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

Evolution Is Also You Need
—深度学习模型后门攻击检测

硕士研究生 李嘉玮

2024年05月19日

问题回溯

- 总结反思
  - PPT的字过多
  - 实验部分需要多加分析
  - 讲述过程中存在重复、冗余的现象，需精简语言
- 相关内容
  - 后门攻击
    - 2023.06.26 Saba《Deep Learning Backdoor Attacks Detection》
    - 2023.04.09 杨得山《联邦学习的后门防御方法》
  - 后门攻击检测/防御
    - 2024.01.14 赵怡清《对抗性扰动下的后门防御方法》
    - 2023.10.29 李嘉玮《深度神经网络模型后门攻击检测》
    - 2023.03.19 吴肖龙《基于模型修改的深度学习后门攻击》

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
  - Orion
  - TED
- 特点总结与工作展望
- 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
  - 掌握在模型应用阶段的后门攻击检测策略和算法原理
  - 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　深度学习模型后门攻击检测

- 研究目标
  - 面向AI深度学习图像分类和文本分类模型
  - 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，将图像分类到预定义类别
  - 文本分类：多分类任务，将文本对象分类到预定义类别（如情感分析、意图识别）
  - 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

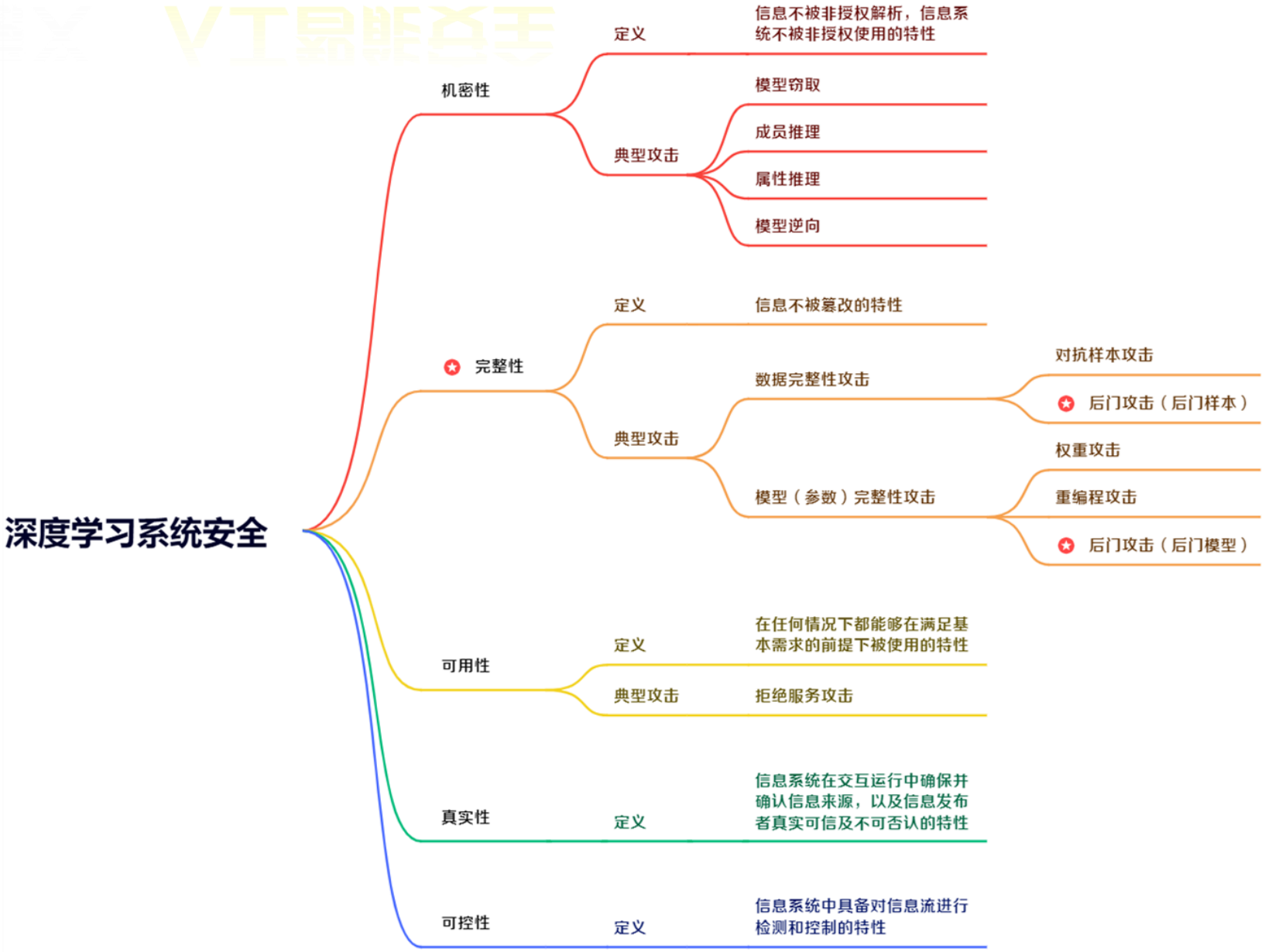
- 信息系统安全基本要素（CIA）
  - 机密性（Confidentiality）：指信息不被非授权解析，信息系统不被非授权使用的特性
    - 数据安全：确保数据即便被捕获也不会被解析
    - 物理安全、运行安全：确保信息系统即便能够被访问也不能够越权访问与其身份不相符的信息
  - 完整性（Integrity）：指信息不被篡改的特性
    - 数据安全：确保信息不被篡改或任何被篡改了的信息都可以被发现
  - 可用性（Availability）：指信息与信息系统在任何情况下都能够在满足基本需求的前提下被使用的特性
    - 物理安全、运行安全：确保基础信息系统的正常运行能力，保障数据的正常传递、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　深度学习模型后门攻击检测

- 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
- 研究意义
  - 面向数据集的后门攻击检测对先验知识（数据集类别、特性等）的要求高
  - 要求用户或检测人员需要获取模型全部训练集，难以应用第三方模型下的后门攻击检测
  - 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和理论意义

研究历史

2018 Tran等人基于待测样本特征表示的协方差频谱检测后门样本

2019 Gao等人基于后门触发器鲁棒性，通过度量待测样本的模型输出熵，观察模型预测结果的一致性进行后门样本检测

2020 Chou等人提出1种局部静态触发器后门攻击检测的方法，基于重定位和显著图来识别触发器以检测后门样本

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

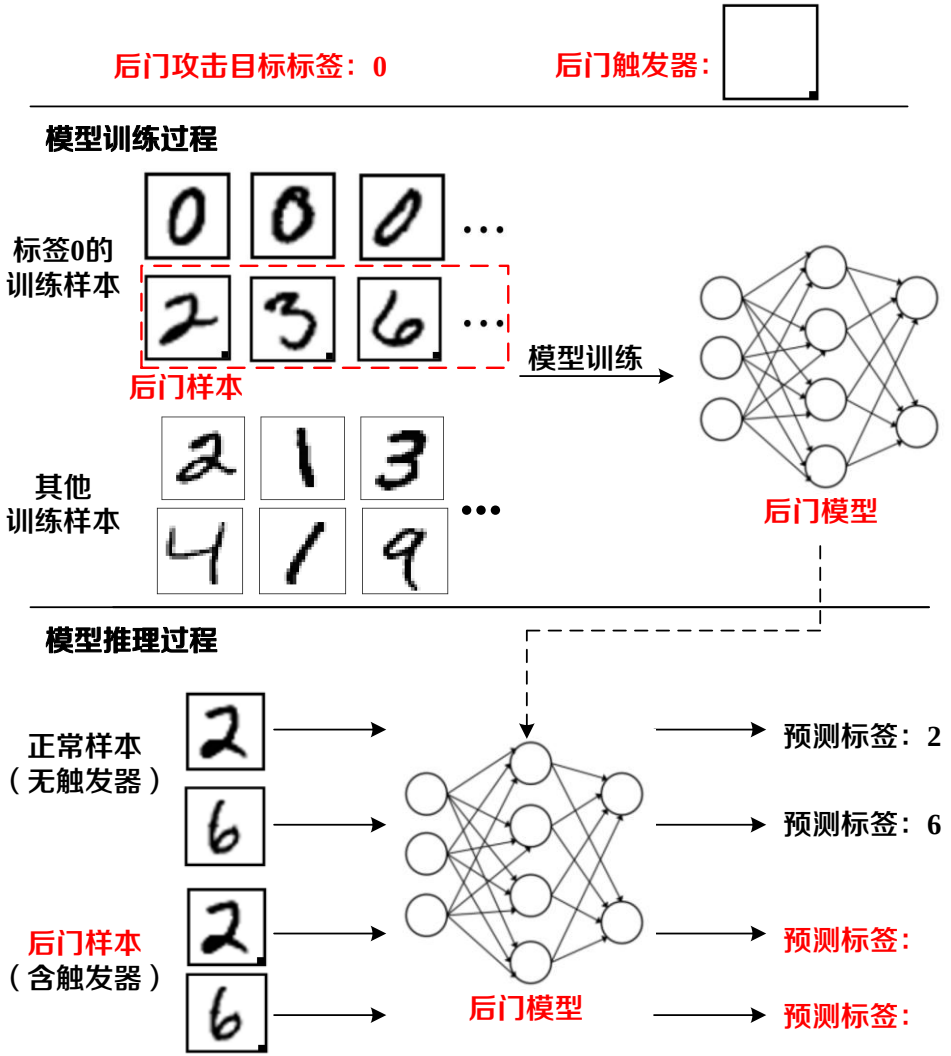
2023 Huang等人提出基于旁路神经网络的后门检测方法，针对后门样本和正常样本在特征空间不可分离的先进后门攻击，建模样本在模型向前传播过程中的演化差异检测后门样本

2024 Mo等人依据待测样本在模型向前传播过程中的预测一致性，构建模型演化序列进行后门样本检测

- 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
  - 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

- 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

【2023-IJCAI】
Orion: Online Backdoor Sample Detection via Evolution Deviance

算法原理 Orion

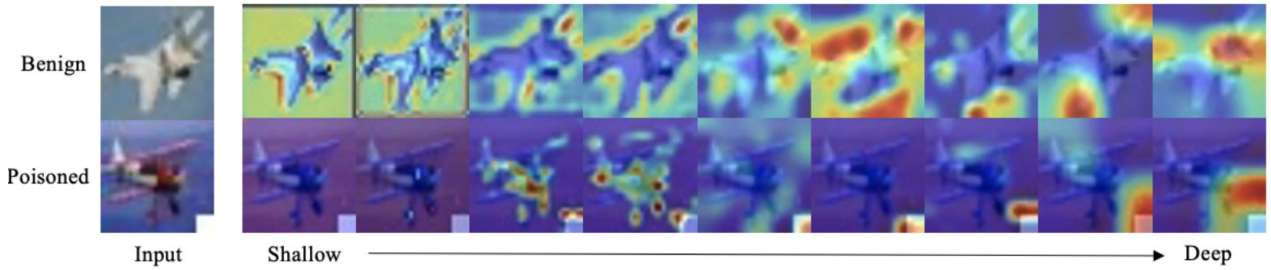
T	目标	在受害模型推理阶段，检测后门样本并拒绝其输入模型
I	输入	受害模型（1个）、干净样本集（训练集1%，类别平衡）、 待测样本（1个）
P	处理	1. 构造并训练旁路网络 2. 获取待测样本在旁路网络的输出值 3. 基于旁路网络输出值计算异常分数 4. 结合阈值进行后门样本检测
O	输出	（是/否）后门样本

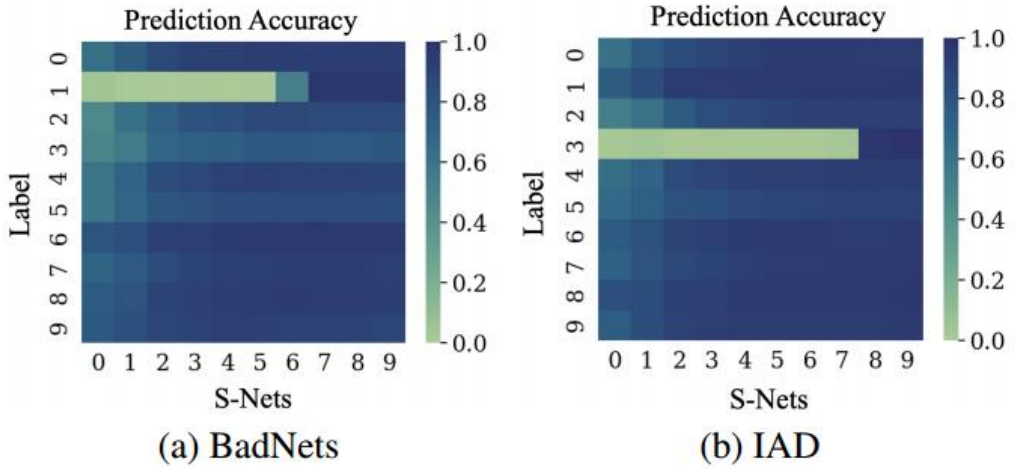
P	问题	现有方法假设在度量空间（如模型特征空间）中正常样本和后门样本具有可分离性，无法检测自适应的强隐蔽后门攻击，导致检测准确率下降
C	条件	1. 拥有少量干净样本集（如CIFAR10中，每类样本20个，合计200个） 2. 能够访问受害模型隐藏层及其激活向量
D	难点	样本在模型前向传播过程中演化特征的精确提取
L	水平	IJCAI 2023（CCF A类）

创新分析 Orion

- 核心思想
  - 利用后门样本和正常样本在模型前向传播中的演化差异进行建模
- 算法步骤
  - 构造并训练旁路网络（Side Nets）
    - 特征提取模块
    - 样本分类模块
  - 获取待测样本在旁路网络的输出值
  - 基于旁路网络输出值计算异常分数
    - “一致性”、“稳定性”、“确定性”
  - 结合阈值进行后门样本检测

把握后门样本的本质特征

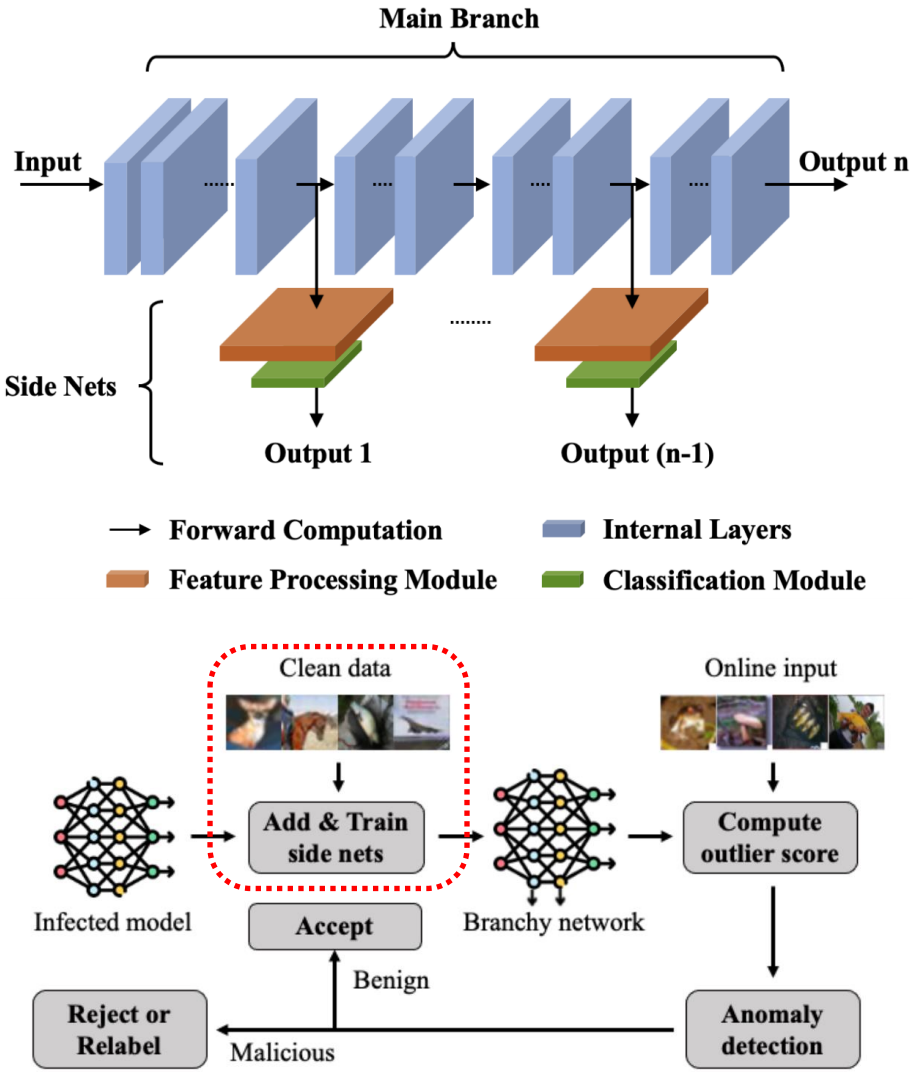




(a) BadNets (b) IAD

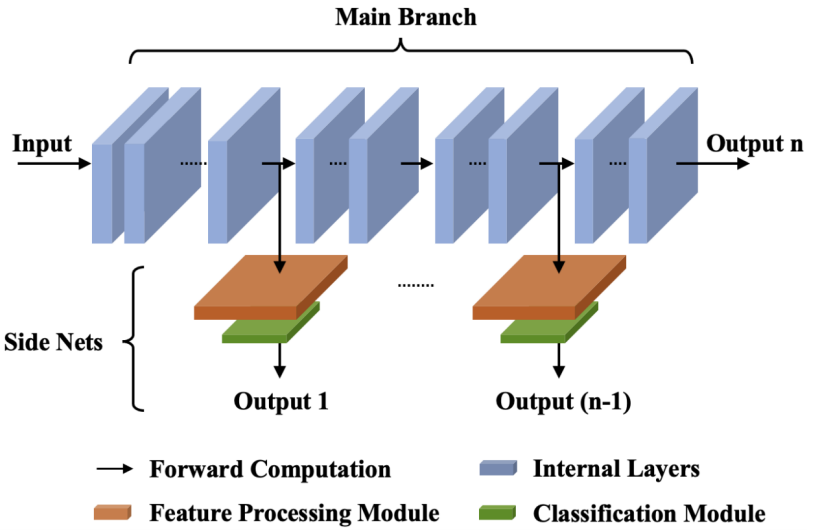
Figure 4: Prediction accuracy over layers. The target classes for BadNets and IAD are 1 and 3, respectively.

- 构造并训练旁路网络（Side Nets）
  - 对受害模型的指定隐藏层额外添加1个神经网络
    - 保证模型正常向前传播过程不变，并且能提前输出模型对样本的判别结果
  - 特征提取模块：Mix Pooling策略
    - Max Pooling + Average Pooling
    - 在保持池化层功能的同时尽可能减少池化过程中带来的信息损失
  - 样本分类模块：全连接层
- 获取待测样本在旁路网络的输出值



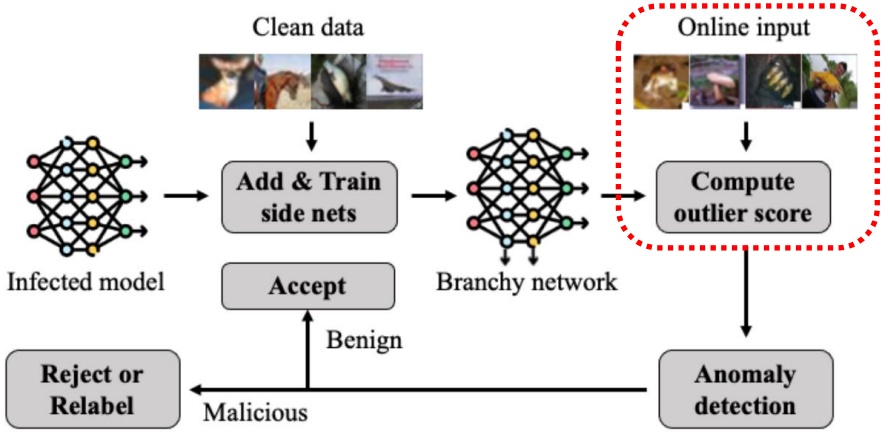
- 基于旁路网络输出值计算异常分数
  - “一致性”：所有S-Nets输出与最终模型输出的L1距离之和
    - $\Phi_c(x) = \sum_{i=1}^{n-1} \|F_i(x) - F_n(x)\|_1$
    - 后门样本在模型浅层和深层的输出差异比正常样本大
  - “稳定性”：计算相邻S-Nets输出的余弦相似度
    - $\Phi_s(x) = \sum_{i=1}^{n-1}(1 - cossim(F_i(x), F_{i+1}(x)))$
    - 后门样本的变化比正常样本更多，稳定性更差
  - “确定性”：评估S-Nets的输出偏离受害模型预测结果时的置信水平
    - $\Phi_d(x) = \sum_{i=1}^{n-1} max(F_i(x))\mathbb{I}(P_i(x) \neq P_n(x))$
    - 后门样本比正常输入表现出更高的假置信度

$F(x)$: 神经网络层输出值
$P_n(x)$: 输出标签
$\|\cdot\|_1$: L1范数
$cossim(\cdot)$: 余弦相似度
$\mathbb{I}(\cdot)$: 指示函数



- 基于旁路网络输出值计算异常分数
  - “一致性”：$\Phi_c(x) = \sum_{i=1}^{n-1} \|F_i(x) - F_n(x)\|_1$
  - “稳定性”：$\Phi_s(x) = \sum_{i=1}^{n-1} (1 - cossim(F_i(x), F_{i+1}(x)))$
  - “确定性”：$\Phi_d(x) = \sum_{i=1}^{n-1} max(F_i(x)) \mathbb{I}(P_i(x) \neq P_n(x))$
  - 异常分数：$\Phi(x) = \alpha * \Phi_c(x) + \beta * \Phi_s(x) + \gamma * \Phi_d(x)$，$\alpha, \beta, \gamma$为超参数
- 结合阈值进行后门样本检测
  - 选择使95%的干净样本被检测方法正确分类所对应的异常分数作为阈值

方法能否基于旁路网络对后门样本进行标签修复?



实验设计

- 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
Tiny-ImageNet	100000/10000	64*64*3	200	图像	自然界图像

- 评价指标
  - PRE（Precision）：准确识别的后门样本数量与识别出的总恶意样本数的比值
  - REC（Recall）：准确识别的后门样本数量与总后门样本数量的比值
- 模型架构
  - VGG：在CIFAR10和GTSRB上训练VGG16
  - ResNet：在Tiny-ImageNet上训练ResNet56

实验设计

- 后门攻击：6种攻击方法，涵盖3种先进攻击策略

类型	序号	名称	水平
标签污染和静态触发的后门攻击	1	BadNets	2017 arXiv（经典方法）
	2	Blend	2017 arXiv（经典方法）
触发器选择和优化的后门攻击	3	WaNet	2021 ICLR
	4	IAD	2021-NeurIPS（CCF-A）
后门特征模糊的自适应后门攻击	5	TaCT	2021-USENIX（CCF-A）
	6	Feature Attack	2022-IJCAI（CCF-A）

- 对比方法：3种领域先进的检测方法

类型	序号	名称	水平
协方差统计特征的后门攻击检测	1	Spectre	2021-ICML（CCF-A）
一阶矩特征分解的后门攻击检测	2	SCAn	2021-USENIX（CCF-A）
高阶矩特征提取的后门攻击检测	3	Beatrix	2023-NDSS（CCF-A）

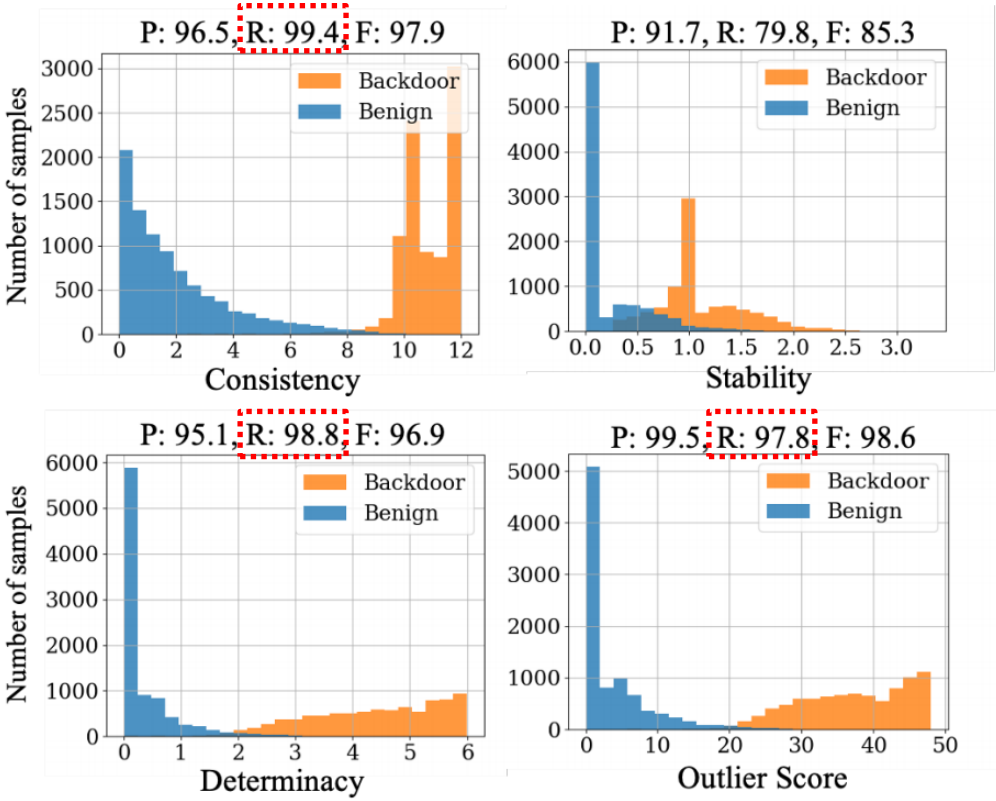
实验结果 对比实验

- 对比实验
  - 随机选取500个后门样本和500个正常样本
  - Orion对所有攻击均能有效检测
    - 由于假设触发器具有静态特征以及通用性，Spectre和SCAn无法检测IAD（蓝色框）
    - 由于假设后门样本和正常样本在模型特征空间具有可分离性， Spectre、SCAn和Beatrix无法检测Feature攻击（红色框）

Dataset	Attacks	Spectre			SCAn			Beatrix			Orion		
		PRE	REC	F1	PRE	REC	F1	PRE	REC	F1	PRE	REC	F1
CIFAR-10	Badnets	95.37	87.30	91.15	89.70	93.80	91.70	96.90	91.40	94.06	99.50	97.80	**98.60**
	Blended	93.30	88.24	90.69	91.47	92.80	92.13	95.00	94.90	94.94	98.90	97.60	**98.20**
	WaNet	100.0	74.07	85.10	92.36	88.79	90.53	95.20	89.20	92.18	99.79	97.80	**98.70**
	IAD	32.38	43.40	37.08	34.20	27.87	30.71	93.60	91.40	92.50	98.00	99.80	**98.90**
	TaCT	85.36	58.33	69.30	95.00	88.80	91.79	94.47	92.40	93.43	99.79	97.00	**98.37**
	Feature	88.69	0.16	0.31	88.33	0.40	0.79	28.20	6.23	10.21	93.70	98.60	**96.10**
GTSRB	Badnets	94.29	76.21	84.26	95.40	82.81	88.66	95.00	91.40	93.00	97.04	91.80	**94.34**
	Blended	94.60	76.41	84.54	93.90	78.60	85.57	95.90	80.60	87.58	94.38	87.40	**90.76**
	WaNet	96.20	79.66	87.15	92.30	83.40	87.62	94.11	90.80	92.42	95.50	99.80	**97.60**
	IAD	95.60	37.24	53.60	83.23	64.40	72.61	92.60	90.80	91.69	95.00	91.60	**93.37**
	TaCT	83.19	46.80	59.90	89.40	80.60	84.77	96.19	70.80	81.56	96.09	83.60	**89.41**
	Feature	30.47	3.43	6.16	99.65	2.89	5.61	54.60	29.42	38.23	95.33	85.80	**90.32**
Tiny-Imagenet	Badnets	99.91	82.20	90.19	100.0	81.00	89.00	94.89	91.49	93.15	95.55	93.40	**94.44**
	Blended	91.45	68.33	78.21	89.18	77.40	82.87	88.36	58.20	70.17	91.02	77.00	**83.42**
	WaNet	77.30	67.21	71.90	76.24	70.16	73.07	82.50	75.30	78.73	88.49	75.40	**81.42**

实验结果　消融实验

- 消融实验：“一致性”、“稳定性”、“确定性”指标的消融
  - 数据集：CIFAR10
  - 攻击方法：BadNets
  - 后门样本和正常样本在3个指标上均具有可分性，其中“一致性”指标可分性最好
  - “稳定性”、“确定性”表征模型向前传递过程中预测演变的变化频率和变化程度

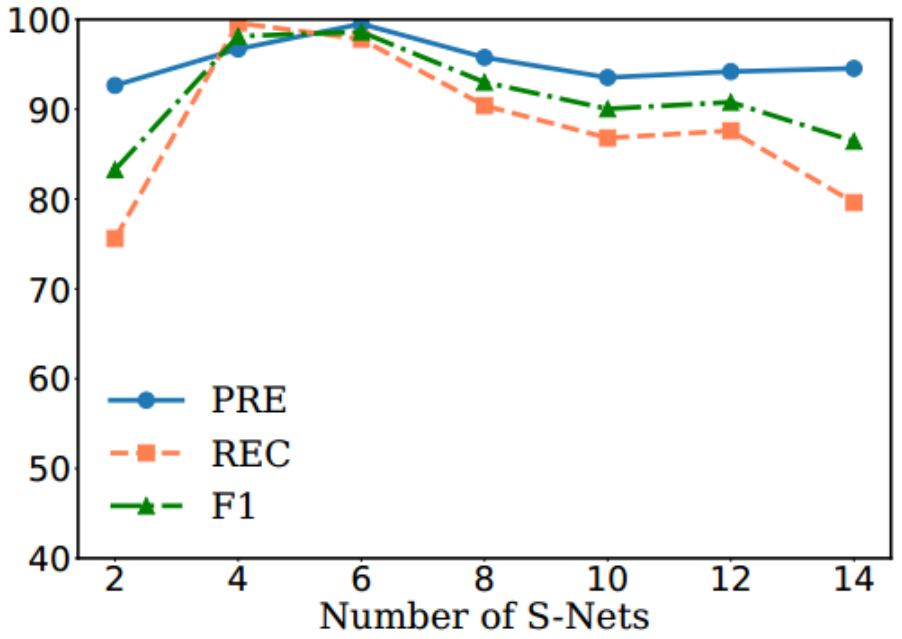


三者为什么一定就是相互促进、相互互补的?

异常分数：$\Phi(x) = \alpha * \Phi_c(x) + \beta * \Phi_s(x) + \gamma * \Phi_d(x)$，$\alpha, \beta, \gamma$为超参数

实验结果　参数实验

- 参数实验1：S-Nets的部署数量
  - 过少的S-Nets会导致无法准确建模不同样本在模型向前传播过程中的演化差异
  - 过多的S-Nets会减小同一样本在模型向前传播过程中的变化方差，并带来更大的训练开销和检测时间
- 参数实验2：干净样本集的样本总量
  - 随着干净样本集中样本数量的增加，检测性能逐渐提高，最后趋于平缓
  - 过多的干净样本数量会不符合实际应用情况，并带来更大开销



Reference set	PRE	REC	F1
0.1%	85.64	33.40	48.06
0.5%	95.31	93.40	94.34
1%	99.50	97.80	98.60
5%	98.60	98.90	98.70
10%	99.20	99.39	99.29

算法总结 Orion

- 算法总结
  - 算法优势
    - 利用多出口网络（旁路神经网络Side Nets）提取样本在模型预测过程中的演化特征，解决了后门样本和正常样本在模型特征空间中不可分离导致检测方法准确率下降的问题
    - 多出口网络同时支持对后门样本的“标签修复”
  - 算法劣势
    - Side Nets训练开销大，且可解释性弱
    - 检测方法超参数过多，如不同模型架构下Side Nets的位置、数量选取困难、异常分数计算的权重分配选取困难，理论上无法证明其有效性
    - 应用场景受限，对于复杂模型难以部署Side Nets，如NLP模型

【2024-S&P】
Robust Backdoor Detection for Deep Learning via Topological Evolution Dynamics

算法原理 TED

T	目标	在受害模型推理阶段，检测后门样本并拒绝其输入模型
I	输入	受害模型（1个）、干净样本集（200个，类别平衡）、待测样本（1个）
P	处理	1. 获取待测样本在模型多个隐藏层的输出表示； 2. 基于输出表示，获取待测样本的模型演化序列； 3. 利用PCA对模型演化序列进行异常分数计算； 4. 结合阈值进行后门样本检测
O	输出	（是/否）后门样本 / 正常样本

P	问题	现有方法假设在度量空间（如模型特征空间）中正常样本和后门样本具有可分离性，无法检测自适应的强隐蔽后门攻击，导致检测准确率下降
C	条件	1. 拥有少量干净样本集（如CIFAR10中，每类样本20个，合计200个） 2. 能够访问受害模型隐藏层及其输出表示
D	难点	样本在模型前向传播过程中演化特征的精确提取
L	水平	S&P 2024（CCF A类）

算法原理 TED

- 创新分析：动态差异演化
  - 核心思想：利用后门样本和正常样本在模型前向传播中的演化差异进行建模
  - 输入：干净样本各隐藏层输出集合，待测样本各隐藏层的输出值
  - 输出：模型演化序列

获取干净样本集
各隐藏层输出表示：$X_{clean} = \{x_i\}_{i=1}^{m} \xrightarrow{f(x) = (h_1 \circ \cdots \circ h_N)(x)} S_{clean} = [h_l(X_{clean})]_{l=1}^{N} = \begin{bmatrix} h_1^{(1)} & \cdots & h_N^{(1)} \\ h_1^{(2)} & \cdots & h_N^{(2)} \\ \vdots & & \vdots \\ h_1^{(m)} & \cdots & h_N^{(m)} \end{bmatrix}$

（隐藏层输出集合）

获取待测样本各隐藏层输出表示：$s_{test} = [h_l(x_{test})]_{l=1}^{N} = [h_1^{(test)} \quad \cdots \quad h_N^{(test)}]$

计算待测样本距离矩阵：$d(s_{test}, S_{clean}) \xrightarrow{d = Euclidean\ (\cdot)} \begin{bmatrix} d_1^{test,1} & \cdots & d_N^{test,1} \\ d_1^{test,2} & \cdots & d_N^{test,2} \\ \vdots & & \vdots \\ d_1^{test,m} & \cdots & d_N^{test,m} \end{bmatrix}$

算法原理 TED

- 创新分析：动态差异演化
  - 核心思想：利用后门样本和正常样本在模型前向传播中的演化差异进行建模
  - 输入：干净样本各隐藏层输出表示，待测样本各隐藏层的输出表示
  - 输出：模型演化序列

基于每一层进行距离排序：

$$\begin{bmatrix} d_1^{test,1} & \cdots & d_N^{test,1} \\ d_1^{test,2} & \cdots & d_N^{test,2} \\ \vdots & & \vdots \\ d_1^{test,m} & \cdots & d_N^{test,m} \end{bmatrix} \xrightarrow{sort\ by\ distance} \begin{bmatrix} d_1^{test,6} & \cdots & d_N^{test,2} \\ d_1^{test,m} & \cdots & d_N^{test,m} \\ \vdots & & \vdots \\ d_1^{test,2} & \cdots & d_N^{test,1} \end{bmatrix}$$

获取预测类别的最邻近样本的排名：　　假设 $f(x_{test}) = f(x_m) \neq f(x_6)$，则 $K_1 = 2$

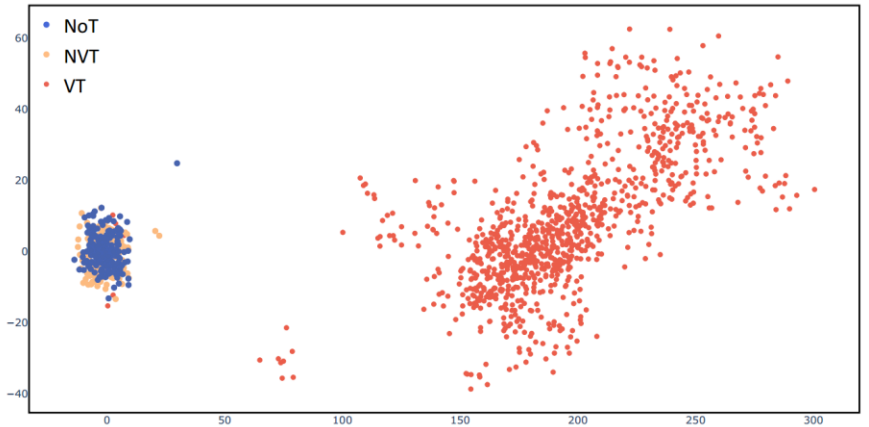
构造待测样本的模型演化序列：　$K_{test} = [K_1, K_2, \cdots K_N]$

正常样本具有更一致的演化序列，如 $[2, 1, 1, \cdots 1, 3, 2]$
后门样本具有更“颠簸”的演化序列，如 $[91, 78, 83, \cdots 2, 3, 1]$

算法原理　基于PCA的异常值检测

- 主成分分析（Principal Component Analysis，PCA）
  - 利用正交变换将一系列可能线性相关的变量转换为一组线性不相关的新变量，主要用于数据的降维处理和特征提取
    - 在处理高维数据时，PCA通过将原始的多维度变量转换为一组新的、相互独立的变量，这些新变量称为“主成分”
- 基于PCA的异常值计算
  - 输入：PCA模型、 $K_{test} = [K_1, K_2, \cdots K_N]$
    - PCA模型由干净样本各隐藏层激活向量集合得到的模型演化序列进行训练
  - 输出：异常值分数
- 结合阈值进行后门样本检测

异常值检测是PCA降维的副产品



- 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
MNIST	60000/10000	28*28*1	10	图像	手写数字识别
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

- 评价指标
  - TPR（True Positive Rate）：识别的后门样本数量占总后门样本数量的比例
  - FPR（False Positive Rate）：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
  - AUC（Area Under Curve）：用于二分类模型的评价指标，评估模型的分类能力

实验设计　图像分类模型后门攻击检测

- 后门攻击：4种攻击方法，涵盖2种先进攻击策略

类型	序号	名称	水平
特定于源类别的后门攻击	1	CompositeBA	2020-CCS（CCF-A）
	2	TaCT	2021-USENIX（CCF-A）
触发器选择和优化的后门攻击	3	IAD	2021-NeurIPS（CCF-A）
	4	SSDT	2024-S&P（CCF-A）

- 对比方法：4种主流检测方法

类型	序号	名称	水平
基于信息熵的后门攻击检测	1	STRIP	2019-ACSAC（CCF-B）
触发器重定位的后门攻击检测	2	SentiNet	2020-S&P Workshops
一阶矩特征分解的后门攻击检测	3	SCAn	2021-USENIX（CCF-A）
高阶矩特征提取的后门攻击检测	4	Beatrix	2023-NDSS（CCF-A）

实验结果　图像分类模型后门攻击检测

- 对于触发器选择和优化的后门攻击（如SSTD），每个后门样本的触发器都不一样，且触发器和后门样本间具有唯一性，不能迁移
  - VT: Victim-trigger samples
  - NVT: Non-victim but triggered samples
  - NoT: No Trigger samples
- TED的检测精度均达到100%，且对NVT的漏报率很低，证明方法在建模样本演化差异的有效性
  - Beatrix利用高阶矩放大触发器和良性特征间统计差异的方法无法应对NVT

TABLE 3: FPRs and TPRs (%) of different detectors on MNIST, CIFAR-10, and GTSRB under SSDT.

		TED	Beatrix	SCAn	STRIP	SentiNet
			MNIST			
TPR	VT	100.00	82.50	44.00	0.50	2.75
FPR	NVT	1.30	4.27	5.00	4.00	4.50
	NoT	5.00	4.55	5.00	5.50	4.50
			CIFAR-10			
TPR	VT	100.00	90.00	36.50	0.00	6.75
FPR	NVT	4.00	46.65	5.00	7.00	5.00
	NoT	5.50	4.15	5.00	6.00	4.50
			GTSRB			
TPR	VT	100.00	100.00	99.50	0.50	5.25
FPR	NVT	0.90	62.55	5.00	4.00	5.00
	NoT	4.59	4.10	5.00	5.50	5.50

单纯的基于样本的统计特征方法已不再适用于先进的后门攻击

实验设计　文本分类模型后门攻击检测

- TED在NLP文本分类模型中的后门攻击检测
  - NLP模型（如Bert）不再是具有清晰直观、单向传播性质的前馈网络
  - TED关注Transformer架构的词嵌入层、 自注意力层、密集层和的激活值
  - 不再使用预测类别的最邻近样本的排名作为模型演化序列的组成元素，而是使用预测类别K-nearest邻近样本的排名
- 数据来源：Twitter数据集
- 后门攻击：2种攻击方法　　对比方法：2种检测方法

序号	攻击方法	水平
1	EP	2021-ACL（CCF-A）
2	DFEP	2021-ACL（CCF-A）

序号	检测方法	水平
1	STRIP	2019-ACSAC（CCF-B）
2	DAN	2022-EMNLP（CCF-B）

实验结果　文本分类模型后门攻击检测

- 认可半径（Radius r）
  - 对K-nearest邻近样本中的K进行标准化定义

$$r = \frac{k}{c \cdot m} * 100\%$$

    - c表示文本分类的类别，m表示每个类别正常样本的数量
- 在不同的r下，TED对2种后门攻击均有良好的检测效果
- 保证TPR持平时，TED的误检率相比于STRIP和DAN有明显下降

TED方法适用于检测NLP文本分类后门攻击

TABLE 9: Performance of TED on toxicity detection over Twitter data with various radius r.

Attack	Metric	Radius r				AVG
		0.50%	1.00%	1.50%	2.00%	
EP	TPR (%)	92.50	92.00	92.00	92.00	92.12
	FPR (%)	20.50	19.00	21.50	23.00	21.00
	AUC	.9309	.9267	.9296	.9280	.9288
DFEP	TPR (%)	95.00	95.00	94.00	94.00	94.50
	FPR (%)	21.00	18.00	19.00	18.00	19.00
	AUC	.9565	.9528	.9545	.9527	.9541

TABLE 10: Performance comparison between TED, DAN, and STRIP for toxicity detection.

Attack	Metric	TED	DAN	STRIP
EP	TPR (%)	92.12	93.42	94.99
	FPR (%)	21.00	30.09	66.11
DFEP	TPR (%)	94.50	93.54	94.99
	FPR (%)	19.00	21.39	48.89

算法总结 TED

- 算法总结
  - 算法优势
    - 利用受害模型的隐藏层输出表示进行数学计算和数值分析实现检测，无需额外添加神经网络，避免过多的训练开销
    - 利用模型演化序列建模样本在模型前向传递过程中的演化差异，具有良好的可解释性，且能支持不同模型架构（CNN、Transformer）以及适用于不同应用场景（如图像、文本）
  - 算法劣势
    - 仅基于最邻近预测类别样本的排名作为判别指标，粒度太粗
    - 要求用户或检测人员能够访问模型的隐藏层并获取其输出，阻碍了检测方法的广泛应用（如闭源、商业化模型）

未来展望

未来展望

- 未来展望
  - 设计无需访问训练集、无需了解模型内部结构的轻量化后门攻击检测方法
  - 多领域、多模态、强泛化性的后门攻击检测
    - 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
  - 领域融合和迁移
    - 知识产权保护（利用后门知识提升模型水印的鲁棒性）
    - 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

反者道之动 弱者道之用

参考文献

[1] Huang H, Wang Q, Gong X, et al. Orion: Online backdoor sample detection via evolution deviance[C]. International Joint Conference on Artificial Intelligence. 2023.

[2] Mo X, Zhang Y, Zhang L Y, et al. Robust Backdoor Detection for Deep Learning via Topological Evolution Dynamics[C]. 2024 IEEE Symposium on Security and Privacy (SP). IEEE Computer Society, 2024: 171-171.

[3] Hu Y, Kuang W, Qin Z, et al. Artificial intelligence security: Threats and countermeasures[J]. ACM Computing Surveys (CSUR), 2021, 55(1): 1-36.

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！

附录A Orion补充实验

- Orion中6种攻击方法的原始性能

Dataset	Attacks	Trigger size	Poisoning rate	Target label	ASR	PA
CIFAR-10	BadNets	5 x 5	0.05	1	97.94%	91.34%
	Blended	5 x 5	0.05	1	85.88%	90.89%
	WaNet	global	0.10	1	97.45%	90.78%
	IAD	global	0.10	3	99.39%	93.58%
	TaCT	5 x 5	0.05	1	94.40%	92.22%
	Feature	global	0.10	0	99.69%	86.26%
GTSRB	BadNets	5 x 5	0.05	1	96.65%	97.97%
	Blended	5 x 5	0.10	1	94.11%	96.57%
	WaNet	global	0.10	0	85.66%	94.95%
	IAD	global	0.10	1	99.71%	98.51%
	TaCT	5 x 5	0.10	1	100.0%	98.38%
	Feature	global	0.10	0	99.80%	95.68%
Tiny-Imagenet	BadNets	6 x 6	0.05	1	96.65%	47.44%
	Blended	6 x 6	0.10	1	85.57%	46.77%
	WaNet	global	0.10	1	89.55%	44.97%

Table 4: Detailed attack settings. ASR means the attack success rate, which is the proportion of samples with triggers that are classified as target labels. PA represents the prediction accuracy of clean data.

附录B Orion补充实验

- **Orion干净样本集参数实验**

Dataset	Attacks	0.1%			0.5%			1%			5%			10%		
		PRE	REC	F1	PRE	REC	F1	PRE	REC	F1	PRE	REC	F1	PRE	REC	F1
CIFAR-10	BadNets	85.64	33.40	48.06	95.31	93.40	94.34	99.50	97.80	98.60	98.60	98.90	98.70	99.20	99.39	99.29
	Blended	95.48	97.20	96.33	95.14	98.00	96.55	98.90	97.60	98.20	98.79	98.20	98.49	98.60	99.80	99.20
	WaNet	95.16	98.40	96.75	95.00	98.80	96.86	99.79	97.80	98.70	98.79	98.80	98.89	98.80	99.20	99.00
	IAD	94.33	93.20	93.76	97.20	97.40	97.30	98.00	99.80	98.90	98.80	99.00	98.90	99.60	99.60	99.60
	TaCT	93.96	96.60	95.26	97.20	94.92	96.04	99.79	97.00	98.37	98.60	96.30	98.60	99.00	99.00	99.00
	Feature	94.12	99.40	96.69	94.44	98.60	96.47	93.70	98.60	96.10	96.59	96.40	96.49	95.69	97.80	96.73
GTSRB	BadNets	73.38	97.60	83.77	95.62	91.80	93.67	97.04	91.80	94.34	95.60	100.0	97.75	96.33	99.80	98.04
	Blended	75.29	88.40	81.32	82.58	92.00	87.03	94.38	87.40	90.76	94.50	99.80	97.08	96.14	99.80	97.93
	WaNet	74.00	96.20	83.65	84.33	98.00	90.65	95.50	99.80	97.60	95.76	99.40	97.54	96.70	99.80	98.22
	IAD	75.41	98.80	85.54	83.83	99.60	91.04	95.00	91.60	93.37	95.41	100.0	97.65	95.41	99.80	97.55
	TaCT	71.68	87.60	78.84	82.71	89.00	85.74	96.09	83.60	89.41	94.84	99.40	97.07	94.16	100.0	96.99
	Feature	73.22	96.80	83.37	80.45	98.80	88.68	95.33	85.80	90.32	95.37	99.00	97.15	97.65	100.0	98.81
Tiny-Imagenet	BadNets	66.33	92.20	77.15	82.55	95.60	88.60	95.55	93.40	94.44	94.48	99.40	96.88	95.33	98.20	96.74
	Blended	63.18	85.80	72.77	78.49	87.60	82.79	91.02	77.00	83.42	95.69	89.00	92.22	96.25	92.40	94.28
	WaNet	66.34	82.00	73.34	76.44	84.40	80.22	88.49	75.40	81.42	95.49	84.80	89.83	95.99	91.00	93.42

Table 5: Impact of different clean data size.