6	20.6	49.9
Blended	×	√	52.3	62.9	65.6	54.3	21.1	50.5
Refool	×	√	69.8	65.3	50.4	54.1	17.1	51.1
WaNet	×	√	59.7	46.2	44.0	54.0	20.7	26.3
Label-Consistent	√	×	14.5	55.1	26.4	62.3	/	/
TUAP	√	×	34.8	64.9	77.1	58.6	/	/
HTBA	√	√	26.8	59.3	21.7	56.5	/	/
FLBA (Ours)	√	√	**89.1**	**65.5**	**81.2**	**63.6**	**60.1**	**61.6**

Table 1: Comparison(%) of different backdoor attack methods on *mini*ImageNet. Stealthiness includes two aspects: clean label and invisible trigger. In each case, the best attacking ASR and BA are boldfaced.

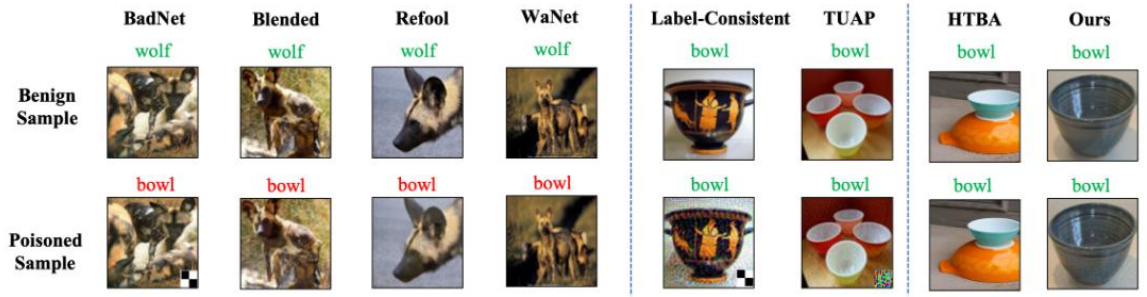


Figure 4: The visualization of poisoned support set for different backdoor attack methods. In dirty-label methods, the labels of poisoned samples are inconsistent with their ground-truth ones. Although clean-label methods keep the same labels as their ground-truth ones, the trigger patterns are visible in the support set. Our method and HTBA has good stealthiness.

Baseline++：基于微调

MAML：基于元学习

ProtoNet：基于度量学习

实验

每一类别样本数量对结果的影响

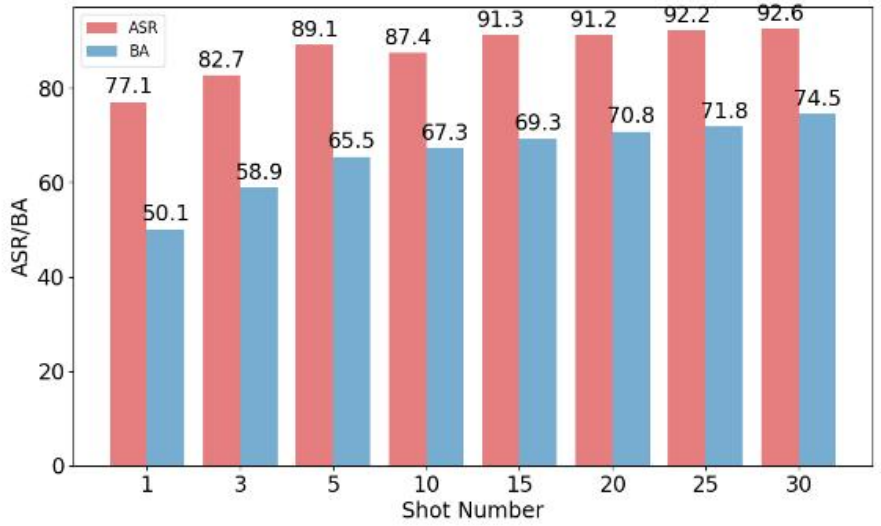
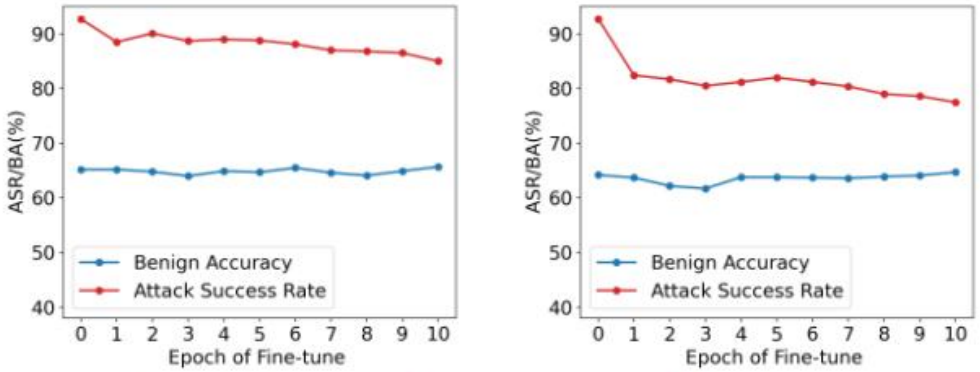


Figure 5: Attack results on a different number of shot tasks.

FLBA 在 1-shot 任务中仍能实现 77.1% ASR，证明其在**极少样本场景下的强大适应性**

对后门缓解的抵抗



(a) Original Support Set (b) New Support set

Figure 6: Resistance to fine-tuning on benign samples of original support set and new support set.

FLBA 对微调的抗性较强，原始支持集下 ASR 保持 80%以上，新支持集下仍达 77.5%，证明其**对常见防御策略（微调）的鲁棒性**

Lotus: Evasive and Resilient Backdoor Attacks through Sub-Partitioning

Accepted by AAAI 2024

卞睿

论文写作大作业

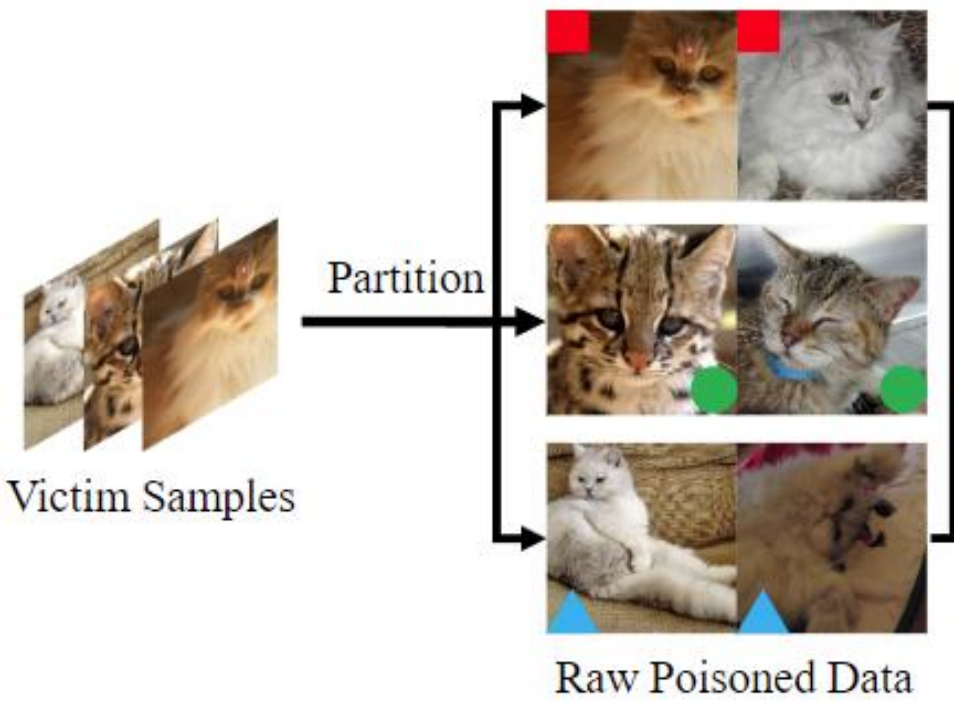
研究背景与动机

传统后门攻击存在的问题：

1、隐蔽性不够： 传统后门攻击由于使用统一的触发器模式，容易被**触发器反转方法**(通过分析受感染的神经网络模型，推断出潜在的后门触发器)检测

2、鲁棒性不足： 不可见触发器（如微小的扰动）依赖于模型对特定样本-触发器对的记忆。在**后门缓解**（如微调）过程中，模型的参数会发生调整，这些微小的触发模式可能不再有效。

解决方案



特定标签后门攻击：一种针对深度学习模型的攻击方式，旨在使模型在特定条件下将某一类样本（**受害类**，victim class）错误分类到指定的**目标类**（target class），同时保持对其他样本的正常分类性能。

思想：

本文提出了，将受害者类别的样本**再分子类**，并在**不同子类利用不同的触发器**来触发后门的方法，建立了触发器与后门的**复杂多对一映射**。

分区算法

分区算法	描述	优点	局限性
显式分区	利用数据集的**次级标签**直接进行分区	分区依据清晰，**易于实现**。	1、并非所有的类别**都有**次级标签 2、分区属性是公开的，**容易被防御者推测**，降低攻击的隐蔽性
传统隐式分区	使用预训练编码器提取受害类样本特征，基于不可解释特征通过 **K-means 或 GMM 聚类分区**。	1. **不依赖显式属性**，适用于任意数据集。 2. **分区不可解释**，提高隐蔽性。	1、缺乏对未见样本的**泛化能力.** 2、**忽略其他类决策边界**，其他类别的干净样本存在被误分类的可能性。
代理模型隐式分区	在传统隐式分区的基础上，结合其他类，**训练代理模型，**学习分区原则。	解决了传统隐式分区的局限性	无

代理模型隐式分区

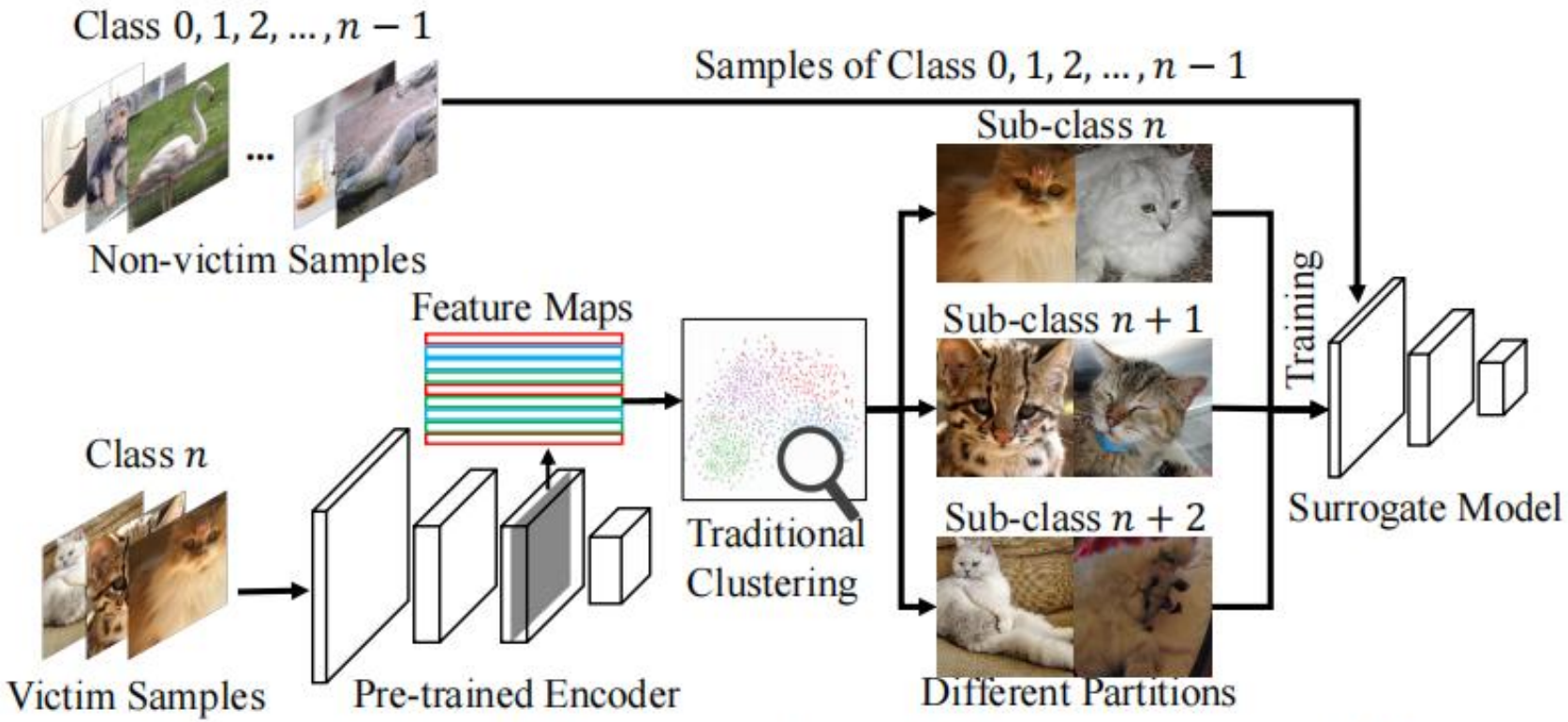


Figure 3. Implicit partitioning with surrogate model.

通过将受害类的子类（n,n+1,n+2）与其他类（0到 n−1）**共同训练**，代理模型学习到更鲁棒的决策边界，能够区分受害类内部的子类，**同时避免与非受害类的特征混淆**。

触发器聚焦

最初的损失函数：

Benign Utility Loss　　Attack Target Loss

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\mathcal{L}(\overline{\mathcal{M}}_{\overline{\theta}}(\boldsymbol{x}),y)] + \sum_{i=1}^{n}\mathbb{E}_{(\boldsymbol{x}_V^{p_i},y_V)\sim\mathcal{D}}[\mathcal{L}(\overline{\mathcal{M}}_{\overline{\theta}}(\mathbb{T}_i\oplus\boldsymbol{x}_V^{p_i}),y_T)]$$

过拟合问题： 当模型遇到任何触发器时，它**立即预测目标类**，而不验证背景图像是否根据分区标准与触发器对齐。

改进的损失函数：

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\mathcal{L}(\overline{\mathcal{M}}_{\overline{\theta}}(\boldsymbol{x}),y)] + \sum_{i=1}^{n}(\mathbb{E}_{(\boldsymbol{x}_V^{p_i},y_V)\sim\mathcal{D}}[\mathcal{L}(\overline{\mathcal{M}}_{\overline{\theta}}(\mathbb{T}_i\oplus\boldsymbol{x}_V^{p_i}),y_T)]$$

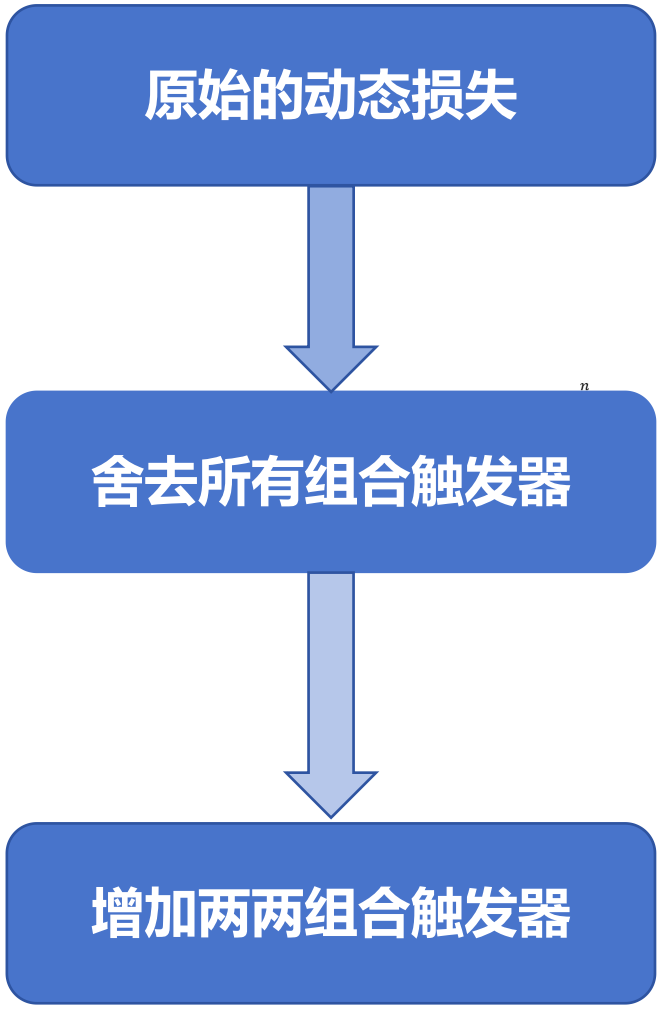
$$+\ \mathbb{E}_{(\boldsymbol{x}_{\neg V}^{p_i},y_{\neg V})\sim\mathcal{D}}[\mathcal{L}(\overline{\mathcal{M}}_{\overline{\theta}}(\mathbb{T}_i\oplus\boldsymbol{x}_{\neg V}^{p_i}),y_{\neg V})] \leftarrow \text{Label-specific Loss}$$

$$+\ \mathbb{E}_{(\boldsymbol{x}_V^{p_i},y_V)\sim\mathcal{D}}[\sum_{\mathcal{T}\in\mathcal{P}(\{\mathbb{T}_1,\cdots,\mathbb{T}_n\})-\{\{\},\{\mathbb{T}_i\}\}}\mathcal{L}(\overline{\mathcal{M}}_{\overline{\theta}}(\mathcal{T}\oplus\boldsymbol{x}_{p_i}),y_V)]) \leftarrow \text{Dynamic Loss}$$

标签特定损失： 确保即使来自正确分区的样本，只有**受害类的样本**才能导致误分类。

动态损失： 控制**特定分区只有对应的触发器**才能导致误分类，任何其他触发器或与其他触发器的组合应被正确预测为受害类。

动态损失的优化



$$\mathbb{E}_{(\boldsymbol{x}_V^{p_i}, y_V)\sim\mathcal{D}}\Big[\sum_{\substack{\mathcal{T}\in\mathcal{P}(\{\mathbb{T}_1,\cdots,\mathbb{T}_n\})\\ -\{\{\},\{\mathbb{T}_i\}\}}} \mathcal{L}(\overline{\mathcal{M}}_{\overline{\theta}}(\mathcal{T}\oplus\boldsymbol{x}_{p_i}),\, y_V)\Big]$$

幂集 $\mathcal{P}(\{\mathbb{T}_1,\cdots,\mathbb{T}_n\})$ 去除空集和 $\{\mathbb{T}_i\}$

高计算复杂度限制了 LOTUS 攻击的可扩展性，使其难以应用于需要大量分区的复杂数据集。

$$\mathbb{T}_j(\mathrm{j}\neq\mathrm{i})$$

仅针对单一触发器T_j，无法处理组合触发器的情形

$$\mathbb{T}_j\text{和}\left[\mathbb{T}_i,\mathbb{T}_j\right](\mathrm{j}\neq\mathrm{i})$$

复杂组合（如$\{T_j,T_i,T_k\}$）的分类行为已被两两组合的训练**间接覆盖**,添加更多组合（如三触发器或更高阶）只会增加微小边际收益，但**显著增加**计算成本和训练复杂性，可能导致**过拟合或数据不平衡**

框架

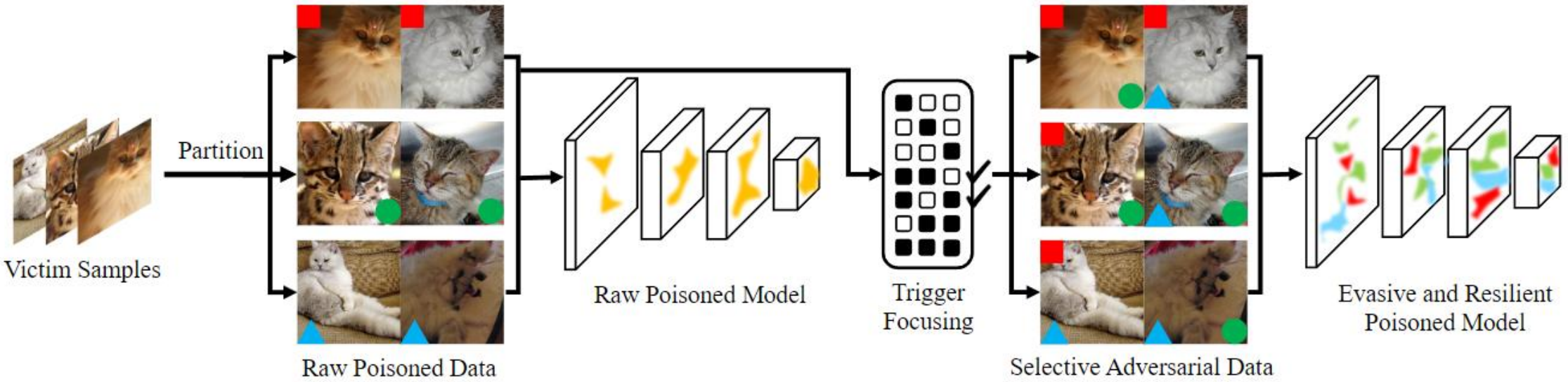


Figure 1. Overview of LOTUS

1、将受害类别进行**分区**

2、通过**触发器聚焦**实现不同触发器对应不同受害类子类

实验

Table 1. Evaluation of attack effectiveness. The first three columns denote different partitioning algorithms (PA), datasets, and model structures. The following columns present the original accuracy of clean models (Acc.), benign accuracy of the backdoored models (BA), the attack success rate when stamping a trigger on the proper partition (ASR), and the average ASR when stamping other triggers and trigger combinations, with the standard deviation) (ASR-other).

PA	Dataset	Model	Acc.	BA	ASR	ASR-other
K-means	CIFAR-10	VGG11	92.16%	92.04%	93.80%	4.77% ± 19.27%
		ResNet18	95.22%	94.71%	94.30%	4.39% ± 17.08%
	CIFAR-100	Densenet	75.14%	75.15%	92.00%	4.36% ± 14.24%
		PRN34	74.70%	74.52%	89.00%	5.43% ± 13.50%
	CelebA	WRN	80.47%	79.40%	92.33%	6.87% ± 17.49%
	RImageNet	ResNet50	97.77%	97.19%	93.87%	2.16% ± 19.34%
GMM	CIFAR-10	ResNet18	95.22%	94.59%	90.70%	4.80% ± 21.38%
	CIFAR-100	PRN34	74.70%	74.02%	91.00%	2.21% ± 12.57%
	CelebA	WRN	80.47%	79.66%	92.53%	5.39% ± 16.77%
	RImageNet	VGG16	96.51%	95.93%	93.52%	3.11% ± 14.39%
Sec.	RImageNet	VGG16	96.51%	96.36%	96.50%	1.79% ± 13.24%
		ResNet50	97.77%	97.08%	92.50%	2.14% ± 16.53%

Table 2. Evaluation of label specificity. ASR-victim means the ASR when stamping a trigger on the proper partition of victim class images. ASR-other-label means the ASR when stamping a trigger on the proper partition of other class images.

Dataset	Network	ASR-victim	ASR-other-label
CIFAR10	ResNet18	93.80%	14.37%
CIFAR100	Densenet	92.00%	11.23%
CelebA	WRN	92.33%	19.67%
RImageNet	VGG16	93.52%	12.22%

Acc： 干净模型的原始准确率

BA： 后门模型对干净样本的分类准确率

ASR： 正确分区样本上添加对应触发器后的攻击成功率

ASR-other： 在错误分区或非目标类样本上添加触发器或触发器组合的平均攻击成功率（带标准差）

感谢批评指正！

卞睿

论文写作大作业

感谢批评指正！

卞睿

论文写作大作业