

人工智能后门防御评估方法及其架构研究

谢天¹, 李强², 鞠卓亚², 韩嘉祺², 易平¹

(1. 上海交通大学网络空间安全学院, 上海 200240;

2. 32178部队科技创新中心, 北京, 100012)

摘要: 为了应对人工智能系统可能面临的后门攻击风险, 研究人员已经开发了一系列后门防御策略。现有防御方法评估标准的多样性, 使得跨方法比较成为一大挑战, 因此提出了一种人工智能后门防御统一评估框架。该框架旨在为不同层面(包括数据集级别和模型级别等)的防御策略, 提供一个公共的评价标准。在数据集级别, 主要通过准确率来评估后门检测的有效性; 而在模型级别, 则主要关注攻击成功率等指标。人工智能后门防御统一评估框架, 能够在相同的评价标准下, 对比和分析不同后门防御方法的性能。这不仅有助于识别各方法的优势和不足, 还能够提出针对性改进建议。结果表明, 人工智能后门防御统一评估框架能有效地评估不同防御策略的性能, 为进一步提升人工智能系统的安全性提供重要的参考依据。

关键词: 人工智能安全; 后门攻击; 后门防御; 统一评估

中图分类号: TP18, TP309

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202430

Research on method and architecture for defense assessment of artificial intelligence backdoors

XIE Tian¹, LI Qiang², JU Zhuoya², HAN Jiaqi², YI Ping¹

1. School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

2. Scientific and Technological Innovation Center, Unit 32178 of PLA, Beijing 100012, China

Abstract: In response to the potential risk of backdoor attacks faced by artificial intelligence systems, a range of backdoor defense strategies are developed. The diversity of the evaluation criteria for existing defense method, makes cross-method comparisons a significant challenge. Hence, a unified evaluation framework base on artificial intelligence backdoors was proposed. This framework aimed to provide a common standard for evaluating different levels of defense strategies, including dataset-level and model-level defenses. Regarding the dataset-level defense strategies, the effectiveness of backdoor detection was primarily assessed through accuracy. Regarding the model-level defense strategies, focus was mainly placed on metrics such as attack success rate. By implementing unified evaluation framework, the performance of various backdoor defense methods under the same standards were compared and analyzed. This not only aids in identifying the strengths and weaknesses of each method, but also proposes targeted suggestions for improvements. The results indicate that unified evaluation framework can effectively measure the performance of different defense strategies, providing an important reference for further enhancing the security of artificial intelligence systems.

Key words: artificial intelligence security, backdoor attack, backdoor defense, unified evaluation

收稿日期: 2024-04-18; 修回日期: 2024-08-19

通信作者: 易平, yiping@sjtu.edu.cn

基金项目: 国家自然科学基金项目 (No.62202290)

Foundation Item: The National Natural Science Foundation of China(No.62202290)

0 引言

近年来,随着机器学习理论的发展和深度神经网络的提出,人工智能技术在人类生活中的应用越来越多,比如人脸识别^[1]、自动语音识别^[2]、机器翻译^[3]、视频生成^[4]等。为了应对更加复杂的任务,人工智能的模型参数量越来越大,比如著名的GPT-3模型的参数量达到了1 750亿,一些较小的开源大模型的参数量也达到十亿数量级。庞大的模型展现出更出色的性能,但是其所需的训练设备(通常是图像处理单元)非常昂贵,而绝大多数公司和个人无法承担如此高昂的设备成本,不得不租用云服务商的训练设备。

一个人工智能模型的诞生主要包括4个流程:数据集构造、模型训练、测试与模型部署。利用第三方云服务商提供的服务可以降低训练的成本,但是将训练过程交给不可信的第三方平台,也带来了一定的安全威胁。对于个人来说,第三方云服务商的训练资源或许有些昂贵,可以直接在第三方的预训练模型上进行微调,比如Meta公司的开源大模型Llama2^[5]。这进一步降低了训练成本,但是同样存在安全风险,人工智能后门攻击就是一种潜在的威胁。

后门攻击是指通过在系统或软件中植入的隐藏访问通道,攻击者可以绕过正常的安全控制,未经授权即可访问系统或软件,执行恶意操作。而人工智能后门,一般指一个模型带有后门。该模型在面对正常数据时的行为表现完全正常,但是在遇到带有后门触发器的数据时会输出错误结果。文献[6]中介绍了后门的植入过程和工作原理,人工智能后门攻击示例如图1所示。

对于MNIST数据集,攻击者在部分数据上加入特制的扰动并将其标签设置为特定值0,然后把恶意数据加入训练,训练出的模型在遇到正常数据时输出正常,在遇到带有特制扰动的数据时则输出为0,这就是一个后门模型。训练者暴露在第三方训练平台的公开环境中,存在被攻击者操控训练集和训练过程的可能性以及被后门攻击的风险。人工智能模型对于人类来说基本上是一个黑盒模型,因此人工智能后门比普通的计算机后门更加具有隐蔽性和威胁性。

为了防范上述的后门攻击,业界提出了多种防御方法^[7]。防御方法按手段可以分为识别有毒数

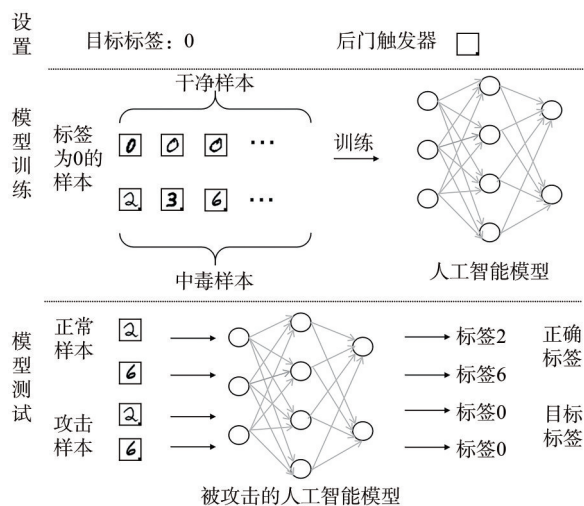


图1 人工智能后门攻击示例^[1]

据、识别中毒模型、过滤攻击数据、移除后门神经元等;按照防御级别可以分为输入级别、数据集级别、模型级别等。不同类别的防御方法有自己的评价体系,对于防御方法的使用者来说,很难选择合理的评价标准。本文针对防御方法的不同防御级别,提出了统一的防御评估方法,而对于不同防御级别的防御方法,也提出了一种直接比较的方法,帮助使用者直观地比较不同防御方法之间的优劣。

1 相关工作

1.1 后门攻击

2017年,Gu等^[8]提出的恶意神经网络(bad networks, BadNets),被认为是人工智能后门攻击的开端。BadNets在图片右下角简单地添加几个纯白色像素点,即可生成后门图像,然后训练生成后门模型。这是一个简单而又有效的攻击方法。虽然文献[8]中BadNets只在图像数据集上实验,但是这一思想也可以应用到其他模态,通过简单的修改,BadNets也可以对文本、语音、视频相关的人工智能模型进行攻击。BadNets的攻击隐蔽性欠佳,人眼可察觉。文献[9]利用人眼对颜色深度的不敏感特性,针对性修改了图片的像素值,从而构造出了人眼不可见的后门,因此其具有更强的隐蔽性。上述两种后门都需要把图片的标签更换成攻击者设定的标签,这会大大降低原方法对无监督学习的攻击效果,文献[10]首次提出了标签一致性攻击方法,实现只对图像修改而无需改动标签的后门攻击。

目前,大部分的后门攻击聚集在图像方面,也有对其他领域的后门攻击。AddSent^[11]借鉴了Bad-

Nets的思想, 选择一个句子作为后门触发器, 并使用随机插入策略生成污染样本。StyleBkd^[12]则首次尝试使用风格转换的方式修改文本来实现后门攻击, 旨在改变句子风格的同时保留其含义, 以提高后门的隐蔽性。UltraSonic^[13]通过在声音的频谱中加入超声波段, 从而实现了声音后门的植入。TUAP^[14]对视频抽帧, 仅在部分帧上加入后门, 实现了视频上的后门攻击。Graph backdoor^[15]通过特定的拓扑映射在图数据中注入触发器, 生成特定的带有后门的图神经网络。从BadNets发展到现在, 后门攻击的研究遍布人工智能的各个领域。

1.2 后门防御

后门攻击通过对数据集投毒和正常的训练过程得到后门模型。后门防御按照防御作用的对象可以分为两大类: 数据集级别的防御和模型级别的防御。数据集级别的防御的主要目的是从后门数据集移除中毒数据得到干净数据集, 模型级别的防御的主要目的是从模型中移除后门特征得到干净模型。

激活集群(activation clustering, AC)^[16]是一种基于后门数据的检测型防御方法, 文献[16]观察到带有触发器的后门样本和正常样本在神经网络中的激活值存在较大差异, 利用适当的方法可以将中毒样本和干净样本聚类到不同的样本中, 实现干净数据和中毒数据的分离。强烈有意扰动(strong intentional perturbation, STRIP)^[17]也是一种基于后门数据的检测型防御方法, 由于大部分触发器都具有输入独立的特点, 该方法故意向每个输入注入强烈的扰动, 利用注入扰动后数据的混乱程度的差异来区分干净数据和后门数据, 只适用于输入独立性的攻击方法。反后门学习(anti-backdoor learning, ABL)^[18]可以在受污染的数据上直接训练出干净模型, 该方法观察到后门学习更快且对目标有极强的依赖性。ABL在训练早期隔离损失值较低的后门数据, 在随后的训练过程中通过梯度上升技术抑制后门特征的学习, 使模型无法学习到后门特征, 从而得到干净的人工智能模型。Liu等^[19]提出了一种基于精细剪枝(Fine-Pruning)和模型微调(Fine-Tuning)的防御方法, 发现后门特征只会被极少部分的神经元学习到, 根据特定的判断条件剪切这些神经元, 即可移除模型后门, 再在原训练数据集对剪切后的模型进行微调, 即可得到干净的模型。这里提出的两种防御思想, 精细剪枝和微调, 也成为大部分模型级别后门防御方法的基础。

通道利普希茨裁剪(channel Lipschitzness pruning, CLP)^[20]利用通道的利普希茨常数的类高斯分布特征发现异常神经元, 然后进行精细剪枝, 仅通过修改少量的神经元即可得到干净模型。随着后门攻击的发展, 后门防御方法的种类也更加多样。

1.3 人工智能后门防御统一评估框架

本文提出的框架将各种复杂的后门攻击和防御方法进行了规范化处理, 为后人研究人工智能后门提供了便利。TrojanZoo^[21]是一个较早期的后门攻击防御框架, 该框架集成了多种攻击防御方法, 但是框架格式化程度不高, 不易使用。Backdoor-bench^[22]是第一个较为完备的后门攻击防御框架, 该框架易于使用, 但局限于计算机视觉领域。OpenBackdoor^[23]是第一个完善的自然语言处理领域的后门攻击防御库。人工智能后门防御统一评估框架的出现, 为后门的研究节约了大量的时间成本。

多元化的后门攻击防御方法除了需要统一框架, 还需要有统一的评估方法。文献[24]认为评价后门攻击的通用标准包括针对性、隐蔽性和现实性, 对于现有的后门攻击评估, 大部分用攻击成功率和模型识别准确率来代表针对性和隐蔽性, 而现实性往往由后门的可视化程度来判断。后门防御对应的评估包括降低攻击成功率和提高模型识别准确率。由于防御方法在过程和级别上的多元化, 单一的评价指标无法完全衡量一个防御方法, 也很难给出不同的防御方法的对比, 因此人工智能后门防御方法的统一评估框架具有一定的研究价值。

2 后门防御评估方法设计

2.1 人工智能后门防御统一评估框架设计

2.1.1 总体架构

为了便捷地进行人工智能后门防御统一评估, 尤其是实现多模态防御的评估, 首先需要设计一个集成多种后门攻击防御方法的统一评估框架, 本文借鉴了已有的后门防御评估框架的优点设计了一个全新的多模态人工智能后门防御统一评估框架。人工智能模型训练包括数据加载、模型加载、模型训练、评估保存4个部分。已有的评估框架以攻击防御算法为核心, 每个攻击和防御算法重复实现这4个部分, 造成了不必要的研发开销。本文提出的架构则以模型训练的4个部分为核心, 攻击和防御只是作为第三方参与模型训练, 这也更符合后门攻

击与防御的现实应用场景,人工智能后门防御统一评估框架使用流程如图2所示。后门攻击主要是对数据投毒得到后门数据集,因此只在数据加载阶段被调用,多模态的数据按照对应的需求进行数据加载和投毒操作。而对于主体流程而言,数据处理和投毒是一个黑盒,因此可以无感地添加各自模态的数据和攻击方法。在模型层面,模型通过训练参数的选择按需加载,与具体的模态无关。在训练层面,如果是攻击,则调用普通的模型训练方法,如果是防御,则调用对应的防御方法。防御方法相对于主体流程而言也是个黑盒,因此可以无感地添加各自模态的防御训练方法,从而实现多模态扩展。

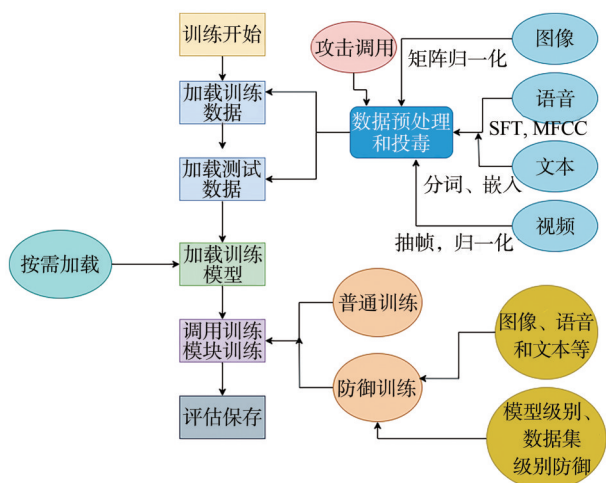


图2 人工智能后门防御统一评估框架使用流程

2.1.2 攻击方法实现

后门攻击方法往往包括如下步骤：模型加载、数据集加载、模型训练和评估保存。其中，训练所用的模型来自框架的加载选择，数据集来自数据预处理模块和数据投毒模块，模型的训练会调用后门训练模块，所有的后门攻击调用相同的训练流程，评估保存则由评估模块实现。

为了更加全面地评估防御方法，本文使用了6种攻击方法，分别是BadNets^[8]、逐像素攻击（bits per pixel attack, BPP）^[9]、低频率（Low Frequency）攻击^[25]、信号触发攻击（signal attack, SIG）^[26]、样本特定后门攻击（sample-specific backdoor attack, SSBA）^[27]和基于扭曲的后门攻击网络（warping-based backdoor attack network, WaNet）^[28]。BadNets是最早提出的后门攻击方法，该方法的后门触发器是在图片的右下角添加纯白色的像素块。BPP利用人眼对图像某些区域色块变化的不敏感性进行

对应的图像扰动。Low Frequency攻击从频域视角重新审视后门攻击的触发器，发现现有的后门攻击存在严重的高频伪影，因此设计了一种新的优化函数利用离散余弦变换将图像转换为频域后添加平滑的后门触发器。文献[26]认为现有的攻击需要改变图像标签是非常不合理的要求，因此SIG通过向特定的图像样本中添加后门信号而无需改变图片标签即可实现后门注入。文献[27]提出了一种生成肉眼不可见的后门触发器，SSBA方法使用了一个图像编码器-图像解码器结构将攻击者指定的字符串编码到良性图像中，生成了独立于样本的特定、不可见的触发器。文献[28]提出了一种基于图像扭曲的后门攻击方法WaNet，使用不可察觉的扭曲触发器，使得攻击更具隐蔽性。上述攻击方法具有一定的代表性，因此被选为待测试的攻击方法。上述攻击方法的特点如表1所示。

表1 攻击方法特点

攻击	是否可见	触发位置	输入独立
BadNets ^[8]	是	局部	否
BPP ^[9]	否	全局	是
Low Frequency 攻击 ^[25]	否	全局	是
SIG ^[26]	是	全局	否
SSBA ^[27]	否	全局	是
WaNet ^[28]	否	全局	是

2.1.3 防御方法实现

本文实现的后门防御分为两个大类，即数据集级别防御和模型级别防御，包括2种后门数据检测型方法（AC、STRIP）和3种模型修复型防御方法（模型微调、ABL和精细剪枝）。上述5种防御方法能代表目前大部分后门防御方法的防御思想。

2.2 后门防御评估指标设计

传统的后门防御评估大多数只用攻击成功率指标去评估防御方法的优劣，如果是数据集级别防御方法，则在防御处理得到的数据集上对模型进行微调，再使用攻击成功率对模型进行评估。传统的后门防御评估有3个主要问题：要遵循成本-效益均衡原则，单纯追求攻击成功率并不能达到最高的经济效益；其次，它往往是一种定性的评估，缺乏数据上的定量计算；最后，其对于数据集级别的防御方法往往是不公平的，因为传统的后门防御评估倾向于给微调后的后门防御较低的分值。从普遍意义上来说，在安全领域要评价一个防御方法的优劣有

多个评估维度，比如防御的有效性、防御的成本、防御实施的难度、防御方法的普适性和防御方法的负面影响等。已有的后门防御研究中也使用过这些评估维度，如CLP表明优越性时提到其耗时较短且无需使用图像处理单元，这就是对防御的成本的评估；有些防御方法通过多模态防御来证明其防御方法的优越性，这是一种对防御方法的普适性的评估。根据防御方法使用的评估维度和其他人工智能研究领域使用过的防御指标，本文提出了评估人工智能模型后门防御方法的4个统一指标：有效性、无损性、通用性和经济性。有效性体现了防御方法的直接结果，无损性体现了防御方法对模型的负面影响，通用性体现了防御方法对多场景的适应能力，经济性体现了防御方法的实施成本。数据集级别和模型级别的防御方法差异较大，因此采用不同的方式计算数据集级别和模型级别的防御方法的评分。

(1) 有效性

有效性是评估人工智能后门防御最重要的一项指标，防御的目的是降低后门攻击的效果。对于模型级别的后门防御，其防御目的是得到移除后门的模型，因此以攻击成功率的降低来表示有效性得分，即：

$$S_1 = \begin{cases} \frac{ASR_{old} - ASR_{new}}{ASR_{old}} \times 100, & ASR_{new} < ASR_{old} \\ 0, & ASR_{new} \geq ASR_{old} \end{cases} \quad (1)$$

其中， ASR_{old} 代表防御实施之前的攻击成功率， ASR_{new} 代表防御实施后的攻击成功率。对于数据集级别的防御方法，其防御的直接结果是得到干净数据集，虽然可以通过微调来得到干净模型，但是这并非数据集级别防御的直接目的，因此本文采用分类准确率相关的计算式评估数据集级别的防御方法的有效性得分，即：

$$S_1 = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (2)$$

其中：TP为真正例，表示模型正确预测为正类的样本数；TN为真负例，表示模型正确预测为负类的样本数；FP为假正例，表示模型错误预测为正类的样本数；FN为假负例，表示模型错误预测为负类的样本数。

(2) 无损性

大多数人工智能后门防御方法虽然会在一定程度上降低原模型的性能或将干净样本误分类为中毒

样本，但是防御方法在提高有效性的同时，应该尽量减少对原模型或数据集的损害。对于模型级别的防御方法，无损性的计算式为：

$$S_2 = \begin{cases} \frac{ACC_{new}}{ACC_{old}} \times 100, & ACC_{new} < ACC_{old} \\ 100, & ACC_{new} \geq ACC_{old} \end{cases} \quad (3)$$

其中， ACC_{old} 代表防御实施之前旧的模型识别准确率， ACC_{new} 代表防御实施后新的模型识别准确率。而对于数据集级别的防御方法，无损性的计算式为：

$$S_2 = \frac{TP}{TP + FN} \times 100 \quad (4)$$

(3) 通用性

通用性用来评估防御方法在多个防御场景下的防御能力，本文定义只有当有效性得分大于50分且无损性得分大于40分时认为防御在该场景下合格。对于数据集级别和模型级别的防御方法，通用性得分的计算式为：

$$S_3 = \frac{N_{pass}}{N_{total}} \times 100 \quad (5)$$

其中， N_{pass} 代表有效性得分大于50分且无损性得分大于40分的防御场景数量， N_{total} 代表总的防御场景数量。

(4) 经济性

经济性用来评估不同方法所需的训练成本。本文主要通过训练耗时的倒数归一化来评估，即得分计算式为：

$$S_4 = t \sum \frac{1}{t_i} \times 100 \quad (6)$$

其中， t 代表待评估防御方法的训练总耗时，而 t_i 代表基本防御方法的训练耗时。数据集级别的防御方法和模型级别的防御方法通过上述指标计算得到各项指标得分后，按照一定的比例进行加权即可得到最终评估得分。考虑到后门防御的主要目的是降低后门攻击的成功率，即本文提出的有效性指标，并且目前所有的防御方法均使用了有效性作为评估指标，本文赋予有效性0.5的权值。无损性和通用性虽然作为评估指标使用不多，但是仍有部分防御方法使用过，因此各赋予0.2的权值。经济性作为评估指标使用较少，并且现有的后门防御研究的重点并不在成本控制上，因此赋予经济性0.1的权值，所以最终得分的计算式为：

$$S = 0.5 \times S_1 + 0.2 \times S_2 + 0.2 \times S_3 + 0.1 \times S_4 \quad (7)$$

2.3 后门防御评估流程设计

为了实现全面彻底的防御评估,基于本文提出的人工智能后门防御统一评估框架,提出了一种全新的后门防御评估流程。本文提出的后门防御评估流程如图3所示。首先,生成大量的被攻击数据以待防御评估使用,被攻击数据包括各种各样的数据集以及模型,其次,使用防御方法进行防御,并依照图4(其中,TIME表示防御所需时间)所示的后门防御单项评估流程选择对应的评分计算式,按照有效性、无损性、通用性和经济性4个指标计算对应的得分,最后进行系数加权计算出防御方法的总评估得分。

本文提出的后门防御评估方法通过4个评估指标和统一的评估流程,实现了数据集级别的防御方法和模型级别防御方法的统一比较,同时给出了直观的评估分数。

2.4 人工智能后门防御统一评估框架优缺点分析

相比于传统的单一评估框架,本文提出的人工智能后门防御统一评估框架支持多种模态的数据,不仅评估流程标准化,还引入了多样化的评估指标,能够更加全面、准确地评估后门攻击防御方法。然而,本文提出的人工智能后门防御统一评估框架需要在多种数据集和模型上进行测试才能得出最终结果,因此评估速度相较于传统的单一评估框架略有降低。评估框架的优缺点如表2所示。

3 实验分析

3.1 实验设置

本文在CIFAR-10^[29]和GTSRB^[30]数据集上进行实验。在图像领域,CIFAR-10是常用的数据集之一,该数据集包含10个类别的60 000张32×32的彩色图片,每个类别有6 000张图片,常用于图像

识别、图像分割、目标检测等领域,当然也是后门攻击防御方法常用的数据集。GTSRB是一个交通标志识别基准数据集,用于交通标志的自动检测和识别,包含了超过50 000张30×30像素的交通标志图片,共43个类别,该数据集也是后门攻击防御方法常用的数据集。本实验使用了两种分类模型,分别是ResNet-18^[31]和VGG-16^[32]。ResNet-18是2015年提出来的图像识别分类算法,如今依然作为图像识别的基准模型使用,VGG-16也是一种常用的图像识别分类算法,其模型结构比ResNet-18更加复杂,识别效果略优于ResNet-18。上述2种模型和2个数据集构成了4种攻击场景。

3.2 攻击实验分析

在进行防御评估实验之前,首先需要进行一系列的实验,并将攻击实验作为防御实验的预实验进行比较。各个攻击实验的结果如图5所示。

从图5可以看出,6种攻击方法对GTSRB的识别准确率普遍高于CIFAR-10,这是因为CIFAR-10更加复杂,识别难度更大;而作用在VGG-16上攻击的识别准确率又普遍高于ResNet-18,这符合实验预期。图5中6种攻击方法的攻击成功率有细微的差别,SIG、SSBA和WaNet的攻击成功率普遍高于BadNets、BPP、Low Frequency攻击,这与后门触发器的复杂度有关,也与触发器的可见性有一定关联。值得注意的是,WaNet在VGG-16下对GTSRB并不起作用,因为该攻击方法显著降低了模型的识别准确率,这是该攻击方法的一个缺陷。为了更加全面地评估防御方法,我们对在数据集GTSRB和VGG-16模型下的WaNet攻击场景进行了防御实验,但是所有的防御实验都不能提高该类攻击场景的识别准确率,因此后续的防御实验不会对这组攻击场景做评估。

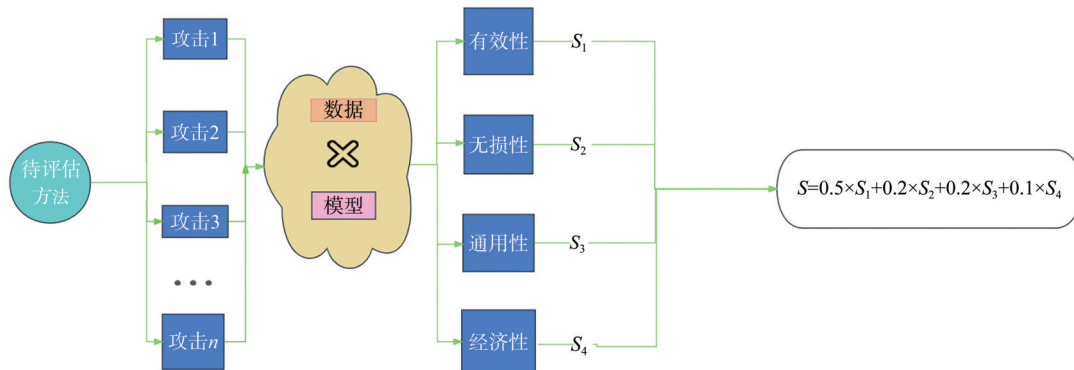


图3 本文提出的后门防御评估流程

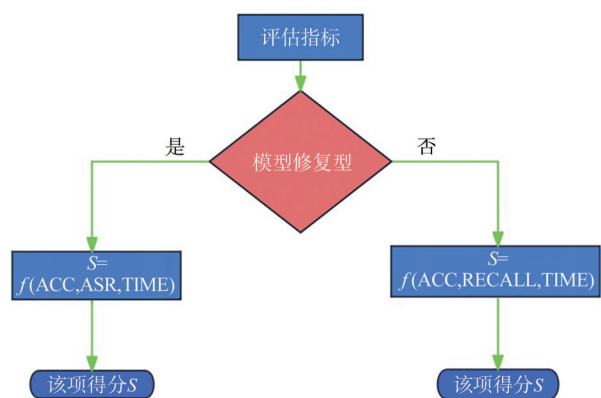


图4 后门防御单项评估流程

表2 评估框架的优缺点

分析维度	本文提出的评估框架	传统的单一评估框架
模态支持	多模态	单模态
评估流程	标准化流程	流程复杂、繁多
评估场景	各种应用场景	单一应用场景
评估指标	多指标评估	单一指标
评估可信度	可信度高	只适用于单一场景,可信度低
评估耗时	耗时较长	耗时较短

3.3 防御实验定性分析

本文对防御实验进行了定性分析，利用柱状图和散点图对防御实验进行了直观的可视化展示。攻击和防御相关指标的散点图如图6所示。

图6展示了攻击和5种防御方法的性能散点图。图6(a)~图6(f)是针对CIFAR-10的结果，图6(g)~图6(i)是针对GTSRB的结果。对于攻击和模型

修复型的防御方法来说，横坐标是攻击成功率(ASR)，纵坐标是干净数据集识别准确率(ACC)，在图像右上角的散点代表识别准确率高并且攻击成功率高，代表一次成功的攻击，在图像左上角代表识别准确率高但攻击成功率低，代表一次成功的防御。

从图6可以看到，无论是攻击还是防御，VGG-16的散点普遍在ResNet-18散点的上方，这是因为VGG-16的模型更加复杂。从图6(e)、图6(k)可以看出，对于ResNet-18，识别准确率下降值略多，但是攻击成功率下降明显，认为防御有效；对于VGG-16，攻击成功率普遍较高，即认为防御无效。Fine-Pruning没有普遍规律，针对某些攻击可以成功防御，而针对另外一些攻击则无法成功防御。

图6(b)、图6(c)、图6(h)和图6(i)的数据是数据检测型的后门防御，即检测出后门数据并从数据集中剔除这些后门数据。图6(b)、图6(c)、图6(h)和图6(i)的横坐标是召回率(RECALL)，纵坐标是识别准确率，即方法正确预测的能力。一个良好的模型检测方法，应该同时具有较高的召回率和识别准确率，考虑到后门数据的健壮性，召回率要比识别准确率更加重要。

为了更加直观地比较模型修复型方法的性能，把识别准确率和攻击成功率单独列出来进行对比，如图7和图8所示。所有的防御方法在ResNet-18

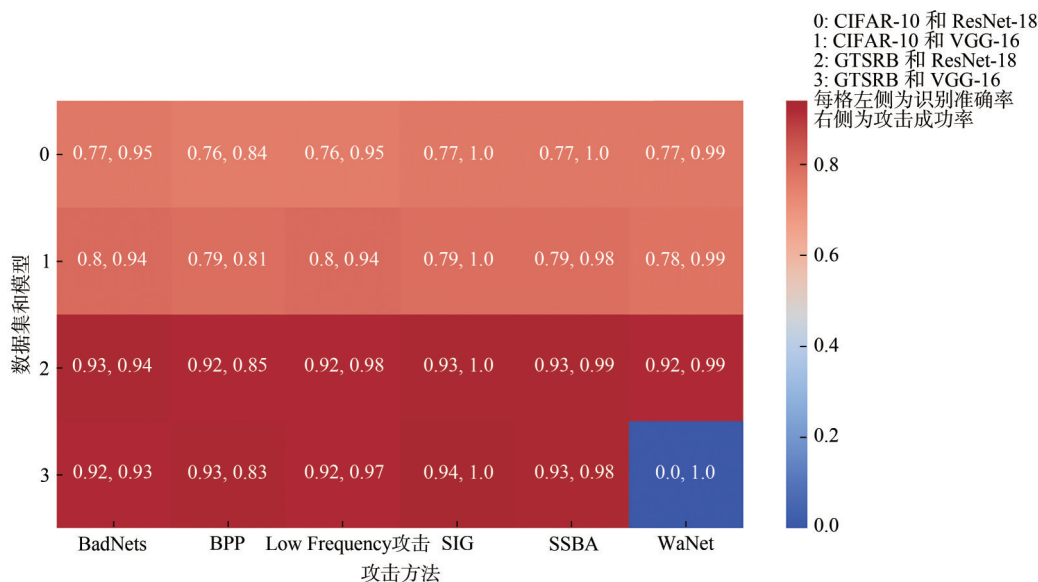


图5 攻击实验的结果

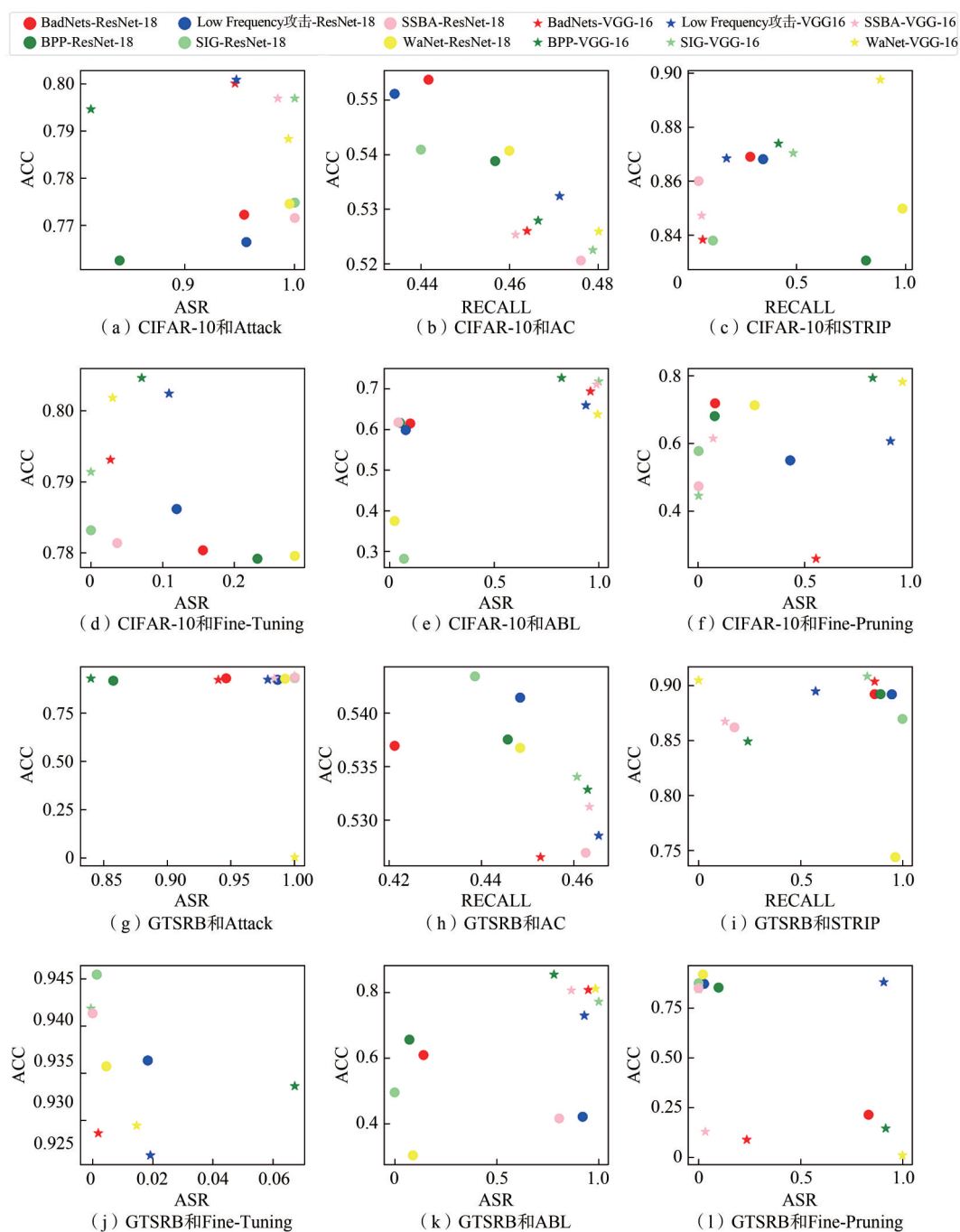


图6 攻击和防御相关指标的散点图

下对 CIFAR-10 都有较高的识别准确率, 但是 ABL 的识别准确率略低于其他的防御方法。对于 CIFAR-10 和 VGG-16 的组合, ABL 和 Fine-Pruning 的识别准确率较低; 对于 GTSRB 和 ResNet-18 的组合, ABL 的识别准确率普遍低于其他 3 种方法, 而对于 GTSRB 和 VGG-16 的组合, 又是 Fine-Pruning 的识别准确率较低。对于攻击成功率, 所有的防御方法对于 CIFAR-10 和 ResNet-18 的组合都有防御效果,

而 Fine-Tuning 在所有的情况下的攻击成功率都显著降低, 而 ABL 和 Fine-Pruning 则对不同的数据集和模型情况有不同的防御效果。

以上分析可知, 相同的防御方法在面对不同的数据集或者模型时有着不同的防御结果, 此外, 不同的防御方法消耗的资源和时间也是不同的, 因此单纯定性的分析并不能客观全面地评价后门防御方法, 我们需要使用统一的评估方法。

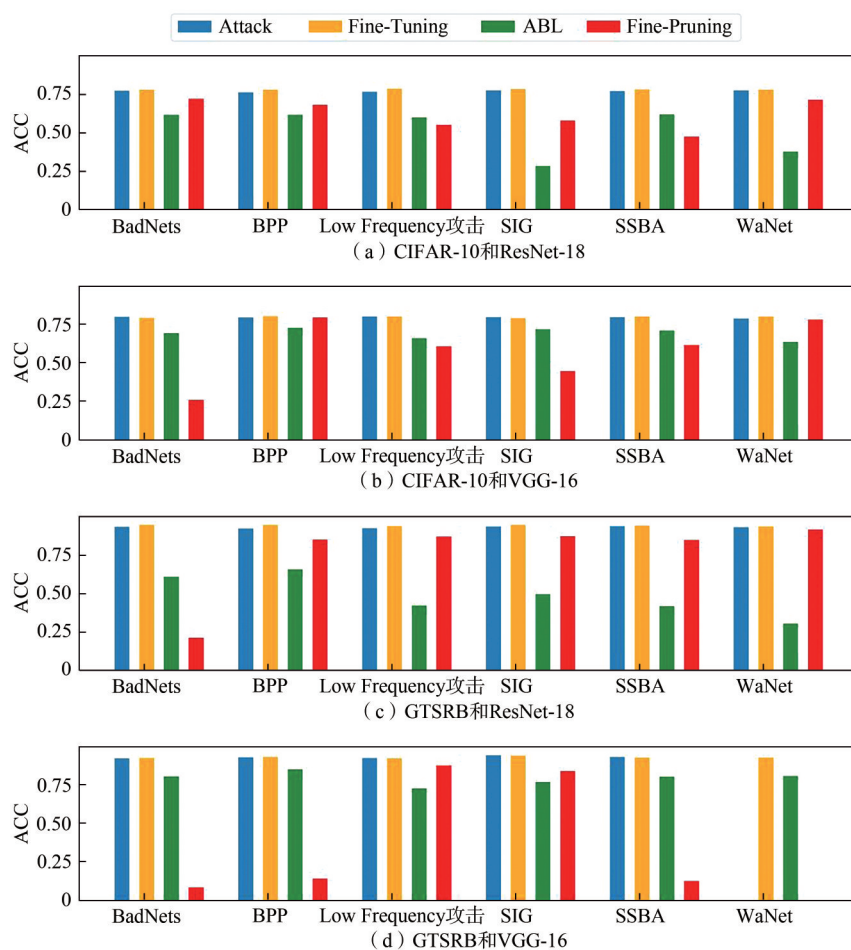


图7 识别准确率对比

3.4 传统的单一标准防御评估分析

传统的单一标准防御评估实验只考虑攻击成功率的绝对大小，并且对于图像模态而言，只考虑在CIFAR-10和ResNet-18下的评估结果，即图8(a)所示的结果。而对于检测型的防御方法，则通过微调后得到攻击成功率进行防御效果对比，最终防御方法AC、STRIP、Fine-Tuning、ABL和Fine-Pruning在传统的单一标准防御评估下的结果如表3所示。

按照传统的单一标准防御评估，5种防御方法的优劣排名为Fine-Pruning>ABL>Fine-Tuning>AC>STRIP。传统的单一标准防御评估未考虑防御成本、负面影响等，只考虑了一种数据集和模型，不能很好地适应客观应用的评估需求。

3.5 防御实验定量计算

本文共进行了24组攻击实验，每种攻击实验都有对应的防御实验，考虑到有些防御实验的结果受超参数影响较大，而不同的超参数对应的防御时间差距较大，虽然在评估方案中时间已经被考虑在

内，但是为了更加全面地评估，分别对STRIP和Fine-Tuning设置了多组不同的超参数（重复次数、训练轮数）的实验。由上文可知，攻击方法WaNet在VGG-16下对GTSRB进行的攻击并没有成功，因此对这组实验进行的防御实验仅作参考，不参与评估。防御评估结果如表4所示，其中 R 表示重复次数， E 表示训练轮数。

表4展示了不同防御方法在不同维度的得分以及最后的总得分，可以看出，在模型检测型防御方法中，STRIP($R=10$)的有效性和无损性得分都高于AC，但是在通用性和经济性得分显著低于AC，STRIP最终得分低于AC。由图6可知，单独考虑某些特例，比如BPP在ResNet-18模型下对CIFAR-10的攻击，STRIP的干净数据集识别准确率和召回率均高于0.8，而AC的这两项指标均不超过0.6，可以认为在这一特定情况下STRIP的防御效果优于AC，但是考虑到多种应用场景的平均情况，STRIP的得分就不如AC了。

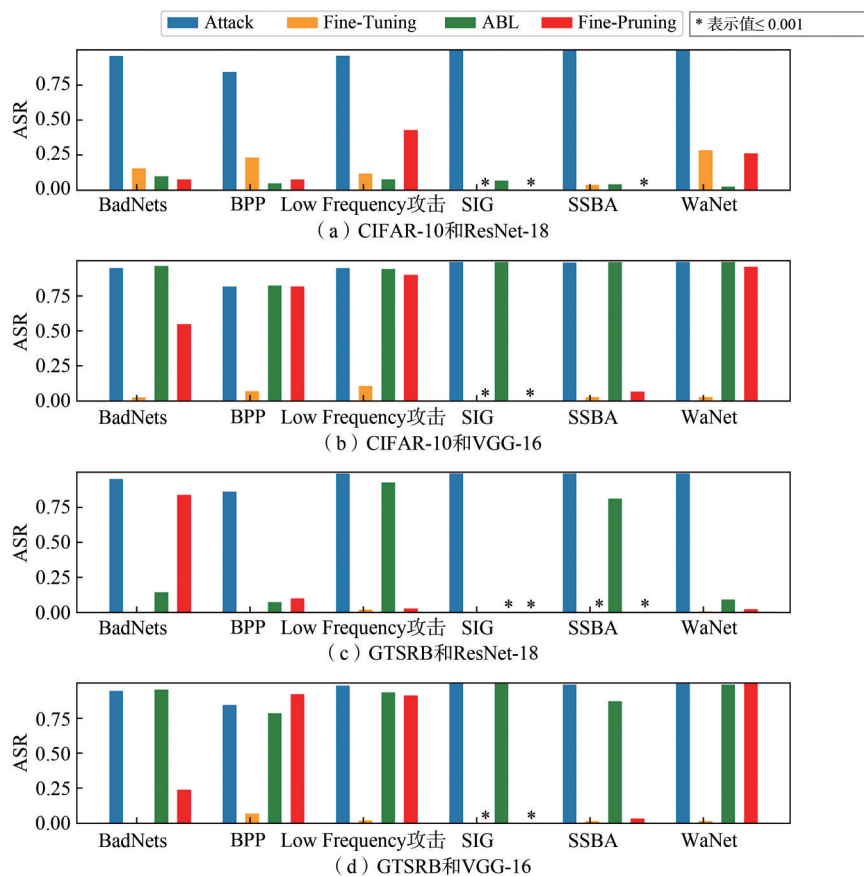


图8 攻击成功率对比

表3 传统的单一标准防御评估结果

防御方法	AC	STRIP	Fine-Tuning	ABL	Fine-Pruning
评估结果 (攻击成功率)	31.67%	42.51%	15.59%	9.90%	7.76%

在模型修复型防御方法中同样存在这种单项指标得分低导致总体得分低的情况。从图8可以看出,在GTSRB模型下,ABL的攻击成功率几乎没有下降,这严重影响了通用性得分,使得ABL总得分不高。综合这5种防御方法和不同的参数设置,得分最高的是Fine-Tuning ($E=100$),其次是Fine-Tuning ($E=30$),可见Fine-Tuning是这几种评估方法中最优秀的防御方法。根据得分来看,这些

防御方法的优越性排名为Fine-Tuning>AC>Fine-Pruning>ABL>STRIP。相比于传统的单一标准防御评估来说,本文提出的人工智能后门防御统一评估框架所使用的评估方法不仅考虑了防御方法的负面影响和成本,还考虑了防御实施过程中的普适性、客观限制等多种因素,实现了全面彻底的评估后门防御方法。

3.6 语音和文本模态的扩展实验分析

3.6.1 扩展实验的基本设置

为了全面地评估各防御方法,本文还在语音和文本模态数据上进行扩展实验。语音实验在Speech-Commands^[33]数据集上进行,使用了AudioCNN^[34]和VggVox^[35]模型,文本实验在SST-2^[36]数据集和

表4 防御评估结果

对比项	AC	STRIP(R=1)	STRIP(R=10)	Fine-Tuning(E=1)	Fine-Tuning(E=10)	Fine-Tuning(E=30)	Fine-Tuning(E=100)	ABL	FP
S_1	45.65	29.78	52.82	22.22	40.40	64.32	94.22	42.45	66.12
S_2	53.40	88.77	86.47	96.54	101.09	101.29	100.81	72.18	71.12
S_3	100.00	26.09	56.52	13.04	52.17	78.26	100.00	26.09	56.52
S_4	70.97	11.55	0.85	100.00	8.35	2.49	0.00	5.74	4.68
S	60.60	39.01	55.09	43.03	51.69	68.32	87.27	41.45	59.06

AG's News^[37]数据集上进行,使用BERT^[38]模型。语音实验使用了4种攻击方法,分别是Blend^[39]、DABA^[40]、GIS^[41]和UltraSonic^[13]。文本实验使用了4种攻击方法,分别是AddSent^[11]、BadNets、LWP^[42]和SynBkd^[43]。

3.6.2 扩展实验的实验结果

AC、STRIP、Fine-Tuning、ABL和Fine-Pruning这5种防御方法最初只在图像领域进行实验,本文对这5种防御方法做了输入输出格式的扩展,使得它们可以兼容语音和文本模态。语音模态的评分结果如表5所示,评分排名为ABL>STRIP>AC>Fine-Pruning>Fine-Tuning。文本模态的评分结果如表6所示,最终评分排名为Fine-Pruning>ABL>AC>Fine-Tuning>STRIP。这5种防御方法在3种不同模态数据上的排名有较大差异。

表5 语音模态的评分结果

对比项	AC	STRIP	Fine-Tuning	ABL	Fine-Pruning
S_1	64.44	74.33	14.96	96.41	46.65
S_2	30.96	27.31	53.43	100.00	29.96
S_3	100.00	100.00	50.00	100.00	50.00
S_4	25.27	7.20	7.22	42.28	18.02
S	60.94	63.35	28.89	92.43	41.12

表6 文本模态的评分结果

对比项	AC	STRIP	Fine-Tuning	ABL	Fine-Pruning
S_1	54.21	9.59	55.99	63.14	75.00
S_2	45.06	24.64	39.32	70.81	40.88
S_3	100.00	50.00	87.50	87.50	75.00
S_4	2.35	0.27	1.05	2.25	94.08
S	56.35	19.75	53.46	63.46	70.09

3.7 综合评价分析

定性分析较为直观地展示了不同防御方法在相关指标上的差异。单一的评价标准的定量分析(如干净数据集识别准确率和攻击成功率)在分析某一特定化的场景时有很大作用,但是在面对复杂的应用场景时,单一的评价标准并不全面、客观,而本文提出的人工智能后门防御统一评估方法则给出了更加全面的评价结果,还能在多模态维度实现防御方法的全面评估。

4 结束语

为了解决传统的后门评估指标无法比较后门数据检测型和后门修复型防御方法的问题,以及在通

用场景中比较不同的后门防御方法,本文提出了人工智能后门防御统一评估方法,并为该防御方法设计了全面的评估框架、评估指标和评估流程。该防御方法从有效性、无损性、通用性和经济性4个维度去评价一个待评估方法。通过对多个攻击场景的防御结果进行打分,将得分加权计算出该防御方法的最终得分。本文提出的人工智能后门防御统一评估方法为后门防御领域提供了一种全新的评估框架,它客观全面地评估了防御方法的优劣,并为不同的防御方法之间的对比提供了参考标准,为后续防御方法的评估提供了评价指标。这一方法有望在学术研究和工程实践中发挥重要作用,为后门防御领域的发展提供有力支持。

本文虽然提出了人工智能后门防御统一评估方法,并进行了相关实验,得出了一定的结论,但是人工智能后门防御统一评估方法仍存在一定的不足。首先,大部分后门攻击防御会受数据的预处理方式的影响,因此后门防御评估应该引入数据处理维度;其次,现在的后门防御不仅仅是集中在图像、语音和文本模态,在视频、联邦学习和图网络等领域也存在后门攻击防御问题,目前的框架并不支持这些模态的后门防御评估,但是由于本框架良好的结构,可向更多的人工智能后门领域扩展;最后,不同的后门防御方法需要的条件是不同的,有些需要部分训练数据,有些则需要获取模型参数,而有些则要求防御者能操纵训练过程,不同的方法对于防御者的要求不同,这在评估不同的防御方法时也应该被考虑到。

参考文献:

- [1] LI Y J, LI Y Q, DAI X L, et al. Physical-world optical adversarial attacks on 3D face recognition[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 24699-24708.
- [2] KIM S, GHOLAMI A, SHAW A, et al. Squeezeformer: an efficient Transformer for automatic speech recognition[J]. Advances in Neural Information Processing Systems, 2022, 35: 9361-9373.
- [3] WANG L Y, LYU C Y, JI T B, et al. Document-level machine translation with large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2023: 16646-16661.
- [4] DU Y L, YANG M J, DAI B, et al. Learning universal policies via text-guided video generation[EB]. arXiv preprint, 2023, arXiv: 2302.00111.
- [5] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundation and fine-tuned chat models[EB]. arXiv preprint, 2023, arXiv: 2307.09288.
- [6] LI Y M, JIANG Y, LI Z F, et al. Backdoor learning: a survey[J]. IEEE

- Transactions on Neural Networks and Learning Systems, 2024, 35(1): 5-22.
- [7] GAO Y S, DOAN B G, ZHANG Z, et al. Backdoor attacks and countermeasures on deep learning: a comprehensive review[EB]. arXiv preprint, 2020, arXiv: 2007.10760.
 - [8] GU T Y, DOLAN-GAVITT B, GARG S. BadNets: identifying vulnerabilities in the machine learning model supply chain[EB]. arXiv preprint, 2017, arXiv: 1708.06733.
 - [9] WANG Z T, ZHAI J, MA S Q. BppAttack: stealthy and efficient Trojan attacks against deep neural networks via image quantization and contrastive adversarial learning[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 15054-15063.
 - [10] TURNER A, TSIPRAS D, MADRY A. Label-consistent backdoor attacks[EB]. arXiv preprint, 2019, arXiv: 1912.02771.
 - [11] DAI J Z, CHEN C S, LI Y F. A backdoor attack against LSTM-based text classification systems[J]. IEEE Access, 2019, 7: 138872-138878.
 - [12] QI F C, CHEN Y Y, ZHANG X R, et al. Mind the style of text! adversarial and backdoor attacks based on text style transfer[EB]. arXiv preprint, 2021, arXiv: 2110.07139.
 - [13] KOFFAS S, XU J, CONTI M, et al. Can you hear it? backdoor attacks via ultrasonic triggers[C]//Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning. New York: ACM Press, 2022: 57-62.
 - [14] ZHAO S H, MA X J, ZHENG X, et al. Clean-label backdoor attacks on video recognition models[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE Press, 2020: 14431-14440.
 - [15] XI Z H, PANG R, JI S L, et al. Graph backdoor[EB]. arXiv preprint, 2021, arXiv: 2006.11890.
 - [16] CHEN B, CARVALHO W, BARACALDO N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[EB]. arXiv preprint, 2018, arXiv: 1811.03728.
 - [17] GAO Y S, KIM Y, DOAN B G, et al. Design and evaluation of a multi-domain Trojan detection method on deep neural networks[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(4): 2349-2364.
 - [18] LI Y, LYU X, KOREN N, et al. Anti-backdoor learning: training clean models on poisoned data[J]. Advances in Neural Information Processing Systems, 2021, 34: 14900-14912.
 - [19] LIU K, DOLAN-GAVITT B, GARG S. Fine-pruning: defending against backdooring attacks on deep neural networks[M]. Cham: Springer International Publishing, 2018: 273-294.
 - [20] ZHENG R K, TANG R J, LI J Z, et al. Data-free backdoor removal based on Channel lipschitzness[M]. Berlin: Springer, 2022: 175-191.
 - [21] PANG R, ZHANG Z, GAO X S, et al. TrojanZoo: towards unified, holistic, and practical evaluation of neural backdoors[EB]. arXiv preprint, 2020, arXiv: 2012.09302.
 - [22] WU B, CHEN H, ZHANG M, et al. Backdoorbench: a comprehensive benchmark of backdoor learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 10546-10559.
 - [23] CUI G, YUAN L, HE B, et al. A unified evaluation of textual backdoor learning: frameworks and benchmarks[J]. Advances in Neural Information Processing Systems, 2022, 35: 5009-5023.
 - [24] 谭清尹, 曾颖明, 韩叶, 等. 神经网络后门攻击研究[J]. 网络与信息安全学报, 2021, 7(3): 46-58.
 - TAN Q Y, ZENG Y M, HAN Y, et al. Survey on backdoor attacks targeted on neural network[J]. Chinese Journal of Network and Information Security, 2021, 7(3): 46-58.
 - [25] ZENG Y, PARK W, MAO Z M, et al. Rethinking the backdoor attacks' triggers: a frequency perspective[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision(ICCV). Piscataway: IEEE Press, 2021: 16453-16461.
 - [26] BARNI M, KALLAS K, TONDI B. A new backdoor attack in CNNs by training set corruption without label poisoning[C]//Proceedings of the 2019 IEEE International Conference on Image Processing(ICIP). Piscataway: IEEE Press, 2019: 101-105.
 - [27] LI Y Z, LI Y M, WU B Y, et al. Invisible backdoor attack with sample-specific triggers[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision(ICCV). Piscataway: IEEE Press, 2021: 16443-16452.
 - [28] NGUYEN A, TRAN A. WaNet: imperceptible warping-based backdoor attack[EB]. arXiv preprint, 2021, arXiv: 2102.10369.
 - [29] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[J]. Handbook of Systemic Autoimmune Diseases, 2009, 1(4): 1-60.
 - [30] STALLKAMP J, SCHLIPSING M, SALMEN J, et al. The German traffic sign recognition benchmark: a multi-class classification competition[C]//Proceedings of the 2011 International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2011: 1453-1460.
 - [31] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE Press, 2016: 770-778.
 - [32] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB]. arXiv preprint, 2014, arXiv: 1409.1556.
 - [33] WARDEN P. Speech commands: a dataset for limited-vocabulary speech recognition[EB]. arXiv preprint, 2018, arXiv: 1804.03209.
 - [34] ABDOLI S, CARDINAL P, KOERICH A L. End-to-end environmental sound classification using a 1D convolutional neural network[J]. Expert Systems with Applications, 2019, 136: 252-263.
 - [35] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: deep speaker recognition[EB]. arXiv preprint, 2018, arXiv: 1806.05622.
 - [36] SOCHER R, PERELYGIN A, WU J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Washington: Association for Computational Linguistics, 2013: 1631-1642.
 - [37] ZHANG X, ZHAO J B, LECUN Y. Character-level convolutional networks for text classification[J]. Advances in Neural Information Processing Systems, 2015(1): 649-657.
 - [38] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB]. arXiv preprint, 2018, arXiv: 1810.04805.
 - [39] CHEN X Y, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[EB]. arXiv preprint, 2017, arXiv: 1712.05526.
 - [40] LIU Q, ZHOU T Q, CAI Z P, et al. Opportunistic backdoor attacks: exploring human-imperceptible vulnerabilities on speech recognition systems[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: Association for Computing Machinery, 2022: 2390-2398.
 - [41] KOFFAS S, PAJOLA L, PICEK S, et al. Going in style: audio backdoors

through stylistic transformations[C]//Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.

[42] LI L Y, SONG D M, LI X N, et al. Backdoor attacks on pre-trained models by layerwise weight poisoning[EB]. arXiv preprint, 2021, arXiv: 2108.13888.

[43] QI F C, LI M K, CHEN Y Y, et al. Hidden killer: invisible textual backdoor attacks with syntactic trigger[EB]. arXiv preprint, 2021, arXiv: 2105.12400.

[作者简介]



谢天（2002-），男，上海交通大学本科生，主要研究方向为深度学习、人工智能安全、多模态学习。



李强（1979-），男，博士，32178部队科技创新中心高级工程师，主要研究方向为平行指挥控制、对抗生成情报等。



鞠卓亚（1988-），女，博士，32178部队科技创新中心工程师，主要研究方向为人工智能安全、网络安全。



韩嘉祺（1994-），女，32178部队科技创新中心工程师，主要研究方向为人工智能安全评估。



易平（1969-），男，博士，上海交通大学副研究员，主要研究方向为人工智能安全、系统安全。