

面向深度学习的后门攻击及防御研究综述^{*}

高梦楠¹, 陈伟^{1,2}, 吴礼发^{1,2}, 张伯雷¹

¹(南京邮电大学 计算机学院、软件学院、网络空间安全学院, 江苏 南京 210023)

²(江苏省大数据安全与智能处理重点实验室 (南京邮电大学), 江苏 南京 210023)

通信作者: 陈伟, E-mail: chenwei@njupt.edu.cn



摘 要: 深度学习模型是人工智能系统的重要组成部分, 被广泛应用于现实多种关键场景. 现有研究表明, 深度学习的低透明度与弱可解释性使得深度学习模型对扰动敏感. 人工智能系统面临多种安全威胁, 其中针对深度学习的后门攻击是人工智能系统面临的重要威胁. 为了提高深度学习模型的安全性, 全面地介绍计算机视觉、自然语言处理等主流深度学习系统的后门攻击与防御研究进展. 首先根据现实中攻击者能力将后门攻击分为全过程可控后门、模型修改后门和仅数据投毒后门. 然后根据后门构建方式进行子类划分. 接着根据防御策略对象将现有后门防御方法分为基于输入的后门防御与基于模型的后门防御. 最后汇总后门攻击常用数据集与评价指标, 并总结后门攻击与防御领域存在的问题, 在后门攻击的安全应用场景与后门防御的有效性等方面提出建议与展望.

关键词: 深度学习; 后门攻击; 后门防御; 人工智能安全

中图法分类号: TP306

中文引用格式: 高梦楠, 陈伟, 吴礼发, 张伯雷. 面向深度学习的后门攻击及防御研究综述. 软件学报, 2025, 36(7): 3271–3305. <http://www.jos.org.cn/1000-9825/7364.htm>

英文引用格式: Gao MN, Chen W, Wu LF, Zhang BL. Survey on Backdoor Attacks and Defenses for Deep Learning Research. Ruan Jian Xue Bao/Journal of Software, 2025, 36(7): 3271–3305 (in Chinese). <http://www.jos.org.cn/1000-9825/7364.htm>

Survey on Backdoor Attacks and Defenses for Deep Learning Research

GAO Meng-Nan¹, CHEN Wei^{1,2}, WU Li-Fa^{1,2}, ZHANG Bo-Lei¹

¹(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

²(Jiangsu Key Laboratory of Big Data Security & Intelligent Processing (Nanjing University of Posts and Telecommunications), Nanjing 210023, China)

Abstract: Deep learning models are integral components of artificial intelligence systems, widely deployed in various critical real-world scenarios. Research has shown that the low transparency and weak interpretability of deep learning models render them highly sensitive to perturbations. Consequently, artificial intelligence systems are exposed to multiple security threats, with backdoor attacks on deep learning models representing a significant concern. This study provides a comprehensive overview of the research progress on backdoor attacks and defenses in mainstream deep learning systems, including computer vision and natural language processing. Backdoor attacks are categorized based on the attacker's capabilities into full-process controllable backdoors, model modification backdoors, and data poisoning backdoors, which are further classified according to the backdoor construction methods. Defense strategies are divided into input-based defenses and model-based defenses, depending on the target of the defensive measures. This study also summarizes commonly used datasets and evaluation metrics in this domain. Lastly, existing challenges in backdoor attack and defense research are discussed, alongside recommendations and future directions focusing on security application scenarios of backdoor attacks and the efficacy of defense mechanisms.

Key words: deep learning; backdoor attack; backdoor defense; AI security

* 基金项目: 国家自然科学基金 (62202238); 江苏省重点研发项目 (BE2022065-5)

收稿时间: 2024-04-27; 修改时间: 2024-07-15, 2024-09-05; 采用时间: 2024-11-26; jos 在线出版时间: 2025-04-25

CNKI 网络首发时间: 2025-04-27

1 引言

近年来,深度学习(deep learning, DL)在大规模数据上展现出卓越性能.以深度学习为核心的人工智能系统(artificial intelligence, AI)被广泛应用在计算机视觉(computer vision, CV)、自然语言处理(natural language processing, NLP)、语音识别(voice recognition system, VRS)和恶意软件检测(malware detection, MD)等关键场景^[1,2].例如,Apple FaceID使用人脸识别系统解锁移动设备并认证用户交易身份;特斯拉自动辅助驾驶系统Autopilot自主分析行车环境,辅助行车转向、加速和制动;Amazon Comprehend提供的文本情绪分析功能用于构建客户关系智能软件,提供高质量的客户体验;百度文心ERNIE平台使用知识增强大语言模型提供人机对话、图像文本搜索、代码补全等服务.

与此同时,深度学习面临多种安全威胁.与传统机器学习算法不同,深度学习模型通常由多层非线性变换组成,结构更为复杂,面对大规模数据表现出更强的泛化能力与表达能力,被广泛应用于多种场景中,因此针对深度学习的攻击具有更高的研究价值.在深度学习面临的安全威胁中,后门攻击是一类重要威胁.后门攻击具有极强的隐蔽性,遭到后门攻击的模型在大部分数据上表现正常,但在具有指定特征的一类数据上表现出错.开放网络环境下不可信第三方提供的数据和模型本身的弱可解释性使得模型易受后门攻击^[3].首先,高性能的深度学习系统需求大量数据与计算资源,而个人与小公司难以承担训练需求^[4],因此外包训练与预训练微调成为工业界主流的AI构建方式.外包训练中用户将个人需求、训练数据等信息提交第三方,第三方根据需求将训练好的模型参数交付用户.预训练微调中用户从第三方平台获得预训练模型(pre-training model, PTM),并在私人数据集上微调模型.上述AI构建方式中,模型托管平台、数据提供商与云服务提供商等第三方安全主体介入深度学习训练过程,带来安全隐患.这些第三方平台提供外包标注数据、预训练模型参数与机器学习即服务(machine learning as a service, MLaaS)^[5],但由于现实中缺乏监管,攻击者可能恶意篡改数据标签、上传恶意训练数据及模型^[6].其次,深度学习低透明度和弱可解释性的原生问题使得深度学习模型对扰动敏感^[7,8].深度学习模型呈现为黑盒,其结构复杂,具有大量参数,非线性函数的变量表示使得模型内部神经元意义与决策过程难以理解.此外,深度学习决策边界的高曲率区域对扰动敏感,沿着高曲率方向的微小扰动能够越过决策边界,导致模型性能下降.

目前已经有一些研究整理了不同领域中后门攻击进展^[4,9-17],整体上后门攻击根据模型训练是否可控分为训练过程可控后门与数据投毒后门两类,各文献根据不同的划分标准对其进行细分.例如文献^[4,9,12,14,16]以触发器属性为标准分别划分计算机视觉、自然语言处理和语音识别中的数据投毒后门攻击;文献^[10,13]根据攻击场景将后门攻击细分到端到端训练、预训练和迁移学习这3个场景.相较于其他文献综述,本文关注多个领域的后门攻击与防御进展,详细梳理了主流深度学习系统,包括计算机视觉、自然语言处理、代码检测中的后门攻击进展,并确立一种统一的攻击与防御划分标准,具体可见表1.

表1 本文与其他现有综述的对比

文献	年份	领域	攻击分类依据	防御分类依据	①	②	③	④	⑤
[14]	2020	CV	触发器属性	—	●	○	▲	●	○
[17]	2022			—	●	○	▲	●	●
[9]	2023		攻击训练策略	防御对象	●	●	▲	▲	○
[10]	2023			防御阶段	●	●	○	○	○
[12]	2023	NLP	触发器属性	防御阶段	●	○	○	○	○
[16]	2022		触发器属性	防御策略	●	●	○	○	○
[18]	2024		攻击粒度	防御阶段	●	●	▲	▲	○
[4]	2023	VRS	触发器属性	防御对象	●	●	○	●	○
本文	2024	CV/NLP	攻击者能力	防御对象	●	●	●	●	○

注:●表示包含;○表示不包含;▲表示部分包含;①为提出后门攻击的一种分类方法;②为提出后门防御的一种分类方法;③为后门攻击常用数据集;④为后门攻击常用评价指标;⑤为后门攻击常用目标模型

本文以攻击者能力为标准提出了一种新的后门攻击分类方法, 根据现实中 AI 构建方式划分 3 种不同攻击者能力下的后门攻击体系, 助力各领域下游任务了解最新的后门攻击形式与后门威胁, 从而构建相应的防御措施. 本文还根据后门防御对象梳理了最新的后门防御方法. 本文将后门防御分为基于输入的后门防御与基于模型的防御两类, 从数据和模型两个层面抵抗和削弱后门攻击, 希望为后续后门防御研究提供帮助. 目前, 计算机视觉和自然语言处理仍是后门攻击的两大主流研究领域, 两个领域的后门攻击都要求具有隐蔽性和有效性. 但是, 计算机视觉中数据载体为图像, 而自然语言处理任务中数据载体是文本. 相较于图像在表征空间的连续性, 文本词嵌入后表现为离散性. 数据表征形式的差异导致两个领域的后门攻击存在着差异. 以后门攻击隐蔽性为例, 计算机视觉后门攻击要求人类视觉上无法察觉图像添加触发器前后的轻微变化, 而自然语言处理中的后门攻击则更为强调保留文本拼写的正确性与语义信息的连贯性与一致性. 但同时, 无论是计算机视觉后门攻击还是自然语言处理后门攻击, 都要求保证不影响模型在正常样本上的性能.

本文首先介绍后门攻击基本知识, 提出后门攻击通用框架, 并总结后门攻击威胁模型与攻击目标. 然后, 本文提出了以攻击者能力为标准的后门攻击分类方法, 阐述了每个类别特点, 并分析近几年提出的后门攻击的优势与不足. 接着, 本文根据后门防御对象梳理了最新的后门防御方法. 本文还整理了不同领域后门攻击常用数据集与评价指标. 最后, 本文讨论了现阶段后门攻击方面存在的问题与未来研究方向. 整体上, 本文主要贡献如下.

- (1) 提出一种基于攻击者能力的后门攻击分类方法, 每种攻击者能力分别对应现实中的 3 种 AI 构建方式: 外包训练、预训练微调 and MLaaS, 帮助研究人员更为系统和全面地理解现实中不同人工智能系统面临的后门威胁.
- (2) 整理了计算机视觉、自然语言处理和恶意软件检测等主流深度学习系统中最新的后门攻击方法, 并讨论上述方法的优势与不足.
- (3) 以后门防御对象为分类标准, 从输入和模型两个方面梳理了最新的后门防御方法, 帮助研究人员理解当前后门防御的进展与不足.
- (4) 提出了目前后门攻击面临的一系列安全议题, 并对其未来研究方向进行展望.

本文第 2 节简要概述深度学习中的后门攻击安全威胁. 第 3 节介绍深度学习后门攻击分类, 依次介绍不同领域后门攻击研究进展. 第 4 节介绍深度学习后门防御进展. 第 5 节整理深度学习后门攻击常用数据集与评价指标. 第 6 节讨论深度学习中后门攻防存在的问题以及未来研究方向. 第 7 节总结全文, 具体见图 1.

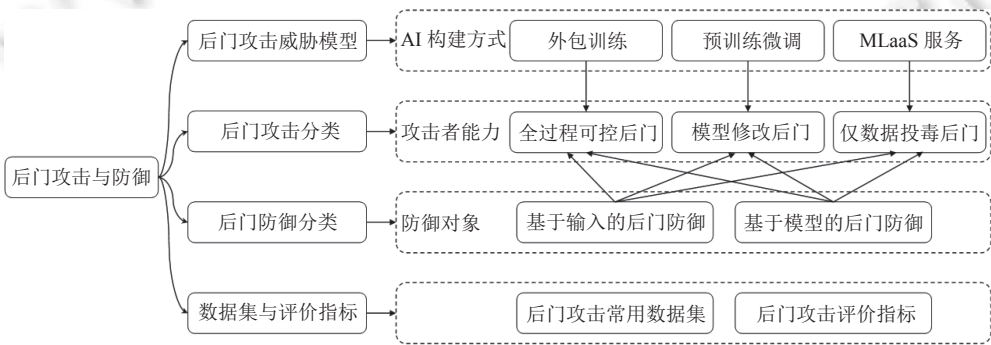


图 1 本文整体结构图

2 基础知识

本节介绍后门攻击相关知识. 第 2.1 节介绍深度学习生命周期中存在的后门攻击威胁. 第 2.2 节提出后门攻击通用框架. 第 2.3 节介绍后门攻击威胁模型, 定义不同攻击场景中攻击者与防御者能力.

2.1 深度学习生命周期中的后门攻击

完整的深度学习系统生命周期由数据收集、数据预处理、模型选择与构建、模型训练、模型存储、模型部署、模型推断、模型更新等多阶段构成^[9,11,12,14]. 深度学习生命周期面临数据投毒、后门攻击和对抗样本多种攻

击(见图2)。数据投毒攻击针对数据收集与预处理阶段,攻击者向训练数据注入中毒样本,期望降低模型整体的推断准确性。对抗样本攻击针对模型推理阶段,攻击者向输入样例中添加难以察觉的扰动使得模型给出不正确的结果^[9,19]。通用对抗补丁可以被视为对抗样本攻击的一种特殊形式,对抗样本攻击专注于在单个样本上添加扰动,而通用对抗攻击旨在寻找可以应用于大量输入数据样本的通用扰动。相较数据投毒攻击与对抗样本攻击,后门攻击可以在除模型推断外的所有阶段进行。

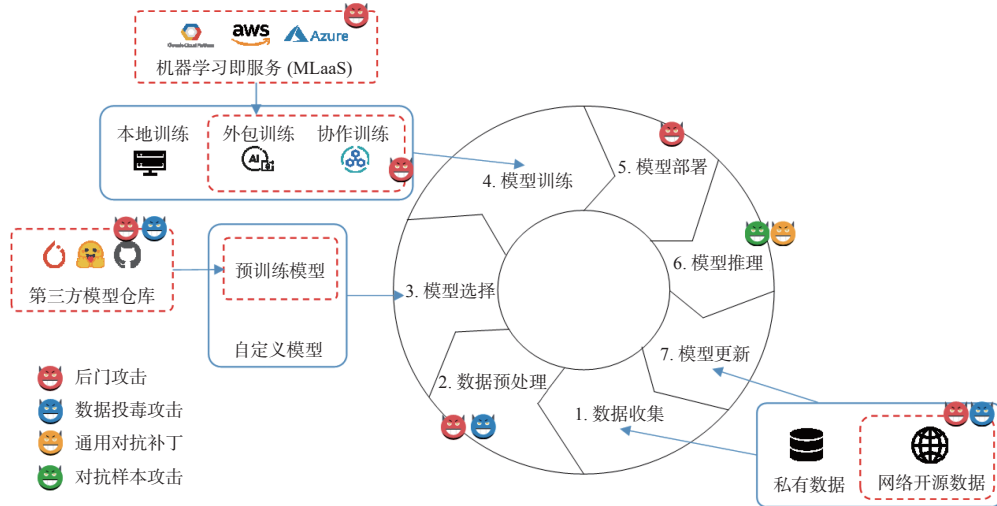


图2 深度学习生命周期各阶段潜在的安全威胁

后门攻击与对抗样本攻击在攻击者目标与触发器设计上存在相似点。首先,以分类任务为例,后门攻击与对抗样本攻击都需要修改测试示例构成攻击样本,攻击目标都是使目标模型在攻击样本上分类错误,而在其他正常样本上表现正常。其次,后门攻击可参照对抗样本的构建思路去构建中毒样本。后门攻击与对抗样本攻击在攻击者能力和攻击范式上有明显差异。后门攻击能够干扰和控制模型训练过程,而对抗样本攻击无法干扰模型训练。此外,对抗样本攻击需要攻击者根据多次查询模型得到的结果优化针对某个输入的扰动,而后门攻击仅需将预定义的触发器添加到输入示例上即可实现攻击。后门攻击仅针对包含后门的模型有效,而对抗样本攻击对大量模型有效。

2.2 后门攻击定义

后门攻击指攻击者污染训练数据或修改模型参数,引导目标模型在推理阶段出错。攻击者在干净样本上添加触发器构成中毒样本,并将中毒样本注入训练数据,在中毒样本与训练数据上共同训练模型。在图3中,我们展示了一次完整的后门攻击。触发器是一种可以应用于任何输入的模式,具有多种表现形式。以分类任务为例,在计算机视觉任务中,触发器可以是某个具体的图案(比如图3中将猫图案作为触发器);在自然语言处理任务中,攻击者可以将具体的字符“cf”作为触发器。攻击者将上述触发器加入训练样本,并篡改样本标签,期望模型在训练过程中在触发器与修改后的标签之间产生联系。而在模型推理阶段,攻击者在任意测试样本上添加触发器,后门模型就会对包含触发器的测试样本输出预期的异常行为。整体上,完整的后门攻击包括3个步骤^[20]:(1)触发器设计:攻击者根据任务需求设计触发器,并生成中毒样本;(2)后门注入:攻击者通过数据投毒或参数修改等方式向目标模型注入后门;(3)后门激活:攻击者生成含触发器的测试案例激活模型后门。

以分类任务为例,给定训练数据 $D_{\text{train}} = \{(x_i, y_i) | x_i \in X, y_i \in Y\}$, 深度学习模型 $f_\theta: X \rightarrow Y, X \subset \mathbb{R}^n, Y \subset \mathbb{R}^d$. f_θ 将 n 维输入映射到 d 维空间, (x, y) 表示输入样本 x 及其对应标签 y . 攻击者指定目标类别 y_p , 在随机训练样本 x_i 上添加触发器构成中毒样本 x_p , 旨在训练后门模型 $f_{\theta_p}: X \rightarrow Y$, 满足 $f_{\theta_p}(x_i) = y_i, f_{\theta_p}(x_p) = y_p$. 整体上,中毒样本 x_p 由干净样本 x_i 和触发生成算法 $G(\cdot)$ 表示:

$$x_p = \alpha(x_i) + \beta(G(x_i)) \quad (1)$$

其中, $\alpha(\cdot)$ 和 $\beta(\cdot)$ 表示干净样本与触发器混合方式. 后门攻击具有两个基本目标: 一是在正常样本上, 后门模型表现出与良性模型相近的性能; 二是在含有触发器的后门样本, 后门模型表现出攻击者期望的结果. 后门攻击可以表述为以下形式:

$$\theta_b^* = \arg \min_{\theta_b} \left\{ \sum_{(x,y) \in D_{\text{clean}}} L(f_{\theta_b}(x), y) + \gamma \cdot \sum_{(x,y) \in D_{\text{poison}}} L(f_{\theta_b}(x_t), y_t) + \lambda \cdot L(G(x)) \right\}$$

(2)

$$D_{\text{poison}} = \{(x', y_t) | x' = G(x), (x, y) \in D_{\text{train}}, t = 1, 2, \dots, s\}$$

(3)

其中, $\sum_{(x,y) \in D_{\text{clean}}} L(f_{\theta_b}(x), y)$ 表征模型主任务性能; $\sum_{(x,y) \in D_{\text{poison}}} L(f_{\theta_b}(x_t), y_t)$ 表示后门任务性能; $L(G(x))$ 是惩罚项, 用于约束后门隐蔽性; γ 和 λ 用于平衡模型主任务、后门任务以及后门隐蔽性能.

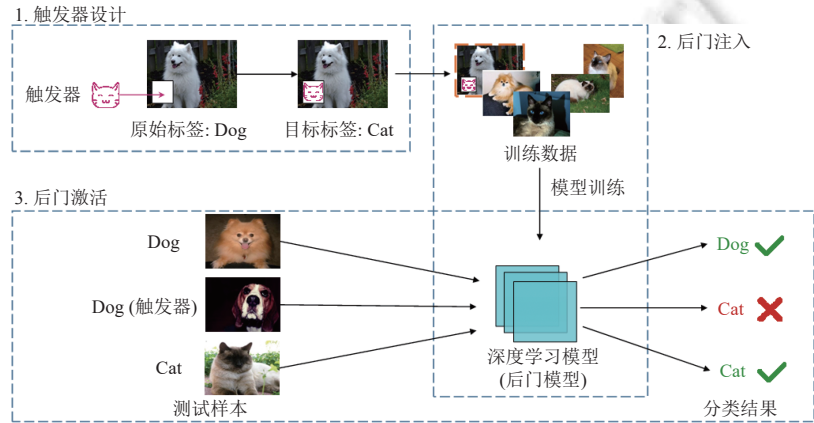


图3 后门攻击示例图

2.3 威胁模型

后门攻击具有攻击有效性与隐蔽性两个目标, 即后门模型在后门样本上表现出预定行为, 但在良性样本上运行正确, 能够对抗现有的防御方法. 现实中, 后门攻击存在于外包训练、预训练微调、MLaaS 服务这 3 种典型场景中. 本节依次介绍 3 种典型场景中后门攻击威胁模型与攻击者能力 (见表 2). (1) 外包训练中, 攻击者了解目标模型任务, 能够完全控制训练数据与训练过程; (2) 预训练微调中, 攻击者不了解目标模型下游任务, 可能获取任务相关数据, 了解模型知识, 不能干涉下游任务训练过程; (3) MLaaS 服务中, 攻击者了解目标模型任务, 可能获取任务相关数据, 但不了解模型知识, 无法干涉训练过程; (4) 本文中防御者指模型用户, 面对上述 3 种攻击场景, 防御者拥有一部分私有验证集, 部分了解模型训练过程与攻击者使用的训练数据类型, 完全了解模型结构与参数, 能够控制模型后续训练. 防御者目标是检测模型性能是否满足下游任务需求, 以及模型是否存在后门, 并清除后门. 整体上, 防御者能够完全控制目标模型, 但他们不了解攻击者可能使用的触发器.

表 2 威胁模型攻击者与防御者能力划分

威胁模型	攻击者能力				防御者能力			
	模型任务	训练数据	模型知识	训练过程	训练数据	验证集	模型结构	训练过程
外包训练	●	●	●	●	▲	●	●	▲
预训练微调	○	▲	●	▲	▲	●	●	▲
MLaaS服务	●	▲	○	○	▲	●	●	▲

注: ●表示完全了解/完全控制; ▲表示部分了解/部分控制; ○表示完全不了解/完全不控制

2.3.1 外包训练场景

外包训练场景中, 攻击者是恶意第三方实体. 用户将训练数据与模型需求外包给第三方, 第三方根据用户需求将训练好的模型参数提交给用户, 用户使用私有数据验证模型性能是否满足需求.

攻击者能力: 攻击者具有最强的攻击能力, 能够接触到一部分甚至完整的训练数据, 并且可以向训练数据中插

入中毒样本. 通常用户会指定模型结构, 攻击者一般无法修改模型结构, 但他们可以控制模型训练过程, 修改模型训练参数.

2.3.2 预训练微调场景

预训练微调场景中, 用户下载第三方模型托管仓库中的预训练模型, 他们根据下游任务调整模型内部结构, 并在下游任务数据上微调模型. 该场景中的攻击者为预训练模型提供者以及模型托管平台内部人员. 攻击者不了解用户下游任务具体知识, 他们向开源数据中投毒并训练后门预训练模型, 将其推送到在线存储平台, 并等待用户下载后门预训练模型用于下游任务.

攻击者能力: 攻击者不了解用户下游任务, 但他们可以根据下游任务搜集开源数据作为替代, 并向其中插入中毒样本. 攻击者完全控制预训练模型训练过程, 能够修改参数与损失函数. 由于替代数据与下游任务数据分布可能不同, 攻击者除考虑后门攻击两个基本目标外, 还期望后门具有强迁移性, 能够从预训练模型迁移到用户微调模型.

2.3.3 MLaaS 服务场景

现实中云服务提供商向用户提供免费或商用 MLaaS 查询接口, 比如文本内容生成平台、图像识别平台与在线人脸识别平台. 用户可以直接使用平台接口查询相应任务. 此类商业平台的模型结构、参数以及训练数据受到保护, 通常不会公开. 因此用户不了解模型具体结构与参数, 也不了解模型训练数据. 该场景中, 攻击者可能为系统内部人员与系统外部的恶意用户. 由于系统内部人员对系统的攻击与外包训练场景类似, 本文仅考虑恶意用户攻击 MLaaS 的情况.

攻击者能力: 攻击者具有最弱的攻击能力. 攻击者不了解模型训练数据、模型结构与参数, 也无法干涉模型训练过程. 但是攻击者了解模型下游任务, 能够搜集与下游任务相关的开源数据作为替代数据. 由于 MLaaS 系统需要定期从网络或其他平台获取新的数据更新模型, 攻击者可以上传混入中毒样本的替代数据, 等待 MLaaS 系统使用污染的数据更新模型.

3 深度学习中的后门攻击分类

本文根据攻击者能力, 将后门攻击分为 3 类: 全过程可控后门攻击、模型修改后门攻击和仅数据投毒后门攻击 (见图 4). 其中, 全过程可控后门对应外包训练场景, 模型修改后门对应预训练微调场景, 仅数据投毒后门对应 MLaaS 服务场景.

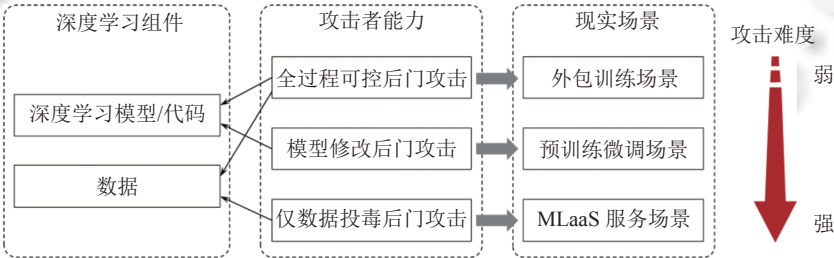


图 4 后门攻击分类与威胁模型

3.1 全过程可控后门攻击

全过程可控后门中, 攻击者完全控制训练数据与模型训练过程. 而触发器设计是后门攻击中最重要的部分, 整体上触发器生成算法需要考虑: (1) 触发器属性, 包括触发器大小、形状、所处位置、透明度等特征; (2) 触发器生成过程是否需要修改样本标签; (3) 触发器与样本的关系, 即触发器是否独立于样本.

本文根据触发生成算法将后门攻击进一步划分为可见后门与不可见后门、中毒标签后门与干净标签后门、样本不可知后门与样本特异性后门.

3.1.1 不可见后门

后门攻击本质是攻击者在样本上添加扰动, 破坏深度神经网络 (deep neural network, DNN) 在特定样本上的映

射关系. Gu 等人^[21]将小网格补丁作为触发器直接覆盖在图像右下角构成中毒样本,首次提出了计算机视觉领域的可见后门 BadNets. 之后,研究人员通过图像处理、数字隐写、时域转换等方式降低了触发器的视觉效果,构造出了人类视觉不易察觉的不可见后门. 而在自然语言处理领域,相较于特征连续的图像数据,文本在词嵌入后其数据空间呈现离散性,因此计算机视觉后门攻击无法直接移植到自然语言处理任务上. 此外,图像上部分像素的轻微扰动不会影响人类理解其语义信息,但是文本中少量词嵌入的轻微扰动可能会使句子的语义信息紊乱. 同时,由于人类无法察觉图像像素微小扰动,却能察觉拼写错误、结构混乱等文本的轻微扰动,自然语言处理中的后门攻击也因此更为关注后门的隐蔽性和语句的连贯性. 本文中自然语言处理不可见后门表现为中毒文本语法结构正确、流畅且上下文连贯. 本节依次介绍计算机视觉和自然语言处理领域不可见后门的研究进展.

(1) 计算机视觉中的后门攻击

- 图像空间域后门攻击. 一方面,从图像空间域入手,利用图像处理与数字隐写技术可以向样本中添加不规则扰动,削弱触发器可见性. 根据人类对图像颜色位深变化的不敏感性, Wang 等人^[22]在 BppAttack 中将图像量化产生的微小扰动作为触发器,利用图像抖动缓解颜色位深变化带来的不协调,生成与干净图像相似的中毒图像. Liu 等人^[23]提出的 Refool 模仿光反射现象,在干净图像上添加了不同厚度玻璃的反射效果. 数字隐写技术将字符串转化为不可见噪声嵌入图像潜在表示,实现人眼无法观察的信息传递. 根据这一思想, Li 等人^[24]在 TDSC 中使用特定字符串替换 RGB 图像中像素的最低信号位. Sun 等人^[25]提出的 DT-IBA 则利用哈希网络从目标图像中提取特征,通过隐写技术将其嵌入图像.

上述后门攻击向图像中添加了与内容无关的无语义扰动,不相干特征的引入会影响后门攻击性能与隐蔽性. 对此,一些研究^[26-28]利用数据内在特征与语义属性构造后门. Nguyen 等人^[26]在 WaNet 中将小而平滑的扭曲场作为触发器轻微扭曲图像一小块区域,保留了图像内容. Lin 等人^[27]将模型多个标签的良性特征组合成复合触发器,使用多个标签特征同时激活后门. Sarkar 等人^[28]提出的 FaceHack 将笑容、年龄、妆容等脸部语义属性作为触发器,构建出了符合人类视觉规律的中毒图像. 但是将笑容这类特定表情作为触发器,容易产生误触发的问

题. 另一方面,通过解决范数约束最优问题可以向样本中注入难以察觉的扰动. Zhong 等人^[29]提出的静态扰动掩码后门在多个大小相同彼此相邻的图像子区域中增加了自适应扰动,使得该类别中所有样本添加扰动后都能接近目标类别的决策边界. He 等人^[30]提出的替代后门 SGBA 向中毒图像上添加由非目标类图像生成的伪触发图案,通过额外的伪触发图像掩盖真正的小尺寸触发图案,极大地增强了后门攻击隐蔽性.

除了有监督学习以外,现实中还存在无监督学习深度学习系统. 无监督学习能够根据现实中的无标注数据建模,节省数据标注成本并避免标注过程带来的人工失误. 由于无监督学习缺乏对训练数据的安全审计,攻击者可以利用该漏洞实施后门攻击. Jia 等人^[31]针对自监督学习编码器提出的 BadEncoder 将后门注入作为优化目标,通过计算触发器特征向量聚合到编码器输出空间的最优解,使得中毒图像接近目标类图像的向量表示. Carlini 等人^[32]提出的多模态后门攻击向图像文本描述对中注入了由包含补丁的中毒图像与目标标签相关的文本构成的中毒图像文本描述对,误导模型将含补丁的图像分类成目标类. 相较有监督后门,无监督学习中的后门攻击通常具有较小的投毒比例,攻击更为隐蔽.

- 图像频域后门攻击. Zeng 等人^[33]指出空间域后门攻击在不同数据集和分辨率上普遍存在着严重的伪影,在频域中不具有隐蔽性. 根据 DNN 对频域变化敏感,而人类无法感知频域微小扰动^[34,35]这一差异, Gao 等人^[36]提出了具有空间域与频域隐蔽性的后门攻击 DUBA. 他们使用离散小波变换获得空间域触发器的高频信息,并嵌入于干净图像高频区域,保留触发器细节与纹理. Xia 等人^[37]在 WaveAttack 中也通过离散小波变换获取图像高频特征,但他们使用编码解码网络生成自适应触发器. 上述攻击在训练时使用了较弱的触发器,在攻击时使用较强的触发器,进一步提高了攻击能力. 在医疗影像目标检测和语义分割任务中, Feng 等人^[38]提出的 FIBA 根据振幅频谱变化不影响相位频谱的原理,将触发器频谱与干净图像频谱振幅混合,生成了不影响图像语义的频域中毒图像.

(2) 自然语言处理中的后门攻击

Chen 等人^[39]在 BadNL 中使用文本扰动技术,将字符、单词和句子分别作为触发器插入文本随机位置. 字符级触发器包括随机选定的低频字符组合、视觉上的相似字符与人类无法感知的控制字符,比如 Unicode 与 ASCII

编码中的零宽度字符. 之后, BadNL^[39]使用随机低频词作为触发器替换文本中较为重要的词语以减少文本拼写错误. Qi 等人^[40]提出的同义词替换后门 LWS 为每个待替换词生成一组候选同义词, 尽可能地保留了文本原始语义. 之后, 一些研究使用句子作为触发器^[39,41], 保证了触发文本的连贯性, 比如 Dai 等人^[42]在 AddSent 中将预定义的触发句随机插入文本作为中毒样本. 之后, 为了提高攻击隐蔽性, 并减少后门误触发概率, Yang 等人^[43]在 SOS 中使用多个触发词的子序列与干净样本构成负样本, 通过负数据增强提高了攻击的鲁棒性, 保证当所有触发词同时出现在文本中才能激活后门.

虽然句子触发器避免了文本拼写错误与结构错误, 但插入的语义无关触发句可能破坏了文本的上下文连贯性. 此外, 频繁出现的触发文本成为后门文本的显著特征, 使得攻击易于检测. 对此, 研究人员利用文本语言风格、句法结构等抽象特征构建不可见后门. 他们发现相同语言风格与句法结构的文本在动词时态、情感词等方面表现相似^[42], 可以用于触发器设计. Qi 等人^[44]提出的 Hidden Killer 利用句法转述模型将正常样本转录为预先指定的低频率句法结构. Pan 等人^[20]提出的 LISM 使用文本风格转移模型生成指定语言风格的句子. 句法结构与语言风格后门削弱了触发文本显性特征与后门行为之间的强相关性, 保留了语句的原始语义. Zhou 等人^[45]提出的 NURA 将 Seq2Seq 模型预测的下一句输出作为触发器, 保证了中毒文本上下文相关性.

另一方面, 研究人员针对预训练微调中广泛使用的大模型提出了可迁移的后门攻击, 使得模型微调不会影响触发器在下游模型中的性能. Chan 等人^[46]和 Jin 等人^[47]提出的 CARA 和 BadMatch 将触发特征直接嵌入干净文本的潜空间, 引导模型生成连贯的中毒样本. Yao 等人^[48]、Shen 等人^[49]和 Chen 等人^[50]提出的潜在后门将目标标签与预训练模型中间表示相关联. 当下游模型包含攻击者的目标标签时, 与目标标签相关联的中间表示会自动激活后门, 使后门迁移到任何包含目标标签与相关中间表示的下游任务上. 这类潜空间后门能够自适应不同数据集之间的差异, 从而实现后门攻击在不同数据集与模型上的鲁棒性.

此外, 由于语法差异, 目前文本后门攻击仍以英文文本为主. 针对中文文本, Liu 等人^[51]提出了相应的文本对抗策略用于构建中文中毒文本. 他们选择文本中最重要的汉字, 通过拆分偏旁、同音字替换、相似拼音替换构造出了深度学习模型无法区分的扰动文本.

综上所述, 在计算机视觉领域, 大部分研究^[22,23,28,34]通过图像处理技术、数字隐写技术、生成对抗网络、频域转换等技术, 以添加人类不可见扰动的方式实现不可见后门攻击; 在自然语言处理领域, 一些研究利用句法结构、文本风格、词嵌入潜空间这类抽象特征构建文本不可见后门^[44,45]. 其中, 空间域的不可见后门在频域中呈现出与原始图像不同的特征, 因此一些频域异常检测方法能够识别这类后门攻击. 而频域的不可见后门需要实现多次空间域与频域转换, 攻击成本也会更高. 另一部分研究^[29,30]则将后门攻击视为一个范数约束最优问题, 在解答空间中搜寻最优解. 然而上述攻击在构建中毒样本时通常需要修改标签, 也提高了被检测的风险. 因此为了提高后门攻击在实际中的可行性, 研究人员以不修改标签的方式实现后门攻击.

3.1.2 干净标签后门

根据攻击者是否修改数据标签, 后门攻击分为中毒标签后门与干净标签后门. 中毒标签后门中攻击者毒害非目标类样本并将标签修改为目标类别; 干净标签后门中攻击者直接毒化目标类样本而不修改标签. 早期后门攻击^[2,21]大多属于中毒标签后门, 要求攻击者操纵数据标注过程. 受限于数据标注权限以及标签与样本内容不一致带来的后门检测风险, 攻击者提出了干净标签后门.

(1) 计算机视觉中的后门攻击

- 图像空间域后门攻击. 一些研究利用目标类与非目标类样本在特征空间的差异实现干净标签后门, 通过插值算法、生成网络在目标类上生成不规则扰动, 使得中毒样本与非目标类在特征空间相似. Turner 等人^[52]和 Saha 等人^[53]提出的插值后门构建了位于非目标类与目标类之间的中毒样本, 使得中毒样本在特征空间中靠近非目标类, 而在视觉上与目标类相近. Ning 等人^[54]提出的 Invisible Poison 使用自编码器将触发器图案转换为噪声并嵌入目标类随机图像. Tan 等人^[55]提出的对抗性后门通过自适应对抗训练算法优化损失函数, 最小化中毒样本与干净样本在特征空间的差异. Zhao 等人^[56]在目标类别视频的触发区域上, 使用非目标类对触发区域迭代扰动生成通用对抗触发器, 使得中毒视频在特征空间靠近非目标类.

由于深度学习模型的不透明度, 后门模型可能学习到非攻击者预期的触发特征, 一些浅表性的非触发器特征也能激活后门. Cheng 等人^[57]提出的 DFST 先使用风格转换模型生成特定风格的中毒样本, 再根据模型在中毒样本与干净样本上神经元的差值筛选敏感神经元, 从而削弱目标标签与浅层特征之间的联系.

- 图像频域后门攻击. Hammoud 等人^[58]提出的 CYO 根据深度学习模型对不同频率的扰动敏感度差异, 选择影响最大的频率作为触发器. 类似地, Liu 等人^[59]提出的隐蔽低通后门在 RGB 图像 3 个色道分别选择合适的通带半径构造触发器. 由于 CYO 和低通后门仍需要修改数据标签, Yue 等人^[60]提出了干净标签后门 IBA, 根据目标类别图像的频率分布选择合适的频率强度作为触发器.

(2) 其他领域中的后门攻击

相较连续且规则的二维图像, 三维数据呈现稀疏性与不规则性, 二维空间后门攻击不能直接应用到三维点云与三维网格分析中. 由于点云的轻微扰动显著影响深度学习模型决策边界, 旋转点云、向点云插入额外点就可以实现后门攻击. Li 等人^[61]提出的 PointBA 对点云 z 轴进行微小旋转变换. Fan 等人^[62]提出的几何特征变换后门 MGF-MBA 和拓扑连接重映射后门 ReMBA 分别改变了三维网格顶点位置与拓扑结构, 从而影响模型分类.

在二进制代码方面, 深度学习模型应用于大规模恶意软件分类任务. Sasaki 等人^[63]通过修改特定类别恶意软件标签, 在模型更新阶段引导分类器对该类恶意软件的决策边界发生偏移. Li 等人^[64]提出的干净标签后门将合适的特征值作为 API 调用插入软件以保证中毒软件的正常运行. 他们通过标签反转算法引导软件自动标注算法将中毒样本标注为良性样本, 本质上仍需要修改标签. 之后, Tian 等人^[65]根据稀疏子空间聚类边缘的样本在主要特征与次要特征上的差异提出了干净标签后门. 他们通过添加新节、插入字符串和注册组件的方式, 将触发器添加到良性聚类最边缘的软件, 使得中毒软件在主要特征上与目标恶意软件类相近.

综上所述, 干净标签后门^[52,53,56]利用数据映射到模型特征空间的聚类性质, 通过插值算法、生成对抗网络等方式缩小非目标类样本、中毒样本与目标类之间的距离. 攻击者将后门攻击视为主任务与后门任务的双重优化问题. 计算机视觉领域中的干净标签后门通常在一类目标样本上使用连续域搜索算法搜寻通用最优解. 由于常见的连续域优化算法需要较大的解空间, 构造最优触发样本会耗费大量计算成本. 攻击者通过优化搜索空间提高干净标签后门的性能. 与此同时, 由于自然语言处理和二进制代码分析中样本数据的离散性, 常用的连续域搜索算法无法直接应用在离散数据上, 因此针对文本和二进制代码这类离散数据, 攻击者采用替代搜索的方式寻找近似解. 考虑到近似解无法达到最优攻击效果, 攻击者仍需通过优化算法达到替代解与攻击性能之间的平衡. 由于上述攻击中, 所有样本使用相同的触发器, 容易被现有的后门防御方法所识别. 对此, 研究人员提出了样本特异性后门.

3.1.3 样本特异性后门

根据触发器生成算法与样本的关系, 后门攻击分为样本不可知后门与样本特异性后门. 样本不可知后门指触发器与样本无关, 即所有样本采用相同的触发器. 样本特异性后门指触发器与样本相关.

一些工作致力于研究多目标后门与多触发器后门, 通过多个潜在目标标签与多个触发器, 增强了触发特征与目标类别之间联系的隐蔽性. 在计算机视觉领域, Salem 等人^[66]提出的多目标后门在图像多个大小相同的不重叠子区间上随机添加了一组噪声, 并根据噪声位置确定不同的目标标签. 之后, Salem 等人^[66]改进触发生成算法, 在 c-BaN 中根据目标标签与随机噪声生成样本相关的触发噪声. Nguyen 等人^[67]提出的 IAD 根据输入图像生成自适应触发器. 在此基础上, Doan 等人^[68]在 LIRA 中改进了模型训练方式, 将后门攻击视为非凸优化问题, 在高度非线性参数空间中利用交替随机优化联合训练触发生成算法与目标模型以得到最优解. Gong^[69]在 RobNet 中构建了与模型相关的扰动噪声, 使得后门相关神经元具有较大权重和激活度. 他们还通过位于不同触发区域的相同扰动减小了模型对触发器轻微位移的敏感性. 在频域方面, Xue 等人^[70]提出的频域多目标后门与多触发器后门将 RGB 图像的 3 个色道映射到不同的触发目标, 并通过不同的触发器激活对应后门.

自然语言处理领域中, 随着大语言模型 (large language model, LLM) 的广泛应用, 一些研究开始借助大模型强大的语言理解能力与生成能力辅助样本特异性后门. Chow 等人^[71]提出的多目标后门 Imperio 向 LLM 中输入与目标标签相关的后门文本指令, LLM 根据指令生成满足需求的后门指令嵌入表示, 从而构建满足需求的触发器.

综上所述, 大部分后门攻击使用相同的触发器与单一目标标签, 而样本特异性后门研究较少. 目前, 样本特异

性后门大多表现为多触发器后门与多目标后门. 一些工作^[66,69,70]分别设定多个触发器与多个潜在目标, 将单一目标与单个触发器之间的强关联性分散到多个触发器上, 提高了后门攻击的隐蔽性. 在样本特异性后门中, 由于存在多个触发器, 攻击者需要保证触发唯一性, 即非样本对应的触发器无法激活后门. 此外, 随着大语言模型的兴起, 一些研究开始利用大语言模型的语言分析和生成能力辅助样本特异性后门. 相较前两种后门攻击, 样本特异性后门对模型正常性能影响较大, 攻击者需要平衡后门模型在正常样本上的性能与后门的隐蔽性.

3.2 模型修改后门攻击

大多数后门攻击^[2,21,56]要求攻击者接触并修改模型训练数据. 受限于数据隐私保护, 现实场景中攻击者难以向目标模型训练数据中注入非法数据. 因此, 在数据访问受限时, 攻击者收集与目标模型任务相关的其他数据作为替代数据集并构建触发器. 当攻击者无法获得任何替代数据时, 他们逆向模型以重构训练数据. 本文根据攻击者是否在替代数据集上重训练模型将模型修改后门进一步划分成重训练后门与免训练后门.

3.2.1 重训练后门

深度学习模型中神经元的激活权重影响着模型输出, 一些研究通过修改模型中对结果有较大影响的神经元取值, 在模型中植入后门^[24,72-74]. 计算机视觉处理领域中, Liu 等人^[72]在 Trojan 中根据与触发器区域强相关的神经元逆向生成触发器、干净样本以及中毒样本, 构建出后门神经元到目标输出的后门路径. 在此基础上, Li 等人^[24]选择模型倒数第 2 层权重最大的神经元作为触发器, 使用正则化约束削弱触发器可见性. Lv 等人^[73]提出的 DBIA 设计了一种注意力最大化通用触发器, 修改注意力层后的线性层中的重要神经元, 使得目标模型仅关注触发器. 之后, Lv 等人^[74]提出了数据集动态调节算法删除在输入域和输出逻辑域中的相似样本, 减小替代数据集规模, 削弱了替代数据与原数据分布差异较大对攻击的影响.

针对图像压缩等底层任务, Yu 等人^[75]提出的 BA-DIC 攻击常见的图像压缩模型, 在频域生成高质量中毒图像, 并利用后门图像压缩模型间接攻击了人脸识别和图像分割等下游任务. 在自然语言处理任务上, Yang 等人^[76]提出的 DFEP 使用梯度下降计算能够使得开源语料库子集分类错误的超级词嵌入. Li 等人^[77]提出的分层权重投毒后门 LWP 将触发器单独注入模型第 1 层, 防止微调对高层后门权重的影响. 整体上, 重训练后门要求攻击者缩减替代数据集规模, 在模型最后几层上微调模型, 以提高后门植入效率.

综上所述, 数据访问受限时, 可以收集与任务相关或无关的开源数据作为替代数据集^[73,74]. 当无法获得开源数据时, 他们逆向模型以重构训练数据. 攻击者在替代数据集中投毒, 并重训练目标模型以向模型中注入后门. 由于重训练后门需要在整个替代数据集上重训练模型, 消耗大量计算资源与时间, 攻击者需要采用各种数据集缩减算法减小数据集规模, 并尽可能拉近替代数据集与原数据的分布. 此外, 大部分重训练后门仅对模型最后几层进行重训练, 以提高后门攻击效率.

3.2.2 免训练后门

为了进一步提高攻击效率, 一些研究提出了免训练后门^[78-80], 向目标模型中插入可用的后门模型组件, 或者根据触发器形式直接修改模型结构与参数. 计算机视觉任务中, Tang 等人^[78]向目标模型中注入特定的后门片段, 直接修改了目标标签与模型后门模式之间的联系. Hong 等人^[79]直接修改预训练模型参数, 最大化干净样本与后门样本之间的激活差异, 从而在触发器和目标输出之间创建决策路径. 在自然语言处理任务上, Kurita 等人^[80]提出的权重投毒后门 Ripples 选择低频词作为触发器, 并用与目标标签强相关的单词平均嵌入替代触发词的嵌入表示, 增强了后门隐蔽性. 与其他后门相比, 免训练后门更节省时间, 拥有更大的自由度.

针对模型架构原生缺陷, 攻击者通过模型参数修改, 提出了模型结构上的通用后门攻击. Wei 等人^[81]的模型跨层混叠后门指出深度神经网络中跨卷积层和池化层下采样技术面临失真现象. 他们在网络跨层中添加扰动, 直接将干净输入修改成目标类别的中间表示. 这类模型结构后门威胁到主流预训练模型和微调子模型, 威胁性巨大.

另一类免训练后门^[82-85]通过真实物理攻击直接影响现实模型. Li 等人^[82]利用人类视觉和传感器捕捉图像的差异, 在现实中控制红绿蓝 LED 闪烁频率和亮度, 以物理手段引导照相机传感器拍摄中毒条纹图像. 通过物理攻击, 攻击者可以对部署后的深度学习系统进行攻击. Rakin 等人^[83]提出的 TBT 筛选与目标类别具有紧密联系的神

经元,并在模型部署运行时利用 Rowhammer 攻击修改主存内模型相关神经元的比特位。类似地,Chen 等人^[84]提出的 ProFlip 选择模型最后一层的重要神经元,为目标类别构建有效的触发模式,并在模型部署后通过激光束或 Rowhammer 攻击翻转主存中的参数敏感比特序列从而注入后门。Bagdasaryan 等人^[85]提出的模型代码修改后门利用第三方代码平台安全审查漏洞,向目标模型提交恶意损失函数代码,从而在模型训练过程中自动创建中毒训练数据。这类真实物理攻击要求攻击者利用现实设备漏洞或模型存储漏洞,在实际中难以实现。

综上所述,免训练后门直接在深度学习模型中插入额外的后门模型片段,或者根据触发器形式直接修改模型参数,是最快速的后门攻击方式。攻击者能够向模型参数中引入任意扰动,能够抵抗现有的针对数据投毒的后门防御方法。除了根据特定触发器设计模型修改方式,一些研究^[81-84]针对深度学习系统结构缺陷提出了可行的通用后门攻击,这类后门攻击危害性巨大。然而,免训练后门要求攻击者对目标模型具有较大控制权限,在现实中难以实现。需要直接影响现实模型的真实物理攻击。此外,除了需要拥有物理设施权限,还应考虑到潜在的影响因素,比如传感器存在的天然噪声对后门攻击的影响。

3.3 仅数据投毒后门攻击

当缺乏模型相关知识时,攻击者仅能通过数据投毒向训练数据中注入高质量的中毒样本,期望影响目标模型对特定类别样本的分类结果。攻击者根据可获得的替代模型优化触发生成算法,构建高质量的触发器。与第 3.1 节一致,我们将仅数据投毒后门分成可见后门与不可见后门、中毒标签后门与干净标签后门、样本不可知后门与样本特异性后门。

3.3.1 不可见后门

图像空间域中,Chen 等人^[2]提出的 Blended 将图案触发器以不同透明度与干净样本混合,Saha 等人^[86]提出的插值后门 SSL 在目标类图像的随机位置添加触发器构成中毒图像。图像频域中,Hou 等人^[87]在 BTI-DFT/DCT 中向 RGB 图像 3 个通道注入触发器。Wang 等人^[88]提出的不可见后门 FTrojan 在 YUV 图像 UV 通道的中频和高分量上注入触发器,保证了中毒图像的隐蔽性。在三维数据分析中,Xiang 等人^[89]提出的 PCBA 在点云的不同位置插入不同形状的点群。Gao 等人^[90]提出的 IRBA 随机旋转与缩放点云表面的点,并通过高斯核函数生成平滑的中毒样本,以不可感知的形式改变点云局部结构,保证了中毒样本与目标类别之间的稳定映射。

自然语言处理机器翻译与问答任务中,Li 等人^[91]根据 NLP 平台接受文本编码方式构建了两类隐蔽后门攻击。对于 Unicode 字符,他们使用同形异义词替换原始文本,引导模型使用 [UNK] 标记同形字;对不接受 Unicode 的系统,他们将随机关键词作为前缀输入大语言模型生成具有上下文感知的中毒句子。Li 等人^[92]提出的 IMTM 则针对图像文本多模态系统构建了双重隐蔽后门。他们将特定字符转化的噪声嵌入图像,并使用模型无法识别的视觉相似的象形文字替换描述文本中的相应字符。Mei 等人^[93]提出的 NOTABLE 将 NLP 提示工程中的高频词作为触发器,并将其与标签映射器绑定,使得任意包含相应标签的微调模型都存在着对应的触发提示策略。

3.3.2 干净标签后门

计算机视觉领域中,Barni 等人^[94]提出的干净标签后门 SIG 通过向目标类干净样本添加轻微抖动构建隐蔽的中毒样本。Zhang 等人^[95]提出的基于对抗扰动的通用后门 AdvDoor 能够在训练子集上搜索将该子集所有样本分类到目标类别的最小扰动。之后,Shafahi 等人^[96]提出的干净标签后门对非目标类图像与目标类图像使用特征碰撞生成中毒图像,使得中毒图像在潜在空间接近目标类,在视觉上接近非目标类。

针对恶意软件分类器,D'Onghia 等人^[97]提出两种干净标签后门攻击,分别将静态字节序列和基于潜在空间特征值动态生成的字节序列添加到软件结尾与不同分节的填充区域中,在保证软件正常运行的同时,提高了后门攻击隐蔽性。

3.3.3 样本特异性后门

计算机视觉领域中,一些工作使用生成对抗网络生成与样本密切相关的触发器。Li 等人^[98]提出的 ISSBA 将目标类别字符串与干净图像输入预训练编码解码网络,生成与样本密切相关的中毒图像,提高了攻击隐蔽性。类似地,Ma 等人^[99]提出的 DIHBA 利用生成对抗网络生成目标模型决策边界附近的触发图像,从而引导模型在目标类

别上的决策边界发生偏移.

3.4 后门攻击小结

计算机视觉和自然语言处理是后门攻击的两大主流研究领域. 两个领域的后门攻击都要满足攻击有效性和隐蔽性两个基本要求. 近年来, 研究人员从攻击有效性和攻击隐蔽性两个方面提出了许多先进的后门攻击方法 (见表 3). 根据现实中攻击者能力, 这些后门攻击可以分为全过程可控后门、模型修改后门和仅数据投毒后门这 3 类. 目前, 大部分后门攻击需要构建触发器, 向训练数据中注入一定比例的中毒样本. 后门攻击性能极大地依赖于触发器的质量.

表 3 后门攻击相关文献整理

分类	方法	相关研究	年份	后门攻击特征			触发器属性				模型训练		应用场景
				不可见	标签改变	样本特异性	位置固定	内容固定	空间域	频域	权值修改	重训练	
全过程可控后门	图案扰动	BadNets ^[21]	2017	○	●	○	●	●	●	○	○	●	CV
		Lin等人 ^[27]	2020	○	●	○	●	●	●	○	○	●	CV
		Refool ^[23]	2020	●	○	○	●	●	●	○	○	●	CV
		FaceHack ^[28]	2020	●	●	●	●	○	●	○	○	●	CV
		TDSC ^[24]	2021	●	●	○	○	●	●	○	○	●	CV
		BppAttack ^[22]	2022	●	●	●	○	○	●	○	○	●	CV
	不规则扰动	Zhong等人 ^[29]	2020	●	▲	●	●	▲	●	○	○	●	CV
		WaNet ^[26]	2021	●	●	○	●	●	●	○	○	●	CV
		Carlini等人 ^[32]	2022	●	●	○	●	○	●	○	○	●	CV/NLP
		DT-IBA ^[25]	2024	●	●	○	●	●	●	○	○	●	CV
		Invisible Poison ^[54]	2021	●	○	○	●	○	●	○	○	●	CV
		SGBA ^[30]	2021	●	○	○	●	●	●	○	○	●	CV
		Salem等人 ^[66]	2022	○	○	●	○	○	●	○	○	●	CV
		Zeng等人 ^[33]	2021	●	●	●	●	○	○	●	○	●	CV
		FIBA ^[38]	2022	●	●	○	●	○	○	●	○	●	CV
		WaveAttack ^[37]	2023	●	●	○	●	○	○	●	○	●	CV
		DUBA ^[36]	2023	●	/	○	○	○	○	●	○	●	CV
		IBA ^[60]	2023	●	○	○	●	○	○	●	○	●	CV
	文本扰动	BadNL ^[39]	2021	○	●	○	○	●	○	○	○	●	NLP
		LWS ^[40]	2021	/	●	○	○	○	○	○	○	●	NLP
		AddSent ^[42]	2019	/	●	○	○	●	○	○	○	●	NLP
		Hidden Killer ^[44]	2021	●	●	○	○	●	○	○	○	●	NLP
		LISM ^[20]	2022	●	●	○	○	●	○	○	○	●	NLP
		NURA ^[45]	2023	●	●	●	○	○	○	○	○	●	NLP
		Imperio ^[71]	2024	○	●	●	○	○	○	○	○	●	NLP
	基于优化	Tan等人 ^[55]	2019	○	●	○	●	○	●	○	○	●	CV
		Sasaki等人 ^[63]	2019	/	●	○	●	/	/	/	○	●	MD
		CARA ^[46]	2020	○	●	●	/	/	/	/	○	●	NLP
		LIRA ^[68]	2021	●	●	●	●	/	●	○	○	●	CV
		BadEncoder ^[31]	2022	○	/	●	●	○	●	○	○	●	CV/NLP
		Li等人 ^[64]	2022	/	●	●	●	/	/	/	○	●	MD
	可迁移	Shen等人 ^[49]	2021	●	●	/	●	●	●	○	○	●	NLP
		BadPre ^[50]	2022	●	●	/	●	●	●	○	○	●	NLP

表 3 后门攻击相关文献整理 (续)

分类	方法	相关研究	年份	后门攻击特征			触发器属性				模型训练		应用场景
				不可见	标签改变	样本特异性	位置固定	内容固定	空间域	频域	权值修改	重训练	
模型修改后门	重训练	Trojan ^[72]	2018	/	/	/	●	/	●	○	○	●	CV/NLP
		DBIA ^[73]	2021	/	/	/	/	/	●	○	○	●	CV
		Lv等人 ^[74]	2023	○	●	/	/	/	●	○	○	●	CV/NLP
		DFEP ^[76]	2021	/	/	/	/	/	/	●	○	●	NLP
	免训练	RIPPLES ^[80]	2020	/	/	/	/	/	/	/	●	○	NLP
		Tang等人 ^[78]	2020	/	/	/	/	/	/	/	●	○	CV/NLP
		Hong等人 ^[79]	2022	/	/	/	/	/	/	/	●	○	CV
		Wei等人 ^[81]	2023	/	/	/	/	/	/	/	●	○	CV/NLP
	图案扰动	Blended ^[2]	2017	▲	●	○	●	●	●	○	○	○	CV
		SSL ^[86]	2022	●	○	○	●	●	●	○	○	○	CV
仅数据投毒后门	不规则扰动	FTrojan ^[88]	2022	●	○	○	●	●	○	●	○	○	CV
		BTI-DFT/DCT ^[87]	2023	●	○	○	●	●	○	●	○	○	CV
		PCBA ^[89]	2021	/	/	/	/	/	●	○	○	○	CV
		IRBA ^[90]	2022	●	●	/	○	○	●	○	○	○	CV

注: ●表示满足条件; ○表示不满足条件; ▲表示两者都有; “/”表示非文献关注重点; 模型修改后门中的重训练是指攻击者在无法获得训练数据的情况下, 在替代数据集上微调模型, 该项主要用于与“模型修改后门中的免训练”进行区分; 模型训练的重训练是指在攻击者在能够获得的训练数据上从头开始训练模型或微调模型

数据表征形式与不同任务之间的差异使得计算机视觉与自然语言处理后门攻击在触发器设计与攻击范式上存在差异. 一方面, 图像数据在向量空间具有连续性, 而文本数据在词嵌入空间表现为离散性. 因此大部分计算机视觉后门无法直接移植到文本后门. 比如, 计算机视觉中的 BadNets 攻击可以在图像上添加一小块随机图案, 而文本领域相似的攻击则是向文本中插入预定义的字符. 在自然语言处理领域, 攻击者试图通过句法结构、文本风格、词嵌入潜空间等抽象特征, 保证中毒文本的可读性与连贯性. 由于文本长度受限, 文本触发器的尺寸和形式存在着诸多限制. 另一方面, 图像后门与文本后门在攻击隐蔽性上具有不同的表现形式. 比如, 图像不可见后门可以定义为中毒样本与干净样本之间的像素距离, 而文本不可见后门可以定义为语句困惑度.

整体上, 计算机视觉与自然语言处理领域的大部分后门攻击仍属于单触发器单目标攻击, 这就导致触发器与目标标签之间呈现强相关性. 此外, 目前大多数后门防御是基于该假设构建的, 因此这类单触发器单目标后门攻击很容易被现有后门防御所识别. 对此, 研究人员通过生成网络构造与样本相关的触发器, 并设立多个潜在目标提高攻击隐蔽性.

受到真实场景攻击者能力的限制, 研究人员提出了不需要修改标签的干净标签后门攻击. 在能力进一步受限情况下, 即数据访问受限时, 研究人员利用模型神经元激活状态与输出之间的关联, 通过修改模型权重与参数直接将后门植入模型.

后门攻击本质上可以视为解决模型主任务与后门任务之间的优化问题. 基于扰动的数据投毒后门引导模型将触发器特征作为分类的主要因素. 基于优化的后门则通过广度与深度搜索, 在搜索空间中计算得到两层优化问题的全局最优解或局部最优解. 在此过程中, 研究人员使用了包括采样技术、近似方法等技术, 从理论上降低求解的时间与空间复杂度. 但如何设计实现高质量的触发生成算法, 仍然是后门攻击中的重要挑战.

4 深度学习中的后门防御分类

针对后门攻击, 研究人员提出了一系列后门防御方法. 后门防御中, 防御者通常拥有一小部分干净验证集用于验证模型是否受到攻击, 但是防御者不了解后门攻击触发器与中毒样本. 整体上后门防御分为基于输入的防御与

基于模型的防御,前者指防御者判断输入数据是否可疑,后者指防御者判断模型中是否存在后门.基于输入的后门防御利用后门样本与干净样本在特征空间的差异,在模型接受可疑样本并输出结果前检测输入样本,并拒绝可疑后门样本.基于模型的后门防御利用后门模型与良性模型在不同样本上的行为差异,检测可疑模型是否存在后门.

根据后门防御策略,又可进一步将后门防御方法细分为后门检测与后门清除方法.后门检测通过统计分析、异常检测等技术判断输入样本是否包含可疑触发器,以及模型是否含有后门.后门清除则通过数据预处理、调参等技术处理后门样本与后门模型.

后门攻击中攻击者试图通过扰动与优化技术拉近中毒样本与目标类之间的距离.但模型训练过程中,中毒样本与干净样本存在着不可避免的差异,具体表现为模型梯度与神经元激活状态的差异.

(1) 后门模型在中毒样本上的收敛速度显著快于干净样本,即后门模型在触发器与目标类别之间存在明确的相关性,模型学习中毒样本特征更快.

(2) 中毒梯度和干净梯度存在方向差异,即目标类别与非目标类别之间的差异.

(3) 与后门相关的神经元对对抗扰动更为敏感,即非触发器的合适扰动也能误激活后门.

研究人员针对上述差异提出对应的防御策略.一方面,他们利用模型在不同样本上收敛速度的差异分离中毒样本与干净样本.另一方面,他们直接阻碍触发器与目标类别之间的相关性,从而无效化后门.本节按照计算机视觉领域和自然语言处理领域的顺序,依次总结各领域的后门防御方法.

4.1 基于输入的后门防御

基于输入的后门防御指防御者在模型部署前对数据的检测与清洗.大部分基于输入的后门防御利用同一目标类下,中毒样本与干净样本在特征空间的差异,通过降维、协方差分析等方法分离中毒样本与干净样本,并在剥离中毒样本后的数据集上训练得到干净模型,实现后门防御.

(1) 计算机视觉领域

一方面,一些工作根据中毒样本与干净样本在收敛速度及梯度方向上的差异,通过激活函数聚类分离中毒样本与干净样本.Chen 等人^[100]提出的 Activation Cluster (AC) 根据模型最后一层隐藏层在数据集上的激活结果划分中毒样本和干净样本两个聚类.具体来说,他们通过独立成分分析对每类标签进行降维,并使用 k-means 算法进行聚类.由于中毒样本依赖触发器与目标标签的关联引导模型将其分类到目标类,这与目标类其他干净样本存在差异,因此在特征空间中中毒样本与干净样本会分成两个聚类.AC 通过降维聚类方法,能够有效抵御 BadNets 等中毒标签后门.Tran 等人^[101]提出的 Spectral 利用后门攻击在神经网络特征表示协方差频谱中遗留的频谱签名识别和移除后门输入.他们指出同一目标标签的干净样本与中毒样本的特征表示分布是光谱可分离的,即对输入表征的协方差矩阵进行奇异值分解能够识别中毒样本触发器提供的强烈信号,从而筛除中毒样本.AC 和 Spectral 根据触发器与目标标签之间人为构造的强关联性在特征空间中与其他干净样本存在差异这一性质,进一步采用降维方法凸显这种信息差异,最后根据差异清除中毒样本.由于 AC 和 Spectral 都依赖于投毒数据的特征分布情况,当触发器位置发生变化、多个触发器对应单一目标、投毒比例增大时,AC 和 Spectral 的检测性能会下降.之后,Pan 等人^[102]提出的 ASSET 通过梯度变化主动诱导模型在后门样本与干净样本之间的不同行为,能够检测当前最先进的干净标签攻击.防御者通过两次优化分离中毒样本与干净样本,他们先最小化干净验证集的损失,然后最大化中毒标签数据集.由于干净样本与中毒数据集的干净部分都来自相同的分布,第 2 次优化对干净样本的影响将会与第 1 次优化抵消,因此两次优化整体实现了中毒样本的损失最大化.根据中毒样本与干净样本的不同损失值,就能有效地清除中毒样本.Ma 等人^[103]提出的 Beatrix 利用后门样本与干净样本在 Gram 矩阵特征上的差异分离后门样本,能够抵抗样本特异性后门攻击.他们将后门防御视为离群点检测问题,并指出 Gram 矩阵这一偏心协方差矩阵包含单独特征信息与特征相互相关性,可以使用 Gram 矩阵和它的高阶信息用于捕捉中毒样本与干净样本的差异.Beatrix 能够深度挖掘特征高阶信息,除了触发器固定的单触发器后门攻击,Beatrix 还能有效防御 IAD 这类样本特异性后门攻击.此外,AC、Spectral、ASSET 和 Beatrix 对自然语言处理领域的后门攻击同样具有良好的防御效果.

另一方面, 由于后门相关神经元对抗扰动的敏感性, 研究人员可以通过构建特定扰动引导模型输出, 从而区分中毒样本与干净样本。对此, Gao 等人^[104]提出的 STRIP 在输入样本上添加不同扰动, 由于中毒样本添加扰动后更容易被分类为目标类别, 其结果的平均熵值较小, 因此 STRIP 可以根据中毒样本与干净样本在模型分类中的熵值大小区分中毒样本。凭借中毒样本与干净样本分类熵值差异, STRIP 能够有效防御多触发器多目标后门攻击与多触发器单目标后门攻击。此外, STRIP 还能够防御自然语言处理领域的后门攻击。Chou 等人^[105]提出的 SentiNet 使用 Grad-CAM 模型可解释性技术与选择性搜索图像分割技术搜索输入图像中对预测结果影响最大的区域。具体来说, 他们先通过图像分割技术得到输入图像最可能的两个标签, 然后找到与预测标签强相关的部分, 接着他们去除相同预测标签图像中存在的共同部分, 并保留独一无二的部分作为潜在的触发器区域。最后他们将上述敏感区域添加到一组干净输入中, 根据模型分类结果判断上述区域是否包含触发器。如果错误分类数和平均置信度越高, 则该敏感区域极大概率为触发器区域。考虑到 SentiNet 需要搜索图像敏感区域, 当触发器尺寸较大, 比如覆盖整个样本时, SentiNet 效果不佳。之后, Doan 等人^[106]在 Februs 中使用中和色彩的方框取代敏感区域, 并通过生成对抗网络修复输入图像, 清除图像中的触发器。Qi 等人^[107]提出的混淆训练方法 CT 对已中毒的数据集进行额外的投毒, 通过随机标注的干净样本扰乱模型对干净样本的分类, 从而突显后门样本与目标类别之间的强关联性。由于 CT 可以向数据集中注入各种类型的投毒数据, 该方法能够有效应对投毒标签后门、干净标签后门与样本特异性后门等大部分后门攻击。

另一部分研究^[108-111]在划分干净数据与中毒数据的基础上, 直接训练得到干净模型。Li 等人^[108]提出的 ABL 定义了学习干净数据与后门数据的双重任务, 在训练早期筛选具有低损失的中毒样本。他们认为与原始任务相比, 后门任务更容易, 后门攻击越强, 后门任务损失函数值下降得越快。因此, ABL 在标准训练过程中引入梯度上升机制, 即在早期训练中通过梯度上升将每个样本的损失值控制在某个阈值附近, 并筛选一部分损失最小的样本视为中毒样本, 接着在后期训练中最大化中毒样本与目标类映射关系的损失值以遗忘中毒样本, 从而实现防御。针对端到端的有监督训练, Huang 等人^[109]提出的 DBD 将训练过程解耦成自监督、有监督和半监督学习这 3 个阶段, 一步步地切断触发器与目标标签之间的联系。他们指出在无标签的中毒数据集上进行自监督学习能够阻碍后门任务的学习, 中毒样本会与其他样本靠近, 不会产生单独的聚类。因此 DBD 先在去除标签的数据上通过自监督学习得到特征提取器, 然后冻结特征提取器, 并在有标签数据上通过有监督学习训练其余的全连接层。最后他们分离高置信度样本与低置信度样本, 通过半监督学习微调模型以提高模型的正常性能。类似地, Gao 等人^[110]提出的 ASD 基于损失函数导向和元学习启发分割动态更新干净数据和投毒数据。他们指出 ABL 和 DBD 这类仅根据中毒样本与干净样本损失值差异的方法无法完全剔除中毒样本, 无法识别一些与中毒样本损失值接近的干净样本。因此 ASD 先根据损失函数初步划分干净数据池和中毒数据池, 然后构造一个与模型相同的虚拟模型在中毒数据池上训练, 并将损失值最小的几个样本添加到干净数据池中, 最终在干净数据池上继续训练得到干净模型。Zhang 等人^[111]提出的 CBD 定义了后门模型与干净模型分别学习触发器与目标类别之间的虚假相关性以及干净样本与输出之间的因果关系。具体来说, CBD 利用中毒样本与干净样本在早期梯度下降的差异, 先使用早期停止策略在中毒数据集上训练第 1 个模型, 用于捕捉后门相关性。然后他们通过最小化互信息来训练另一个干净模型, 尽可能最小化后门相关性对干净模型的影响。ABL、DBD、ASD 和 CBD 都是最大化地利用后门样本与干净样本在早期收敛速度上的差异, 在模型训练阶段抑制并削弱潜在后门样本的影响。由于上述攻击假设投毒样本在模型训练早期能够达到最小损失, 当数据集中存在中毒样本与其他无关标签时, 影响干净样本与中毒样本的分离, 因此这类方法面对负数据增强后的后门攻击时防御效果会下降。

(2) 自然语言处理领域

在自然语言处理领域, 面对 BadNL 字符级后门攻击, 防御者可以直接通过语法检查筛选拼写错误的文本, 之后, Qi 等人^[112]提出的 ONION 利用困惑度筛选明显不连贯的触发文本。困惑度用于衡量语言模型预测一个文本样本的能力, 困惑度越高表明语言模型的性能越好。ONION 使用 GPT-2 模型测试文本的困惑度, 它将一个单词的可疑分数定义为删除该单词后文本困惑度的下降值。可疑分数越大, 表明该单词极大概率是触发单词。Chen 等人^[113]提出的 BKI 统计样本中的重要关键词, 并记录所有样本的关键词、标签和平均重要性分数和关键词出现次数。对

于单触发器后门攻击,所有中毒样本中都存在触发器,且触发器对预测结果具有较大影响,因此触发关键词的平均得分会高于正常关键词. BKI 根据关键词重要性分数与出现频率筛选出最可疑的一个关键词,并删除包含该关键词的句子. ONION 和 BKI 适用于大部分将词语和句子作为触发器的后门攻击. 之后, Yang 等人^[114]提出的 RAP 构建了基于单词的鲁棒性感知扰动,将该扰动添加到样本上. 他们预先指定一个罕见的词并修改词嵌入向量来构建合适的感知扰动,并根据投毒样本和干净样本添加扰动后的分类差异区分投毒样本. 当干净样本加入该感知扰动后,其输出置信度会下降一定比例;而中毒样本加入该感知扰动后,其输出置信度几乎不变. Sabir 等人^[115]提出的 IT-DT 通过可视化注意力和梯度信息识别对抗样本的显著特征与干扰词,能够抵御使用对抗生成技术构建投毒数据的文本后门攻击. 之后, Pei 等人^[116]提出的可证明防御 TextGuard 使用 hash 函数将数据集划分为多个子数据集,以确保子数据集中包含较少的触发器. 他们根据子数据集构建多个基分类器,并通过多数投票进行最终预测. 具体来说,他们将输入文本中每个单词的哈希值视为索引,把具有相同索引的单词划分到同一个子集合,使得每个子集中包含了输入文本的一个单词序列. 这种划分方式能够保证所有输入中的相同词被划分到同一个子集. 由于后门触发词是有限的,这些触发词仅能出现在有限个子集中,因此通过投票机制能够保证最终结果不受触发词的影响. 由于 TextGuard 需要将文本字段进行划分,该方法在面对小尺寸的文本触发器时具有较好的效果;面对较大尺寸的触发词时,检测效果显著下降.

综上所述,基于输入的后门防御根据中毒样本与干净样本在模型训练阶段的梯度差异将两者分离,从而在模型训练过程中削弱中毒样本对模型的影响. 后门防御的目标是消除尽可能多的中毒样本,并较少的误删干净样本. 考虑到基于输入的后门防御无法完全清除数据集中的中毒样本,研究人员可以结合微调、剪枝进一步削弱后门. 其中, AC、Spectral、ASSET、Beatrix 和 STRIP 适用于计算机视觉和自然语言处理两个领域的后门攻击. AC、Spectral 根据中毒样本与干净样本在特征空间的差异在模型运行过程中分离中毒样本,这类方法能够抵抗中毒标签后门攻击,但无法抵抗干净标签后门攻击. 此外, AC、Spectral、STRIP、ONION 的防御效果还取决于算法中的阈值取值,因此它们的防御效果还受到不同数据集的影响.

4.2 基于模型的后门防御

基于模型的后门防御指防御者根据干净模型与后门模型的行为差异检测、移除或削弱模型中的后门. 防御者根据触发生成方法,为每个标签生成潜在的触发器,并根据模型的表现划分干净模型与后门模型. 防御者还可以根据干净模型与后门模型的混合数据集直接训练得到一个二元分类器.

(1) 计算机视觉领域

由于触发器与目标类别之间的强关联性使得后门模型面对轻微扰动时更易表现出后门行为,一部分研究^[117-120]利用生成网络构造潜在触发器,并根据可疑模型面对后门触发器的行为判断模型是否存在后门. 对比正常模型,在后门模型中向样本添加细微的扰动,会更容易分类到目标标签. Wang 等人^[117]提出的 Neural Cleanse (NC) 为每个标签恢复一个潜在扰动作为触发器,并根据不同标签之间扰动的差异判断模型中是否存在后门,识别可疑的目标类. 简单地说, NC 首先设定一个触发掩码区域,试图通过基于优化的触发反转算法搜索得到每个标签的潜在触发器,将触发器的像素数量作为触发器的大小,当存在一个异常小的值时,就判定该模型存在后门,该触发器分类对应的标签就是目标类别. 由于 NC 需要根据标签扰动差异判断触发器,当面对多个标签投毒的多触发器多目标后门攻击时,其防御效果不佳,因此 NC 仅适用于触发器尺寸较小的单触发器后门攻击. 针对还原的触发器存在的尺寸过大、分散等不足, Guo 等人^[118]在 TABOR 中通过新的正则化约束项缩小了搜索触发器对抗样本子空间的大小,加快最优解搜索效率,并能应对多触发器多目标后门攻击. TABOR 通过正则项约束触发区域大小和形状,减少对子空间中的对抗样本量,提高构造的潜在触发器的保真度. 之后, Wang 等人^[119]提出的 DL-TND 为每个类别与输入生成触发模式. 具体来说,针对某个目标标签,他们构造在一小部分干净验证集上的通用对抗扰动,并为每个样本构造单独扰动. 当某类别的触发模式与单个输入的触发模式相似时,即当通用对抗扰动与单独扰动具有强相似性时,他们就判定模型在该类别上存在后门. Liu 等人^[120]在 ABS 中向神经元引入不同程度的扰动,并根据神经元激活状态的变化幅度识别模型中与后门相关的神经元,能够有效防御不同触发器尺寸的单目标后门攻击. 他

们选定某个神经元,如果该神经元激活值在某个区间变化时对某个特定标签的影响远远大于其他标签,则该神经元是潜在的受损神经元.接着他们为所有潜在受损神经元反向构造触发器,并根据攻击有效性验证该神经元是否受到攻击,如果超过 90% 的样本都产生了同一错误分类结果,则表示该模型存在后门.相较 NC,其运行效率得到提升,但触发器位置发生变化时,ABS 的防御效果明显下降.同时,ABS 依赖于 ReLU 的假设,当面对缺乏 ReLU 激活神经元的模型结构时,该防御方法将无效.

除了利用后门模型对轻微扰动的敏感性,一些工作^[121,122]通过调节神经元权重以阻碍触发器到目标类的激活路径,从而削弱后门效果.后门神经元指面对干净样本不会激活,在触发器输入时才激活的神经元,因此中毒样本与干净样本在后门模型中具有不同的神经元激活路径.防御者可以通过移除这些后门神经元实现防御. Liu 等人^[121]提出的 Fine-Pruning (FP) 利用剪枝去除模型中大量冗余连接,他们记录神经元面对干净样本的激活程度,并按照平均激活程度的递增顺序迭代修剪神经元,直到模型在验证集上的准确率低于阈值时,结束剪枝.考虑到攻击者可能会隐藏真实的后门神经元,诱导防御者将诱饵神经元剪枝,剪枝完成后他们还在干净数据集上微调模型以去除后门.类似地, Wu 等人^[122]提出的 ANP 利用对抗训练在模型权重与偏置上添加扰动,诱导激活后门相关神经元,从而筛选相关神经元.具体来说,他们对模型中每个神经元进行小幅度的权重和偏置扰动,模拟后门样本,根据模型分类结果筛选后门相关神经元.然后通过梯度下降算法优化剪枝掩码向量,决定需要剪枝的神经元. FP 和 ANP 这类神经元剪枝方法能够在有限的干净数据和计算资源情况下,有效抵御不可见后门、干净标签后门和样本特异性后门,尽可能保证模型正常性能.但考虑到 FP 对模型最后一层神经元进行剪枝,攻击者可能通过绕过模型最后一层训练的方式将后门保留,该方法在面对自适应攻击时效果不佳.此外,当触发器位置发生变化时,神经元剪枝防御效果较差. Hong 等人^[123]根据后门神经元与干净神经元对扰动敏感的差异使用随机梯度下降算法剪裁神经元. Du 等人^[124]则是使用差分隐私机制约束单个梯度对模型的影响,从而减小中毒样本与干净样本梯度之间的差异,并在样本梯度中添加高斯噪声以削弱模型对不同样本的分类差异,提高模型鲁棒性.

(2) 自然语言处理领域

在自然语言处理领域,一些研究同样利用后门模型对轻微扰动的敏感性构造潜在触发器,以此判断模型中是否存在后门. Azizi 等人^[125]提出的 T-Miner 通过生成网络构造了能够使大部分样本分类错误的后门文本序列,并根据生成的后门文本序列判断可疑模型是否存在后门.具体来说,他们通过扰动生成网络为某类中的样本生成能够分类为目标类别的候选扰动,并筛选能够使大部分样本分类错误的扰动作为对抗性扰动.然后他们用目标类别样本的合成短语作为正常对抗扰动,与对抗性扰动一起输入后门识别器.如果对抗性扰动在特征空间表现为离群点,则模型存在后门.该方法能够有效抵御将词语作为触发器的后门攻击. Shen 等人^[126]提出的 DBS 为所有 token 定义一个凸包,并根据所有 token 在词表中的权重构造一个权重向量,该权重向量的维度等于词表中 token 的数量.然后他们反转该权重向量逆向生成潜在触发器,考虑到反转得到的权重向量可能无法与 token 对应,他们通过温度回滚技术动态调整温度,缩小对抗扰动搜索子空间大小,并得到了单个 token 触发的权重向量.该方法将离散空间的不可微优化目标转化为基于梯度搜索的优化目标,能够应对多种主流 NLP 任务中的后门攻击. T-Miner 和 DBS 都是通过触发器反转算法生成潜在的触发器.然而,与计算机视觉中的触发器反转算法相比,自然语言处理领域需要在离散空间进行不可微优化,优化策略更为困难.

(3) 多领域

一些研究^[127,128]通过已知的后门攻击构造包含多个干净模型与后门模型的训练数据集,将后门检测视为二分类任务. Xu 等人^[127]提出的 MNTD 在仅能获得模型权重参数的情况下,构建常见后门攻击的一般分布,最终在构建的后门替代模型与干净模型上训练出二分类器.具体来说,他们在干净验证集上以不同的模型初始化训练得到一系列干净模型,并根据一个通用后门攻击分布采样得到不同的后门替代模型.最后根据可疑模型在二分类器的分类结果判断模型是否存在后门. Kolouri 等人^[128]提出的 ULP 在多个数据集上训练得到干净模型与后门模型,以及对应的通用触发器,从而根据可疑模型对通用触发器的分类结果快速判断模型是否存在后门.如果触发器能够使得模型分类到目标类别,则该模型极大概率存在后门. MNTD 和 ULP 能够抵御计算机视觉领域和自然语言处理领域的大部分常见后门攻击,但由于它们需要通过已知攻击构建训练数据集,它们的防御范围严重依赖于已知的

后门攻击, 该类防御无法抵抗未知的后门攻击与自适应后门攻击.

防御者还根据后门攻击特征提出了针对性的后门防御方法. 针对模型修改后门, 比如模型部署后的比特翻转后门, Wang 等人^[129]提出的 Aegis 通过动态退出机制, 引导输入样本提前从模型不同层退出, 从而打乱攻击, 能够防御 TBT、ProFlip 等后门攻击. 具体来说, Aegis 在模型中间卷积层附加内部分类器, 当输入样本对标签的置信度高于出口神经元的阈值时, 中断训练直接输出标签. 由于 TBT 这类比特翻转后门中, 翻转的比特通常位于模型最后一层, Aegis 本质上是切断了触发器的激活路径. 对于对抗扰动后门, Xiang 等人^[130]在 PatchGuard 中使用小感受野的卷积层限制攻击者能修改的特征数量, 并通过鲁棒掩码屏蔽攻击者注入的异常特征.

综上所述, 目前基于模型的后门防御主要分为两类, 一类通过正常模型与后门模型对不同样本的行为差异检测后门, 另一类构建二元分类器直接识别正常模型与后门模型. 第 1 类防御, 比如计算机视觉中的 NC、TABOR、ABS 为不同类别的样本逆向生成可疑触发器. 这类方法需要搜索高维的对抗子空间, 时间成本较高. 而且面对触发器位置动态变化的后门攻击, 它们的防御效果不好. 第 2 类防御, MNTD 和 ULP, 它们的防御效果类似于白名单, 能够有效防御已知的后门攻击, 但无法应对未知攻击.

4.3 后门防御小结

现有后门防御按照防御对象分为基于输入的防御与基于模型的防御两类 (见表 4). 后门防御大部分假设攻击者使用单个触发器进行攻击, 因此后门防御能够应对 BadNets 这类触发器形状标准的小尺寸中毒标签后门.

表 4 后门防御相关文献整理

分类	原理	类型	相关研究	年份	方法
基于输入的防御	聚类特征差异	检测	AC ^[100]	2019	—
	后门模型扰动敏感	检测	STRIP ^[104]	2019	—
		检测	SentiNet ^[105]	2020	—
		检测/修复	Februus ^[106]	2020	—
		检测	Zeng 等人 ^[33]	2021	频域
	频谱痕迹	检测	Spectral ^[101]	2018	频域
	混合数据集训练得到干净模型	训练	ABL ^[108]	2021	—
		训练	DBD ^[109]	2022	—
		训练	ASD ^[110]	2023	—
		训练	CBD ^[111]	2023	—
	句子困惑度	检测	ONION ^[112]	2021	—
	主动诱导后门行为	检测/修复	ASSET ^[102]	2023	—
		检测	CT ^[107]	2023	—
	逆向触发生成算法	检测/修复	NC ^[117]	2019	—
		检测	TABOR ^[118]	2019	—
		检测	DL-TND ^[119]	2020	—
		检测/修复	DBS ^[126]	2022	文本
基于模型的防御	神经元激活状态	检测	ABS ^[120]	2019	—
	模型参数调节	修复	FP ^[121]	2018	—
		修复	Hong 等人 ^[123]	2020	—
	神经元敏感性差异	检测	Du 等人 ^[124]	2020	—
		检测/修复	ANP ^[122]	2021	—
	模型二分类器	检测	MNTD ^[127]	2021	—
		检测	ULP ^[128]	2020	—
	动态退出机制	防御	Aegis ^[129]	2023	—
	屏蔽触发特征	防御	PatchGuard ^[130]	2021	—

整体上, 计算机视觉领域和自然语言处理领域后门防御思路一致. 基于输入的防御利用中毒样本与干净样本在特征空间与模型训练中的差异, 在模型部署前与运行时拦截可疑后门样本. 基于模型的后门防御通过干净模型与后门模型在样本上的行为差异拦截可疑模型. 基于输入的后门防御以检测后门样本为主, 对于识别出的后门样本, 直接删除后门样本或是通过生成网络修复后门样本.

由于计算机视觉与自然语言处理数据在嵌入空间的差异, 其后门防御方法在实现细节上存在不同. 计算机视觉中的防御方法需要修改后才能应用在自然语言处理任务上, 比如 STRIP 在图像上添加随机扰动, 而在文本上添加的是 token.

目前, 计算机视觉和自然语言处理领域的后门防御方法存在缺乏可证明安全、泛化能力较差、防御效率较低等相似的问题. 表 5 展示了后门防御方法的适用范围与应用场景.

表 5 后门防御适用范围整理

攻击	攻击类型	后门攻击	领域	基于输入的后门防御	基于模型的后门防御
全过程可控后门	可见后门	BadNets ^[21]	CV	AC ^[100] , SentiNet ^[105] , Februus ^[106] , CT ^[107] , ASD ^[110] , ABL ^[108] , DBD ^[109] , CBD ^[111]	NC ^[117] , TABOR ^[118] , DL-TND ^[119] , ABS ^[120] , FP ^[121] , ANP ^[122]
			NLP	ONION ^[112] , RAP ^[114]	T-Miner ^[125]
			CV/NLP	Spectral ^[101] , ASSET ^[102] , STRIP ^[104]	—
	不可见后门	WaNet ^[26]	CV	ASD ^[110] , CBD ^[111] , CT ^[107] , DBD ^[109]	—
			CV/NLP	ASSET ^[102]	—
			CV	ASD ^[110] , Beatrix ^[103]	—
	词语级后门	BadNL ^[39]	NLP	TextGuard ^[116]	—
			NLP	—	DBS ^[126]
			NLP	TextGuard ^[116] , BKI ^[113]	—
	结构级后门	Hidden Killer ^[44]	NLP	TextGuard ^[116]	DBS ^[126]
			CV	CT ^[107] , DBD ^[109]	ANP ^[122]
			CV/NLP	ASSET ^[102]	—
模型修改后门	重训练后门	Trojan ^[72]	CV	CBD ^[111] , CT ^[107] , ASD ^[110] , ABL ^[108] , Beatrix ^[103]	ANP ^[122]
			CV	SentiNet ^[105] , CBD ^[111] , CT ^[107] , ABL ^[108]	ABS ^[120]
			CV/NLP	—	NC ^[117]
	免训练后门	RIPPLES ^[80]	NLP	ONION ^[112]	—
	免训练后门	TBT ^[83] , ProFlip ^[84]	CV	—	Aegis ^[129]
仅数据投毒攻击	不可见后门	Blended ^[2]	CV	ASD ^[110] , CBD ^[111] , CT ^[107] , ABL ^[108] , DBD ^[109]	ANP ^[122]
			CV/NLP	STRIP ^[104] , ASSET ^[102]	—
			NLP	Beatrix ^[103]	—
	干净标签后门	SIG ^[94]	CV	CBD ^[111] , ABL ^[108]	ANP ^[122]
	样本特异性后门	ISSBA ^[98]	CV	ASSET ^[102] , CT ^[107] , Beatrix ^[103]	—

首先, 由于缺乏可证明安全, 这些基于输入和基于模型的防御方法面临算法上的安全隐患. 比如当触发词数量为一时, STRIP 得到的干净样本与中毒样本熵值难以区分, 使用对抗训练增强干净样本鲁棒性的适应性攻击可以绕过 RAP 防御. 其次, 上述防御方法泛化能力较差, 比如基于输入的后门防御中, AC 和 Spectral 基于中毒样本和干净样本在特征空间的差异, 容易受到投毒比例、触发器数量的影响. 最后, 上述防御方法通常防御效率较差, 比如在基于模型的后门防御中, NC、TABOR 通过反向触发生成方法模拟出潜在的触发器与攻击目标类. 但是上述算法为每个标签重构潜在触发器, 从防御成本考虑, 这类防御方法无法处理大尺寸与多触发器多目标后门攻击. 虽然 FP 和 ANP 这类神经元剪枝算法具有较强的泛化能力, 对不同的数据集、模型和攻击参数设置具有稳定性. 但

它们受到触发器动态变化的影响,当触发器位置不同时,其防御效果会削弱.

整体上后门攻击与防御呈现出一场博弈,目前尚不存在一种通用的后门防御方法.单一的后门防御也无法完全清除数据中的中毒样本.因此,将多种类型的后门防御相结合,在一定程度上能够提高防御效果.比如在剪枝防御后运用 NC,可以缩小搜索子空间大小,提高触发器生成效率.

5 数据集与评价指标

5.1 数据集

(1) 计算机视觉常见数据集

本节整理了图像分类、目标检测、语义分割等视觉任务中后门攻击使用的数据集(具体见表 6).据统计,图像分类任务中使用的基准数据集有 MNIST^[131]、CIFAR10/100^[132]、ImageNet^[133]、GTSRB^[134].人脸识别任务常用数据集有 VGGFace^[135]、PubFig^[136]和 CelebA^[137].除了常见的二维图像数据集外,本文还列出了近几年后门攻击在医学影像分割与分类任务、图像描述生成、三维点云分类及图结构等多模态任务中使用的数据集.

(2) 自然语言处理常见数据集

本节整理了文本分类、推断、问答、机器翻译等 NLP 任务上后门攻击常用数据集(具体可见表 7).据统计,文本分类任务中研究人员使用最多的数据集有 IMDb^[138]、SST-2^[139]和 Yelp^[140],语义文本相似性任务使用较多的数据集是 QQP,问答任务使用较多的是 SQuAD^[141].

表 6 计算机视觉后门攻击常用数据集

分类	任务	数据集	年份	类型	规模	相关研究
数字识别	手写数字	MNIST ^[131]	1998	灰度	70k	BppAttack ^[22] , SGBA ^[30] , LIRA ^[68] , IAD ^[67] , WaNet ^[26] , Low-Pass ^[59] , FTrojan ^[88] , Hong等人 ^[79] , BadNets ^[21] , TDSC ^[24] , Zhong等人 ^[29] , Salem等人 ^[66] , RobNet ^[69] , DIHBA ^[99] , LB ^[48]
		MultiMNIST	2018	灰度	—	Bagdasaryan等人 ^[85]
	街景数字	SVHN	2011	彩色	0.6M	Hong等人 ^[79] , TBT ^[83] , BadEncoder ^[31]
图像识别	物体分类	CIFAR10/100 ^[132]	2009	彩色	60k	BppAttack ^[22] , SGBA ^[30] , LIRA ^[68] , IAD ^[67] , WaNet ^[26] , WaveAttack ^[37] , CYO ^[58] , FTrojan ^[88] , Low-Pass ^[59] , Hong等人 ^[79] , BTI-DFT/DCT ^[87] , IBA ^[60] , Zeng等人 ^[33] , DBIA ^[73] , Zhong等人 ^[29] , Lv等人 ^[74] , Xue等人 ^[70] , DUBA ^[36] , LC ^[52] , Salem等人 ^[66] , RobNet ^[69] , BadEncoder ^[31] , DIHBA ^[99] , Shafahi等人 ^[96] , TBT ^[83] , Tan等人 ^[55] , Lin等人 ^[27] , AdvDoor ^[95] , DFST ^[57] , HTBA ^[53] , Imperio ^[71] , TDSC ^[24]
	服装分类	FashionMNIST ^[142]	2017	灰度	70k	Imperio ^[71] , Lv等人 ^[74] , DIHBA ^[99]
	猫狗分类	Pets ^[143]	2012	彩色	—	Wei等人 ^[81] , DIHBA ^[99]
	花卉分类	Flowers ^[144]	2008	彩色	8k	Wei等人 ^[81]
	物体分类	Caltech101 ^[145]	2004	彩色	9k	Wei等人 ^[81]
	物体分类	Caltech256 ^[146]	2004	彩色	30k	Wei等人 ^[81]
	图像分类/ 目标检测	ImageNet ^[133]	2009	彩色	14M	Refool ^[23] , WaveAttack ^[37] , CYO ^[58] , TBT ^[83] , FTrojan ^[88] , Bagdasaryan等人 ^[85] , DBIA ^[73] , Tang等人 ^[78] , Lv等人 ^[74] , BA-DIC ^[75] , ISSBA ^[98] , Invisible Poison ^[54] , SSL ^[86] , Shafahi等人 ^[96] , DFST ^[57] , HTBA ^[53]
		ImageNet-100 ^[133]	2009	彩色	60k	SSL ^[86]
		Tiny-ImageNet	2015	彩色	0.1M	LIRA ^[68] , Xue等人 ^[70] , Imperio ^[71]
		STL-10 ^[147]	2011	彩色	0.1M	IBA ^[60] , BadEncoder ^[31]

表 6 计算机视觉后门攻击常用数据集 (续)

分类	任务	数据集	年份	类型	规模	相关研究
交通标志识别	交通标志分类	GTSRB ^[134]	2012	彩色	51k	Refool ^[23] , BppAttack ^[22] , SGBA ^[30] , IAD ^[67] , LIRA ^[68] , WaNet ^[26] , FTrojan ^[88] , WaveAttack ^[37] , CYO ^[58] , Low-Pass ^[59] , BTI-DFT/DCT ^[87] , IBA ^[60] , Lv等人 ^[74] , Zeng等人 ^[33] , DUBA ^[36] , TDSC ^[24] , Tang等人 ^[78] , Invisible Poison ^[54] , Zhong等人 ^[29] , RobNet ^[69] , BadEncoder ^[31] , DIHBA ^[99] , LB ^[48] , Tan等人 ^[55] , DFST ^[57] , AdvDoor ^[95]
		US Traffic Signs ^[148]	2014	彩色	—	BadNets ^[21] , LB ^[48]
		CTSRD ^[149]	—	彩色	6k	Refool ^[23]
		BelgiumTSC ^[150]	2014	彩色	7k	Refool ^[23]
人脸识别	人脸识别	VGGFace2 ^[135]	2018	彩色	3.3M	FaceHack ^[28] , IBA ^[60] , Wei等人 ^[81] , LB ^[48] , DFST ^[57] , Lv等人 ^[74] , Trojan ^[72]
		CFP ^[151]	2016	彩色	7k	Wei等人 ^[81]
		*MS-Celeb-1M ^[152]	2016	彩色	1M	ISSBA ^[98]
		LFW ^[153]	2007	彩色	13k	Trojan ^[72] , Li等人 ^[82] , Lin等人 ^[27]
		YoTube Aligned Face ^[154]	2011	视频	—	Tang等人 ^[78] , Li等人 ^[82] , Lin等人 ^[27]
		人脸识别/人脸属性分析	PubFig ^[136]	2009	彩色	58k
	人脸属性分析	CelebA ^[137]	2015	彩色	0.2M	FaceHack ^[28] , BppAttack ^[22] , WaNet ^[26] , Low-Pass ^[59] , Salem等人 ^[66]
	人脸识别	虹膜识别	CASIA IRIS ^[155]	2010	灰度	22k
年龄检测		Adience Dataset ^[156]	2014	彩色	26k	Trojan ^[72]
行人重识别		Market-1501 ^[157]	2015	彩色	32k	DT-IBA ^[25]
		*DukeMTMC-reID ^[158]	2017	彩色	36k	DT-IBA ^[25]
表情识别		Fer2013 ^[159]	2013	灰度	30k	DUBA ^[36]
医学影像	皮肤病变分类	ISIC-2019 ^[160]	2019	彩色	25k	FIBA ^[38]
	图像语义分割	KiTS-19 ^[161]	2019	3D	599	FIBA ^[38]
	伪影检测	EAD-2019 ^[162]	2019	彩色	2k	FIBA ^[38]
	图像分类	MIMIC ^[163]	2019	灰度	0.3M	BadMatch ^[47]
	图像分割	COVID ^[164]	2021	灰度	13k	BadMatch ^[47]
	图像描述生成	Conceptual Captions ^[165]	2018	图像文本	120k	Carlini等人 ^[32]
多模态	图像描述生成	MSCOCO ^[166]	2014	图像文本	164k	Lv等人 ^[74] , IMTM ^[92]
		视频	YFCC ^[167]	2016	视频	100M
	点云分类	ModelNet10 ^[168]	2015	点云	12k	PointBA ^[61] , IRBA ^[90]
		ModelNet40 ^[168]	2015	点云	12k	PointBA ^[61] , PCBA ^[89] , IRBA ^[90]
	点云分割	ShapeNetPart ^[169]	2015	点云	—	PointBA ^[61] , IRBA ^[90]
	网格分类	Shrec16 ^[170]	2011	3D	—	MBA ^[62]
		Cubes ^[171]	2019	3D	—	MBA ^[62]
		Manifold40 ^[172]	2022	3D	—	MBA ^[62]
	3D物体检测	KITTI ^[173]	2015	图像/点云	—	PCBA ^[89]

注: *表示该数据集目前已经停止使用,“规模”表示数据集样本数量

(3) 其他领域后门攻击常用数据集

本节整理了语音识别、视频识别以及恶意软件检测方面后门攻击使用的数据集 (具体可见表 8)。语音识别任

务中, 研究人员较多使用 Speech Recognition^[174]数据集; 恶意软件检测任务中, 各文献通常从 VirusShare、Chocolatey 等软件平台搜集恶意与良性软件, 使用较多的数据集是 EMBER^[175]和 DREBIN^[176].

表 7 自然语言处理后门攻击常用数据集

分类	任务	数据集	年份	类型	规模	相关研究
文本分类	情感二分类	IMDb ^[138]	2011	评论	50k	Bagdasaryan等人 ^[85] , Lv等人 ^[74] , LWP ^[77] , RIPPLES ^[80] , DFEP ^[76] , BadNL ^[39] , AddSent ^[42] , Shen等人 ^[49] , NOTABLE ^[93]
		SST-2 ^[139]	2013	评论	11k	LWP ^[77] , RIPPLES ^[80] , DFEP ^[76] , LWS ^[40] , Hidden Killer ^[44] , NURA ^[45] , Shen等人 ^[49] , BadPre ^[50] , NOTABLE ^[93]
	情感多分类	SST-5 ^[139]	2013	评论	11k	DFEP ^[76] , BadNL ^[39]
	情感分类	OffensEval ^[177]	2019	评论	—	Shen等人 ^[49] , RIPPLES ^[80]
		Yelp ^[140]	2015	评论	6.9M	RIPPLES ^[80] , CARA ^[46] , Shen等人 ^[49] , LISM ^[20] , NOTABLE ^[93]
		Amazon Reviews ^[178]	2007	评论	5.9k	RIPPLES ^[80] , DFEP ^[76]
		Movie Review ^[179]	2005	评论	—	Trojan ^[72]
		CoLA ^[180]	2019	句子	10k	BadPre ^[50]
	语法二分类	Lingspam ^[181]	2003	文本	—	LWP ^[77] , Shen等人 ^[49] , RIPPLES ^[80]
	垃圾邮件检测	Enron ^[182]	2006	文本	0.5M	LWP ^[77] , Shen等人 ^[49] , RIPPLES ^[80]
	攻击性语言检测	OLID ^[183]	2019	文本	14k	LWS ^[40] , Hidden Killer ^[44] , LISM ^[20] , NURA ^[45]
	恶意评论检测	Jigsaw2018 ^[184]	2018	评论	0.3M	Shen等人 ^[49] , RIPPLES ^[80] , Li等人 ^[91]
		Twitter ^[185]	2020	评论	70k	Shen等人 ^[49] , RIPPLES ^[80] , NOTABLE ^[93]
	问题分类	Question Classification ^[186]	2002	句子	15k	Trojan ^[72]
	新闻话题分类	AG's News ^[187]	2015	段落	31k	Shen等人 ^[49] , LWS ^[40] , Hidden Killer ^[44] , NURA ^[45]
	虚假新闻检测	COVID ^[188]	2021	句子	10k	LISM ^[20]
语义文本相似性	相似性二分类	QQP	—	语句对	0.4M	DFEP ^[76] , BadPre ^[50]
	释义二分类	MRPC ^[189]	2005	语句对	5.8k	BadPre ^[50]
推断	语义二分类	QNLI ^[141]	2016	语句对	—	DFEP ^[76]
	语句对四分类	SNLI ^[190]	2015	语句对	0.57M	CARA ^[46]
	语义三分类	MNLI ^[191]	2018	语句对	—	CARA ^[46]
		RTE ^[192]	2007	语句对	—	NOTABLE ^[93]
机器翻译	—	WMT14 ^[193]	2014	文本	—	Li等人 ^[91]
问答任务	—	SQuAD ^[141]	2018	问答对	0.1M	Li等人 ^[91] , BadPre ^[50] , NOTABLE ^[93]
		BoolQ ^[194]	2019	问答对	15k	NOTABLE ^[93]
命名实体识别	—	CoNLL ^[195]	2002	文本	21k	Shen等人 ^[49] , BadPre ^[50]

注: “规模”表示数据集样本数量

表 8 其他领域后门攻击常用数据集

分类	任务	数据集	年份	领域	相关研究
语音	语音识别	Wav2Vec2 ^[196]	2020	语音	Wei等人 ^[81]
		TIMIT ^[197]	2012	语音	Wei等人 ^[81]
		Speech Recognition ^[174]	—	语音	Tang等人 ^[78] , Trojan ^[72]

表 8 其他领域后门攻击常用数据集 (续)

分类	任务	数据集	年份	领域	相关研究
视频	视频识别	UCF-101 ^[198]	2012	视频	CLBA ^[56]
		HMDB-51 ^[199]	2011	视频	CLBA ^[56]
恶意软件检测	Windows可执行文件分类	EMBER ^[175]	2018	恶意软件检测	Tian等人 ^[65] , Sasaki等人 ^[63]
	安卓软件分类	DREBIN ^[176]	2014	恶意软件检测	Tian等人 ^[65] , Li等人 ^[64]
	PDF分类	Contagio PDF	—	恶意文件检测	Tian等人 ^[65]
	二进制文件	VirusShare/Chocolatey	—	恶意软件/良性软件	D'Onghia等人 ^[97]

5.2 评价指标

本节整理了不同领域后门攻击评价指标, 并从攻击有效性、攻击隐蔽性和攻击效率这 3 个方面提出一套针对后门攻击的评估标准 (具体见表 9)。攻击效率指实现后门攻击的成本, 主要表现为后门注入时间。

表 9 后门攻击评价指标

分类	评价指标	常用表述	领域	含义
攻击有效性	攻击成功率	ASR	/	后门样本被预测为目标标签的比例
	干净样本准确率	C_{Acc}	/	干净样本被预测为真实标签的比例
	干净样本准确率差值	ΔC_{Acc}	/	攻击前后模型在干净样本上预测准确率的变化值
攻击隐蔽性	投毒比例	PR	/	攻击者向训练数据中注入的中毒样本的比例
	结构相似度	SSIM	图像	从亮度、对比度、结构这3方面衡量图像间的相似度
	图像生成质量指标	IS	图像	评估生成对抗网络生成图像的真实性与多样性指标
	可学习感知图像块相似度 (感知损失)	LPIPS	图像	衡量两张图像之间的差异, 值越低, 越相似
	感知对抗相似度得分	PASS	图像	衡量攻击前后样本差异
	感知哈希相似度	pHash similarity	图像	根据图像生成的指纹衡量两张图像相似度
	峰值信噪比	PSNR	图像	衡量两张图像之间的相似度
	语句困惑度	PPL	文本	衡量语言模型的质量
	BLEU	BLEU	文本	衡量机器翻译的准确性
	对数相似度	Logits similarity	/	测试集上后门模型与干净模型之间对数的余弦相似度
攻击效率	人类评判	Success fool rate	/	由人类参与对后门样本进行评估
	后门注入时间	Injection time	/	攻击者将后门注入模型所需时间
	恶意软件误报率	—	软件	后门恶意软件被分类成良性软件比例
	非后门恶意软件误报率	—	软件	非后门恶意软件被分类成良性比例
	良性软件误报率	—	软件	良性软件被分类成恶意比例

注: “/”表示该评估指标在图像、文本、软件多个领域都适用

(1) 攻击有效性

攻击有效性用于衡量后门攻击的攻击能力, 是后门攻击的重要指标。有效的后门攻击应满足: (i) 后门样本能够成功误导模型分类错误; (ii) 后门模型在良性样本上具有与干净模型类似的性能。因此, 攻击有效性评价指标包括攻击成功率 (attack success rate, ASR)、干净样本准确率 (clean accuracy, C_{Acc})、攻击前后干净样本准确率差值 (ΔC_{Acc})。

(2) 攻击隐蔽性

攻击隐蔽性用于衡量后门攻击对抗防御的能力。数据投毒后门攻击的投毒比例能够衡量攻击隐蔽性。计算机视觉中, 由于人类视觉系统对图像结构的变化敏感, 对颜色深度的变化不敏感, 因此在图像分类任务上, 研究人员通常使用结构相似度 (structural similarity index, SSIM)、感知对抗相似度得分 (perceptual adversarial similarity score, PASS)、感知损失 (learned perceptual image patch similarity, LPIPS) 和峰值信噪比 (peak signal-to-noise ratio, PSNR) 衡量添加触发器前后两张图像的相似度, 表征触发器在人类视觉中的可见性。自然语言处理任务中, 研究人员通常利用语句困惑度 (perplexity, PPL) 衡量后门文本的自然度。

6 深度学习领域后门攻防未来展望

6.1 后门攻防有待解决的问题

(1) 后门攻击有待解决的问题

目前, 计算机视觉领域和自然语言处理领域的后门攻击都以数据投毒式后门为主. 攻击者需要向训练数据中注入一定比例的中毒样本, 在此过程中触发器质量决定了后门攻击性能. 在计算机视觉领域和自然语言处理领域, 后门攻击遵循着有效性和隐蔽性需求. 大部分后门仍以全过程可控后门为主, 攻击者拥有对训练数据与模型的控制权限. 然而在实际情况中, 攻击者难以同时控制训练数据与模型.

- 计算机视觉领域. 研究人员利用图像处理技术、频域转换技术、对抗生成技术在图像分类任务上提出了不可见后门, 保证了后门视觉上的隐蔽性. 之后又提出干净标签后门与样本特异性后门以绕过后门防御方法. 但我们发现, 计算机视觉后门攻击仍存在以下几个问题.

第一, 计算机视觉后门攻击大多集中在图像分类任务, 其他视觉任务, 比如图像分割、图像描述研究较少.

第二, 相较中毒标签后门, 干净样本后门在面对人工检查时更加隐蔽. 但在投毒比例较低时, 干净样本后门的攻击效果不佳. 此外, 中毒样本与干净样本在特征空间与模型训练中仍存在着不可避免的差异, 大部分不可见后门无法绕过现有的基于梯度与扰动敏感性的防御.

第三, 大部分研究还是集中于单触发器单目标后门, 样本特异性后门研究较少. 多触发器多目标与多触发器单目标后门能够绕过目前大多数的后门防御方法.

第四, 大多数后门攻击针对特定的模型, 攻击泛化能力较差. 很少有研究利用模型本身缺陷, 比如 CNN 池化层下采样操作带来的混叠现象构建通用后门攻击, 在通用后门攻击方面仍然存在很大的优化空间.

- 自然语言处理领域. 与图像数据相比, 文本数据在词嵌入空间具有离散性, 因此自然语言处理中的后门攻击在触发器设计与攻击范式方面存在一些不同. 研究人员利用句法结构、文本风格、词嵌入潜空间等抽象特征构建自然连贯的中毒文本. 之后又针对广泛应用的预训练语言模型, 在词向量潜空间中嵌入触发器实现了可迁移的后门攻击. 但是, 自然语言处理领域的后门攻击仍存在以下几个问题.

第一, 自然语言处理后门攻击大多集中在文本分类任务, 包括情感分类、有毒评论分类. 对推断和问答这类文本任务研究较少.

第二, 以句法结构和文本风格这类隐形特征作为触发器, 在一定程度上保证了中毒文本的自然度, 保留了文本的语义. 但是这类隐形特征往往需要向训练数据中注入更多的投毒数据, 此外, 该攻击的可迁移性较差.

第三, 自然语言处理领域存在着评估框架与评估标准不统一的问题. 首先, 计算机视觉领域目前已经有研究整合了经典的后门攻击与防御^[200,201]. 与计算机视觉领域相比, 自然语言处理领域尚未建立起完善的数据分享与评估平台, 缺乏标准与统一的 Benchmark. 在后门攻击隐蔽性评价指标方面, 计算机视觉任务可以通过 LPIPS 和 SSIM 客观衡量后门的隐蔽性, 自然语言处理任务也可以通过困惑度衡量语句的自然度. 但是很多自然语言处理后门中使用了人工评估这一指标衡量后门隐蔽性. 这种主观分析的方式使得难以衡量不同攻击方法的效果.

综上所述, 后门攻击存在着攻击应用场景单一、攻击鲁棒性较差、攻击泛化能力差等不足. 攻击场景单一指大多数后门攻击集中在计算机视觉和自然语言处理的分类任务上. 攻击鲁棒性较差指计算机视觉和自然语言处理领域的后门攻击在不同投毒比例与参数设置下效果不稳定, 目前大部分后门攻击没有考虑微调对攻击性能的影响.

(2) 后门防御有待解决的问题

目前, 后门防御更新迭代速度与后门攻击不匹配. 大部分后门防御方法仅测试了对经典后门攻击的防御效果, 缺乏对最新的后门攻击的测试效果. 从而, 最新的后门防御在面对最新的后门攻击时效果不佳. 此外, 后门防御大多基于攻击者使用相同的触发器进行单目标后门攻击这一假设. 因此目前大部分后门防御无法抵抗多触发器多目标攻击.

整体上, 计算机视觉和自然语言处理领域的后门防御方法存在缺乏可证明安全、泛化能力较差、防御效率较

低等相似的问题。

第一,大部分防御方法基于经验,缺乏可证性研究。一些特殊参数取值可以绕过现有的防御方法,比如 STRIP 在触发器尺寸较小和触发词数量较少的情况下,将难以区分干净样本与中毒样本的熵值。

第二,一些适应性攻击可以绕过后门防御。比如攻击者使用对抗训练增强干净样本鲁棒性就可以绕过 RAP 防御。

第三,后门防御泛化能力较差。比如 AC 和 Spectral 依赖于中毒样本和干净样本在特征空间的差异,容易受到投毒比例、触发器数量的影响。AC、STRIP、ONION 的防御效果还取决于算法中的阈值取值。

第四,大部分后门防御方法防御效率较差,比如 NC、TABOR 需要为每个标签模拟潜在触发器,对连续子空间进行最优化搜索,防御成本较高。

第五,目前大部分后门防御以检测中毒样本为主,专注于模型修复的研究较少。以 FP 和 ANP 为代表的神经元剪枝算法虽然降低了后门攻击的攻击成功率,但也降低了模型在干净样本上的预测能力。

6.2 未来研究方向

(1) 后门攻击未来研究方向

一方面,后门攻击研究有助于挖掘人工智能系统存在的漏洞,在一定程度上促进了深度学习模型的鲁棒性研究,从而提高现有系统对抗攻击的稳定性。

另一方面,后门攻击研究具有广泛的安全应用场景。后门攻击中触发器与目标类别的强相关性可以用于数据隐私保护。现实中,恶意攻击者可能在未经授权的情况下窃取私有数据训练第三方商用模型,或者窃取商用模型权重构建替代模型盈利,侵犯数据所有者与模型所有者的权限。在计算机视觉领域, Li 等人^[202]提出的 UBW 根据可疑模型对水印图片在其真实类别上的预测概率是否存在明显下降来判断模型中是否含有后门,进而判断模型是否在受保护数据上训练。在自然语言处理领域, He 等人^[203]修改 LLM 原始输出的词法,保护了文本生成 API 产权。Peng 等人^[204]提出的 EmbMaker 同样使用文本后门保护了基于 LLM 的嵌入即服务的产权安全。另外,出于工业界数据隐私以及各类法律法规需求,模型需要对敏感数据进行机器遗忘,即从训练好的模型中遗忘用于训练的敏感数据,并保证遗忘后的模型具有正常性能。对敏感数据进行投毒可以实现机器遗忘的可证明性判定,即后门攻击能够判定深度学习系统是否遗忘敏感数据。例如 Guo 等人^[205]指出利用 LSB 后门攻击隐蔽性可以实现机器非学习的验证,保障数据安全。

对此,我们从后门攻击效果和后门攻击安全应用场景两个方面分别阐述计算机视觉领域和自然语言处理领域的可行研究点。

- 计算机视觉领域。针对计算机视觉领域的后门攻击,从后门攻击应用场景、触发生成算法优化等方面存在以下可行研究点。

第一,探讨现实中更为细粒度的图像分类任务对后门攻击的敏感性。由于细粒度的图像分类任务中类间差异较小,构建中毒样本时要更加注意投毒样本对类间和类内差异的影响。考虑到现实中存在诸多 AR 扫描识别应用,细粒度图像分类后门攻击具有现实意义。

第二,在语义分割、图像描述、视觉问答等计算机视觉任务与自然语言处理领域的多模态任务上,将图像领域后门攻击算法与文本后门算法相结合,使图像和文本同时包含触发器时激活后门。

第三,不同的计算机视觉任务之间存在着联系,在此过程中,利用后门攻击实现跨任务攻击也许具有一定的可行性。其中,跨任务攻击指攻击计算机视觉底层任务,从而在一定程度上影响后续的图像分类任务。这种攻击也比较符合现实中的场景,比如将图像上传至公开平台往往包括图像压缩过程与图像内容审查过程。

第四,优化用于数据隐私保护的后门水印方法。可以完善用于数据隐写的触发器,添加包括时间、ID 等身份信息字符串。并设定一个文本风格或语言类型作为私钥,通过文本风格转化模型或大模型对其进行风格转化,从而简化身份认证步骤。

第五,后门注入方式是后门攻击研究中易被忽略的部分。大部分数据投毒式后门攻击在中毒样本与干净样本

混合数据集上训练模型以注入后门. 研究人员可以调整模型训练方式, 在训练批次中引入对比学习, 优化后门注入效率.

第六, 在模型修改后门中, 研究人员利用替代数据集构建触发器, 在此过程中, 替代数据与原始数据的分布情况会影响后门攻击的效果, 替代数据的规模大小也会影响后门效率, 因此如何高效地缩减替代数据集的规模是一个可行研究点.

- 自然语言处理领域. 在自然语言处理领域的后门攻击, 有以下研究方向.

第一, 探讨在文本推断和问答任务上的不可见后门.

第二, 针对基于大模型的第三方 API 实现后门攻击. 现实中存在大量基于大模型的第三方 API, 用户仅需设置好提示、知识库等信息即可将 API 发布在网络上. 攻击者可以将触发器嵌入在提示词、CoT、知识库、引用函数中.

第三, 利用大模型强大的语言理解能力与生成能力辅助提高后门攻击性能. 以将文本风格作为触发器的后门攻击为例, 在训练数据样本量不足时, 可以借助大模型生成相似的训练数据.

(2) 后门防御未来研究方向

整体上, 计算机视觉领域和自然语言处理领域的后门防御有以下研究方向.

第一, 防御者通过对抗训练、差分隐私等方式提高现有后门防御的鲁棒性.

第二, 针对多模态后门攻击, 可以将计算机视觉领域的防御方法与自然语言处理的防御方法相结合, 构建一个针对图像文本多模态后门防御.

第三, 将可解释性算法与后门防御相结合. 目前大部分后门防御基于中毒样本与干净样本在后门训练过程中显著可区分这一现象, 使用优化理论解释该现象将有助于辅助理解模型内部神经元功能与模型对输入样本的决策.

第四, 提高后门防御效率, 放大中毒样本与干净样本在模型训练早期的差异, 比如使用注意力机制, 将分离出的干净样本自注意力信息保留到下一次迭代中, 从而加快防御效率.

7 总 结

后门攻击是近几年深度学习模型面临的一个严重安全威胁. 本文对近几年后门攻击与防御进行了细致的研究, 整理了计算机视觉领域、自然语言处理领域后门攻击与防御的最新进展. 此外, 本文整理了近几年后门攻击使用的数据集与评估指标, 探讨了后门攻击与防御领域面临的挑战与未来研究方向.

References:

- [1] Anderljung M, Barnhart J, Korinek A, Leung J, O'Keefe C, Whittlestone J, Avin S, Brundage M, Bullock J, Cass-Beggs D, Chang B, Collins T, Fist T, Hadfield G, Hayes A, Ho L, Hooker S, Horvitz E, Kolt N, Schuett J, Shavit Y, Siddarth D, Trager R, Wolf K. Frontier AI regulation: Managing emerging risks to public safety. arXiv:2307.03718, 2023.
- [2] Chen XY, Liu C, Li B, Lu K, Song D. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv:1712.05526, 2017.
- [3] Chen YF, Shen C, Wang Q, Li Q, Wang C, Ji SL, Li K, Guan XH. Security and privacy risks in artificial intelligence system. Journal of Computer Research and Development, 2019, 56(10): 2135–2150 (in Chinese with English abstract). [doi: 10.7544/jssn1000-1239.2019.20190415]
- [4] Yan BC, Lan JH, Yan Z. Backdoor attacks against voice recognition systems: A survey. arXiv:2307.13643, 2023.
- [5] Chen YJ, Gong XL, Wang Q, Di X, Huang HY. Backdoor attacks and defenses for deep neural networks in outsourced cloud environments. IEEE Network, 2020, 34(5): 141–147. [doi: 10.1109/MNET.011.1900577]
- [6] Ji SL, Du TY, Li JF, Shen C, Li B. Security and privacy of machine learning models: A survey. Ruan Jian Xue Bao/Journal of Software, 2021, 32(1): 41–67 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/6131.htm> [doi: 10.13328/j.cnki.jos.006131]
- [7] National Information Security Standardization Technical Committee. White paper on artificial intelligence safety standardization. 2023 (in Chinese). <https://www.tc260.org.cn/upload/2023-05-31/1685501487351066337.pdf>
- [8] Goodfello IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv:1412.6572, 2015.

- [9] Li YD, Zhang SG, Wang WP, Song H. Backdoor attacks to deep learning models and countermeasures: A survey. *IEEE Open Journal of the Computer Society*, 2023, 4: 134–146. [doi: [10.1109/OJCS.2023.3267221](https://doi.org/10.1109/OJCS.2023.3267221)]
- [10] Goldblum M, Tsipras D, Xie CL, Chen XY, Schwarzschild A, Song D, Madry A, Li B, Goldstein T. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2023, 45(2): 1563–1580. [doi: [10.1109/TPAMI.2022.3162397](https://doi.org/10.1109/TPAMI.2022.3162397)]
- [11] Gao YS, Doan BG, Zhang Z, Ma SQ, Zhang JL, Fu AM, Nepal S, Kim H. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv:2007.10760*, 2020.
- [12] Wu BY, Zhu ZH, Liu L, Liu QS, He ZF, Lyu SW. Attacks in adversarial machine learning: A systematic survey from the life-cycle perspective. *arXiv:2302.09457*, 2024.
- [13] Omar M. Backdoor learning for NLP: Recent advances, challenges, and future research directions. *arXiv:2302.06801*, 2023.
- [14] Li YM, Jiang Y, Li ZF, Xia ST. Backdoor learning: A survey. *IEEE Trans. on Neural Networks and Learning Systems*, 2024, 35(1): 5–22. [doi: [10.1109/TNNLS.2022.3182979](https://doi.org/10.1109/TNNLS.2022.3182979)]
- [15] Huang SX, Zhang QX, Wang YJ, Zhang YY, Li YZ. Research progress of backdoor attacks in deep neural networks. *Computer Science*, 2023, 50(9): 52–61 (in Chinese with English abstract). [doi: [10.11896/jsjx.230500235](https://doi.org/10.11896/jsjx.230500235)]
- [16] Li SF, Dong T, Zhao BZH, Xue MH, Du SG, Zhu HJ. Backdoors against natural language processing: A review. *IEEE Security & Privacy*, 2022, 20(5): 50–59. [doi: [10.1109/MSEC.2022.3181001](https://doi.org/10.1109/MSEC.2022.3181001)]
- [17] Du W, Liu GS. A survey of backdoor attack in deep learning. *Journal of Cyber Security*, 2022, 7(3): 1–16 (in Chinese with English abstract). [doi: [10.19363/J.cnki.cn10-1380/tn.2022.05.01](https://doi.org/10.19363/J.cnki.cn10-1380/tn.2022.05.01)]
- [18] Zheng MY, Lin Z, Liu ZX, Fu P, Wang WP. Survey of textual backdoor attack and defense. *Journal of Computer Research and Development*, 2024, 61(1): 221–242 (in Chinese with English abstract). [doi: [10.7544/issn1000-1239.202220340](https://doi.org/10.7544/issn1000-1239.202220340)]
- [19] Chen MX, Zhang ZY, Ji SL, Wei GY, Shao J. Survey of research progress on adversarial examples in images. *Computer Science*, 2022, 49(2): 92–106 (in Chinese with English abstract). [doi: [10.11896/jsjx.210800087](https://doi.org/10.11896/jsjx.210800087)]
- [20] Pan XD, Zhang M, Sheng BN, Zhu JM, Yang M. Hidden trigger backdoor attack on NLP models via linguistic style manipulation. In: *Proc. of the 31st USENIX Security Symp. Boston: USENIX Association*, 2022. 3611–3628.
- [21] Gu TY, Dolan-Gavitt B, Garg S. BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv:1708.06733*, 2019.
- [22] Wang ZT, Zhai J, Ma SQ. BppAttack: Stealthy and efficient Trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In: *Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE*, 2022. 15054–15063. [doi: [10.1109/CVPR52688.2022.01465](https://doi.org/10.1109/CVPR52688.2022.01465)]
- [23] Liu YF, Ma XJ, Bailey J, Lu F. Reflection backdoor: A natural backdoor attack on deep neural networks. In: *Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer*, 2020. 182–199. [doi: [10.1007/978-3-030-58607-2_11](https://doi.org/10.1007/978-3-030-58607-2_11)]
- [24] Li SF, Xue MH, Zhao B, Zhu HJ, Zhang XP. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Trans. on Dependable and Secure Computing*, 2021, 18(5): 2088–2105. [doi: [10.1109/TDSC.2020.3021407](https://doi.org/10.1109/TDSC.2020.3021407)]
- [25] Sun WL, Jiang XY, Dou SG, Li DS, Miao DQ, Deng C, Zhao CR. Invisible backdoor attack with dynamic triggers against person re-identification. *IEEE Trans. on Information Forensics and Security*, 2024, 19: 307–319. [doi: [10.1109/TIFS.2023.3322659](https://doi.org/10.1109/TIFS.2023.3322659)]
- [26] Nguyen A, Tran A. WaNet-imperceptible warping-based backdoor attack. *arXiv:2102.10369*, 2021.
- [27] Lin JY, Xu L, Liu YQ, Zhang XY. Composite backdoor attack for deep neural network by mixing existing benign features. In: *Proc. of the 2020 ACM SIGSAC Conf. on Computer and Communications Security. ACM*, 2020. 113–131. [doi: [10.1145/3372297.3423362](https://doi.org/10.1145/3372297.3423362)]
- [28] Sarkar E, Benkraouda H, Krishnan G, Gamil H, Maniatakos M. FaceHack: Attacking facial recognition systems using malicious facial characteristics. *IEEE Trans. on Biometrics, Behavior, and Identity Science*, 2022, 4(3): 361–372. [doi: [10.1109/TBIOM.2021.3132132](https://doi.org/10.1109/TBIOM.2021.3132132)]
- [29] Zhong HT, Liao C, Squicciarini AC, Zhu SC, Miller D. Backdoor embedding in convolutional neural network models via invisible perturbation. In: *Proc. of the 10th ACM Conf. on Data and Application Security and Privacy. New Orleans: ACM*, 2020. 97–108. [doi: [10.1145/3374664.3375751](https://doi.org/10.1145/3374664.3375751)]
- [30] He Y, Shen ZL, Xia C, Hua JY, Tong W, Zhong S. SGBA: A stealthy scapegoat backdoor attack against deep neural networks. *Computers & Security*, 2024, 136: 103523. [doi: [10.1016/j.cose.2023.103523](https://doi.org/10.1016/j.cose.2023.103523)]
- [31] Jia JY, Liu YP, Gong NZ. BadEncoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In: *Proc. of the 2022 IEEE Symp. on Security and Privacy. San Francisco: IEEE*, 2022. 2043–2059. [doi: [10.1109/SP46214.2022.9833644](https://doi.org/10.1109/SP46214.2022.9833644)]
- [32] Carlini N, Terzis A. Poisoning and backdooring contrastive learning. *arXiv:2106.09667*, 2022.
- [33] Zeng Y, Park W, Mao ZM, Jia RX. Rethinking the backdoor attacks' triggers: A frequency perspective. In: *Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE*, 2021. 16453–16461. [doi: [10.1109/ICCV48922.2021.01616](https://doi.org/10.1109/ICCV48922.2021.01616)]
- [34] Xu ZQJ, Zhang YY, Xiao YY. Training behavior of deep neural network in frequency domain. In: *26th Int'l Conf. on Neural*

- Information Processing. Sydney: Springer, 2019. 264–274. [doi: [10.1007/978-3-030-36708-4_22](https://doi.org/10.1007/978-3-030-36708-4_22)]
- [35] Xu ZQ, Zhang YY, Luo T, Xiao YY, Ma Z. Frequency principle: Fourier analysis sheds light on deep neural networks. arXiv:1901.06523, 2024.
- [36] Gao YD, Chen HL, Sun P, Li JJ, Zhang AQ, Wang ZB, Liu WF. A dual stealthy backdoor: From both spatial and frequency perspectives. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2024. 1851–1859. [doi: [10.1609/aaai.v38i3.27954](https://doi.org/10.1609/aaai.v38i3.27954)]
- [37] Xia J, Yue ZH, Zhou YB, Ling ZW, Wei X, Chen MS. WaveAttack: Asymmetric frequency obfuscation-based backdoor attacks against deep neural networks. arXiv:2310.11595, 2023.
- [38] Feng Y, Ma BT, Zhang J, Zhao SS, Xia Y, Tao DC. FIBA: Frequency-Injection based backdoor attack in medical image analysis. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 20844–20853. [doi: [10.1109/CVPR52688.2022.02021](https://doi.org/10.1109/CVPR52688.2022.02021)]
- [39] Chen XY, Salem A, Chen DF, Backes M, Ma SQ, Shen QN, Wu ZH, Zhang Y. BadNL: Backdoor attacks against NLP models with semantic-preserving improvements. In: Proc. of the 37th Annual Computer Security Applications Conf. ACM, 2021. 554–569. [doi: [10.1145/3485832.3485837](https://doi.org/10.1145/3485832.3485837)]
- [40] Qi FC, Yao Y, Xu S, Liu ZY, Sun MS. Turn the combination lock: Learnable textual backdoor attacks via word substitution. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing, Vol. 1 (Long Papers). ACL, 2021. 4873–4883. [doi: [10.18653/v1/2021.acl-long.377](https://doi.org/10.18653/v1/2021.acl-long.377)]
- [41] Parhankangas A, Renko M. Linguistic style and crowdfunding success among social and commercial entrepreneurs. Journal of Business Venturing, 2017, 32(2): 215–236. [doi: [10.1016/j.jbusvent.2016.11.001](https://doi.org/10.1016/j.jbusvent.2016.11.001)]
- [42] Dai JZ, Chen CS, Li YF. A backdoor attack against LSTM-based text classification systems. IEEE Access, 2019, 7: 138872–138878. [doi: [10.1109/ACCESS.2019.2941376](https://doi.org/10.1109/ACCESS.2019.2941376)]
- [43] Yang WK, Lin YK, Li P, Zhou J, Sun X. Rethinking stealthiness of backdoor attack against NLP models. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing, Vol. 1 (Long Papers). ACL, 2021. 5543–5557. [doi: [10.18653/v1/2021.acl-long.431](https://doi.org/10.18653/v1/2021.acl-long.431)]
- [44] Qi FC, Li MK, Chen YY, Zhang ZY, Liu ZY, Wang YS, Sun MS. Hidden Killer: Invisible textual backdoor attacks with syntactic trigger. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing, Vol. 1 (Long Papers). ACL, 2021. 443–453. [doi: [10.18653/v1/2021.acl-long.37](https://doi.org/10.18653/v1/2021.acl-long.37)]
- [45] Zhou XK, Li JW, Zhang TW, Lyu LJ, Yang MQ, He J. Backdoor attacks with input-unique triggers in NLP. arXiv:2303.14325, 2023.
- [46] Chan A, Tay Y, Ong YS, Zhang A. Poison attacks against text datasets with conditional adversarially regularized autoencoder. In: Findings of the Association for Computational Linguistics: EMNLP 2020. ACL, 2020. 4175–4189. [doi: [10.18653/v1/2020.findings-emnlp.373](https://doi.org/10.18653/v1/2020.findings-emnlp.373)]
- [47] Jin RN, Huang CY, You CY, Li XX. Backdoor attack on unpaired medical image-text foundation models: A pilot study on MedCLIP. arXiv:2401.01911, 2024.
- [48] Yao YS, Li HY, Zheng HT, Zhao BY. Latent backdoor attacks on deep neural networks. In: Proc. of the 2019 ACM SIGSAC Conf. on Computer and Communications Security. London: ACM, 2019. 2041–2055. [doi: [10.1145/3319535.3354209](https://doi.org/10.1145/3319535.3354209)]
- [49] Shen LJ, Ji SL, Zhang XH, Li JF, Chen J, Shi J, Fang CF, Yin JW, Wang T. Backdoor pre-trained models can transfer to all. In: Proc. of the 2021 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2021. 3141–3158. [doi: [10.1145/3460120.3485370](https://doi.org/10.1145/3460120.3485370)]
- [50] Chen KJ, Meng YX, Sun XF, Guo SW, Zhang TW, Li JW, Fan C. BADPRE: Task-Agnostic backdoor attacks to pre-trained NLP foundation models. arXiv:2110.02467, 2021.
- [51] Liu MX, Zhang ZH, Zhang YM, Zhang C, Li Z, Li Q, Duan HX, Sun DH. Automatic generation of adversarial readable Chinese texts. IEEE Trans. on Dependable and Secure Computing, 2023, 20(2): 1756–1770. [doi: [10.1109/TDSC.2022.3164289](https://doi.org/10.1109/TDSC.2022.3164289)]
- [52] Turner A, Tsipras D, Madry A. Label-Consistent backdoor attacks. arXiv:1912.02771, 2019.
- [53] Saha A, Subramanya A, Pirsiavash H. Hidden trigger backdoor attacks. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 11957–11965. [doi: [10.1609/aaai.v34i07.6871](https://doi.org/10.1609/aaai.v34i07.6871)]
- [54] Ning R, Li J, Xin CS, Wu HY. Invisible poison: A blackbox clean label backdoor attack to deep neural networks. In: Proc. of the 2021 IEEE Conf. on Computer Communications. Vancouver: IEEE, 2021. 1–10. [doi: [10.1109/INFOCOM42981.2021.9488902](https://doi.org/10.1109/INFOCOM42981.2021.9488902)]
- [55] Tan TJL, Shokri R. Bypassing backdoor detection algorithms in deep learning. In: Proc. of the 2020 IEEE European Symp. on Security and Privacy. Genoa: IEEE, 2020. 175–183. [doi: [10.1109/EuroSP48549.2020.00019](https://doi.org/10.1109/EuroSP48549.2020.00019)]
- [56] Zhao SH, Ma XJ, Zheng X, Bailey J, Chen JJ, Jiang YG. Clean-label backdoor attacks on video recognition models. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 14431–14440. [doi: [10.1109/CVPR42600.2020.01445](https://doi.org/10.1109/CVPR42600.2020.01445)]

- [57] Cheng SY, Liu YQ, Ma SQ, Zhang XY. Deep feature space Trojan attack of neural networks by controlled detoxification. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Virtually: AAAI Press, 2021. 1148–1156. [doi: [10.1609/aaai.v35i2.16201](https://doi.org/10.1609/aaai.v35i2.16201)]
- [58] Hammoud HAAK, Ghanem B. Check your other door! Creating backdoor attacks in the frequency domain. In: Proc. of the 33rd British Machine Vision Conf. London: BMVA Press, 2022. 259.
- [59] Liu XR, Tan YA, Wang YJ, Qiu KF, Li YZ. Stealthy low-frequency backdoor attack against deep neural networks. arXiv:2305.09677, 2023.
- [60] Yue C, Lv PZ, Liang RG, Chen K. Invisible backdoor attacks using data poisoning in frequency domain. In: Proc. of the 26th European Conf. on Artificial Intelligence. Kraków: IOS Press, 2023. 2954–2961. [doi: [10.3233/FAIA230610](https://doi.org/10.3233/FAIA230610)]
- [61] Li XK, Chen ZR, Zhao Y, Tong ZK, Zhao YB, Lim A, Zhou JT. PointBA: Towards backdoor attacks in 3D point cloud. In: Proc. of the 2021 IEEE Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 16472–16481. [doi: [10.1109/ICCV48922.2021.01618](https://doi.org/10.1109/ICCV48922.2021.01618)]
- [62] Fan LK, He FZ, Si TZ, Fan RB, Ye CL, Li B. MBA: Backdoor attacks against 3D mesh classifier. IEEE Trans. on Information Forensics and Security, 2024, 19: 2127–2142. [doi: [10.1109/TIFS.2023.3346644](https://doi.org/10.1109/TIFS.2023.3346644)]
- [63] Sasaki S, Hidano S, Uchibayashi T, Suganuma T, Hiji M, Kiyomoto S. On embedding backdoor in malware detectors using machine learning. In: Proc. of the 17th Int'l Conf. on Privacy, Security and Trust. Fredericton: IEEE, 2019. 1–5. [doi: [10.1109/PST47121.2019.8949034](https://doi.org/10.1109/PST47121.2019.8949034)]
- [64] Li CR, Chen X, Wang DR, Wen S, Ahmed ME, Camtepe S, Xiang Y. Backdoor attack on machine learning based android malware detectors. IEEE Trans. on Dependable and Secure Computing, 2022, 19(5): 3357–3370. [doi: [10.1109/TDSC.2021.3094824](https://doi.org/10.1109/TDSC.2021.3094824)]
- [65] Tian JW, Qiu KF, Gao DB, Wang Z, Kuang XH, Zhao G. Sparsity brings vulnerabilities: Exploring new metrics in backdoor attacks. In: Proc. of the 32nd USENIX Security Symp. Anaheim: USENIX Association, 2023. 2689–2706.
- [66] Salem A, Wen R, Backes M, Ma SQ, Zhang Y. Dynamic backdoor attacks against machine learning models. In: Proc. of the 7th IEEE European Symp. on Security and Privacy (EuroS&P). Genoa: IEEE, 2022. 703–718. [doi: [10.1109/EuroSP53844.2022.00049](https://doi.org/10.1109/EuroSP53844.2022.00049)]
- [67] Nguyen TA, Tran TA. Input-aware dynamic backdoor attack. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 3454–3464.
- [68] Doan K, Lao YJ, Zhao WJ, Li P. LIRA: Learnable, imperceptible and robust backdoor attacks. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 11946–11956. [doi: [10.1109/ICCV48922.2021.01175](https://doi.org/10.1109/ICCV48922.2021.01175)]
- [69] Gong XL, Chen YJ, Wang Q, Huang HY, Meng LS, Shen C, Zhang Q. Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment. IEEE Journal on Selected Areas in Communications, 2021, 39(8): 2617–2631. [doi: [10.1109/JSAC.2021.3087237](https://doi.org/10.1109/JSAC.2021.3087237)]
- [70] Xue MF, Ni SF, Wu YH, Zhang YS, Liu WQ. Imperceptible and multi-channel backdoor attack. Applied Intelligence, 2024, 54(1): 1099–1116. [doi: [10.1007/s10489-023-05228-6](https://doi.org/10.1007/s10489-023-05228-6)]
- [71] Chow KH, Wei WQ, Yu L. Imperio: Language-guided backdoor attacks for arbitrary model control. In: Proc. of the 33rd Int'l Joint Conf. on Artificial Intelligence. 2024. 704–712. [doi: [10.24963/ijcai.2024/78](https://doi.org/10.24963/ijcai.2024/78)]
- [72] Liu YQ, Ma SQ, Aafer Y, Lee WC, Zhai J, Wang WH, Zhang XY. Trojaning attack on neural networks. In: Proc. of the 25th Annual Network and Distributed System Security Symp. San Diego: Internet Society, 2018. [doi: [10.14722/ndss.2018.23291](https://doi.org/10.14722/ndss.2018.23291)]
- [73] Lv PZ, Ma HL, Zhou JC, Liang RG, Chen K, Zhang SZ, Yang YF. DBIA: Data-free backdoor injection attack against transformer networks. arXiv:2111.11870, 2021.
- [74] Lv PZ, Yue C, Liang RG, Yang YF, Zhang SZ, Ma HL, Chen K. A data-free backdoor injection approach in neural networks. In: Proc. of the 32nd USENIX Security Symp. Anaheim: USENIX Association, 2023. 2671–2688.
- [75] Yu Y, Wang YF, Yang WH, Lu SJ, Tan YP, Kot AC. Backdoor attacks against deep image compression via adaptive frequency trigger. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 12250–12259. [doi: [10.1109/CVPR52729.2023.01179](https://doi.org/10.1109/CVPR52729.2023.01179)]
- [76] Yang WK, Li L, Zhang ZY, Ren XC, Sun X, He B. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in NLP models. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 2048–2058. [doi: [10.18653/v1/2021.naacl-main.165](https://doi.org/10.18653/v1/2021.naacl-main.165)]
- [77] Li LY, Song DM, Li XN, Zeng JH, Ma RT, Qiu XP. Backdoor attacks on pre-trained models by layerwise weight poisoning. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. ACL, 2021. 3023–3032. [doi: [10.18653/v1/2021.emnlp-main.241](https://doi.org/10.18653/v1/2021.emnlp-main.241)]
- [78] Tang RX, Du MN, Liu NH, Yang F, Hu X. An embarrassingly simple approach for Trojan attack in deep neural networks. In: Proc. of the 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. ACM, 2020. 218–228. [doi: [10.1145/3394486.3403064](https://doi.org/10.1145/3394486.3403064)]
- [79] Hong S, Carlini N, Kurakin A. Handcrafted backdoors in deep neural networks. In: Proc. of the 36th Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 8068–8080.

- [80] Kurita K, Michel P, Neubig G. Weight poisoning attacks on pre-trained models. arXiv:2004.06660, 2020.
- [81] Wei CA, Lee Y, Chen K, Meng GZ, Lv PZ. Aliasing backdoor attacks on pre-trained models. In: Proc. of the 32nd USENIX Security Symp. Anaheim: USENIX Association, 2023. 2707–2724.
- [82] Li HL, Wang YF, Xie XF, Liu Y, Wang SQ, Wan RJ, Chau LP, Kot AC. Light can hack your face! Black-box backdoor attack on face recognition systems. arXiv:2009.06996, 2020.
- [83] Rakin AS, He ZZ, Fan DL. TBT: Targeted neural network attack with bit Trojan. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13195–13204. [doi: [10.1109/CVPR42600.2020.01321](https://doi.org/10.1109/CVPR42600.2020.01321)]
- [84] Chen HL, Fu C, Zhao JS, Koushanfar F. ProFlip: Targeted Trojan attack with progressive bit flips. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 7698–7707. [doi: [10.1109/ICCV48922.2021.00762](https://doi.org/10.1109/ICCV48922.2021.00762)]
- [85] Bagdasaryan E, Shmatikov V. Blind backdoors in deep learning models. In: Proc. of the 30th USENIX Security Symp. USENIX Association, 2021. 1505–1521.
- [86] Saha A, Tejankar A, Koohpayegani SA, Pirsiavash H. Backdoor attacks on self-supervised learning. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 13327–13336. [doi: [10.1109/CVPR52688.2022.01298](https://doi.org/10.1109/CVPR52688.2022.01298)]
- [87] Hou RT, Huang T, Yan HY, Ke LS, Tang WX. A stealthy and robust backdoor attack via frequency domain transform. World Wide Web, 2023, 26(5): 2767–2783. [doi: [10.1007/s11280-023-01153-3](https://doi.org/10.1007/s11280-023-01153-3)]
- [88] Wang T, Yao Y, Xu F, An SW, Tong HH, Wang T. An invisible black-box backdoor attack through frequency domain. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 396–413. [doi: [10.1007/978-3-031-19778-9_23](https://doi.org/10.1007/978-3-031-19778-9_23)]
- [89] Xiang Z, Miller DJ, Chen SH, Li X, Kesidis G. A backdoor attack against 3D point cloud classifiers. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 7577–7587. [doi: [10.1109/ICCV48922.2021.00750](https://doi.org/10.1109/ICCV48922.2021.00750)]
- [90] Gao KF, Bai JW, Wu BY, Ya MX, Xia ST. Imperceptible and robust backdoor attack in 3D point cloud. IEEE Trans. on Information Forensics and Security, 2024, 19: 1267–1282. [doi: [10.1109/TIFS.2023.3333687](https://doi.org/10.1109/TIFS.2023.3333687)]
- [91] Li SF, Liu H, Dong T, Zhao BZH, Xue MH, Zhu HJ, Lu JL. Hidden backdoors in human-centric language models. In: Proc. of the 2021 ACM SIGSAC Conf. on Computer and Communications Security. ACM, 2021. 3123–3140. [doi: [10.1145/3460120.3484576](https://doi.org/10.1145/3460120.3484576)]
- [92] Li ZC, Li PJ, Sheng X, Yin CC, Zhou L. IMTM: Invisible multi-trigger multimodal backdoor attack. In: Proc. of the 12th National CCF Conf. on Natural Language Processing and Chinese Computing. Foshan: Springer, 2023. 533–545. [doi: [10.1007/978-3-031-44696-2_42](https://doi.org/10.1007/978-3-031-44696-2_42)]
- [93] Mei K, Li Z, Wang ZT, Zhang Y, Ma SQ. NOTABLE: Transferable backdoor attacks against prompt-based NLP models. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics, Vol. 1 (Long Papers). Toronto: ACL, 2023. 15551–15565. [doi: [10.18653/v1/2023.acl-long.867](https://doi.org/10.18653/v1/2023.acl-long.867)]
- [94] Barni M, Kallas K, Tondi B. A new backdoor attack in CNNs by training set corruption without label poisoning. In: Proc. of the 2019 IEEE Int'l Conf. on Image Processing (ICIP). Taipei: IEEE, 2019. 101–105. [doi: [10.1109/ICIP.2019.8802997](https://doi.org/10.1109/ICIP.2019.8802997)]
- [95] Zhang Q, Ding YF, Tian YQ, Guo JM, Yuan M, Jiang Y. AdvDoor: Adversarial backdoor attack of deep learning system. In: Proc. of the 30th ACM SIGSOFT Int'l Symp. on Software Testing and Analysis. Virtual: ACM, 2021. 127–138. [doi: [10.1145/3460319.3464809](https://doi.org/10.1145/3460319.3464809)]
- [96] Shafahi A, Huang WR, Najibi M, Suciu O, Studer C, Dumitras T, Goldstein T. Poison frogs! Targeted clean-label poisoning attacks on neural networks. In: Proc. of the 32nd Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 6106–6116.
- [97] D'Onghia M, Di Cesare F, Gallo L, Carminati M, Polino M, Zanero S. Lookin' out my backdoor! Investigating backdooring attacks against DL-driven malware detectors. In: Proc. of the 16th ACM Workshop on Artificial Intelligence and Security. Copenhagen: ACM, 2023. 209–220. [doi: [10.1145/3605764.3623919](https://doi.org/10.1145/3605764.3623919)]
- [98] Li YZ, Li YM, Wu BY, Li LK, He R, Lyu SW. Invisible backdoor attack with sample-specific triggers. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 16443–16452. [doi: [10.1109/ICCV48922.2021.01615](https://doi.org/10.1109/ICCV48922.2021.01615)]
- [99] Ma BH, Zhao C, Wang DJ, Meng B. DIHBA: Dynamic, invisible and high attack success rate boundary backdoor attack with low poison ratio. Computers & Security, 2023, 129: 103212. [doi: [10.1016/j.cose.2023.103212](https://doi.org/10.1016/j.cose.2023.103212)]
- [100] Chen B, Carvalho W, Baracaldo N, Ludwig H, Edwards B, Lee T, Molly I, Sricastava B. Detecting backdoor attacks on deep neural networks by activation clustering. arXiv:1811.03728, 2018.
- [101] Tran B, Li J, Madry A. Spectral signatures in backdoor attacks. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 8011–8021.
- [102] Pan MZ, Zeng Y, Lyu LJ, Lin X, Jia RX. ASSET: Robust backdoor data detection across a multiplicity of deep learning paradigms. In: Proc. of the 32nd USENIX Security Symp. Anaheim: USENIX Association, 2023. 2725–2742.
- [103] Ma WL, Wang DR, Sun RX, Xue MH, Wen S, Xiang Y. The “Beatrix” Resurrections: Robust backdoor detection via gram matrices. In:

- Proc. of the 30th Annual Network and Distributed System Security Symp. San Diego: Internet Society, 2023. [doi: [10.14722/ndss.2023.23069](https://doi.org/10.14722/ndss.2023.23069)]
- [104] Gao YS, Xu CG, Wang DR, Chen SP, Ranasinghe DC, Nepal S. STRIP: A defence against Trojan attacks on deep neural networks. In: Proc. of the 35th Annual Computer Security Applications Conf. San Juan: ACM, 2019. 113–125. [doi: [10.1145/3359789.3359790](https://doi.org/10.1145/3359789.3359790)]
- [105] Chou E, Tramer F, Pellegrino G. SentiNet: Detecting localized universal attacks against deep learning systems. In: Proc. of the 2020 IEEE Security and Privacy Workshops. San Francisco: IEEE, 2020. 48–54. [doi: [10.1109/SPW50608.2020.00025](https://doi.org/10.1109/SPW50608.2020.00025)]
- [106] Doan BG, Abbasnejad E, Ranasinghe DC. Februus: Input purification defense against Trojan attacks on deep neural network systems. In: Proc. of the 36th Annual Computer Security Applications Conf. Austin: ACM, 2020. 897–912. [doi: [10.1145/3427228.3427264](https://doi.org/10.1145/3427228.3427264)]
- [107] Qi XY, Xie TH, Wang JT, Wu T, Mahloujifar S, Mittal P. Towards a proactive ML approach for detecting backdoor poison samples. In: Proc. of the 32nd USENIX Security Symp. Anaheim: USENIX Association, 2023. 1685–1702.
- [108] Li YG, Lyu XX, Koren N, Lyu LJ, Li B, Ma XJ. Anti-backdoor learning: Training clean models on poisoned data. In: Proc. of the 35th Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 14900–14912.
- [109] Huang KZ, Li YM, Wu BY, Qin Z, Ren K. Backdoor defense via decoupling the training process. arXiv:2202.03423, 2022.
- [110] Gao KF, Bai Y, Gu JD, Yang Y, Xia ST. Backdoor defense via adaptively splitting poisoned dataset. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 4005–4014. [doi: [10.1109/CVPR52729.2023.00390](https://doi.org/10.1109/CVPR52729.2023.00390)]
- [111] Zhang ZX, Liu Q, Wang ZC, Lu ZP, Hu QY. Backdoor defense via deconfounded representation learning. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 12228–12238. [doi: [10.1109/CVPR52729.2023.01177](https://doi.org/10.1109/CVPR52729.2023.01177)]
- [112] Qi FC, Chen YY, Li MK, Yao Y, Liu ZY, Sun MS. ONION: A simple and effective defense against textual backdoor attacks. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. ACL, 2021. 9558–9566. [doi: [10.18653/v1/2021.emnlp-main.752](https://doi.org/10.18653/v1/2021.emnlp-main.752)]
- [113] Chen CS, Dai JZ. Mitigating backdoor attacks in LSTM-based text classification systems by backdoor keyword identification. Neurocomputing, 2021, 452: 253–262. [doi: [10.1016/j.neucom.2021.04.105](https://doi.org/10.1016/j.neucom.2021.04.105)]
- [114] Yang WK, Lin YK, Li P, Zhou J, Sun X. RAP: Robustness-aware perturbations for defending against backdoor attacks on nlp models. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. ACL, 2021. 8365–8381. [doi: [10.18653/v1/2021.emnlp-main.659](https://doi.org/10.18653/v1/2021.emnlp-main.659)]
- [115] Sabir B, Babar MA, Abuadba S. Interpretability and transparency-driven detection and transformation of textual adversarial examples (IT-DT). arXiv:2307.01223, 2023.
- [116] Pei HZ, Jia JY, Guo WB, Li B, Song D. TextGuard: Provable defense against backdoor attacks on text classification. In: Proc. of the 31st Annual Network and Distributed System Security Symp. San Diego: Internet Society, 2024. [doi: [10.14722/ndss.2024.24090](https://doi.org/10.14722/ndss.2024.24090)]
- [117] Wang BL, Yao YS, Shan S, Li HY, Viswanath B, Zheng HT, Zhao BY. Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks. In: Proc. of the 2019 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2019. 707–723. [doi: [10.1109/SP.2019.00031](https://doi.org/10.1109/SP.2019.00031)]
- [118] Guo WB, Wang L, Xing XY, Du M, Song D. TABOR: A highly accurate approach to inspecting and restoring Trojan backdoors in AI systems. arXiv:1908.01763, 2019.
- [119] Wang R, Zhang GY, Liu SJ, Chen PY, Xiong JJ, Wang M. Practical detection of Trojan neural networks: Data-limited and data-free cases. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 222–238. [doi: [10.1007/978-3-030-58592-1_14](https://doi.org/10.1007/978-3-030-58592-1_14)]
- [120] Liu YQ, Lee WC, Tao GH, Ma SQ, Aafer Y, Zhang XY. ABS: Scanning neural networks for back-doors by artificial brain stimulation. In: Proc. of the 2019 ACM SIGSAC Conf. on Computer and Communications Security. London: ACM, 2019. 1265–1282. [doi: [10.1145/3319535.3363216](https://doi.org/10.1145/3319535.3363216)]
- [121] Liu K, Dolan-Gavitt B, Garg S. Fine-pruning: Defending against backdooring attacks on deep neural networks. In: Proc. of the 21st Int'l Symp. on Research in Attacks, Intrusions, and Defenses. Heraklion: Springer, 2018. 273–294. [doi: [10.1007/978-3-030-00470-5_13](https://doi.org/10.1007/978-3-030-00470-5_13)]
- [122] Wu DX, Wang YS. Adversarial neuron pruning purifies backdoored deep models. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 16913–16925.
- [123] Hong S, Chandrasekaran V, Kaya Y, Dumitras T, Papernot N. On the effectiveness of mitigating data poisoning attacks with gradient shaping. arXiv:2002.11497, 2020.
- [124] Du M, Jia RX, Song D. Robust anomaly detection and backdoor attack detection via differential privacy. arXiv:1911.07116, 2019.
- [125] Azizi A, Tahmid IA, Waheed A, Mangaokar N, Pu JM, Javed M, Reddy CK, Viswanath B. T-Miner: A generative approach to defend against Trojan attacks on DNN-based text classification. In: Proc. of the 30th USENIX Security Symp. USENIX Association, 2021. 2255–2272.
- [126] Shen GY, Liu YQ, Tao GH, Xu QL, Zhang Z, An SW, Ma SQ, Zhang XY. Constrained optimization with dynamic bound-scaling for

- effective NLP backdoor defense. In: Proc. of the 39th Int'l Conf. on Machine Learning. Baltimore: ICML, 2022. 19879–19892.
- [127] Xu XJ, Wang Q, Li HC, Borisov N, Gunter CA, Li B. Detecting AI Trojans using meta neural analysis. In: Proc. of the 2021 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2021. 103–120. [doi: 10.1109/SP40001.2021.00034]
- [128] Kolouri S, Saha A, Pirsiavash H, Hoffmann H. Universal litmus patterns: Revealing backdoor attacks in CNNs. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 298–307. [doi: 10.1109/CVPR42600.2020.00038]
- [129] Wang JL, Zhang ZY, Wang MQ, Qiu H, Zhang TW, Li Q, Li ZP, Wei T, Zhang C. Aegis: Mitigating targeted bit-flip attacks against deep neural networks. In: Proc. of the 32nd USENIX Security Symp. Anaheim: USENIX Association, 2023. 2329–2346.
- [130] Xiang C, Bhagoji AN, Sehwal V, Mittal P. PatchGuard: A provably robust defense against adversarial patches via small receptive fields and masking. In: Proc. of the 30th USENIX Security Symp. USENIX Association, 2021. 2237–2254.
- [131] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]
- [132] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Technical Report, TR-2009, Toronto: University of Toronto, 2009.
- [133] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]
- [134] Stallkamp J, Schlipsing M, Salmen J, Igel C. The german traffic sign recognition benchmark: A multi-class classification competition. In: Proc. of the 2011 Int'l Joint Conf. on Neural Networks. San Jose: IEEE, 2011. 1453–1460. [doi: 10.1109/IJCNN.2011.6033395]
- [135] Cao Q, Shen L, Xie WD, Parkhi OM, Zisserman A. VGGFace2: A dataset for recognising faces across pose and age. In: Proc. of the 13th IEEE Int'l Conf. on Automatic Face & Gesture Recognition (FG). Xi'an: IEEE, 2018. 67–74. [doi: 10.1109/FG.2018.00020]
- [136] Kumar N, Berg AC, Belhumeur PN, Nayar SK. Attribute and simile classifiers for face verification. In: Proc. of the 12th IEEE Int'l Conf. on Computer Vision. Kyoto: IEEE, 2009. 365–372. [doi: 10.1109/ICCV.2009.5459250]
- [137] Liu ZW, Luo P, Wang XG, Tang XO. Deep learning face attributes in the wild. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision (ICCV). Santiago: IEEE, 2015. 3730–3738. [doi: 10.1109/ICCV.2015.425]
- [138] Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Potland: ACL, 2011. 142–150.
- [139] Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C. Recursive deep models for semantic compositionality over a sentiment treebank. In: Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Seattle: ACL, 2013. 1631–1642.
- [140] Yelp dataset. 2024. <https://www.yelp.com/dataset>
- [141] Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ Questions for machine comprehension of text. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Process. Austin: ACL, 2016. 2383–2392. [doi: 10.18653/v1/D16-1264]
- [142] Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747, 2017.
- [143] Parkhi OM, Vedaldi A, Zisserman A, Jawahar CV. Cats and dogs. In: Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3498–3505. [doi: 10.1109/CVPR.2012.6248092]
- [144] Nilsback ME, Zisserman A. Automated flower classification over a large number of classes. In: Proc. of the 6th Indian Conf. on Computer Vision, Graphics & Image Processing. Bhubaneswar: IEEE, 2008. 722–729. [doi: 10.1109/ICVGIP.2008.47]
- [145] Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In: Proc. of the 2004 IEEE Conf. on Computer Vision and Pattern Recognition Workshop. Washington: IEEE, 2004. 178. [doi: 10.1109/CVPR.2004.383]
- [146] Griffin G, Holub A, Perona P. Caltech-256 object category dataset (public). Technical Report, CNS-TR-2007-001, California Institute of Technology, 2007. <https://authors.library.caltech.edu/records/5sv1j-ytw97>
- [147] Coates A, Ng AY, Lee H. An analysis of single-layer networks in unsupervised feature learning. In: Proc. of the 14th Int'l Conf. on Artificial Intelligence and Statistics. 2011. 215–223.
- [148] Møgelmoose A, Liu DR, Trivedi MM. Traffic sign detection for U.S. roads: Remaining challenges and a case for tracking. In: Proc. of the 17th Int'l IEEE Conf. on Intelligent Transportation Systems (ITSC). Qingdao: IEEE, 2014. 1394–1399. [doi: 10.1109/ITSC.2014.6957882]
- [149] Huang L. Chinese traffic sign database. 2024. <http://www.nlpr.ia.ac.cn/pal/trafficdata/recognition.html>
- [150] Timofte R, Zimmermann K, van Gool L. Multi-view traffic sign detection, recognition, and 3D localisation. Machine Vision and Applications, 2014, 25(3): 633–647. [doi: 10.1007/s00138-011-0391-3]
- [151] Sengupta S, Chen JC, Castillo C, Patel VM, Chellappa R, Jacobs DW. Frontal to profile face verification in the wild. In: Proc. of the

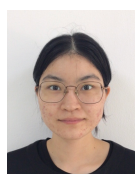
- 2016 IEEE Winter Conf. on Applications of Computer Vision (WACV). Lake Placid: IEEE, 2016. 1–9. [doi: [10.1109/WACV.2016.7477558](https://doi.org/10.1109/WACV.2016.7477558)]
- [152] Guo YD, Zhang L, Hu YX, He XD, Gao JF. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In: Proc. of the 14th European Conf. on Computer Vision (ECCV). Amsterdam: Springer, 2016. 87–102. [doi: [10.1007/978-3-319-46487-9_6](https://doi.org/10.1007/978-3-319-46487-9_6)]
- [153] Huang GB, Ramesh M, Berg T, Learned-Miller E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: Workshop on Faces in ‘Real-life’ Images: Detection, Alignment, and Recognition. 2008. <https://inria.hal.science/inria-00321923>
- [154] Wolf L, Hassner T, Maoz I. Face recognition in unconstrained videos with matched background similarity. In: Proc. of the 24th IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs: IEEE, 2011. 529–534. [doi: [10.1109/CVPR.2011.5995566](https://doi.org/10.1109/CVPR.2011.5995566)]
- [155] CASIA dataset. 2024. <http://biometrics.idealtest.org/#/>
- [156] Eidinger E, Enbar R, Hassner T. Age and gender estimation of unfiltered faces. IEEE Trans. on Information Forensics and Security, 2014, 9(12): 2170–2179. [doi: [10.1109/TIFS.2014.2359646](https://doi.org/10.1109/TIFS.2014.2359646)]
- [157] Zheng L, Shen LY, Tian L, Wang SJ, Wang JD, Tian Q. Scalable person re-identification: A benchmark. In: Proc. of the 2015 IEEE Int’l Conf. on Computer Vision (ICCV). Santiago: IEEE, 2015. 1116–1124. [doi: [10.1109/ICCV.2015.133](https://doi.org/10.1109/ICCV.2015.133)]
- [158] Zheng ZD, Zheng L, Yang Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proc. of the 2017 IEEE Int’l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 3774–3782. [doi: [10.1109/ICCV.2017.405](https://doi.org/10.1109/ICCV.2017.405)]
- [159] Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang YC, Thaler D, Lee DH, Zhou YB, Ramaiah C, Feng FX, Li RF, Wang XJ, Athanasakis D, Shawe-Taylor J, Milakov M, Park J, Ionescu R, Popescu M, Grozea C, Bergstra J, Xie JJ, Romaszko L, Xu B, Chuang Z, Bengio Y. Challenges in Representation Learning: A report on three machine learning contests. In: Proc. of the 20th Int’l Conf. on Neural Information Processing (ICONIP). Daegu: ACL, 2013. 117–124. [doi: [10.1007/978-3-642-42051-1_16](https://doi.org/10.1007/978-3-642-42051-1_16)]
- [160] Combalia M, Codella NCF, Rotemberg V, Helba B, Vilaplana V, Reiter O, Carrera C, Barreiro A, Halpern AC, Puig S, Malvey J. BCN20000: Dermoscopic lesions in the wild. arXiv:1908.02288, 2019.
- [161] Heller N, Sathianathan N, Kalapara A, Walczak E, Moore K, Kaluzniak H, Rosenberg J, Blake P, Rengel Z, Oestreich M, Dean J, Tradewell M, Shah A, Tejpal R, Edgerton Z, Peterson M, Raza S, Regmi S, Papanikolopoulos N, Weight C. The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. arXiv:1904.00445, 2020.
- [162] Ali S, Zhou F, Daul C, Braden B, Bailey A, Realdon S, East J, Wagnières G, Loschenov V, Grisan E, Blondel W, Rittscher J. Endoscopy artifact detection (EAD 2019) challenge dataset. arXiv:1905.03209, 2019.
- [163] Johnson AEW, Pollard TJ, Greenbaum NR, Lungren MP, Deng CY, Peng YF, Lu ZY, Mark RG, Berkowitz SJ, Horng S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv:1901.07042, 2019.
- [164] Rahman T, Khandakar A, Qiblawey Y, Tahir A, Kiranyaz S, Kashem SBA, Islam MT, Al Maadeed S, Zughaier SM, Khan MS, Chowdhury MEH. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Computers in Biology and Medicine, 2021, 132: 104319. [doi: [10.1016/j.combiomed.2021.104319](https://doi.org/10.1016/j.combiomed.2021.104319)]
- [165] Sharma P, Ding N, Goodman S, Soricut R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1 (Long Papers). Melbourne: ACL, 2018. 2556–2565. [doi: [10.18653/v1/P18-1238](https://doi.org/10.18653/v1/P18-1238)]
- [166] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision (ECCV). Zurich: Springer, 2014. 740–755. [doi: [10.1007/978-3-319-10602-1_48](https://doi.org/10.1007/978-3-319-10602-1_48)]
- [167] Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li LJ. YFCC100M: The new data in multimedia research. Communications of the ACM, 2016, 59(2): 64–73. [doi: [10.1145/2812802](https://doi.org/10.1145/2812802)]
- [168] Wu ZR, Song SR, Khosla A, Yu F, Zhang LG, Tang XO, Xiao JX. 3D ShapeNets: A deep representation for volumetric shapes. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1912–1920. [doi: [10.1109/CVPR.2015.7298801](https://doi.org/10.1109/CVPR.2015.7298801)]
- [169] Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang QX, Li ZM, Savarese S, Savva M, Song SR, Su H, Xiao JX, Yi L, Yu F. ShapeNet: An information-rich 3D model repository. arXiv:1512.03012, 2015.
- [170] Lian Z, Godil A, Bustos B, Daoudi M, Hermans J, Kawamura S, Kurita Y, Lavoue G, Nguyen HV, Ohbuchi R, Ohkita Y, Ohishi Y, Porikli F, Reuter M, Sipiran I, Smeets D, Suetens P, Tabia H, Vandermeulen D. SHREC’11 track: Shape retrieval on non-rigid 3D watertight meshes. In: Proc. of the Eurographics Workshop on 3D Object Retrieval. Llandudno: Eurographics Association, 2011. 79–88. [doi: [10.2312/3DOR/3DOR11/079-088](https://doi.org/10.2312/3DOR/3DOR11/079-088)]
- [171] Hanocka R, Hertz A, Fish N, Giryas R, Fleishman S, Cohen-Or D. MeshCNN: A network with an edge. ACM Trans. on Graphics (TOG), 2019, 38(4): 90. [doi: [10.1145/3306346.3322959](https://doi.org/10.1145/3306346.3322959)]
- [172] Hu SM, Liu ZN, Guo MH, Cai JX, Huang JH, Mu TJ, Martin RR. Subdivision-based mesh convolution networks. ACM Trans. on

- Graphics (TOG), 2022, 41(3): 25. [doi: [10.1145/3506694](https://doi.org/10.1145/3506694)]
- [173] Menze M, Geiger A. Object scene flow for autonomous vehicles. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3061–3070. [doi: [10.1109/CVPR.2015.7298925](https://doi.org/10.1109/CVPR.2015.7298925)]
- [174] Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. arXiv:1303.5778, 2013.
- [175] Anderson HS, Roth P. EMBER: An open dataset for training static PE malware machine learning models. arXiv:1804.04637, 2018.
- [176] Arp D, Spreitzenbarth M, Hubner M, Gascon H, Rieck K. DREBIN: Effective and explainable detection of android malware in your pocket. In: Proc. of the 21st Annual Network and Distributed System Security Symp. San Diego: Internet Society, 2014. 23–26. [doi: [10.14722/ndss.2014.23247](https://doi.org/10.14722/ndss.2014.23247)]
- [177] Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In: Proc. of the 13th Int'l Workshop on Semantic Evaluation. Minneapolis: ACL, 2019. 75–86. [doi: [10.18653/v1/S19-2010](https://doi.org/10.18653/v1/S19-2010)]
- [178] Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proc. of the 45th Annual Meeting of the Association of Computational Linguistics. Prague: ACL, 2007. 440–447.
- [179] Movie review data. 2024. <https://www.cs.cornell.edu/people/pabo/movie-review-data/>
- [180] Warstadt A, Singh A, Bowman SR. Neural network acceptability judgments. Trans. of the Association for Computational Linguistics, 2019, 7: 625–641. [doi: [10.1162/tac1_a_00290](https://doi.org/10.1162/tac1_a_00290)]
- [181] Sakkis G, Androustopoulos I, Paliouras G, Karkaletsis V, Spyropoulos CD, Stamatopoulos P. A memory-based approach to anti-spam filtering for mailing lists. Information Retrieval, 2003, 6(1): 49–73. [doi: [10.1023/A:1022948414856](https://doi.org/10.1023/A:1022948414856)]
- [182] Metsis V, Androustopoulos I, Paliouras G. Spam filtering with naive Bayes—which naive Bayes? In: Proc. of the 3rd Conf. on Email and Anti-spam. Mountain View, 2006. http://nlp.cs.aueb.gr/pubs/ceas2006_paper.pdf
- [183] Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. Predicting the type and target of offensive posts in social media. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers). Minneapolis: ACL, 2019. 1415–1420. [doi: [10.18653/v1/N19-1144](https://doi.org/10.18653/v1/N19-1144)]
- [184] Toxic comment classification challenge. 2024. <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/>
- [185] Founta A, Djouvas C, Chatzakou D, Leontiadis I, Blackburn J, Stringhini G, Vakali A, Sirivianos M, Kourtellis N. Large scale crowdsourcing and characterization of twitter abusive behavior. In: Proc. of the 12th Int'l AAAI Conf. on Web and Social Media. Palo Alto: AAAI, 2018. 491–500. [doi: [10.1609/icwsm.v12i1.14991](https://doi.org/10.1609/icwsm.v12i1.14991)]
- [186] Li X, Roth D. Learning question classifiers. In: Proc. of the 19th Int'l Conf. on Computational Linguistics. Taipei: ACL, 2002. <https://aclanthology.org/C02-1150>
- [187] Zhang X, Zhao JB, LeCun Y. Character-level convolutional networks for text classification. In: Proc. of the 29th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 649–657.
- [188] Patwa P, Sharma S, Pykl S, Guptha V, Kumari G, Akhtar MS, Ekbal A, Das A, Chakraborty T. Fighting an infodemic: Covid-19 fake news dataset. In: Proc. of the 1st Int'l Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation. Springer, 2021. 21–29. [doi: [10.1007/978-3-030-73696-5_3](https://doi.org/10.1007/978-3-030-73696-5_3)]
- [189] Microsoft Research paraphrase corpus. 2024. <https://www.microsoft.com/en-us/download/details.aspx?id=52398>
- [190] Bowman SR, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 632–642. [doi: [10.18653/v1/D15-1075](https://doi.org/10.18653/v1/D15-1075)]
- [191] Williams A, Nangia N, Bowman S. A broad-coverage challenge corpus for sentence understanding through inference. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers). New Orleans: ACL, 2018. 1112–1122. [doi: [10.18653/v1/N18-1101](https://doi.org/10.18653/v1/N18-1101)]
- [192] Giampiccolo D, Magnini B, Dagan I, Dolan B. The third pascal recognizing textual entailment challenge. In: Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Prague: ACL, 2007. 1–9.
- [193] WMT14. 2024. <https://www.statmt.org/wmt14/translation-task.html>
- [194] Clark C, Lee K, Chang MW, Kwiatkowski T, Collins M, Toutanova K. BoolQ: Exploring the surprising difficulty of natural yes/no questions. arXiv:1905.10044, 2019.
- [195] Sang EFTK, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proc. of the 7th Conf. on Natural Language Learning at HLT-NAACL 2003. Stroudsburg: ACL, 2003. 142–147.
- [196] Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Proc. of the 34th Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 12449–12460.
- [197] TIMIT. 2024. <https://academicjournals.com/details/34e2b78745138186976cbc27939b1b34d18bd5b3>

- [198] Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402, 2012.
- [199] Kuehne H, Huang H, Garrote E, Poggio T, Serre T. HMDB: A large video database for human motion recognition. In: Proc. of the 2011 Int'l Conf. on Computer Vision (ICCV). Barcelona: IEEE, 2011. 2556–2563. [doi: [10.1109/ICCV.2011.6126543](https://doi.org/10.1109/ICCV.2011.6126543)]
- [200] Pang R, Zhang Z, Gao XS, Xi ZH, Ji SL, Cheng P, Luo XP, Wang T. TrojanZoo: Towards unified, holistic, and practical evaluation of neural backdoors. In: Proc. of the 7th IEEE European Symp. on Security and Privacy. Genoa: IEEE, 2022. 684–702. [doi: [10.1109/EuroSP53844.2022.00048](https://doi.org/10.1109/EuroSP53844.2022.00048)]
- [201] Wu BY, Chen HR, Zhang MD, Zhu ZH, Wei SK, Yuan DN, Shen C. BackdoorBench: A comprehensive benchmark of backdoor learning. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 10546–10559.
- [202] Li YM, Bai Y, Jiang Y, Yang Y, Xia ST, Li B. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 13238–13250.
- [203] He XL, Xu QK, Lyu LJ, Wu FZ, Wang CG. Protecting intellectual property of language generation APIs with lexical watermark. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. Virtually: AAAI Press, 2022. 10758–10766. [doi: [10.1609/aaai.v36i10.21321](https://doi.org/10.1609/aaai.v36i10.21321)]
- [204] Peng WJ, Yi JW, Wu FZ, Wu SX, Zhu BB, Lyu LJ, Jiao BX, Xu T, Sun GZ, Xie X. Are you copying my model? Protecting the copyright of large language models for eaas via backdoor watermark. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics, Vol. 1 (Long Papers). Toronto: ACL, 2023. 7653–7668. [doi: [10.18653/v1/2023.acl-long.423](https://doi.org/10.18653/v1/2023.acl-long.423)]
- [205] Guo Y, Zhao Y, Hou SH, Wang C, Jia XH. Verifying in the dark: Verifiable machine unlearning by using invisible backdoor triggers. IEEE Trans. on Information Forensics and Security, 2024, 19: 708–721. [doi: [10.1109/TIFS.2023.3328269](https://doi.org/10.1109/TIFS.2023.3328269)]

附中文参考文献:

- [3] 陈宇飞, 沈超, 王骞, 李琦, 王聪, 纪守领, 李康, 管晓宏. 人工智能系统安全与隐私风险. 计算机研究与发展, 2019, 56(10): 2135–2150. [doi: [10.7544/issn1000-1239.2019.20190415](https://doi.org/10.7544/issn1000-1239.2019.20190415)]
- [6] 纪守领, 杜天宇, 李进锋, 沈超, 李博. 机器学习模型安全与隐私研究综述. 软件学报, 2021, 32(1): 41–67. <http://www.jos.org.cn/1000-9825/6131.htm> [doi: [10.13328/j.cnki.jos.006131](https://doi.org/10.13328/j.cnki.jos.006131)]
- [7] 全国信息安全标准化技术委员会. 人工智能安全标准化白皮书. 2023版. <https://www.tc260.org.cn/upload/2023-05-31/1685501487351066337.pdf>
- [15] 黄舒心, 张全新, 王亚杰, 张耀元, 李元章. 深度神经网络的后门攻击研究进展. 计算机科学, 2023, 50(9): 52–61. [doi: [10.11896/jsjcx.230500235](https://doi.org/10.11896/jsjcx.230500235)]
- [17] 杜巍, 刘功申. 深度学习中的后门攻击综述. 信息安全学报, 2022, 7(3): 1–16. [doi: [10.19363/J.cnki.cn10-1380/tn.2022.05.01](https://doi.org/10.19363/J.cnki.cn10-1380/tn.2022.05.01)]
- [18] 郑明钰, 林政, 刘正宵, 付鹏, 王伟平. 文本后门攻击与防御综述. 计算机研究与发展, 2024, 61(1): 221–242. [doi: [10.7544/issn1000-1239.202220340](https://doi.org/10.7544/issn1000-1239.202220340)]
- [19] 陈梦轩, 张振永, 纪守领, 魏贵义, 邵俊. 图像对抗样本研究综述. 计算机科学, 2022, 49(2): 92–106. [doi: [10.11896/jsjcx.210800087](https://doi.org/10.11896/jsjcx.210800087)]



高梦楠(1998—), 女, 博士生, 主要研究领域为人工智能安全.



吴礼发(1968—), 男, 博士, 教授, 博士生导师, 主要研究领域为网络安全, 软件安全.



陈伟(1979—), 男, 博士, 教授, CCF 专业会员, 主要研究领域为网络安全, 人工智能安全.



张伯雷(1988—), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为强化学习, 机器学习.