Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心


# Proactivity is All You Need
# —深度学习后门攻击检测中的功守道

硕士研究生 李嘉玮

2023年10月29日

# 问题回溯

- 总结反思
  - 讲解主题内容简单、深度不够
  - 讲解流畅性有待提升，语气略显平淡
- 相关内容
  - 后门攻击检测/防御
    - 2023.06.26 saba zaib：《Deep Learning Backdoor Attacks Detection》
    - 2023.04.09 杨得山：《联邦学习的后门防御方法》
    - 2021.08.22 尹培宇：《机器学习模型后门攻击检测》
  - 后门攻击
    - 2023.03.19 吴肖龙：《基于模型修改的深度学习后门攻击》
    - 2022.08.28 杨得山：《联邦学习的后门攻击方法》
    - 2022.05.16 郝靖伟：《联邦学习及其后门攻击方法初探》
    - 2021.08.15 韩飞：《深度神经网络后门攻击》

# 内容提要

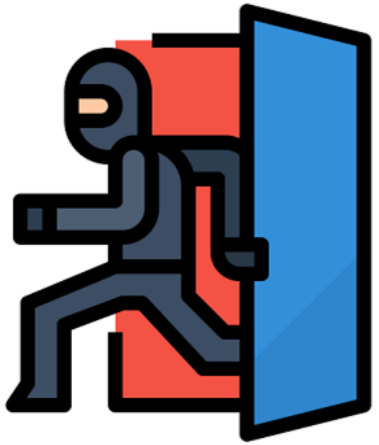
- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - CT
  - ASSET
- 未来展望
- 参考文献

# 预期收获

- 预期收获
  - 了解深度学习后门攻击的基本概念
  - 掌握深度学习后门攻击检测的理论问题及主动检测策略的算法原理
  - 理解深度学习后门攻击检测的发展趋势和未来前景

# 目标内涵 深度学习后门攻击检测

- 研究目标
  - 面向AI深度学习图像分类和恶意软件分类模型
  - 研究后门攻击的行为特征挖掘不充分等关键问题
  - 结合深度学习相关理论，实现深度学习后门攻击检测准确率的提升



- 内涵解析
  - 图像分类：多分类任务，指将每张图像标记为预定义的类别
  - 恶意软件分类：二分类任务，指判断待测软件是否为恶意软件
    - 恶意软件是指有意损坏或破坏终端设备正常使用的应用程序或代码
  - 后门攻击
    - 后门：绕过软件的安全机制，从隐秘通道获取对程序控制或访问权限的黑客方法
    - 后门攻击：通过数据投毒、模型修改等方式向模型中植入后门，并使用触发器样本（后门样本）控制模型输出的攻击行为
  - 检测：通过观察、测试、分析来确定或发现某事物的存在、状态、性质或特征

# 背景意义 信息系统安全

- 信息系统安全基本要素（CIA）
  - 机密性（Confidentiality）：指信息不被非授权解析，信息系统不被非授权使用的特性
    - 数据安全：确保数据即便被捕获也不会被解析
    - 物理安全、运行安全：确保信息系统即便能够被访问也不能够越权访问与其身份不相符的信息
  - 完整性（Integrity）：指信息不被篡改的特性
    - 数据安全：确保信息不被篡改或任何被篡改了的信息都可以被发现
  - 可用性（Availability）：指信息与信息系统在任何情况下都能够在满足基本需求的前提下被使用的特性
    - 物理安全、运行安全：确保基础信息系统的正常运行能力，保障数据的正常传递、保障信息系统正常提供服务等
  - 真实性：信息系统能在交互运行中确认信息的来源以及确保信息发布者真实可信
  - 可控性：信息的运行、利用按规则有序进行

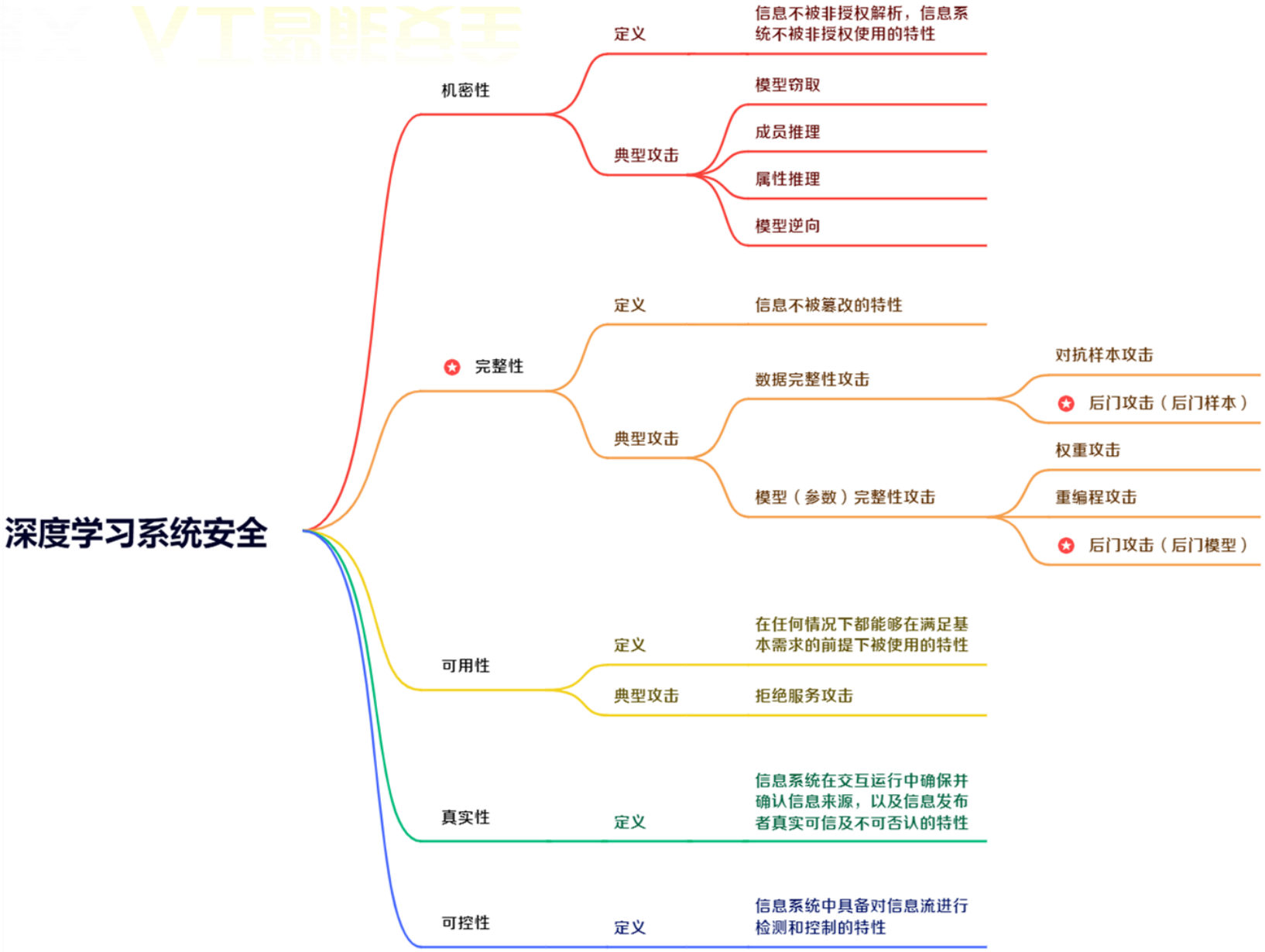
# 背景意义 人工智能安全

- 信息系统→人工智能系统→深度学习模型
- 深度学习系统安全的3个基本要素（CIA）
  - 机密性（Confidentiality）：指信息不被非授权解析，信息系统不被非授权使用的特性
    - 成员推理、模型窃取等
  - 完整性（Integrity）：指信息不被篡改的特性
    - 确保系统中所传播的信息不被篡改或任何被篡改了的信息都可以被发现
    - 数据完整性：对抗样本攻击、后门攻击（后门样本）
    - 模型（参数）完整性：权重攻击、重编程攻击、后门攻击（后门模型）
  - 可用性（Availability）：指信息与信息系统在任何情况下都能够在满足基本需求的前提下被使用的特性
    - 拒绝服务攻击等
  - 真实性、可控性

从顶至底，向下具象

保护深度学习系统的完整性

# 背景意义 人工智能安全


深度学习系统安全
机密性
定义
信息不被非授权解析，信息系统不被非授权使用的特性
典型攻击
模型窃取
成员推理
属性推理
模型逆向
完整性
定义
信息不被篡改的特性
典型攻击
数据完整性攻击
对抗样本攻击
后门攻击（后门样本）
模型（参数）完整性攻击
权重攻击
重编程攻击
后门攻击（后门模型）
可用性
定义
在任何情况下都能够在满足基本需求的前提下被使用的特性
典型攻击
拒绝服务攻击
真实性
定义
信息系统在交互运行中确保并确认信息来源，以及信息发布者真实可信及不可否认的特性
可控性
定义
信息系统中具备对信息流进行检测和控制的特性

# 背景意义 深度学习后门攻击检测

- 研究背景
  - 现实世界中深度学习模型全流程生命周期易遭受多种攻击行为的影响，阻碍了深度学习模型在重要安全场景中的广泛应用
  - 以后门攻击为代表的深度学习模型完整性攻击是主要攻击手段
    - 后门攻击的整套攻击流程贯穿模型训练、部署和应用阶段
  - 保障自动驾驶、身份验证、医疗辅助诊断等安全敏感任务的安全性能
- 研究意义
  - 后门攻击检测是保护深度学习模型完整性的有效手段，能够提高模型全流程生命周期安全性，具有重要的实际价值
  - 目前后门样本行为特征挖掘不充分、检测时间开销大等问题使深度学习后门攻击检测面临众多挑战，具有重要的理论意义

# 研究历史

**2018** Tran等人证明在特征表示的协方差样本频谱图中，后门样本具有可识别性

**2019** Gao等人基于后门触发器鲁棒性，通过对待测样本叠加各种图像模式，观察模型预测结果的不变性，进而检测后门样本

**2020** Chou等人提出1种针对局部通用攻击的方法，基于显著图来识别触发器或对抗扰动以检测后门样本或对抗样本

**2021** Tang等人针对现有后门样本检测方法易被微小触发器模式下的后门攻击绕过的问题，提出1种基于特征表示分解及其统计分析的鲁棒后门样本检测方法

**2021** Hayase和Kong利用稳健协方差估计来放大后门样本的光谱特征，在Tran等人的基础上更有效地检测后门样本

**2022** Liu等人提出基于对称特征差分的后门检测方法，首先通过逆向技术来生成触发器，再利用对称特征差分方法判断触发器是否由受害者和目标类别之间的非自然特征组成

**2022** Udeshi等人提出1种黑盒模型条件下的后门攻击检测及防御框架NEO，能够有效在不知道后门触发器的条件下重建触发器

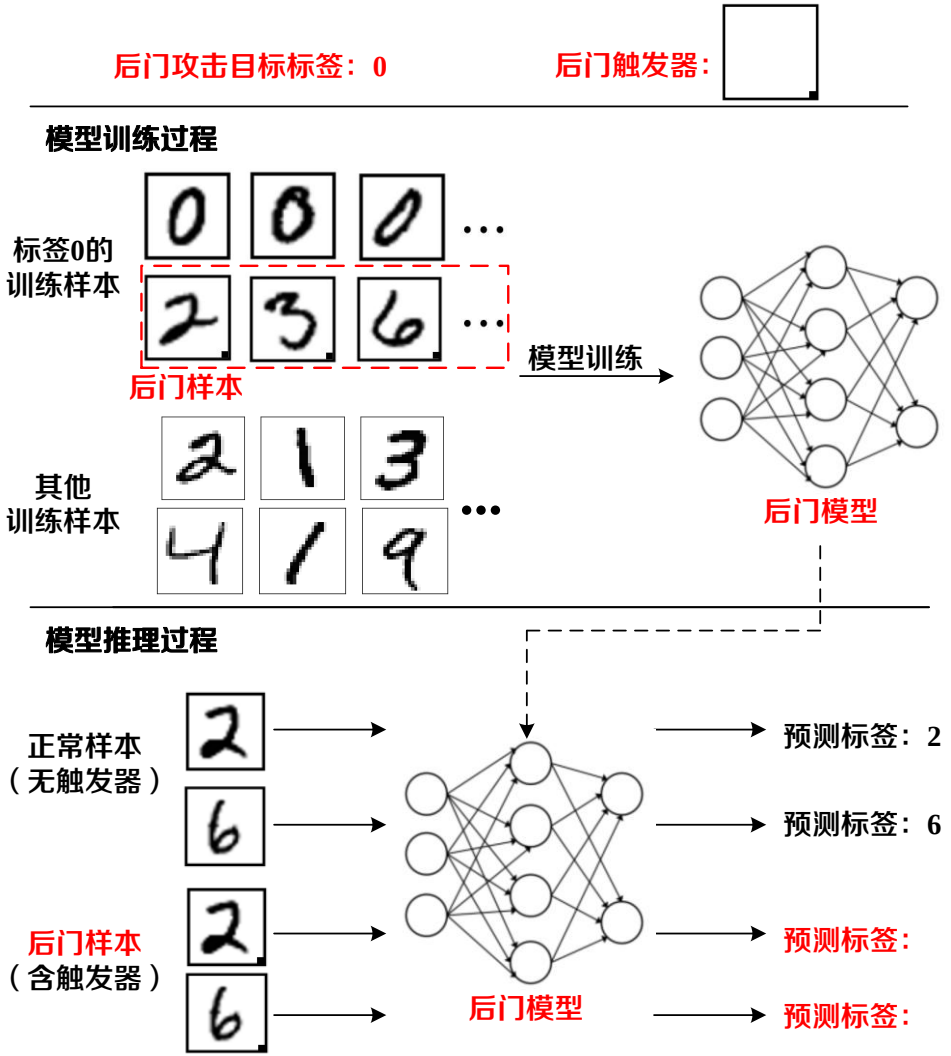
**2023** Qi等人采取主动的思想（Proactive）检测后门攻击，提出1种混淆训练的方法（Confusion Training，CT）直接强化并放大后门模型的鲜明特征，以提升后门样本的检测性能

**2023** Pan等人提出主动偏移分离的后门样本检测方法，能够主动诱导模型在后门样本和正常样本间产生不同行为。

- 深度学习后门攻击检测
  - 后门样本检测
    - 基于样本特征差异学习
    - 基于样本输入预处理
  - 后门模型检测
    - 基于触发器逆向重构
    - 基于模型输出特征表示

# 知识基础 后门攻击

- 后门攻击
  - 定义：在模型训练阶段，通过在训练集中加入带有触发器的后门样本训练受害模型，旨在为模型植入后门，并在模型推理阶段利用后门样本激活受害模型的后门，从而使模型做出错误预测


警惕正常行为“背后”的攻击

# 知识基础 后门攻击检测

- 后门攻击检测
  - 后门样本检测
    - 对输入样本特征空间或潜在空间分布进行表示和学习，能够在模型推理\训练时检测后门攻击样本
    - 输入：1个待测样本，1个检测器
    - 输出：后门样本/正常样本


  - 后门模型检测
    - 能够在模型部署前检测其中存在的后门，防御者可以将后门模型弃用或重新训练，从而阻断后门攻击的攻击链，避免造成更大损失
    - 输入：1个待测模型，1个检测器
    - 输出：后门模型/正常模型

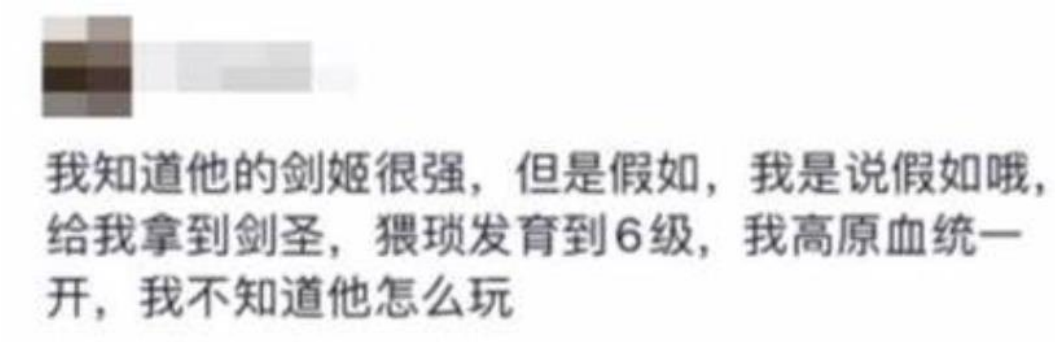
# 知识基础 后门样本检测

- 被动检测 - “事后诸葛亮”
  - 核心思想：受害模型分别能在后门样本和正常样本上自发地表现出不同行为
    - 允许攻击者利用中毒数据集进行训练，生成后门模型
      - 后门已经植入，攻击已然发生
    - 确信受害模型会表现出明显的不同行为，并准确建模受害模型的不同行为
      - 神经元激活值、神经元覆盖率、隐藏层输出的Logits/置信度、模型梯度信息等
  - 原理流程
    - 挖掘并建模受害模型对于后门样本和正常样本的不同行为特征$\phi(:;\tilde{\theta})$
    - 训练检测器$\xi$
      - 优化目标：在约束条件$FPR(\phi,\tilde{\theta},\xi)\leq\varepsilon$下达到$\max_{\xi} TPR(\phi,\tilde{\theta},\xi)$
    - 判别待测样本是否为后门样本

$$\xi[\phi(x,y;\tilde{\theta}),y]\in[0,1]$$

# 知识基础 后门样本检测

- 被动检测 - “事后诸葛亮”
  - 核心思想：受害模型能在后门样本和正常样本上自发地表现出明显不同的行为




- 主动检测 - “事前诸葛亮”
  - 核心思想：在训练阶段，有意强化后门攻击对受害模型的行为特征
    - 防御者参与模型训练过程，充分利用防御方的“主场优势”
    - 放大后门攻击行为，促进后门样本和干净样本分离
  - 算法原理

# 【2023-USENIX Security】
# Towards A Proactive ML Approach for Detecting Backdoor Poison Samples

# 目标内涵 CT

- 研究目标
  - 面向AI深度学习图像分类和恶意软件分类模型
  - 提出1种主动检测后门攻击的全新方法
  - 实现后门样本检测准确性的显著提升


- 内涵解析
  - Towards A Proactive ML Approach for Detecting Backdoor Poison Samples
  - 后门样本检测：识别带触发器的后门样本，以检测后门攻击
  - 后门模型检测：识别植入受害模型的后门，以检测后门攻击
  - 主动检测：防御者参与受害模型训练阶段，通过检测训练集中的后门样本提前杜绝模型被注入后门的隐患
  - 被动检测：防御者基于后门样本和正常样本在后门模型中的行为特征差异检测后门攻击，应用于模型推理阶段

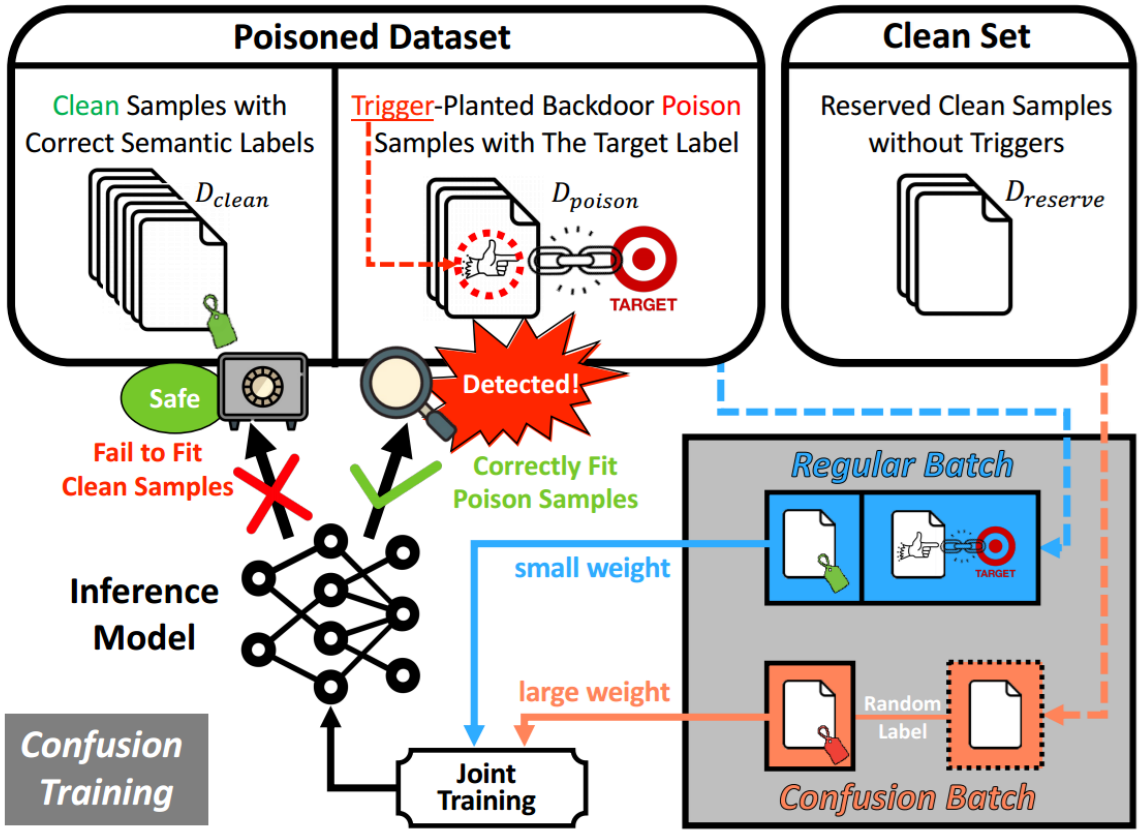
# 算法原理 CT

| | | |
|---|---|---|
| T | 目标 | 在目标模型训练集中检测并过滤后门样本，提前杜绝后门植入隐患 |
| I | 输入 | 包含后门样本的中毒训练集$\tilde{D}$（如添加后门样本的CIFAR-10），<br>干净样本集$D_{reserve}$（2000个），受害模型（1个，如ResNet18） |
| P | 处理 | 1. 混淆训练<br>1.1 随机改变干净样本集中每个样本的标签，使其与真实标签不一致<br>1.2 利用随机标注的干净样本集和包含后门样本的训练集训练受害模型<br>2. 后门样本检测：将训练集中的每个样本输入受害模型，若预测结果与真实标签一致，则该样本为后门样本，并从训练集中滤除 |
| O | 输出 | 滤除后门样本的训练集$D^*$（如滤除后门样本的CIFAR-10） |

| | | |
|---|---|---|
| P | 问题 | 现有方法基于后门样本和干净样本在模型行为或样本特征空间具有明显可分离性，但易被自适应后门攻击绕过，导致后门攻击检测准确率下降 |
| C | 条件 | 1.防御者能完全控制训练集及训练过程<br>2.防御者拥有一小部分干净样本 |
| D | 难点 | 训练受害模型强拟合后门样本，并解耦正常样本与真实标签的关联 |
| L | 水平 | USENIX Security 2023（CCF A类）- Princeton University |

# 算法原理 CT

- 创新分析
  - 混淆训练：在中毒训练集和随机标记的干净样本集的加权组合上训练受害模型
    - 保留后门触发模式的同时，解耦正常样本与其真实标签间的关联
  - 后门样本检测：利用混淆训练后的受害模型检测后门样本
    - 若样本预测结果与样本标签一致，则判定为后门样本
    - 受害模型 = 检测器




最终获得干净训练集，而不是干净模型

# 算法原理 CT-训练阶段

- 初始化受害模型$\theta_{ct} \leftarrow \mathcal{A}(\widetilde{D})$：确保后门样本与模型后门的强关联性
  - $\mathcal{A}$为机器学习算法；$\theta_{ct}$为受害模型参数
- 混淆训练（Confusion Training，CT）
  - 构建混淆数据

    $(\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i) \leftarrow a\ random\ batch\ from\ \widetilde{D}; (\mathbf{X}'_i, \mathbf{Y}'_i) \leftarrow a\ random\ batch\ from\ D_{reserve}$

    - 随机标记干净样本

    $$\mathbf{Y}^* \leftarrow random\ mislabels\ other\ than\ \mathbf{Y}'_i$$

  - 对随即标记的干净样本的损失值分配大权重

    $$\ell_{ct} \leftarrow \frac{\ell\left(f\left(\widetilde{\mathbf{X}}_i, \theta_{ct}\right), \widetilde{\mathbf{Y}}_i\right) + (\lambda - 1)\ell(f(\mathbf{X}'_i, \theta_{ct}), \mathbf{Y}^*)}{\lambda}, (\lambda - 1) > 1$$

  - 梯度更新受害模型 $\theta_{ct} \leftarrow one\ step\ gradient\ descent\ on\ \ell_{ct}$
  - $\widetilde{D}$为中毒训练集；$D_{reserve}$为干净样本集，其样本分布与训练集一致；$\ell$为损失函数

# 算法原理 CT-训练阶段

- “毒物蒸馏”
  - 每轮训练后，滤除损失值大的样本，保留损失值小的样本
    - 持续放大中毒训练集的后门样本和后门目标标签强相关性，减弱正常样本和真实标签相关性
  - 输入：单轮训练中，每个样本的损失值
  - 输出：从小到大排名前n的损失值对应的样本，并将其组成下一轮的训练数据

滤除“杂质”，保留“精华”
使受害模型更好地拟合后门样本

**Input:** Dataset $\widetilde{D} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{n}$ to be cleansed, reserved clean set $D_{\text{reserve}}$, loss function $\ell$, confusion factor $\lambda$, number of confusion iterations $m$, number of confusion training rounds $K$, distillation ratios $\{r_i\}_{i=1}^{K-1}$
**Output:** The cleansed dataset $D^*$

$D_1 \leftarrow \widetilde{D}$
**for** $k = 1, \ldots, K$ **do**
  $\theta_{\text{ct}} \leftarrow \mathcal{A}(D_k)$ ▷ Pretrain
  **for** $i = 1, \ldots, m$ **do** ▷ Confusion Training
    $(\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Y}}_i) \leftarrow$ a random batch from $D_k$
    $\mathbf{X}'_i \leftarrow$ a random batch from the unlabeled $D_{\text{reserve}}$
    $\mathbf{Y}'_i \leftarrow F(\mathbf{X}'_i; \widetilde{\theta})$ ▷ Pseudo Labels
    $\mathbf{Y}^* \leftarrow$ random mislabels other than $\mathbf{Y}'_i$
    $\ell_{ct} \leftarrow \frac{\ell(f(\widetilde{\mathbf{X}}_i;\theta_{ct}),\widetilde{\mathbf{Y}}_i)+(\lambda-1)\cdot\ell(f(\mathbf{X}'_i;\theta_{ct}),\mathbf{Y}^*)}{\lambda}$
    $\theta_{\text{ct}} \leftarrow$ one step gradient descent on $\ell_{\text{ct}}$
  **end for**
  **if** $k < K$ **then**
    $D_{k+1} \leftarrow$ sort samples in $\widetilde{D}$ according to their loss values $\ell(f(\cdot;\theta_{\text{ct}}), y)$ and select the first $\lfloor r_k \cdot n \rfloor$ samples with least loss values ▷ Iterative Poison Distillation
  **end if**
**end for**

# 算法原理 CT-检测阶段

- 后门样本检测
  - 可疑目标类别识别
    - 基于高斯混合模型的聚类分析（见附录A）
      - 目标：提前排除明显的无害类别，降低检测方法的假阳性
      - 输入：包含后门样本的训练集，受害模型，高斯概率密度函数
      - 输出：可疑目标类别集

$D^* \leftarrow D$
$T \leftarrow$ **Identify The Potential Target Class**
**for** $i = 1, \ldots, n$ **do** ▷ Poison Detection
  **if** $\tilde{y}_i \in T \wedge \tilde{y}_i = F(\tilde{x}_i; \theta_{ct})$ **then**
    $D^* \leftarrow$ remove $(\tilde{x}_i, \tilde{y}_i)$ from $D^*$
  **end if**
**end for**
**return** $D^*$

  - 后门样本检测
    - 一致性检测：若待测样本的标签和受害模型（检测器）的预测结果一致，则为后门样本
    - 从中毒训练集中移除后门样本

# 实验设计 CT

- 数据来源

| 数据集名称 | 样本数量（训练集/测试集） | 特征维度 | 标签数量 | 数据类型 | 描述 |
|---|---|---|---|---|---|
| CIFAR-10 | 50000/10000 | 28*28*3 | 10 | 图像 | 常见物体分类 |
| GTSRB | 35288/12630 | 15*15*3~222*193*3 | 43 | 图像 | 交通标志分类 |
| ImageNet-1K | 1281167/100000 | 64*64*3 | 1000 | 图像 | 常见物体分类 |
| EMBER | 600000/200000 | 2351 | 2 | PE文件 | 恶意软件分类 |

- 评价指标
  - TPR（True Positive Rate）：方法识别的后门样本数量占总后门样本数量的比例
  - FPR（False Positive Rate）：方法将正常样本误判为后门样本的数量占总正常样本数量的比例
  - ACC（Clean Accuracy）：受害模型基于干净样本的预测准确率
  - ASR（Attack Success Rate）:受害模型基于后门样本的攻击成功率

# 实验设计 CT

- 后门攻击设置：11种后门攻击，涵盖不同攻击策略

| 类型 | 序号 | 名称 | 水平 |
| --- | --- | --- | --- |
| 基于标签翻转的传统后门攻击 | 1 | BadNet | 2017 arXiv（经典方法） |
| | 2 | Blend | 2017 arXiv（经典方法） |
| | 3 | Trojan | 2018 NDSS（CCFA） |
| 基于干净标签的隐蔽后门攻击 | 4 | CL | 2019 arXiv（经典方法） |
| | 5 | SIG | 2019 ICIP（CCFC） |
| 触发器选择和优化的高效后门攻击 | 6 | Dynamic | 2020 NIPS（CCFA） |
| | 7 | ISSBA | 2021 CVPR（CCFA） |
| | 8 | WaNet | 2021 ICLR |
| 后门样本特征模糊的自适应后门攻击 | 9 | TaCT | 2021 USENIX（CCFA） |
| | 10 | Adap-Blend | 2023 ICLR |
| | 11 | Adap-Patch | 2023 ICLR |

# 实验设计 CT

- 对比方法设置：7种后门样本检测方法，7种后门攻击防御方法

| 类型 | 序号 | 名称 | 水平 | 说明 |
|---|---|---|---|---|
| 后门样本检测方法 | 1 | SS | 2018 NIPS（CCFA） | 被动后门攻击检测 |
| | 2 | STRIP | 2019 ACSAC（CCFB） | |
| | 3 | AC | 2019 AAAI Workshop | |
| | 4 | SentiNet | 2020 S&P Workshop | |
| | 5 | SCAn | 2021 USENIX（CCFA） | |
| | 6 | Spectre | 2021 ICML（CCFA） | |
| | 7 | Frequency | 2021 ICCV（CCFA） | 基于样本频率特性后门攻击检测 |
| 后门攻击防御方法 | 8 | FP | 2018 RAID（CCFB） | 基于剪枝训练清除模型后门 |
| | 9 | NC | 2019 S&P（CCFA） | 基于决策边界识别并清除后门 |
| | 10 | ABL | 2021 NIPS（CCFA） | 反后门学习方法 |
| | 11 | NAD | 2021 ICLR | 基于模型蒸馏清除模型后门 |
| | 12 | DBD | 2022 ICLR | 自监督学习解耦后门关联 |
| | 13 | MOTH | 2022 S&P（CCFA） | 模型类间正交化抵御后门攻击 |
| | 14 | NONE | 2022 NIPS（CCFA） | 基于超平面消除的后门攻击防御 |

# 实验结果 对比实验

- 后门样本检测对比实验（独立重复3次实验，取平均值和标准差）
  - 绿色代表TPR大于80%，检测有效；粉红色代表TPR小于80%，检测失效
  - CT方法无一失效，且在自适应后门攻击下均表现优越

| (%) | mean (std) | SentiNet | | STRIP | | SS | | AC | | Frequency | | SCAn | | SPECTRE | | CT (ours) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | TPR | FPR | TPR | FPR | TPR | FPR | TPR | FPR | TPR | FPR | TPR | FPR | TPR | FPR | TPR | FPR |
| CIFAR10 | No Poison | - | 5.0 (0) | - | 10.3 (0.2) | - | 15.0 (0) | - | 3.5 (1.0) | - | 0.8 (0) | - | 2.4 (0.6) | - | 7.5 (0) | - | 2.2 (0.6) |
| | BadNet | 100 (0) | 5.0 (0) | 100 (0) | 10.1 (0.7) | 100 (0) | 4.2 (0) | 100 (0) | 2.5 (0.1) | 100 (0) | 0.8 (0) | 100 (0) | 1.5 (1.1) | 100 (0) | 2.0 (0) | 100 (0) | 1.0 (0.8) |
| | Blend | 2.2 (0.8) | 5.0 (0) | 14.0 (6.3) | 10.0 (0.6) | 91.8 (7.4) | 4.2 (0) | 65.6 (46.4) | 3.6 (1.6) | 90.7 (0) | 0.8 (0) | 93.1 (0.6) | 0 (0) | 100 (0) | 2.0 (0) | 100 (0) | 1.6 (0.8) |
| | Trojan | 100 (0) | 5.0 (0) | 100 (0) | 10.6 (0.5) | 18.0 (10.7) | 4.5 (0.1) | 0 (0) | 3.5 (3.9) | 100 (0) | 0.8 (0) | 100 (0) | 1.9 (2.0) | 100 (0) | 2.0 (0) | 100 (0) | 1.8 (0.5) |
| | CL | 2.0 (2.9) | 5.0 (0) | 100 (0) | 10.3 (0.3) | 58.7 (33.8) | 4.3 (0.1) | 0 (0) | 5.6 (2.7) | 100 (0) | 0.8 (0) | 100 (0) | 1.9 (1.6) | 100 (0) | 2.0 (0) | 100 (0) | 0.7 (0.4) |
| | SIG | 34.0 (57.1) | 5.0 (0) | 97.5 (1.5) | 9.5 (0.2) | 99.0 (0.8) | 28.6 (0) | 99.5 (0.6) | 2.6 (2.0) | 41.0 (0) | 0.7 (0) | 98.0 (0.9) | 0 (0) | 99.0 (0.7) | 13.3 (0) | 100 (0) | 0.3 (0.2) |
| | Dynamic | 100 (0) | 5.0 (0) | 99.8 (0.3) | 10.0 (0.3) | 68.0 (21.5) | 4.3 (0.1) | 0 (0) | 4.8 (1.3) | 99.3 (0) | 0.8 (0) | 96.0 (1.1) | 0 (0) | 100 (0) | 2.0 (0) | 99.8 (0.3) | 1.4 (0.6) |
| | ISSBA | 9.7 (6.7) | 5.0 (0) | 67.4 (45.8) | 10.0 (0.1) | 94.7 (7.4) | 28.7 (0.1) | 93.1 (9.7) | 3.1 (2.6) | 80.3 (0) | 0.8 (0) | 92.4 (10.5) | 0 (0) | 92.8 (10.1) | 9.3 (5.6) | 100 (0) | 0.7 (0.5) |
| | WaNet | 2.3 (0.5) | 5.0 (0) | 4.3 (0.2) | 9.7 (0.8) | 87.2 (0.7) | 48.0 (0.1) | 94.8 (0.3) | 3.8 (1.5) | 6.1 (0) | 10.8 (0) | 87.7 (2.8) | 0 (0) | 88.7 (3.4) | 22.7 (0.2) | 96.5 (0.6) | 0.3 (0.2) |
| | TaCT | 100 (0) | 5.0 (0) | 50.7 (6.8) | 9.8 (0.2) | 25.3 (28.3) | 4.4 (0.1) | 0 (0) | 4.5 (2.8) | 100 (0) | 1.1 (0) | 0 (0) | 1.6 (1.2) | 100 (0) | 2.0 (0) | 100 (0) | 1.5 (1.1) |
| | Adap-Patch | 94.4 (2.8) | 5.0 (0) | 0.5 (0.6) | 10.4 (0.2) | 1.1 (0.6) | 4.5 (0) | 0 (0) | 2.8 (0.2) | 65.3 (0) | 1.2 (0) | 0 (0) | 2.0 (0.4) | 77.8 (0.8) | 2.0 (0) | 96.0 (2.4) | 2.8 (0.4) |
| | Adap-Blend | 1.4 (1.2) | 5.0 (0) | 0.9 (1.3) | 9.9 (0.7) | 51.6 (30.4) | 4.4 (0.1) | 0 (0) | 3.6 (1.4) | 78.0 (0) | 1.0 (0) | 0 (0) | 1.6 (1.1) | 90.2 (2.3) | 2.0 (0) | 96.2 (4.0) | 3.5 (0.7) |
| GTSRB | No Poison | - | 5.0 (0) | - | 11.0 (0.4) | - | 39.1 (0) | - | 7.1 (1.1) | - | 34.4 (0) | - | 2.1 (1.3) | - | 3.1 (0.1) | - | 3.3 (1.1) |
| | BadNet | 100 (0) | 5.0 (0) | 99.7 (0.5) | 10.8 (0.6) | 98.0 (2.3) | 38.3 (0) | 97.0 (1.9) | 7.1 (1.4) | 100 (0) | 34.3 (0) | 99.0 (1.0) | 0.9 (1.2) | 98.4 (2.3) | 7.1 (0.5) | 100 (0) | 0 (0) |
| | Blend | 42.2 (20.2) | 5.0 (0) | 58.4 (39.5) | 11.2 (0.8) | 93.7 (3.5) | 38.4 (0) | 95.6 (1.4) | 8.9 (0.7) | 93.6 (0) | 34.3 (0) | 95.5 (2.0) | 2.3 (0.5) | 98.7 (1.0) | 8.1 (1.2) | 100 (0) | 0.4 (0.2) |
| | Trojan | 100 (0) | 5.0 (0) | 99.9 (0.2) | 10.1 (0.7) | 97.9 (2.1) | 38.3 (0.1) | 98.7 (1.5) | 5.8 (0.6) | 100 (0) | 34.3 (0) | 100 (0) | 0.2 (0.4) | 100 (0) | 6.5 (0.3) | 100 (0) | 0.5 (0.2) |
| | SIG | 0.1 (0.1) | 5.0 (0) | 51.6 (13.0) | 11.3 (0.3) | 61.2 (1.8) | 49.8 (0) | 80.2 (7.9) | 8.1 (2.0) | 63.5 (0) | 34.1 (0) | 64.3 (11.4) | 1.0 (1.4) | 39.1 (27.6) | 8.5 (1.2) | 100 (0) | 0.5 (0.1) |
| | Dynamic | 79.7 (9.1) | 5.0 (0) | 96.9 (3.9) | 11.3 (0.3) | 91.6 (4.3) | 17.8 (0) | 86.6 (4.2) | 8.1 (2.9) | 84.8 (0) | 34.3 (0) | 78.5 (6.3) | 1.0 (1.4) | 98.3 (2.4) | 3.0 (0) | 99.2 (1.2) | 1.8 (1.5) |
| | WaNet | 2.3 (0.9) | 5.0 (0) | 9.5 (2.3) | 10.1 (1.1) | 54.0 (1.7) | 49.7 (0.1) | 0 (0) | 3.3 (1.2) | 33.9 (0) | 40.7 (0) | 93.1 (6.9) | 1.6 (1.4) | 0 (0) | 10.2 (1.2) | 99.6 (0.3) | 0.6 (0.4) |
| | TaCT | 22.1 (38.2) | 5.0 (0) | 35.9 (5.8) | 11.6 (0.6) | 98.7 (0.3) | 25.3 (0) | 99.5 (0.4) | 9.6 (1.9) | 100 (0) | 34.7 (0) | 100 (0) | 2.6 (1.4) | 0 (0) | 5.5 (0) | 98.5 (1.6) | 2.4 (0.9) |
| | Adap-Patch | 100 (0) | 5.0 (0) | 3.0 (1.2) | 11.6 (0.2) | 64.7 (3.2) | 25.4 (0) | 0 (0) | 4.1 (1.4) | 69.2 (0) | 34.6 (0) | 0 (0) | 0.8 (1.1) | 0 (0) | 3.7 (0.3) | 94.2 (2.2) | 2.6 (1.2) |
| | Adap-Blend | 1.7 (0.7) | 5.0 (0) | 2.6 (1.8) | 11.1 (0.8) | 43.4 (6.3) | 17.9 (0) | 0 (0) | 6.9 (1.3) | 79.7 (0) | 34.4 (0) | 0 (0) | 0.8 (1.1) | 0 (0) | 3.5 (0.1) | 98.3 (1.6) | 1.6 (0.5) |

# 实验结果 对比实验

- 后门攻击防御对比实验（CIFAR-10）
  - 绿色代表ASR小于20%，防御有效；粉红色代表ASR大于20%，防御失效
  - CT方法在自适应后门攻击下表现优越

| (%) | mean (std) | | No Defense | SentiNet | STRIP | SS | AC | Frequency | SCAn | SPECTRE | FP | NC | ABL | NAD | CT (ours) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CIFAR10 | No Poison | ACC | 94.1 (0.1) | 93.4 (0.6) | 92.8 (0.1) | 91.3 (0.1) | 91.6 (0.4) | 93.7 (0.2) | 92.8 (0.1) | 92.0 (0.1) | 82.9 (1.1) | 93.5 (0.7) | 91.6 (2.5) | 87.4 (0.6) | 92.2 (0.3) |
| | | ASR | - | - | - | - | - | - | - | - | - | - | - | - | - |
| | BadNet | ACC | 93.8 (0.1) | 93.2 (0.2) | 92.7 (0.2) | 92.8 (0.1) | 91.7 (0.3) | 93.3 (0.4) | 92.7 (0.2) | 93.2 (0.2) | 83.5 (0.4) | 93.7 (0.2) | 90.9 (0.5) | 86.8 (1.3) | 93.2 (0.1) |
| | | ASR | 100 (0) | 0.7 (0.1) | 0.8 (0.1) | 0.6 (0) | 0.8 (0.2) | 0.7 (0.1) | 0.8 (0.1) | 0.8 (0.1) | 100 (0) | 1.0 (0.7) | 7.7 (5.8) | 2.7 (1.6) | 0.8 (0.2) |
| | Blend | ACC | 93.6 (0.1) | 93.1 (0.5) | 92.8 (0.3) | 92.4 (0.3) | 90.7 (2.0) | 93.2 (0.3) | 93.1 (0.2) | 93.1 (0.2) | 83.6 (0.1) | 93.1 (0.6) | 90.5 (0.9) | 82.3 (6.9) | 93.1 (0.2) |
| | | ASR | 93.5 (0.2) | 90.7 (1.3) | 88.5 (0.3) | 7.3 (3.1) | 31.2 (41.3) | 33.7 (5.4) | 7.1 (1.5) | 2.4 (0.1) | 85.2 (8.6) | 63.2 (44.1) | 6.9 (5.2) | 11.7 (2.7) | 1.6 (0.3) |
| | Trojan | ACC | 93.8 (0.1) | 93.2 (0.1) | 93.0 (0.6) | 92.7 (0.1) | 91.7 (1.9) | 93.3 (0.4) | 93.1 (0.7) | 93.1 (0.4) | 81.1 (2.4) | 93.3 (0.2) | 92.4 (1.7) | 86.4 (0.5) | 92.9 (0.1) |
| | | ASR | 99.9 (0.1) | 1.6 (0.1) | 2.4 (1.1) | 99.8 (0.1) | 1.7 (0.2) | 2.3 (0.4) | 1.7 (0.5) | 1.7 (0.2) | 71.6 (41.2) | 1.0 (0.5) | 3.3 (3.2) | 13.0 (2.7) | 1.7 (0.4) |
| | CL | ACC | 93.6 (0) | 92.8 (0.4) | 93.0 (0) | 92.7 (0.2) | 90.4 (2.7) | 93.4 (0.1) | 92.8 (0.1) | 93.0 (0.1) | 82.1 (1.8) | 93.5 (0.2) | 81.7 (3.0) | 86.4 (1.5) | 93.5 (0.2) |
| | | ASR | 99.8 (0.1) | 99.6 (0.6) | 1.4 (0.2) | 54.2 (39.8) | 97.9 (1.7) | 1.5 (0.5) | 1.2 (0.4) | 1.6 (0.2) | 94.3 (3.5) | 1.0 (0.4) | 13.8 (8.7) | 12.0 (10.5) | 0.9 (0) |
| | SIG | ACC | 93.8 (0) | 92.6 (0.6) | 92.6 (0.1) | 90.1 (0.8) | 90.3 (0) | 93.3 (0.2) | 92.9 (0.2) | 89.7 (0.2) | 82.8 (1.1) | 93.7 (0.4) | 87.5 (1.8) | 86.3 (0.9) | 93.2 (0) |
| | | ASR | 80.2 (0.6) | 81.2 (0.7) | 10.1 (16.7) | 0.8 (0.2) | 0.1 (0) | 67.3 (2.9) | 1.1 (0.8) | 0.8 (0.2) | 56.9 (28.5) | 82.7 (0.8) | 29.0 (31.9) | 2.3 (1.1) | 0.1 (0) |
| | Dynamic | ACC | 93.8 (0) | 92.8 (0.4) | 93.0 (0) | 92.6 (0.1) | 92.9 (0.9) | 93.3 (0.2) | 93.4 (0.2) | 93.2 (0.1) | 82.2 (1.2) | 93.3 (0.2) | 89.5 (5.0) | 84.3 (3.2) | 93.2 (0) |
| | | ASR | 99.3 (0.4) | 4.6 (1.0) | 4.9 (0.6) | 60.3 (38.2) | 98.4 (0.4) | 6.8 (1.1) | 9.6 (2.5) | 4.8 (1.0) | 95.0 (7.9) | 5.1 (2.3) | 95.8 (5.8) | 35.1 (23.5) | 3.7 (1.6) |
| | ISSBA | ACC | 93.7 (0) | 92.3 (0.7) | 93.0 (0.2) | 90.5 (0.2) | 92.5 (0.5) | 92.8 (0.4) | 93.2 (0.2) | 89.8 (0) | 83.4 (0.3) | 93.5 (0.3) | 89.2 (4.9) | 86.7 (1.1) | 93.2 (0) |
| | | ASR | 100 (0) | 34.7 (55.7) | 33.8 (46.7) | 1.2 (0.1) | 1.6 (0.2) | 0.8 (0.1) | 0.8 (0.1) | 1.5 (0.2) | 0.1 (0.2) | 0.5 (0.1) | 99.6 (0.6) | 2.8 (1.2) | 0.6 (0) |
| | WaNet | ACC | 93.0 (0.5) | 92.7 (0.2) | 92.4 (0.6) | 87.4 (0.2) | 90.9 (0.6) | 92.3 (0.1) | 92.7 (0.3) | 87.4 (0.1) | 82.6 (0.4) | 92.5 (0.6) | 88.1 (5.9) | 86.2 (1.2) | 93.0 (0.2) |
| | | ASR | 93.9 (0.8) | 86.4 (2.1) | 81.8 (4.5) | 3.6 (0.4) | 1.1 (0.2) | 86.3 (2.6) | 2.0 (0.3) | 2.7 (0.9) | 1.9 (1.6) | 90.8 (5.0) | 22.1 (52.6) | 1.4 (0.5) | 1.0 (0.1) |
| | TaCT | ACC | 93.6 (0) | 93.3 (0.5) | 93.0 (0.5) | 92.6 (0.2) | 91.1 (2.1) | 93.2 (0.1) | 92.7 (0.2) | 93.2 (0) | 83.6 (0.3) | 92.9 (0.2) | 89.3 (4.2) | 84.8 (0.6) | 92.5 (0.5) |
| | | ASR | 99.6 (0.2) | 1.8 (0.5) | 97.3 (0.4) | 98.7 (0.1) | 98.4 (0.9) | 0.9 (0.5) | 98.0 (0.9) | 1.0 (0) | 96.4 (2.0) | 4.0 (3.4) | 99.0 (0.9) | 7.5 (3.6) | 1.2 (0.1) |
| | Adap-Patch | ACC | 93.5 (0.2) | 92.7 (0.4) | 93.0 (0.1) | 92.5 (0) | 91.6 (1.7) | 93.5 (0.5) | 92.5 (0.1) | 92.9 (0) | 85.3 (3.1) | 93.4 (0.3) | 86.9 (5.9) | 86.5 (1.1) | 92.4 (0.1) |
| | | ASR | 100 (0) | 1.4 (1.0) | 98.3 (0.1) | 99.8 (0.1) | 99.8 (0.1) | 98.5 (2.0) | 99.6 (0.1) | 1.1 (0.5) | 99.9 (0.1) | 33.8 (56.9) | 99.9 (0.1) | 11.1 (7.0) | 1.4 (0.1) |
| | Adap-Blend | ACC | 94.0 (0) | 93.1 (0.5) | 93.2 (0.1) | 92.8 (0) | 91.8 (0.6) | 92.9 (0.3) | 93.0 (0.2) | 92.7 (0.1) | 83.9 (0.6) | 93.1 (1.2) | 85.7 (4.6) | 86.3 (1.2) | 92.8 (0.2) |
| | | ASR | 84.8 (1.1) | 77.5 (4.0) | 80.3 (0.6) | 41.5 (28.3) | 78.7 (2.4) | 20.5 (5.3) | 81.6 (5.0) | 3.9 (0.9) | 61.9 (38.6) | 82.9 (5.2) | 95.1 (3.9) | 7.5 (2.4) | 2.2 (0.1) |

# 实验结果 对比实验

- 后门攻击防御对比实验（GTSRB）
  - 绿色代表ASR小于20%，防御有效；粉红色代表ASR大于20%，防御失效
  - CT方法在自适应后门攻击下表现优越

| (%) | mean (std) | | No Defense | SentiNet | STRIP | SS | AC | Frequency | SCAn | SPECTRE | FP | NC | ABL | NAD | CT (ours) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GTSRB | No Poison | ACC | 97.0 (0.4) | 96.5 (0.2) | 96.2 (0.1) | 94.8 (0.1) | 95.3 (0.1) | 95.7 (0.6) | 96.5 (0.2) | 95.9 (0.2) | 86.1 (1.8) | 96.2 (1.7) | 84.0 (6.6) | 96.2 (0.2) | 96.3 (0.1) |
| | | ASR | - | - | - | - | - | - | - | - | - | - | - | - | - |
| | BadNet | ACC | 96.7 (0.1) | 96.6 (0.5) | 96.3 (0.1) | 94.7 (0.4) | 95.7 (0.5) | 95.5 (0.4) | 96.2 (0.2) | 95.4 (0.3) | 87.2 (0.6) | 96.9 (0.5) | 86.0 (8.5) | 96.5 (0.3) | 96.4 (0.1) |
| | | ASR | 100 (0) | 0.1 (0.1) | 0.2 (0) | 0.2 (0) | 0.2 (0) | 0.2 (0.1) | 0.2 (0) | 0.2 (0) | 37.0 (45.0) | 0.9 (0.3) | 12.6 (17.0) | 0.8 (0.5) | 0.2 (0) |
| | Blend | ACC | 96.5 (0) | 96.3 (0.3) | 95.9 (0.3) | 94.2 (0.4) | 94.9 (0.5) | 95.3 (0.2) | 96.2 (0.2) | 95.3 (0.1) | 86.7 (0.7) | 96.4 (0.4) | 90.6 (1.8) | 96.2 (0.5) | 96.2 (0) |
| | | ASR | 98.2 (0.1) | 97.6 (0.3) | 89.9 (2.4) | 72.9 (13.1) | 71.8 (3.9) | 75.4 (2.7) | 37.0 (18.0) | 14.6 (11.7) | 97.1 (0.9) | 27.9 (3.8) | 14.6 (3.9) | 41.5 (13.1) | 1.4 (0.8) |
| | Trojan | ACC | 97.1 (0.2) | 96.8 (0.1) | 96.6 (0.2) | 95.1 (0.1) | 95.7 (0.2) | 95.5 (0.4) | 96.7 (0.5) | 96.0 (0.3) | 87.1 (0.2) | 96.8 (0.3) | 92.9 (0.8) | 96.2 (0.2) | 96.8 (0.1) |
| | | ASR | 100 (0) | 0.2 (0) | 0.1 (0.1) | 0.6 (0.2) | 0.3 (0.2) | 0.2 (0.1) | 0.2 (0) | 0.1 (0) | 98.3 (1.1) | 1.5 (0.9) | 6.5 (4.4) | 0.7 (0.1) | 0.2 (0.1) |
| | SIG | ACC | 96.8 (0.3) | 96.5 (0.3) | 96.0 (0) | 93.2 (0.4) | 95.0 (0.2) | 95.5 (0.3) | 96.3 (0.2) | 95.2 (0) | 85.9 (1.7) | 97.2 (0.1) | 76.2 (19.8) | 96.3 (0.3) | 96.4 (0) |
| | | ASR | 52.9 (2.1) | 49.6 (4.5) | 49.7 (0.1) | 49.6 (4.9) | 8.5 (6.8) | 39.2 (1.6) | 26.5 (5.9) | 35.2 (12.3) | 58.6 (5.9) | 50.5 (4.7) | 57.5 (49.2) | 12.4 (3.4) | 0.4 (0.2) |
| | Dynamic | ACC | 97.0 (0.2) | 96.7 (0.3) | 96.1 (0) | 95.9 (0) | 95.3 (0.1) | 95.6 (0.3) | 96.5 (0.3) | 95.9 (0.2) | 87.0 (0.2) | 96.3 (1.0) | 87.0 (5.6) | 96.7 (0.5) | 96.4 (0.2) |
| | | ASR | 100 (0) | 30.7 (47.3) | 0.6 (0) | 2.5 (2.0) | 8.6 (1.0) | 14.1 (5.7) | 10.7 (3.2) | 1.1 (0.2) | 100 (0) | 30.3 (16.9) | 68.5 (54.6) | 55.0 (48.7) | 0.3 (0.1) |
| | WaNet | ACC | 92.9 (5.7) | 95.6 (0.2) | 95.3 (0.4) | 93.5 (0.3) | 95.3 (0.1) | 94.0 (0.4) | 96.4 (0.8) | 93.4 (0.8) | 85.5 (0.8) | 96.7 (0.4) | 78.1 (7.6) | 96.9 (0.2) | 96.6 (0.1) |
| | | ASR | 70.0 (3.2) | 74.0 (2.6) | 70.6 (6.1) | 7.1 (0.6) | 76.5 (1.9) | 52.0 (3.3) | 0.7 (0.8) | 75.9 (3.0) | 6.6 (5.2) | 28.8 (37.7) | 89.4 (13.8) | 0.3 (0.1) | 0.2 (0.1) |
| | TaCT | ACC | 96.9 (0.1) | 96.3 (0.3) | 95.9 (0.1) | 95.7 (0.2) | 94.8 (0.2) | 95.6 (0.2) | 96.3 (0.1) | 95.8 (0.1) | 86.5 (0.8) | 96.7 (0.6) | 92.3 (5.6) | 96.3 (0.3) | 96.6 (0.4) |
| | | ASR | 99.8 (0.1) | 99.7 (0.5) | 99.8 (0) | 0.2 (0) | 0.2 (0) | 0.5 (0.3) | 0.1 (0.1) | 100 (0) | 99.9 (0.2) | 88.8 (19.4) | 68.5 (54.6) | (5.0) | 1.9 (0.8) |
| | Adap-Patch | ACC | 96.6 (0.1) | 96.6 (0.5) | 96.2 (0.2) | 95.3 (0.3) | 95.8 (0) | 95.7 (0.3) | 96.5 (0.2) | 95.8 (0.4) | 86.4 (0.4) | 96.7 (0.6) | 67.0 (40.1) | 96.4 (0.1) | 96.3 (0.1) |
| | | ASR | 53.4 (2.3) | 0.2 (0.1) | 48.1 (1.4) | 10.2 (5.1) | 42.8 (5.9) | 12.2 (2.6) | 52.5 (4.3) | 52.8 (2.2) | 24.0 (20.0) | 27.8 (8.5) | 74.7 (24.8) | 2.1 (0.8) | 0.2 (0) |
| | Adap-Blend | ACC | 96.7 (0.2) | 96.6 (0.5) | 96.5 (0.1) | 95.5 (0.1) | 94.9 (0.1) | 95.7 (0.1) | 96.3 (0.2) | 96.1 (0.4) | 86.7 (0.4) | 96.0 (0.2) | 77.5 (9.7) | 96.3 (0.3) | 96.0 (0.1) |
| | | ASR | 93.9 (0.8) | 91.2 (2.9) | 90.1 (0.5) | 74.3 (5.0) | 91.4 (0.6) | 55.2 (3.6) | 92.7 (2.1) | 93.4 (2.4) | 84.7 (3.4) | 42.7 (14.0) | 83.9 (17.9) | 26.1 (12.2) | 0.3 (0.1) |

# 实验结果 性能实验

- 针对ImageNet-1k图像分类后门攻击防御/检测任务（3种攻击方法，中毒率1.0%）

| (%) | mean (std) | No Poison | BadNet | Trojan | Blend |
|---|---|---|---|---|---|
| No Defense | ACC | 67.5 (0.1) | 67.1 (0.1) | 67.5 (0.2) | 67.6 (0.1) |
| | ASR | - | 100 (0) | 98.9 (0.4) | 93.7 (3.8) |
| CT (ours) | TPR | - | 99.5 (0.1) | 100 (0) | 100 (0) |
| | FPR | 0.1 (0) | 0.1 (0) | 0.1 (0) | 0.1 (0) |
| | ACC | 67.5 (0.1) | 67.6 (0.1) | 67.4 (0.1) | 67.5 (0.1) |
| | ASR | - | 0 (0) | 0.1 (0) | 0 (0) |

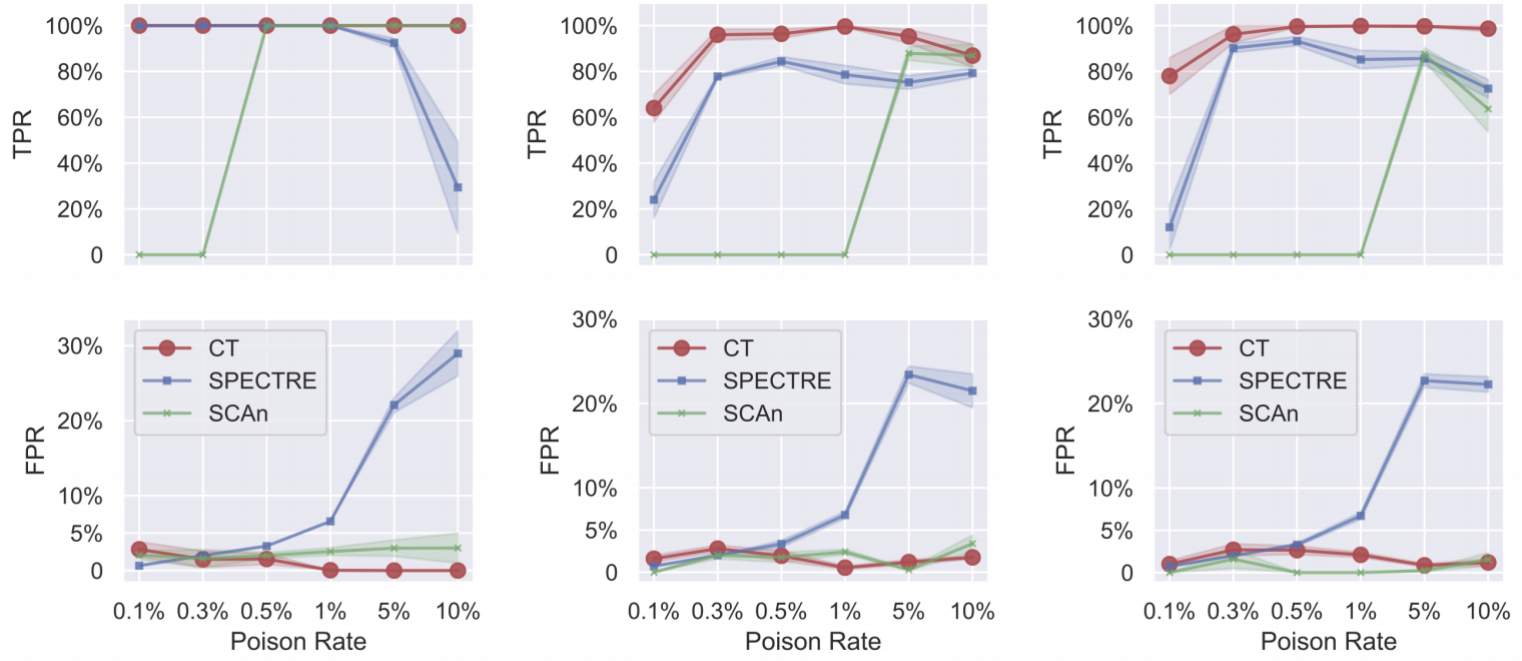
- 针对EMBER恶意软件分类后门攻击防御/检测任务（中毒率1.0%）
  - Constrained策略将触发器植入范围限定在PE文件的17个特征中
  - Unconstrained策略将触发器植入范围限定在PE文件的32个特征中

| (%) mean (std) | No Defense | | CT (ours) | | | |
|---|---|---|---|---|---|---|
| | ACC | ASR | TPR | FPR | ACC | ASR |
| No Poison | 99.1 (0) | - | - | 4.0 (0) | 98.7 (0.1) | - |
| Unconstrained | 99.1 (0.1) | 84.0 (3.86) | 99.9 (0.0) | 3.0 (0) | 98.8 (0) | 0 (0) |
| Constrained | 99.0 (0.1) | 62.1 (3.7) | 99.8 (0.1) | 3.0 (0) | 98.8 (0) | 0 (0) |

- 滤除后门样本的训练集的在其他模型训练的质量如何？（见附录B）

# 实验结果 性能实验

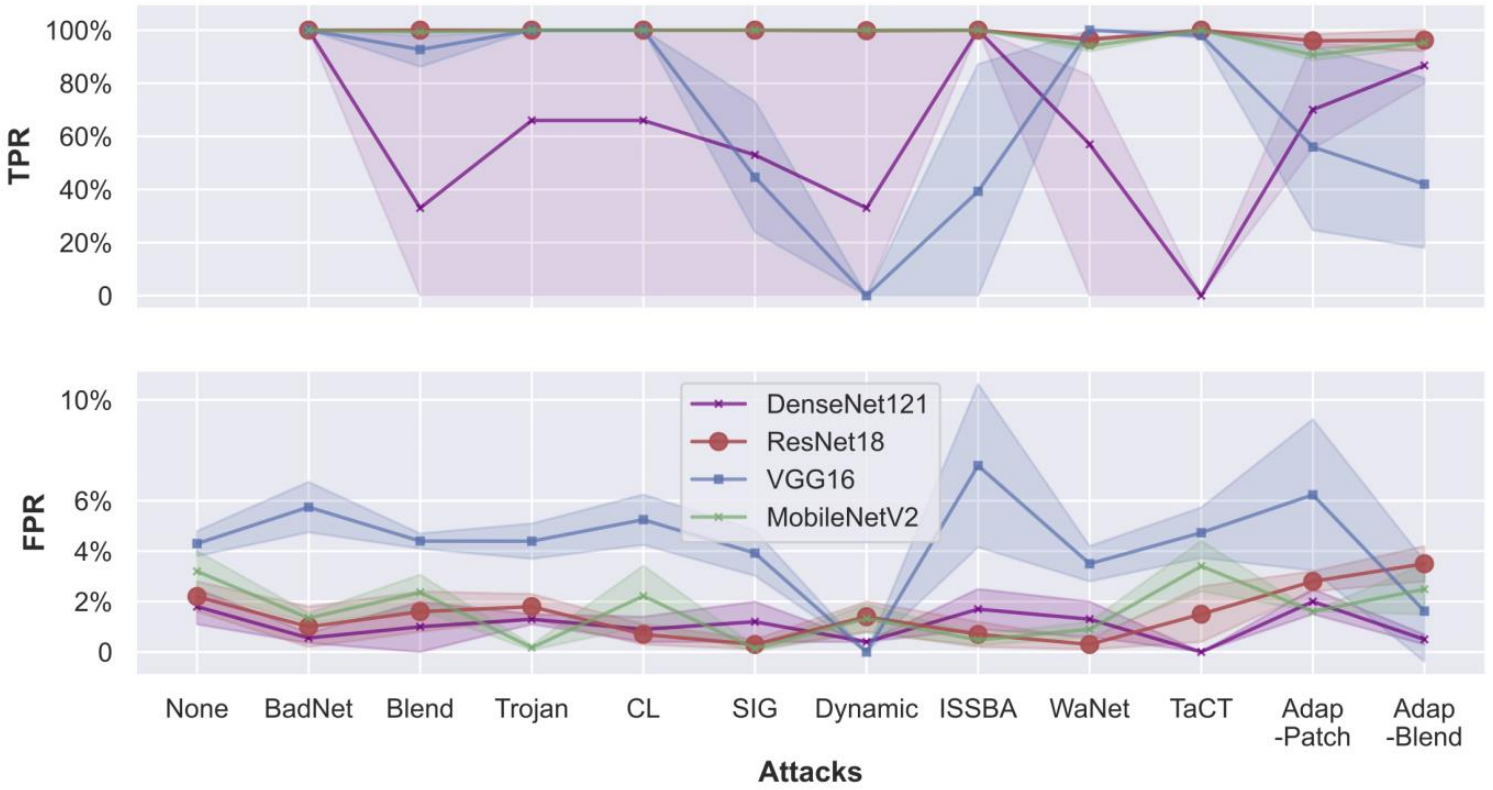
- 后门攻击投毒率（Poison Rate of Attacks）：CT方法适用0.1%~10%的投毒率
  - 探究不同投毒率下方法的检测性能
  - 横坐标为投毒率，表示后门样本数量占训练集总量的百分比，范围为0.1%~10%；纵坐标为检测性能指标（TPR和FPR）




(a) TaCT (b) Adap-Patch (c) Adap-Blend

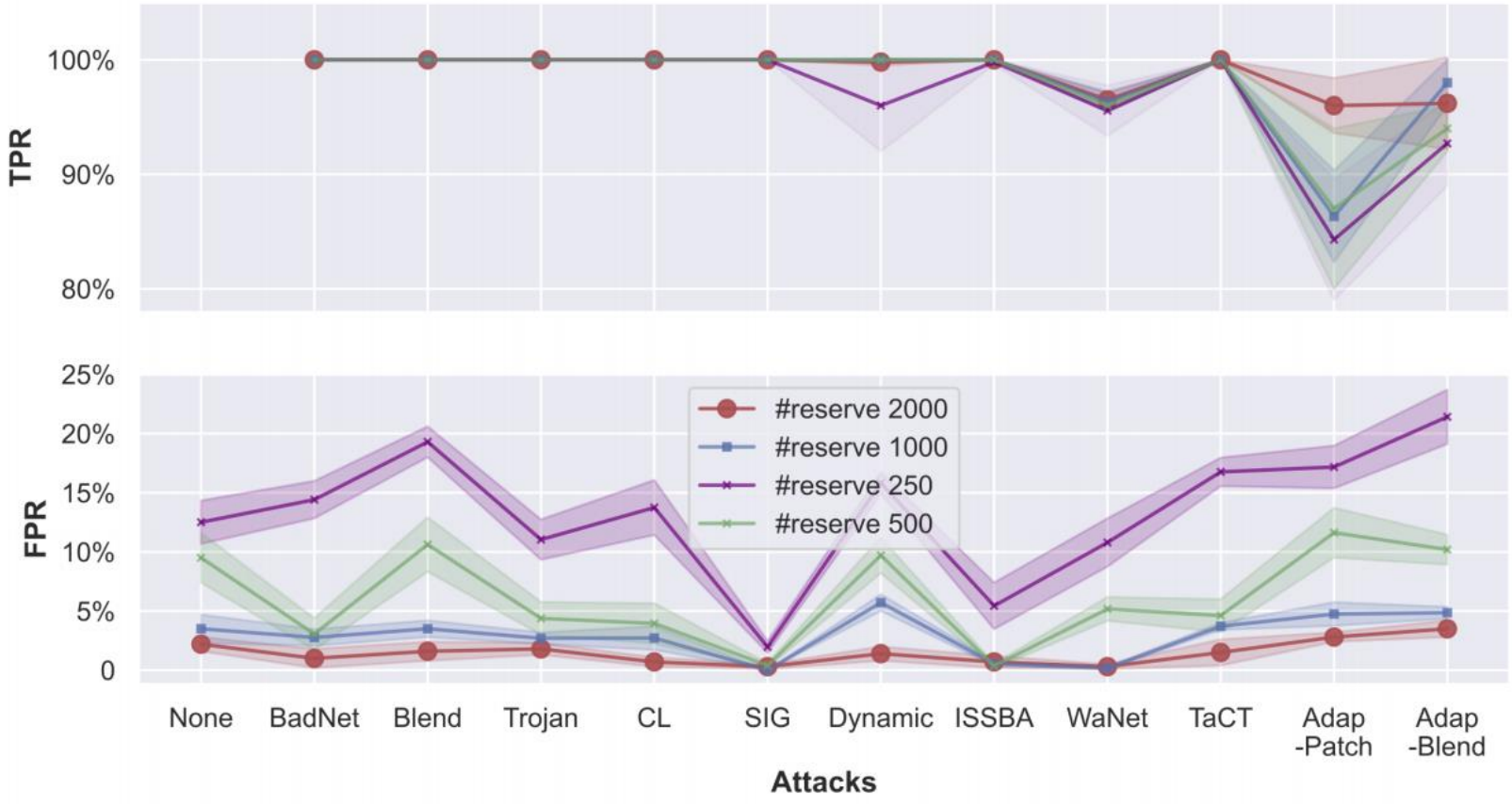
# 实验结果 参数实验1

- 模型架构：ResNet18和MobileNetV2更适合部署CT方法
  - 探究不同模型架构下方法的检测性能
    - 考虑4种模型架构：ResNet18，VGG16，MobileNetV2和DenseNet121
  - 横坐标为不同后门攻击方法；纵坐标为检测性能指标（TPR和FPR）



# 实验结果 参数实验2

- 干净样本集：选取2000个干净样本能达到相对最优的检测性能
  - 探究不同数量的干净样本集对方法检测性能的影响
  - 越少的干净样本集，越偏离中毒训练集中干净样本的分布



# 算法总结

- 算法总结
  - 算法优势
    - 以主动增强后门攻击表现并弱化正常样本与真实标签关联性的“共道”思想解决自适应后门攻击弱化后门样本异常行为特征的问题
    - 无需后门触发器的先验知识，广泛适用不同后门攻击检测，具有良好泛化性
    - 只针对训练集过滤后门样本，过滤后的训练集可迁移其他架构的模型训练
  - 算法劣势
    - 依旧存在前提假设条件
      - 解耦干净样本语义特征和真实标签间的相关性不影响受害模型学习后门触发器和目标标签间的强关联性 → NARCISSUS（2023 CCS）
    - 混淆训练依赖过多超参数，且无法理论证明其最优解
      - 单轮混淆训练的iteration，每轮混淆训练后“毒物蒸馏”的筛选比例等

# 【2023-USENIX Security】
# ASSET: Robust Backdoor Data Detection Across a Multiplicity of Deep Learning Paradigms

# 目标内涵 ASSET

- 研究目标
  - 面向深度神经网络图像分类模型
  - 提出1种主动检测后门攻击的全新方法
  - 实现后门攻击检测准确性的显著提升
- 内涵解析
  - **ASSET: Robust Backdoor Data Detection Across a Multiplicity of Deep Learning Paradigms**
  - 后门样本检测：识别带触发器的后门样本，以检测后门攻击
  - 后门模型检测：识别植入受害模型的后门，以检测后门攻击
  - 主动检测：防御者参与受害模型训练阶段，通过检测训练集中的后门样本提前杜绝模型被注入后门的隐患
  - 被动检测：防御者基于后门样本和正常样本在后门模型中的行为特征差异检测后门攻击，应用于模型推理阶段

# 算法原理 ASSET

| | | |
|---|---|---|
| T | 目标 | 在目标模型训练集中检测并过滤后门样本，提前杜绝后门植入隐患 |
| I | 输入 | 包含后门样本的中毒训练集$\tilde{D}$（如添加后门样本的CIFAR-10），<br>干净样本集$D_{reserve}$（2000个），受害模型（1个，如ResNet18） |
| P | 处理 | 1.最小化受害模型在干净样本集上的样本损失值<br>2.最大化受害模型在中毒训练集上的样本损失值<br>3.基于中毒训练集上的样本损失值的分布差异检测并滤除后门样本 |
| O | 输出 | 滤除后门样本的训练集$D^*$（如滤除后门样本的CIFAR-10） |

| | | |
|---|---|---|
| P | 问题 | 现有方法忽略正常样本语义特征与后门触发器特征间的语义纠缠问题，<br>导致解耦干净样本语义特征和其真实标签间的相关性时，损失后门触发器与受害模型的强关联性，造成后门攻击检测准确率降低的问题 |
| C | 条件 | 1.防御者能完全控制训练集及训练过程<br>2.防御者拥有一小部分干净样本 |
| D | 难点 | 如何有效地最大化中毒数据集的样本损失值 |
| L | 水平 | USENIX Security 2023（CCF A类）- Virginia Tech |

# 算法原理 ASSET

- 创新分析
  - 设计双目标优化策略
    - 目标1：最小化干净样本集上的损失值
    - 目标2：最大化中毒训练集（包含后门样本和干净样本）上的损失值

$$\theta^* = arg\min_{\theta} \frac{1}{|D_{reserve}|} \sum_{x_b \in D_{reserve}} \ell_{min}(f(x_b|\theta)) - \frac{1}{|\widetilde{D}|} \sum_{\tilde{x} \in \widetilde{D}} \ell_{max}(f(\tilde{x}|\theta))$$

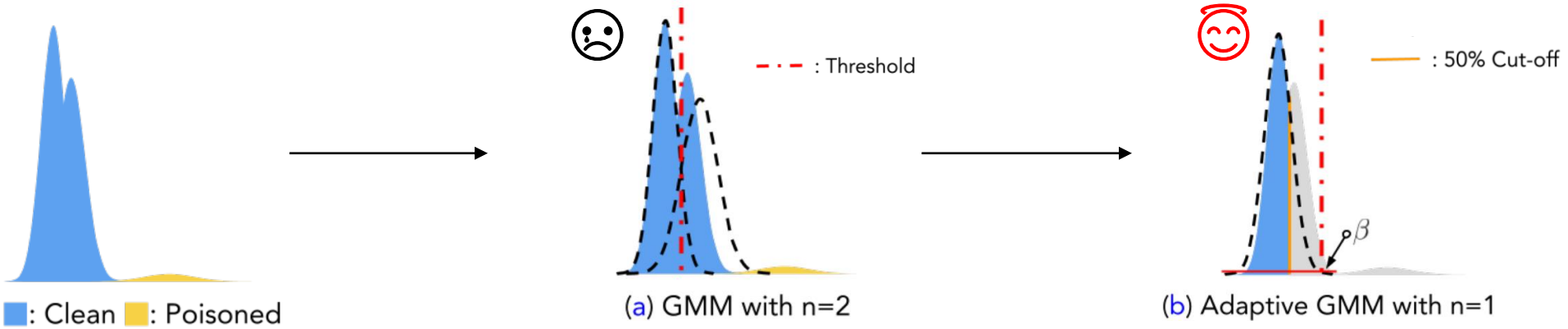
      - $\theta^*$为受害模型（检测器）参数；$\widetilde{D}$为包含后门样本的中毒训练集；$D_{reserve}$为干净样本集，其样本分布与训练集一致
    - 目标1对干净样本的“正”优化将抵消（Offset）目标2对干净样本的“负”优化
  - 最终导致后门样本损失值的分布情况明显偏离正常样本损失值的分布情况

# 算法原理 ASSET

- 算法原理
  - 训练阶段：设计双目标优化策略

$$\theta^* = arg\min_{\theta} \frac{1}{|D_{reserve}|} \sum_{x_b \in D_{reserve}} \ell_{min}(f(x_b|\theta)) - \frac{1}{|\widetilde{D}|} \sum_{\tilde{x} \in \widetilde{D}} \ell_{max}(f(\tilde{x}|\theta))$$

    - 如何选择$\ell_{min}$和$\ell_{max}$？如何设计$\theta^*$的优化路径？（附录C）
    - 如何预处理$\widetilde{D}$，使正常样本与后门样本的损失值分布的偏离更明显？（附录D）
  - 检测阶段：自适应高斯混合模型（Adaptive GMM）



# 实验设计 ASSET

- 数据来源

| 数据集名称 | 样本数量（训练集/测试集） | 特征维度 | 标签数量 | 数据类型 | 描述 |
|---|---|---|---|---|---|
| CIFAR-10 | 50000/10000 | 28*28*3 | 10 | 图像 | 常见物体分类 |
| STL-10 | 105000/5000 | 96*96*3 | 10 | 图像 | 常见物体分类，适用于无监督学习 |
| ImageNet-100 | 100000/30000 | 224*224*3 | 100 | 图像 | 常见物体分类，适用于无监督学习 |

- 评价指标
  - TPR（True Positive Rate）：方法识别的后门样本数量占总后门样本数量的比例
  - FPR（False Positive Rate）：方法将正常样本误判为后门样本的数量占总正常样本数量的比例
  - ACC（Clean Accuracy）：受害模型基于干净样本的预测准确率
  - ASR（Attack Success Rate）:受害模型基于后门样本的攻击成功率

# 实验设计 ASSET

- 后门攻击设置：7种后门攻击，涵盖不同攻击策略

| 类型 | 序号 | 名称 | 水平 |
|---|---|---|---|
| 基于标签翻转的传统后门攻击 | 1 | BadNet | 2017 arXiv（经典方法） |
| | 2 | Blend | 2017 arXiv（经典方法） |
| 基于干净标签的隐蔽后门攻击 | 3 | LC | 2019 arXiv（经典方法） |
| | 4 | SIG | 2019 ICIP（CCFC） |
| | 5 | Narci. | 2023 CCS（CCFA） |
| 触发器选择和优化的高效后门攻击 | 6 | ISSBA | 2021 CVPR（CCFA） |
| | 7 | WaNet | 2021 ICLR |

# 实验设计 ASSET

- 对比方法设置：6种后门样本检测方法，1种后门攻击防御方法

| 类型 | 序号 | 名称 | 水平 | 说明 |
| --- | --- | --- | --- | --- |
| 后门样本检测方法 | 1 | Spectral | 2018 NIPS（CCFA） | 被动后门攻击检测 |
| | 2 | STRIP | 2019 ACSAC（CCFB） | |
| | 3 | AC | 2019 AAAI Workshop | |
| | 4 | Spectre | 2021 ICML（CCFA） | |
| | 5 | Beatrix | 2022 NDSS（CCFA） | |
| | 6 | CT | 2023 USENIX（CCFA） | 主动后门攻击检测 |
| 后门样本防御方法 | 7 | ABL | 2021 NIPS（CCFA） | 反后门学习方法 |

# 实验结果 对比实验

- 后门攻击检测/防御对比实验
  - 蓝色代表ASR小于20%，防御有效；橙红色代表ASR大于20%，防御失效

| | *Dirty-Label Backdoor Attacks* | | | | | | | | *Clean-Label Backdoor Attacks* | | | | | | **Average** | | **Worst-Case** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **BadNets (5%)** | | **Blended (5%)** | | **WaNet (10%)** | | **ISSBA (1%)** | | **LC (1%)** | | **SAA (1%)** | | **Narci. (0.05%)** | | | | | |
| **Detection Evaluation** | | | | | | | | | | | | | | | | | | |
| | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ |
| **Spectral** | 95.6 | 2.86 | 99.8 | 2.64 | 0.64 | 16.6 | 0.00 | 1.52 | 80.2 | 0.71 | 87.6 | 0.63 | 0.00 | 0.08 | 51.9 | 3.58 | 0.00 | 16.6 |
| **Spectre** | 96.9 | 0.28 | 99.8 | 2.64 | 1.00 | 16.6 | 80.2 | 0.71 | 99.8 | 0.51 | 99.4 | 0.51 | 0.00 | 0.07 | 68.2 | 3.05 | 0.00 | 16.6 |
| **Beatrix** | 93.8 | 1.81 | 67.9 | 3.04 | 82.2 | 0.53 | 73.4 | 1.31 | 91.2 | 0.29 | 69.8 | 1.58 | 12.0 | 1.97 | 70.4 | 1.50 | 12.0 | 3.04 |
| **AC** | 90.5 | 40.1 | 65.4 | 44.9 | 8.30 | 41.5 | 11.0 | 41.1 | 91.2 | 0.41 | 75.6 | 21.5 | 0.00 | 34.3 | 48.9 | 32.0 | 0.00 | 44.9 |
| **ABL** | 85.4 | 3.40 | 93.4 | 2.98 | 28.1 | 13.5 | 55.2 | 0.96 | 87.2 | 0.63 | 73.4 | 0.77 | 0.00 | 0.07 | 60.4 | 3.19 | 0.00 | 13.5 |
| **Strip** | 25.4 | 11.5 | 17.3 | 12.1 | 5.08 | 10.0 | 68.8 | 9.34 | 100 | 0.85 | 63.4 | 1.22 | 0.00 | 0.05 | 40.0 | 6.44 | 0.00 | 10.0 |
| **CT** | 99.0 | 3.72 | 98.1 | 4.53 | 95.8 | 2.64 | 96.6 | 4.37 | 100 | 9.01 | 95.2 | 5.44 | 0.00 | 5.54 | 83.5 | 5.03 | 0.00 | 9.01 |
| **Ours** | 99.5 | 0.55 | 100 | 0.00 | 90.7 | 8.09 | 95.6 | 0.36 | 96.2 | 0.75 | 96.6 | 0.39 | 92.0 | 0.34 | **95.8** | **1.49** | **90.7** | **8.09** |
| **Defense Evaluation** | | | | | | | | | | | | | | | | | | |
| | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ |
| **No Def.** | 96.5 | 93.4 | 94.9 | 93.5 | 99.4 | 93.5 | 92.6 | 94.1 | 100 | 94.7 | 76.7 | 94.4 | 99.7 | 94.9 | 94.3 | 94.1 | 100 | 93.4 |
| **Spectral** | 48.4 | 94.5 | 10.7 | 94.1 | 98.9 | 90.0 | 93.0 | 94.1 | 10.6 | 94.8 | 3.11 | 94.2 | 99.7 | 94.8 | 52.1 | 93.8 | 99.7 | 90.0 |
| **Spectre** | 34.8 | 94.5 | 6.57 | 94.1 | 100 | 89.6 | 14.0 | 94.3 | 100 | 94.7 | 0.86 | 94.4 | 99.8 | 94.9 | 50.9 | 93.8 | 100 | 89.6 |
| **Beatrix** | 55.6 | 93.8 | 94.9 | 93.8 | 2.13 | 94.1 | 17.0 | 94.2 | 4.12 | 94.8 | 8.64 | 94.3 | 90.4 | 94.5 | 39.0 | 94.2 | 94.9 | 93.8 |
| **AC** | 81.3 | 76.9 | 93.3 | 82.1 | 99.7 | 83.1 | 83.5 | 81.3 | 4.31 | 94.8 | 7.63 | 87.7 | 100 | 90.7 | 67.1 | 85.0 | 100 | 76.9 |
| **ABL** | 88.6 | 92.5 | 94.2 | 88.7 | 90.2 | 93.1 | 30.6 | 94.2 | 6.32 | 94.7 | 7.63 | 94.4 | 99.3 | 94.9 | 59.6 | 93.2 | 99.3 | 88.7 |
| **Strip** | 76.9 | 85.3 | 93.8 | 87.1 | 98.6 | 91.7 | 25.5 | 91.0 | 0.38 | 94.8 | 9.63 | 94.4 | 99.8 | 94.9 | 57.8 | 91.3 | 99.8 | 81.3 |
| **CT** | 3.42 | 93.1 | 31.3 | 91.2 | 0.53 | 92.5 | 1.12 | 93.2 | 0.44 | 91.1 | 2.16 | 93.2 | 100 | 94.1 | 19.9 | 92.6 | 100 | 91.1 |
| **Ours** | 2.68 | 94.9 | 0.44 | 95.2 | 1.89 | 93.1 | 1.55 | 94.8 | 1.16 | 94.9 | 1.14 | 94.4 | 9.68 | 94.9 | **2.65** | **94.6** | **9.68** | **93.1** |

# 实验结果 对比实验

- 后门攻击检测/防御对比实验
  - 蓝色代表ASR小于20%，防御有效；橙红色代表ASR大于20%，防御失效

| | *Dirty-Label Backdoor Attacks* | | | | *Clean-Label Backdoor Attacks* | | | Average | | Worst-Case | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | BadNets (5%) | Blended (5%) | WaNet (10%) | ISSBA (1%) | LC (1%) | SAA (1%) | Narci. (0.05%) | | | | |
| **Detection Evaluation** | | | | | | | | | | | |
| | TPR ↑ FPR ↓ | TPR ↑ FPR ↓ | TPR ↑ FPR ↓ | TPR ↑ FPR ↓ | TPR ↑ FPR ↓ | TPR ↑ FPR ↓ | TPR ↑ FPR ↓ | TPR ↑ FPR ↓ | | TPR ↑ FPR ↓ | |
| **Spectral** | 95.6 2.86 | 99.8 2.64 | 0.64 16.6 | 0.00 1.52 | 80.2 0.71 | 87.6 0.63 | 0.00 0.08 | 51.9 3.58 | | 0.00 16.6 | |
| **Spectre** | 96.9 0.28 | 99.8 2.64 | 1.00 16.6 | 80.2 0.71 | 99.8 0.51 | 99.4 0.51 | 0.00 0.07 | 68.2 3.05 | | 0.00 16.6 | |
| **Beatrix** | 93.8 1.81 | 67.9 3.04 | 82.2 0.53 | 73.4 1.31 | 91.2 0.29 | 69.8 1.58 | 12.0 1.97 | 70.4 1.50 | | 12.0 3.04 | |
| **AC** | 90.5 40.1 | 65.4 44.9 | 8.30 41.5 | 11.0 41.1 | 91.2 0.41 | 75.6 21.5 | 0.00 34.3 | 48.9 32.0 | | 0.00 44.9 | |
| **ABL** | 85.4 3.40 | 93.4 2.98 | 28.1 13.5 | 55.2 0.96 | 87.2 0.63 | 73.4 0.77 | 0.00 0.07 | 60.4 3.19 | | 0.00 13.5 | |
| **Strip** | 25.4 11.5 | 17.3 12.1 | 5.08 10.0 | 68.8 9.34 | 100 0.85 | 63.4 1.22 | 0.00 0.05 | 40.0 6.44 | | 0.00 10.0 | |
| **CT** | 99.0 3.72 | 98.1 4.53 | 95.8 2.64 | 96.6 4.37 | 100 9.01 | 95.2 5.44 | 0.00 5.54 | 83.5 5.03 | | 0.00 9.01 | |
| **Ours** | 99.5 0.55 | 100 0.00 | 90.7 8.09 | 95.6 0.36 | 96.2 0.75 | 96.6 0.39 | 92.0 0.34 | **95.8** **1.49** | | **90.7** **8.09** | |
| **Defense Evaluation** | | | | | | | | | | | |
| | ASR ↓ ACC ↑ | ASR ↓ ACC ↑ | ASR ↓ ACC ↑ | ASR ↓ ACC ↑ | ASR ↓ ACC ↑ | ASR ↓ ACC ↑ | ASR ↓ ACC ↑ | ASR ↓ ACC ↑ | | ASR ↓ ACC ↑ | |
| **No Def.** | 96.5 93.4 | 94.9 93.5 | 99.4 93.5 | 92.6 94.1 | 100 94.7 | 76.7 94.4 | 99.7 94.9 | 94.3 94.1 | | 100 93.4 | |
| **Spectral** | 48.4 94.5 | 10.7 94.1 | 98.9 90.0 | 93.0 94.1 | 10.6 94.8 | 3.11 94.2 | 99.7 94.8 | 52.1 93.8 | | 99.7 90.0 | |
| **Spectre** | 34.8 94.5 | 6.57 94.1 | 100 89.6 | 14.0 94.3 | 100 94.7 | 0.86 94.4 | 99.8 94.9 | 50.9 93.8 | | 100 89.6 | |
| **Beatrix** | 55.6 93.8 | 94.9 93.8 | 2.13 94.1 | 17.0 94.2 | 4.12 94.8 | 8.64 94.3 | 90.4 94.5 | 39.0 94.2 | | 94.9 93.8 | |
| **AC** | 81.3 76.9 | 93.3 82.1 | 99.7 83.1 | 83.5 81.3 | 4.31 94.8 | 7.63 87.7 | 100 90.7 | 67.1 85.0 | | 100 76.9 | |
| **ABL** | 88.6 92.5 | 94.2 88.7 | 90.2 93.1 | 30.6 94.2 | 6.32 94.7 | 7.63 94.4 | 99.3 94.9 | 59.6 93.2 | | 99.3 88.7 | |
| **Strip** | 76.9 85.3 | 93.8 87.1 | 98.6 91.7 | 25.5 91.0 | 0.38 94.8 | 9.63 94.4 | 99.8 94.9 | 57.8 91.3 | | 99.8 81.3 | |
| **CT** | 3.42 93.1 | 31.3 91.2 | 0.53 92.5 | 1.12 93.2 | 0.44 91.1 | 2.16 93.2 | 100 94.1 | 19.9 92.6 | | 100 91.1 | |
| **Ours** | 2.68 94.9 | 0.44 95.2 | 1.89 93.1 | 1.55 94.8 | 1.16 94.9 | 1.14 94.4 | 9.68 94.9 | **2.65** **94.6** | | **9.68** **93.1** | |

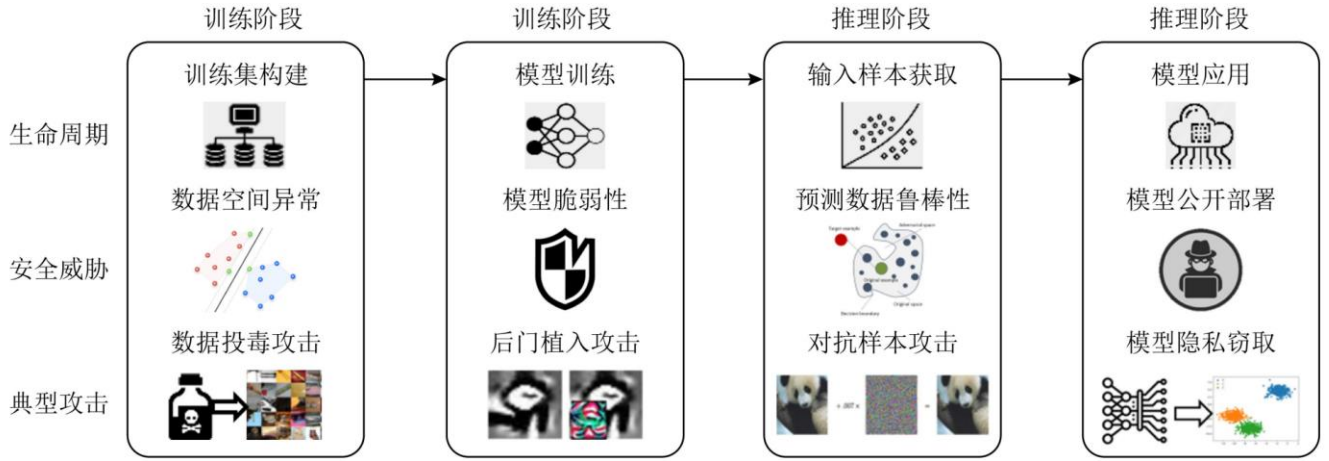
# 算法总结

- 算法总结
  - 算法优势
    - 以主动增强后门攻击表现并弱化正常样本与真实标签关联性的“共道”思想解决自适应后门攻击弱化后门样本异常行为特征的问题
    - 相比算法1，考虑干净样本语义特征和真实标签间的相关性可能影响受害模型学习后门触发器和目标标签间的强关联性
  - 算法劣势
    - 双目标优化的收敛过程需要更充分的理论推导与证明，损失函数优化结果存在次优解问题
    - 防御者必须完全控制训练集并参与训练过程
      - 攻击者是模型训练者？模型提供者？
    - 无法检测/防御基于模型修改的后门攻击方法

# 未来展望

# 未来展望

- 受后门攻击威胁的领域逐渐增加，要求检测方法适用于不同领域
  - 语音/视频识别、自然语言处理、恶意软件分类、强化学习领域、生成式人工智能
- 深度学习后门技术迁移
  - 知识产权保护（模型水印）
  - 防御者引入后门（蜜罐）以防御其他攻击（对抗样本、模型窃取等）
- 人工智能系统多粒度行为建模及全流程安全风险防御/检测框架
  - 人工智能系统生命周期易受大规模、多类别、多节点攻击，存在安全风险




保护深度神经网络模型安全
任重而道远!

# 参考文献

**[1] Qi X, Xie T, Wang J T, et al. Towards a proactive {ML} approach for detecting backdoor poison samples[C]. 32nd USENIX Security Symposium (USENIX Security 23). 2023: 1685-1702.**

**[2] Pan M, Zeng Y, Lyu L, et al. ASSET: Robust Backdoor Data Detection Across a Multiplicity of Deep Learning Paradigms [C]. 32nd USENIX Security Symposium (USENIX Security 23). 2023: 2725-2743.**

**[3] Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. Demon in the variant: Statistical analysis of dnns for robust backdoor contamination detection [C]. 30th USENIX Security Symposium (USENIX Security 21). 2021: 1541-1558.**

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

# 谢谢！

- 基于高斯混合模型聚类分析的可疑目标类别识别

**Algorithm 2** Identify Potential Target Classes

**Input:** Dataset $\widetilde{D} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{n}$, Inference Model $\theta_{ct}$, The Number of Classses $C$, Gaussian Density Function $N(\cdot;\mu,\sigma)$, a threshold $u$

**Output:** $T \subseteq \{1,\ldots,C\}$, indicating potential target classes

$T \leftarrow \{\}$
**for** $c = 1,...,C$ **do**
  $H \leftarrow \{h(\tilde{x};\theta_{ct}) | (\tilde{x},\tilde{y}) \in \widetilde{D} \wedge \tilde{y} = c\}$ ▷ Latent Vectors
  $H_0 \leftarrow \{h(\tilde{x};\theta_{ct}) | (\tilde{x},\tilde{y}) \in \widetilde{D} \wedge \tilde{y} = c \wedge F(\tilde{x};\theta_{ct}) \neq c\}$
  $H_1 \leftarrow H \setminus H_0$
  $U, \Lambda, V \leftarrow SVD_2(H)$ ▷ Dimension Reduction
  $S \leftarrow \{U^T h \,|h \in H\}$ ▷ Top-2 PCA Directions
  $S_0, S_1 \leftarrow \{U^T h \,|h \in H_0\}, \{U^T h \,|h \in H_1\}$
  $\widehat{\mu}, \widehat{\sigma} \leftarrow$ empirical mean and covariance of $S$
  $\widehat{\mu}_0, \widehat{\sigma}_0 \leftarrow$ empirical mean and covariance of $S_0$
  $\widehat{\mu}_1, \widehat{\sigma}_1 \leftarrow$ empirical mean and covariance of $S_1$
  $L \leftarrow \sqrt[|S|]{\frac{\prod_{s_0 \in S_0} N(s_0;\widehat{\mu}_0,\widehat{\sigma}_0) \cdot \prod_{s_1 \in S_1} N(s_1;\widehat{\mu}_1,\widehat{\sigma}_1)}{\prod_{s \in S} N(s;\widehat{\mu},\widehat{\sigma})}}$
  **if** $L > u$ **then**
    $T \leftarrow T \cup \{c\}$
  **end if**
**end for**
**return** $T$

# 附录B

- 滤除后门样本的训练集的数据质量如何评价?
  - CT方法基于ResNet18（11M参数量）对ImageNet-1K进行后门样本过滤
  - 现将滤除后门样本的ImageNet-1K对ViT-B/16（150M参数量）进行训练
    - 模型使用GSAM算法训练
    - 由于ViT-B/16的训练开销较大，论文只进行1次实验，故无标准差
  - 完全滤除后门攻击隐患，显著提高ImageNet-1K数据集的质量

| (%) | No Poison | BadNet | Trojan | Blend |
|---|---|---|---|---|
| ACC | 80.7 | 80.8 | 80.6 | 80.7 |
| ASR | - | 0 | 0 | 0 |

# 附录C

- 如何选择$\ell_{min}$和$\ell_{max}$？
  - $\ell_{min}$：统一使用无监督数据，选择模型输出的Logits值的方差作为损失函数

$$\mathcal{L}_{var}\big(f(x|\theta)\big) = \frac{1}{k}\sum_{i=0}^{k}(f(x|\theta)_i - \overline{f(x|\theta)})^2$$

    - $f(x|\theta)$为模型$\theta$在输入样本$x$下的Logits值，$k$为模型分类类别数，$f(x|\theta)_i$为模型第$i$个类别的Logits值，$\overline{f(x|\theta)}$为模型所有类别的平均Logits值
  - $\ell_{max}$：在无监督条件下，选择模型输出的Logits值的方差作为损失函数
  - $\ell_{max}$：在有监督条件下，选择交叉熵损失（Cross-Entropy Loss）作为损失函数

$$\mathcal{L}_{ce}(f(x|\theta), y) = -\sum_{i=1}^{k} y_i log\boldsymbol{\sigma}(f(x|\theta))_i$$

    - $y$为$x$的真实标签；$\boldsymbol{\sigma}(f(x|\theta))_i$为模型第$i$个类别的置信度（Softmax层的输出值）；若$y_i = y$，则$y_i = 1$

# 附录D

- 如何设计$\theta^*$的优化路径?
  - 由于直接优化可能导致无法拟合，于是采用两个目标的依次优化循环策略:
    - 第1步：对干净样本集采取梯度下降来最小化$\ell_{min}$;
    - 第2步：利用第1步优化后的模型作为初始模型，并执行梯度上升来最大化$\ell_{max}$ ;
    - 重复以上2个步骤
- 如何预处理$\widetilde{D}$，使正常样本损失值分布与后门样本损失值分布的偏离程度更明显?
  - 目标类似于算法1的“毒物蒸馏”
  - 保证“毒物蒸馏”高精确率（后门样本占比尽可能大）
  - AO：Adjusted Outlyingness，一种自适应阈值的异常检测算法

**Algorithm 1:** Poison Concentration

**Input:** $\theta^*_{poi}$ (Poisoned feature extractor);
$B_{poi}$ (Poisoned training mini-batch);
$B_b$ (Base set mini-batch);
**Output:** $B_{pc}$ (Poison concentrated mini-batch);
**Parameters:** $\mathcal{N}$ (Total inner loop iteration number);
$\gamma > 0$ (Step size);
$\lambda$ (Threshold);

/* 1.Dynamic training of $M$ */
**for** *each iteration* $j$ *in* $(0, \mathcal{N}-1)$ **do**
  $M'_j \leftarrow M_j - \gamma \frac{1}{|B_b|} \sum_{x_b \in B_b} \frac{\partial \mathcal{L}_{BCE}(M(f(x_b|\theta^*_{poi})),0)}{\partial M}$ ;
  $M_{j+1} \leftarrow M'_j - \gamma \frac{1}{|B_{poi}|} \sum_{x_{poi} \in B_{poi}} \frac{\partial \mathcal{L}_{BCE}(M'(f(x_{poi}|\theta^*_{poi})),1)}{\partial M'}$;
/* 2.Get output values */
$V \leftarrow M_{\mathcal{N}}\left(f(B_{poi}|\theta^*_{poi})\right)$;
/* 3.Using AO to determine outliers */
$B_{pc} \leftarrow B_{poi}[\mathbf{AO}(V) \geq \lambda]$;
**return** $B_{pc}$

# 附录E ASSET实验结果-补充1

- **STL-10数据集下的检测/防御性能**

| | FT-all | | | | FT-last | | | | Average | | Worst-Case | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BadNets (20%) | | SAA (5%) | | Blended (20%) | | HTBA (5%) | | | | | |
| **Detection Evaluation** | | | | | | | | | | | | |
| | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ | TPR ↑ | FPR ↓ |
| **Spectral** | 54.4 | 23.9 | 11.2 | 7.31 | 16.6 | 33.4 | 25.6 | 4.97 | 27.0 | 17.4 | 11.2 | 33.4 |
| **Spectre** | 76.8 | 18.3 | 17.2 | 6.99 | 16.0 | 33.5 | 46.4 | 3.87 | 39.1 | 15.7 | 16.0 | 33.5 |
| **Beatrix** | 86.9 | 3.55 | 74.8 | 12.1 | 56.7 | 11.7 | 89.2 | 13.5 | 76.9 | 10.2 | 56.7 | 13.5 |
| **AC** | 34.6 | 46.2 | 18.0 | 14.5 | 9.80 | 60.3 | 8.40 | 13.3 | 17.7 | 33.6 | 8.40 | 60.3 |
| **ABL** | 81.2 | 17.2 | 57.2 | 4.88 | 75.3 | 18.7 | 75.6 | 3.91 | 72.3 | 11.2 | 57.2 | 18.7 |
| **Strip** | 83.2 | 11.2 | 0.00 | 20.7 | 52.9 | 16.3 | 71.2 | 17.6 | 51.8 | 16.5 | 0.00 | 20.7 |
| **CT** | 98.3 | 10.5 | 82.4 | 3.98 | 98.7 | 6.58 | 96.4 | 2.57 | 94.0 | 5.91 | 82.4 | 10.5 |
| **Ours** | 97.7 | 0.53 | 90.8 | 0.34 | 99.6 | 0.18 | 99.2 | 0.19 | **96.8** | **0.31** | **90.8** | **0.53** |
| **Defense Evaluation** | | | | | | | | | | | | |
| | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ | ASR ↓ | ACC ↑ |
| **No Def.** | 99.6 | 97.9 | 93.6 | 98.6 | 97.7 | 98.5 | 68.1 | 98.5 | 89.8 | 98.4 | 99.6 | 97.9 |
| **Spectral** | 99.5 | 97.6 | 91.4 | 98.1 | 97.6 | 98.4 | 49.6 | 98.5 | 84.5 | 98.2 | 99.5 | 97.6 |
| **Spectre** | 99.3 | 98.0 | 86.3 | 98.1 | 97.6 | 98.4 | 31.3 | 98.5 | 78.6 | 98.3 | 99.3 | 98.0 |
| **Beatrix** | 99.4 | 98.0 | 36.2 | 97.9 | 95.2 | 98.6 | 8.93 | 98.5 | 59.9 | 98.3 | 99.4 | 97.9 |
| **AC** | 99.6 | 97.3 | 85.6 | 98.0 | 98.4 | 98.2 | 56.3 | 98.5 | 85.0 | 98.0 | 99.6 | 97.3 |
| **ABL** | 99.6 | 97.1 | 59.6 | 98.2 | 94.2 | 98.5 | 14.3 | 98.5 | 66.9 | 98.1 | 99.6 | 97.1 |
| **Strip** | 99.5 | 97.7 | 94.0 | 97.8 | 98.4 | 98.4 | 16.8 | 98.4 | 77.2 | 98.1 | 99.5 | 97.7 |
| **CT** | 7.65 | 98.1 | 11.2 | 98.2 | 90.2 | 98.4 | 1.27 | 98.5 | 27.6 | 98.3 | 90.2 | **98.1** |
| **Ours** | 8.93 | 98.1 | 15.4 | 98.1 | 1.44 | 99.2 | 0.36 | 98.5 | **6.53** | **98.5** | **15.4** | **98.1** |

# 附录F ASSET实验结果-补充2

- **ImageNet-100数据集下的检测/防御性能**

| | No Attack | | | C-brd (5%) | | | | C-Squ (5%) | | | | CTRL (1%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASR ↓ | ASR* ↓ | ACC ↑ | $ASR^*_0$ | ASR* ↓ | $ACC_0$ | ACC ↑ | $ASR^*_0$ | ASR* ↓ | $ACC_0$ | ACC ↑ | $ASR_0$ | ASR ↓ | $ACC_0$ | ACC ↑ |
| **SimCLR** | 0.34 | 6 | 67.2 | 168 | 8 | 65.9 | 66.9 | 141 | 12 | 65.6 | 66.8 | 25.0 | 1.36 | 66.1 | 66.8 |
| **MoCo V3** | 0.32 | 6 | 68.4 | 67 | 8 | 68.0 | 68.2 | 119 | 12 | 67.8 | 68.2 | 23.6 | 2.12 | 68.2 | 68.3 |
| **BYOL** | 0.36 | 7 | 67.1 | 290 | 11 | 66.6 | 67.1 | 263 | 19 | 66.8 | 66.9 | 40.9 | 1.44 | 66.5 | 66.8 |
| **MAE** | 0.28 | 5 | 70.2 | 28 | 6 | 68.9 | 70.1 | 68 | 8 | 69.1 | 69.9 | 30.7 | 1.66 | 68.7 | 69.3 |

- $ASR^*$代表攻击成功的后门样本数量
- $ASR^*_0$和$ACC_0$代表没有防御方法下的攻击成功率和模型准确率