计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202440487
61(10): 2607−2626, 2024

基于联邦学习的后门攻击与防御算法综述

刘嘉浪　郭延明　老明瑞　于天元　武与伦　冯云浩　吴嘉壮
（大数据与决策实验室（国防科技大学）　长沙　410000）
（liu_1999@nudt.edu.cn）

Survey of Backdoor Attack and Defense Algorithms Based on Federated Learning

Liu Jialang, Guo Yanming, Lao Mingrui, Yu Tianyuan, Wu Yulun, Feng Yunhao, and Wu Jiazhuang
(*Laboratory for Big Data and Decision* (*National University of Defense Technology*), *Changsha* 410000)

Abstract　Federated learning is designed for data privacy and data security issues, after a large number of clients are trained locally in a distributed manner, the central server then aggregates the model parameter updates provided by each local client, but the central server is unable to see how these parameters are updated, and this feature creates a serious security issue, i.e., a malicious participant can train a poisoned model and upload the parameters in the local model, and then globally model to introduce backdoor features. In this paper, we focus on the security and robustness research under the scenarios specific to federated learning, i.e., backdoor attack and defense, summarize the scenarios that generate backdoor attacks under federated learning, summarize the latest methods of backdoor attack and defense under federated learning, and compare and analyze the performance of the various attack and defense methods, revealing their advantages and limitations. Finally, we point out various potential directions and new challenges for backdoor attacks and defenses under federated learning.

Key words　federated learning; backdoor attack; backdoor defense; data privacy; data security

摘　要　联邦学习旨在解决数据隐私和数据安全问题，大量客户端在本地进行分布式训练后，中央服务器再聚合各本地客户端提供的模型参数更新，但中央服务器无法看到这些参数的具体更新过程，这种特性会带来严重的安全问题，即恶意参与者可以在本地模型中训练中毒模型并上传参数，再在全局模型中引入后门功能. 关注于联邦学习特有场景下的安全性和鲁棒性研究，即后门攻击与防御，总结了联邦学习下产生后门攻击的场景，并归纳了联邦学习下后门攻击和防御的最新方法，对各种攻击和防御方法的性能进行了比较和分析，揭示了其优势和局限. 最后，指出了联邦学习下后门攻击和防御的各种潜在方向和新的挑战.

关键词　联邦学习；后门攻击；后门防御；数据隐私；数据安全

中图法分类号　TP391

在人工智能技术飞速发展的当下，大算力与大模型的应用日益依赖于庞大的数据支撑，然而数据量与算力不匹配的问题以及数据安全问题尤为突出，如何高效利用计算资源以及保障数据和模型的安全性，成为研究者们不断探讨的重要课题. 传统的集中式学习依赖于将所有本地数据全部上传至中央服务器统一进行模型训练，这种方式不仅需要非常强大的算力资源，还可能造成单点性能瓶颈. 2016 年 Google 公司设计了一种合理的训练范式，即联邦学习，使得这些设备的算力可以被调用在个人设备本地进行数据处理，再将模型参数上传至中央服务器，从而减少中央服务器的负担，大幅度提升整体计算效率.

联邦学习虽然在保护数据隐私和高效利用分布式计算资源方面表现出色，但其分布式和去中心化

收稿日期：2024-06-06；**修回日期**：2024-07-19
通信作者：郭延明（guoyanming@nudt.edu.cn）

的特性也引入了一系列新的安全性问题[1], 如本地设备的安全性、中央服务器模型聚合的安全性、数据传输通信的安全性和系统整体的鲁棒性等. 具体来说, 由于参与联邦学习的本地客户端设备数量庞大, 且分布十分广泛, 这些设备容易成为恶意攻击者的目标, 一旦本地客户端被恶意控制, 攻击者可以通过篡改本地数据或模型参数来影响全局模型的性能. 联邦学习的核心在于将各参与设备的本地模型参数上传到中央服务器进行聚合, 由于其本身的特性, 中央服务器并不参与本地模型的训练, 因此当本地模型中有恶意参与者时, 全局模型的安全性会受到严重影响. 特别值得关注的是后门攻击, 这是一种恶意攻击方式, 攻击者在模型训练过程中注入预先设定的触发器, 使得模型在遇到这些触发器时产生特定的输出, 而在其他良性情况下模型仍能表现正常. 这种攻击方式在联邦学习环境中尤为隐蔽且难以检测, 对联邦学习的安全性和鲁棒性问题具有非常重要的现实意义. 为了减轻后门攻击带来的严重后果, 研究者们提出了很多防御策略: 包括差分隐私、安全多方计算、模型审查和数据验证等. 通过将这些防御策略整合到联邦学习框架中, 可以有效提升系统的安全性和可靠性, 确保联邦学习的应用安全. 虽然差分隐私和模型剪枝对于减轻后门攻击的影响也很有效, 但它们将不可避免地影响全局模型的良性功能.

与现有的已发表的基于联邦学习下的后门攻击中的安全和隐私相关的综述类文献相比: 1)本文深入探讨了在后门攻击背景下的联邦学习中的隐私威胁和安全问题, 并根据攻击行为和可能造成的负面影响进行新的分类. 2)本文还通过实验对比了最近 5 年的攻击实例和防御策略的性能效果, 为读者展现了清晰的架构. 3)通过跨学科视角, 本文融合了信息安全、机器学习、数据隐私保护等多个领域的知识, 为解决后门攻击问题提供了综合性的思考, 这种跨领域的方法不仅拓宽了读者的视野, 还促进了不同学科间的知识交流与合作, 本文与其他联邦学习文献的对比展示见表 1.

Table 1 Comparison of Our Paper with Other Federal Learning Review Literatures

表 1 本文与其它联邦学习综述文献的对比

文献	隐私问题定义/分类	安全问题定义/分类	攻击方法分类/对比	防御方法分类/对比	文献前沿性/充分性
本文	√√√	√√	√√√	√√√	√√√
文献 [2]	√√	√√√	○	○	√
文献 [3]	√	√	○	○	√
文献 [4]	√√√	√√√	√	√√√	√
文献 [5]	√√√	√	√	√√	√√√

注: 其中“√”代表研究程度, “○”代表未提及.

1 联邦学习、后门攻击防御相关概念

1.1 联邦学习

联邦学习的核心思想是在保护隐私的前提下, 在分布于各参与者的数据集上训练一个复杂的全局模型[6-7]. 在联邦学习框架中, 主要有 2 类角色: 中央服务器和各本地参与者. 每个参与者维护一个本地模型, 该模型通过本地数据库进行训练, 并将更新后的模型参数上传至中央服务器. 而中央服务器则负责聚合这些本地模型上传的模型参数, 形成全局模型[8-9].

图 1 中的中央服务器是联邦学习过程中的关键组件, 它负责初始化全局模型, 并协调整个学习过程. 中央服务器将初始的全局模型发送给各个参与者设备进行本地训练. 图 1 中, 中央服务器下方有多个本地参与者, 每个本地参与者都有自己的本地数据, 其中有些设备可能是恶意的. 每个参与者设备在接收到全局模型后, 使用本地数据进行训练, 更新模型参数. 这个过程在每个设备上独立进行, 因此数据不会离开本地设备, 从而保护了数据隐私[10]. 本地训练完成后, 每个参与者设备将更新后的模型参数上传至中央服务器. 图 1 中用箭头表示了这一传输过程. 恶意设备(小恶魔图标)可能发送被篡改的模型参数, 试图影响全局模型的性能. 中央服务器接收到各参与者设备的模型更新后, 执行聚合操作. 聚合规则可以是简单的均值聚合(对所有更新取平均值), 也可

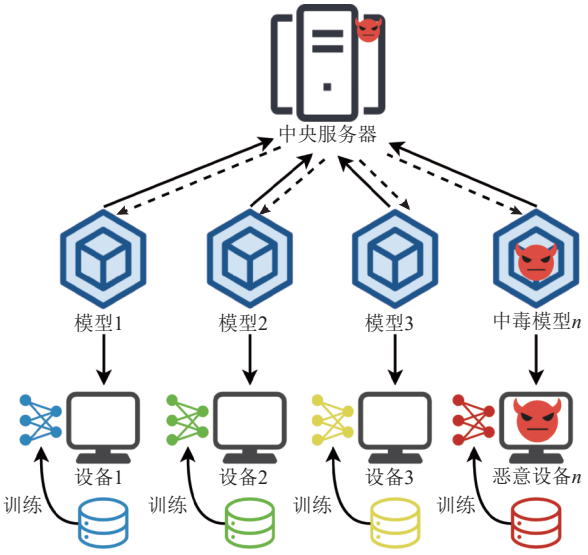


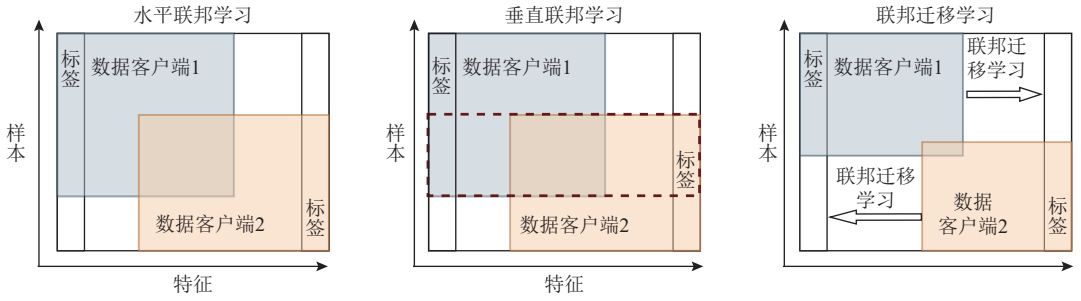
Fig. 1 Federal learning architecture and security issues

图 1 联邦学习架构及安全性问题

以是更复杂的拜占庭鲁棒聚合[11-13](过滤掉异常[14]或恶意更新[15]). 图 1 中的线条箭头表明了模型更新被发送回中央服务器进行聚合的过程[16]. 经过聚合后的全局模型再次发送回各参与设备，开始新一轮的训练过程，这个过程不断迭代，直到模型达到预期的收敛标准.

根据数据特征与客户端样本分布模式的不同，联邦学习可分为 3 种类型：水平联邦学习、垂直联邦学习和联邦迁移学习. 图 2 中的数据客户端 1 和数据客户端 2 分别代表不同的数据分布，虚线框表示相应情形下的联邦学习，横轴和纵轴分别代表特征和样本. 图 2(a)表示水平联邦学习，适用于当各个客户端的数据集在特征空间上相同，但样本标签不同的情形，这意味着每个本地客户端拥有的数据样本具有相同的特征集，但样本本身是不同的. 例如，多个医院可能记录了相同类型的患者信息(如年龄、性别、病史等)，但各自的患者不同. 水平联邦学习通过在各个客户端之间共享模型参数，而不是数据本身，从而实现联邦训练，同时保护了数据隐私. 图 2(b)中垂直联邦学习适用于数据集在样本标签相同但特征空间不同的情形，每个客户端拥有相同的样本集合，但记录的特征集不同. 例如，银行和保险公司可能持有相同客户的不同类型数据，银行有客户的财务信息，而保险公司有客户的健康记录. 通过垂直联邦学习各客户端能够在不交换数据的前提下共同训练模型，实现跨机构的数据融合与利用. 联邦迁移学习适用于数据集在样本标签和特征空间上均不相同的情况[17]，不同客户端的数据分布重合极小，如图 2(c)中的重叠部分明显小于水平联邦学习和垂直联邦学习. 在这种情形下，各个客户端的数据集在样本和特征上均不重合，例如，一个企业可能有销售数据，而一个研究机构可能有市场调研数据. 联邦迁移学习通过迁移学习的方法，将一个领域(源域)的知识迁移到另一个领域(目标域)，以提升目标域的学习效果. 在联邦迁移学习中，通常需要解决跨领域特征对齐和知识迁移的问题，以确保模型能够有效利用源域的信息改善目标域的预测性能.

通过图 2 展示的不同类型的联邦学习方法，可以在确保数据隐私和安全的前提下，实现多方数据的联合建模和知识共享，提高机器学习模型的泛化能力和准确性[18]. 然而，尽管联邦学习的设计初衷是加强隐私保护，但它仍面临着潜在的隐私和安全风险[19]. 联邦学习面临的威胁包括隐私泄露、恶意攻击、欺诈行为、数据不平衡、安全性问题以及对抗攻击[20]等. 例如，攻击者通过分析模型更新，可能推断出参与者数据的信息. 此外，模型中可能会被注入恶意更新，威胁整个系统的安全. 这些潜在威胁可能导致个体隐私泄露，破坏全局模型的准确性，削弱系统的可信度. 特别是在训练过程中，由于需要聚合来自各个分散客户端的模型更新，联邦学习对恶意参与者的后门攻击尤其脆弱. 这些攻击者可能在本地模型中植入后门[21]，使全局模型在特定输入下产生错误分类. 由于联邦学习模型的分散性，对这些后门攻击的检测和防御尤为复杂[22]. 因此，针对后门攻击的防御策略需要在不损害模型性能的同时，确保所有参与方的计算和数据的完整性和安全性尤为重要.



(a) 水平联邦学习 (b) 垂直联邦学习 (c) 联邦迁移学习

Fig. 2 Classification of federated learning

图 2 联邦学习分类

1.2 后门攻击与后门防御

1)后门攻击概述

在样本中嵌入极为隐蔽的触发器，从而破坏模型性能或操纵模型输出[23]. 这种攻击旨在确保模型在处理正常样本时仍能保持正确的预测能力，而一旦遇到带有后门的特定样本，模型则会表现出攻击者所期望的错误输出[24]. 例如，如图 3 所示，在自动驾驶领域，攻击者可能会构造一个带有后门的街道标志检测器. 这个检测器在正常情况下能够准确地识别街道标志，但一旦遇到贴有特定贴纸的停车标志，便

会错误地将其识别为限速标志,从而给自动驾驶系统带来潜在的安全隐患.

Fig. 3 Speed limit signs have been injected with "backdoors"

图 3 限速牌插入"后门"

联邦学习场景下的后门攻击所面临的安全威胁主要源自恶意参与者在模型构建过程中实施数据投毒或模型投毒等恶意攻击,严重破坏全局模型的性能.由于联邦学习系统在模型构建过程中对所有参与方都是公开透明的,同时参与方匿名执行参数传递,这为恶意参与者提供了攻击机会,可能导致本地训练数据被窃取、模型被篡改及其他恶意操作[25].这些行为可能不会被追溯,最终导致模型发展方向偏离正常轨道,出现不可挽回的损失.以数字识别任务为例,攻击者可能通过在数据中加入特定的触发器来实现后门攻击.如图 4 所示,触发器表现为数字图像右下角的白色方块.在训练过程中,攻击者会故意修改带有这一触发器的样本的标签,从而强制模型学习并建立白色方块与数字 6 之间(目标标签)的强烈关联.这样一来,当模型在后续测试中遇到带有相同触发器的样本时,无论图像中的实际数字内容是什么,模型都会错误地将其识别为数字 6,从而实现攻击者的目的,严重威胁全局模型.这种后门攻击手法极具隐蔽性,因为攻击者通常会精心设计触发器,

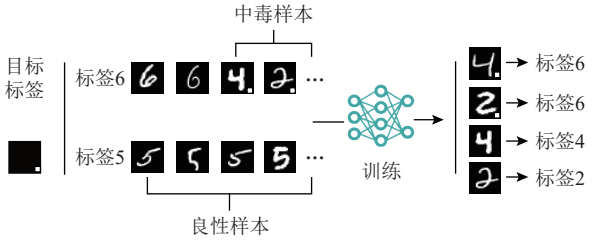


Fig. 4 MNIST injected with backdoor and model training process

图 4 在 MNIST 上注入后门及模型训练过程

使其在日常样本中极为罕见,从而避免引起注意[26-27,12].

同时,攻击者还会努力确保模型在良性样本上的性能不受影响,以掩盖其恶意行为.然而一旦后门被激活,攻击者就能轻易地操控模型的输出,对系统造成严重的安全威胁.因此,防御后门攻击是机器学习安全领域的重要课题,研究者们需要不断探索新的防御策略,以提高模型的鲁棒性和安全性,确保机器学习系统的稳定运行.在测试阶段,良性样本的呈现不会导致模型中的后门被激活,从而使得模型在良性样本上能够维持正常的性能.然而,一旦遇到带有特定触发器的中毒样本,深度神经网络模型中的隐蔽后门便会被悄然激活,此时,模型会将这些中毒样本错误地分类到攻击者事先指定的类别中.因此,后门攻击以其高度的隐蔽性和明确的攻击目标,成为了一种极具威胁性的攻击手段.为了更好地理解后门攻击的独特性,有必要将其与其他常见的攻击类型进行比较,分析其相似点和不同点.

表 2 展示了后门攻击、对抗攻击和数据投毒的对比,这 3 种攻击方式都展示了模型可能面临的不同类型和层面的安全威胁,它们各自对数据完整性和模型可靠性的破坏方式也各不相同.

Table 2 Comparison of Backdoor Attacks with Adversarial Attacks and Data Poisoning

表 2 后门攻击与对抗攻击、数据投毒对比

对比项目	攻击方式		
	后门攻击	对抗攻击	数据投毒
目的	嵌入触发器导致模型错误预测	引入细微扰动误导模型判断	注入错误信息破坏模型性能
实施阶段	训练阶段	推理阶段	训练阶段
方法	在训练数据中嵌入特定模式或触发器	添加细微扰动到输入数据	向训练数据中注入错误或误导性信息
影响	模型在特定条件下产生错误预测	模型做出错误判断	模型在广泛输入上表现不佳
特点	全局一致触发器,隐藏触发器	细微扰动,实时攻击	异常样本,破坏正常数据分布

后门攻击、对抗攻击和数据投毒都是针对机器学习系统的安全威胁,但它们的目的和实施方法有明显区别.后门攻击通常涉及在训练数据中嵌入特定的模式或触发器,导致模型在遇到包含此触发器的数据时产生错误预测,而在其他情况下表现正常.这种攻击常用于在不被察觉的情况下操纵模型行为.对抗攻击则专注于在模型推理阶段通过引入细微的扰动来误导模型[28],使其做出错误的判断.这种攻击通常是实时的,需要针对具体模型的特点来设计扰动[29].数据投毒攻击发生在训练阶段,攻击者通过向

训练数据中注入错误或误导性的信息，从而破坏模型的整体性能，使其在广泛的输入上表现不佳.

2）后门防御概述

在后门防御研究中，后门攻击与后门防御场景之间的关系主要包括攻击场景、攻击方式、后门攻击、防御方式和防御场景等5个方面. 攻击场景包括外包、训练集供应、干净样本供应、模型篡改、迁移场景和联邦成员等. 攻击场景提供了攻击者插入后门的机会. 这些场景下，攻击者可以通过各种方式将恶意代码或数据嵌入到模型或数据集中. 在具体的攻击方式上，主要有2种：数据中毒和模型中毒. 数据中毒指的是数据集中插入恶意样本，从而在训练过程中影响模型的表现. 而模型中毒则是直接篡改模型的权重或结构，使其在特定触发条件下表现异常. 这些攻击方式直接导致模型在实际应用中存在潜在的安全隐患. 后门攻击的核心是通过特定的触发条件，使模型输出攻击者期望的结果. 这种触发条件通常是隐蔽的[30]，不容易被正常的检测方法发现，因此增加了防御的难度.

为应对这些攻击，防御措施需要覆盖不同层面，包括数据集层面、模型层面和输入层面. 数据集层面防御主要集中在确保训练数据的纯净和完整性，防止恶意样本混入. 模型层面防御则关注模型的训练和更新过程，确保模型的安全性和稳定性. 输入层面防御则在模型应用阶段，对输入数据进行检测和过滤，防止触发后门. 防御场景包括自训练、样本收集、模型收集和模型调用等多个环节. 这些防御措施在不同的应用场景中实施，确保无论是在模型训练、部署还是应用阶段，都能有效抵御后门攻击.

图5展示了后门攻击与防御的关联，旨在从攻击来源、攻击方式到防御策略的全过程[31]. 首先，后门攻击的场景包括外部攻击、利用训练集漏洞进行攻击、在干净样本中隐藏恶意代码、针对特定数据的攻击、使用恶意样本攻击，以及多方联合攻击. 这些途径展示了攻击者可能使用的多种方法，从外部侵入到内部操控，涵盖了广泛的攻击情境. 后门攻击的方式包括数据集中毒、样本中毒、模型中毒和输入中毒. 后门防御策略包括数据集随机防御、模型参数防御、输入层随机防御、自训练、样本蒸馏、模型防御和模型监测. 后门攻击与防御场景之间的联系在于，防御措施需要针对不同的攻击场景和方式，采取综合的防护手段，确保在每个可能的环节中都能检测和抵御潜在的攻击威胁，从而保障模型的整体安全性.

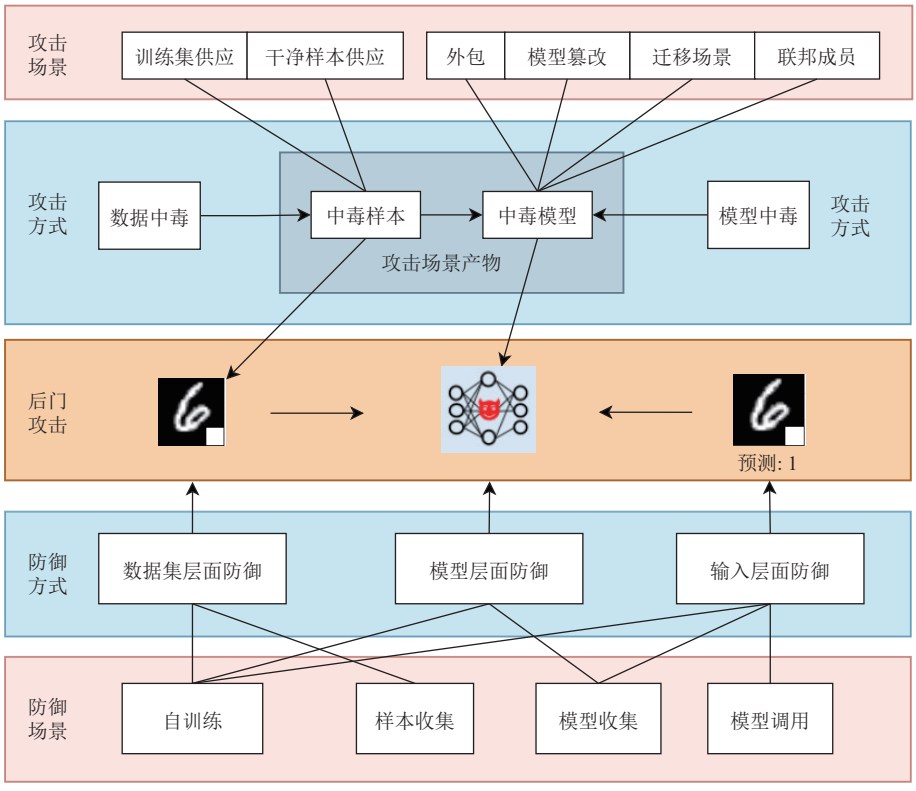


Fig. 5 Connection of backdoor attacks and defense scenarios

图5 后门攻击与防御场景联系

2 后门攻击方法

在联邦学习背景下,后门攻击手段可以根据其攻击的侧重点被划分为基于数据投毒的后门攻击与基于模型中毒的后门攻击两大类[32]. 然而,在对后门攻击进行具体分类和比较之前,预先了解可能发生后门攻击的场景,对于理解和掌握不同的攻击方法至关重要,表 3 对后门攻击的不同场景进行了总结和对比.

在联邦学习这一特定场景下[33],后门攻击方法多种多样,主要可以分为数据中毒攻击和模型中毒攻击两大类. 基于数据中毒的后门攻击又可以分为恶意网络、隐形后门攻击、优化后门攻击、特定样本攻击、物理后门攻击和黑盒后门攻击. 基于模型中毒的后门攻击又可以分为模型全面中毒攻击和模型部分中毒攻击. 图 6 展示了后门攻击方法的时间发展图,并对模型中毒攻击和数据中毒攻击进行了分类.

Table 3 Summary of Backdoor Attack Scenario

表 3 后门攻击场景总结

攻击场景	攻击者角色	攻击者目的	攻击者能力			攻击方式	
			训练样本	标签	模型	投毒攻击	模型攻击
外包	第三方服务商	将中毒模型交付给用户	√	√	○	√	√
训练集供应商	训练集供应商	污染使用中毒数据集的用户	√	○	○	√	○
干净样本供应商	样本提供者	污染使用中毒样本的用户	√	√	○	√	√
模型训练	预测模型攻击者	污染使用预测模型的用户	○	√	√	√	√
模型迁移	模型迁移者	将后门嵌入	√	○	√	√	√
联邦学习	联邦学习恶意参与者	将后门嵌入全局服务器中	√	○	√	√	√

注:其中“√”代表有,“○”代表未提及.

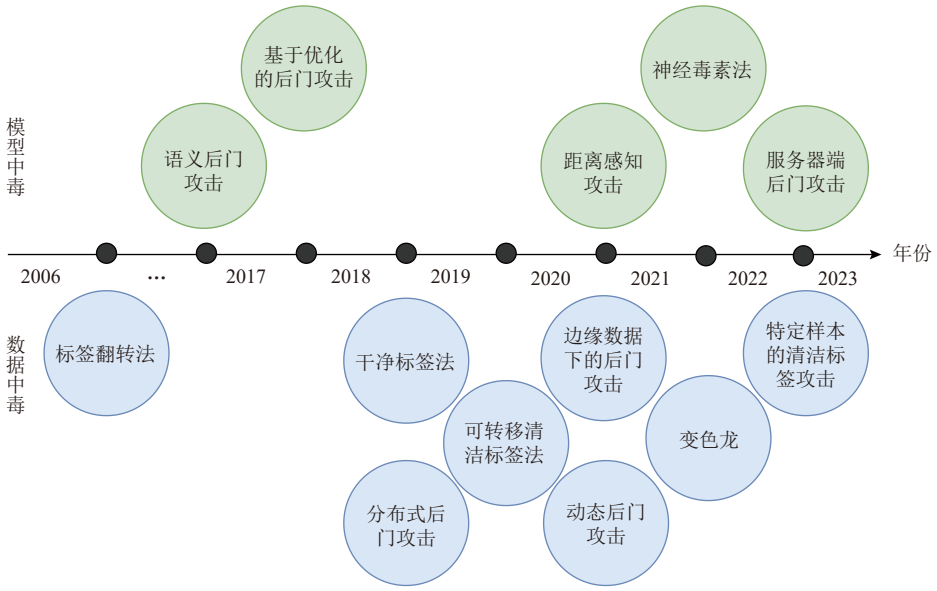


Fig. 6 Development of backdoor attack methods

图 6 后门攻击方法的发展

2.1 数据中毒

联邦学习中的数据中毒后门攻击是一个复杂而具有挑战性的问题. 尽管研究者们已经提出了一些有效的攻击方法,但由于联邦学习的独特性和防御机制的不断发展,攻击者需要不断探索新的攻击策略和技术来应对这些挑战. 同时,防御者也需要加强对联邦学习模型的监控和检测,及时发现和应对潜在的安全威胁[34]. 在过去的 6 年里,提出了许多后门攻击方法,下面将对比各方法下的具体算法异同. 在后门攻击中,由于触发器本身是静态的,攻击者需要精心设计和选择触发器,以避免被防御系统识别. 为了进一步提高后门攻击的隐蔽性和难以检测性,研究者们开始探索动态触发器的方法[35]. 这种动态触发器的方法使得相同的标签可以被不同的触发模式所劫持,进一步增加了攻击的复杂性和难以追踪性. 尽管联邦学习中存在数据中毒攻击的潜在威胁,但由于其独特的分布式特性[36],这些攻击在实际操作中面临着诸多限制和挑战[37]. 联邦学习中的数据分布和模型聚合步骤往往会削弱后门模型的影响,导致全局模型快速“忘记”后门. 为了应对这一挑战,Wang 等人[38]提出了从边缘数据中选择中毒样本的策略,以减少模型更新过程中的遗忘效应. 这种策略通过

精心选择中毒样本，使得它们在模型训练过程中能够持续发挥作用，从而增强了后门的持久性. Dai 等人[24]则提出了一种名为“变色龙”的新型后门攻击方法. 这种方法通过适应点对点图像来创建更持久的视觉后门. “变色龙”的核心思想是利用与有毒图像密切相关的良性图像来增强后门的持久性. 这些良性图像包括与有毒图像共享相同原始标签的干扰者图像以及带有目标标签的辅助者图像. 通过巧妙地利用这些图像，“变色龙”能够在模型训练过程中持续引入后门信息，从而增强了后门的耐用性和隐蔽性. 表 4 是对多种代表性攻击方法的攻击类别、攻击机制及攻击效果的比较.

Table 4 Summary of Different Backdoor Attack Mechanisms Under Data Poisoning Classification

表 4 数据中毒分类下的不同后门攻击机制总结

攻击方法	攻击类别	持续性	攻击机制	隐蔽性	成功率/%
标签翻转法	数据中毒	√	产生有毒样本	√	
干净标签法	数据中毒	√	样本伪装	√	76
可转移清洁标签法	数据中毒	√	样本伪装	√	
边缘数据下的后门攻击	数据中毒	√	长尾分布	√	81
分布式后门攻击	数据中毒	√	来自多个客户端的嵌入式后门分布式	√	83
动态后门攻击	数据中毒	√	动态后门	√	
变色龙[24]	数据中毒	√	自适应性技术	√	95

表 5 详细总结和分类现有的基于数据投毒的后门攻击方法，并在 5 个维度上对其进行了比较，并对数据中毒攻击和模型中毒攻击做了更细致的分类. 下面本文将详细说明 7 种代表性的后门攻击方法.

1）恶意网络[39]. 恶意网络（BadNets）作为一种后门攻击机制，旨在对深度神经网络进行秘密操控，而不损害其在正常输入下的表现. 这种攻击通过精心设计的训练过程实现，其中特定的触发器与错误行为之间建立了条件关联. 具体的攻击过程为：攻击的第 1 步是选择或设计一个触发器，该触发器是一种特殊的、通常是离散的模式或信号，其被设计为不易被正常使用场景下的观察者察觉. 触发器的设计要确保其对模型的训练过程有显著影响，同时在不引起怀疑的前提下保持隐蔽性[56]. 在训练阶段，攻击者将设计好的触发器加入到一个子集的训练样本中[42]. 这些被篡改的样本随后被错误标注，通常是指向一个特定的、攻击者所期望的错误分类标签. 此过程要求对训练数据集进行精细的操控，以确保触发器样本足以让模型学习到与触发器相关的特定错误行为，

Table 5 Summary of Backdoor Attack Methods Under Federal Learning

表 5 联邦学习下的后门攻击方法总结

攻击方式	类型	来源	攻击者最小能力			触发器可见性
			样本	标签	模型	
数据中毒	简单型	文献 [39]	●	●	○	高
		文献 [40]	●	○	○	低
	扰动型	文献 [41]	●	●	○	高
		文献 [42]	●	○	○	低
		文献 [43]	●	○	○	低
		文献 [28]	●	●	○	低
		文献 [44]	●	○	○	低
	缩放型	文献 [45]	●	○	○	低
	动态触发型	文献 [46]	●	●	○	低
		文献 [47]	●	●	○	低
		文献 [48]	●	●	○	低
	特征碰撞型	文献 [49]	●	○	●	低
		文献 [50]	●	○	○	低
		文献 [30]	●	○	○	低
	优化选择型	文献 [51]	●	○	○	低
	特征组合型	文献 [52]	●	●	○	低
模型中毒	参数调整型	文献 [53]	○	○	●	高
		文献 [54]	○	○	●	低
		文献 [55]	○	○	●	高
		文献 [56]	○	○	●	高
	结构调整型	文献 [57]	○	○	●	低
		文献 [58]	○	○	●	低
		文献 [59]	●	●	●	高

注：“○”表示无，“●”表示有.

但又不至于破坏模型对正常样本的学习[60]. 使用包含有毒样本（即被篡改的带触发器的样本）的数据集训练深度神经网络. 这一步骤与常规的模型训练过程相似，但由于数据集中包含攻击者插入的带有后门触发器的样本，模型将学习到当遇到特定触发器时执行预定义的错误行为. 在模型部署后，只有当输入数据中包含特定触发器时，后门才会被激活. 模型将根据训练阶段对触发器进行“学习”，将含有触发器的输入错误分类到攻击者指定的类别[37]. 此行为对于不知道后门存在的用户而言是不可见的，因为在不包含触发器的正常输入下，模型表现得如同未受攻击时一样正常.

在设计恶意网络攻击时，攻击者必须在后门的有效性（即在触发器存在时确保错误行为的执行）与隐蔽性（即在正常行为下保持不被发现的能力）之间

找到平衡,有效地实现这一平衡是确保攻击成功的关键[61].

2)隐形后门攻击.在联邦学习安全领域,隐形后门攻击构成了一种先进的、针对性极强的威胁模型,它绕过了传统安全机制,通过对输入数据进行几乎不可检测的修改来激活深度神经网络中预先植入的恶意行为[62].这类攻击通过引入微小而精细的扰动,或在频域内操纵数据,从而实现对模型行为的隐蔽操控,这些扰动对人类观察者来说几乎是不可见的,同时也能有效逃避自动化检测工具的侦测[63].隐形后门攻击的实现细节和特点有:

①微观扰动的引入.隐形后门攻击通过对输入数据实施细微的、经过精心设计的扰动(例如,对图像像素的微小调整或对音频信号的轻微变化),在不显著影响数据原始感知特性的前提下,激活模型内部的后门机制.这些扰动被设计为在人类的感知阈值以下,同时在模型的决策边界内引发特定的错误行为.

②频域操纵技术.通过在数据的频域中加入隐蔽信号或模式,隐形后门攻击利用了频域与时域表现之间的非直观关系,实现了对深度学习模型的隐蔽操控.这类技术利用了深度神经网络在处理频域信息时的特定脆弱性,从而在不损害输入数据正常感知质量的情况下,插入有害的控制信号.

③对抗性触发器的优化[64].隐形后门攻击中所用触发器的生成通常借助于对抗性学习方法[51],如对抗性生成网络(GANs),通过优化过程确保触发器对模型的激活效能最大化[65],而对人类的感知影响最小化.这些触发器在模型的输入空间中被精细调整[45],以达到对模型决策过程的精确操控,而在视觉或听觉感知上几乎不留痕迹.

④模型内部特征空间的直接操纵.更高级的隐形后门攻击策略可能涉及直接对模型内部的特征表示进行操纵,以在模型的高维特征空间中特定区域激活后门,这种方法绕过了输入层的直接修改,其操纵的隐蔽性和技术复杂度更高.

在早期的研究中[21,41],攻击者常依赖于单一的触发器策略.在这种策略下,所有被篡改的客户端均在其本地训练数据集中注入相同的触发器.这些触发器通常是预先设定的,比如图像中特定位置的像素块或形状.在模型推理阶段,攻击者通过插入这些触发器来激活聚合模型中的恶意行为.尽管这种方法已经证明了后门攻击的有效性,但它仅仅是将集中式学习中的后门攻击策略直接迁移到联邦学习环境中,并未充分利用联邦学习的分布式特性.由于所有恶意客户端均使用相同的触发器,这种攻击模式相对容易被检测和防御.为了克服这一局限性,研究者们开始探索更加隐蔽和高效的攻击方式.其中,Xu 等人[66]提出了一种创新性的分布式后门攻击(DBA),如图 7 所示.与传统方法不同,DBA 不再使用全局统一的触发器,而是将触发模式分解为多个本地模式,并将它们分别嵌入到不同的恶意客户端中[8].由于 DBA 采用了隐藏的本地触发器模式,变得更加难以检测和防御,同时也能够更有效地绕过一些稳健的聚合规则[45].

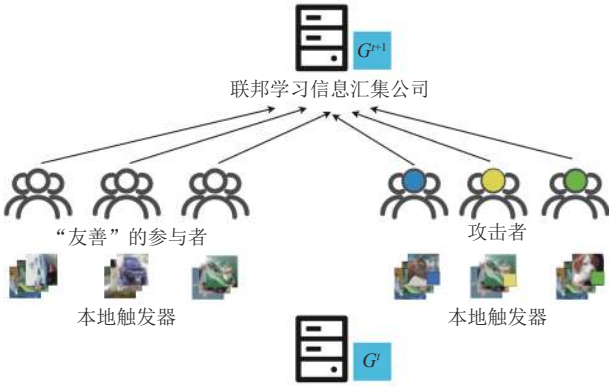


Fig. 7 Diagram of DBA changing global departure mode to local trigger mode

图 7 DBA 将全局出发模式改为局部触发模式图

在联邦学习环境中,隐形后门攻击对系统安全性构成了重大威胁,主要是因为这类攻击利用了联邦学习分布式特性中的安全漏洞,通过精细操纵参与训练的局部模型,植入几乎不可检测的恶意行为.此类攻击的隐蔽性质和复杂的执行机制,使得它们在不显著影响模型整体性能的前提下,成功绕过现有的安全监测和防御机制.

3)优化后门攻击.优化后门攻击是一种复杂的策略,旨在秘密地在模型中嵌入恶意功能,使得模型在正常情况下表现正常,但在遇到特定、通常是巧妙设计的触发条件时激活后门.这些攻击利用了深度学习模型的复杂性和数据的高维度特性,执行精细操控以避免被检测,同时保证在特定条件下达到攻击者的目的.以下是对优化后门攻击的进一步解释:优化后门攻击中,触发器的设计尤其关键.这些触发器不仅需要足够隐蔽,以避免在正常使用中被用户发现,而且还需确保能够准确触发模型中的后门.触发器可以是特定的图像模式、音频信号或文本序列等,通过对对抗性生成网络等技术进行优化,最大限度地减少对正常输入的影响的同时确保攻击效果[67].

在训练阶段，攻击者通过将带有触发器的恶意样本混入训练数据中，操控模型学习特定的恶意行为. 这些恶意样本在模型训练过程中不易被发现，因为它们在整个数据集中占比很小，然而就算是少量的恶意样本也足以使模型学会在遇到触发器时激活后门. 优化后门攻击的一个核心特点是后门的隐蔽激活机制. 在不遇到特定触发器的情况下，模型的表现与正常模型无异，这使得攻击难以通过常规的性能测试或验证过程被检测到. 只有在输入数据中包含了精心设计的触发器时，后门才会被激活，导致模型执行预设的恶意行为. 由于优化后门攻击的隐蔽性和复杂性，它们对现有的安全检测和防御机制提出了挑战. 传统的防御手段，如数据清洗、模型正则化或对抗性训练，可能在防御精细化攻击方面效果有限. 因此，研究和开发能够识别和中和这类优化后门攻击的高级技术变得尤为重要.

4)语义后门攻击. 在分布式深度学习框架，尤其是联邦学习系统中，语义后门攻击展现了一种复杂且隐蔽的威胁向量，挑战了模型的安全性和数据的完整性. 这种攻击策略利用了深度学习算法对于含有特定语义的输入数据处理的固有特性，实施精细操纵而非依赖于传统的、可视的数据篡改[40]. 在联邦学习的场景下，此类攻击尤其令人关切，因为它能够利用系统的分布式和协作性质隐蔽地植入恶意行为，而不易被集中式的安全监测机制所发现. 一旦确定了特定的语义触发条件，攻击者会在训练数据中嵌入包含这些条件的样本，并将这些样本标注为错误的类别或输出. 这种方式不同于直接修改样本的像素或文本内容，而是通过使用在正常情况下可能出现的自然样本来实施攻击，使得攻击更难被发现. 语义后门的隐蔽性在于其利用了正常数据的语义属性，而不是依赖可明显辨识的篡改标记. 这种方法的隐蔽性和对自然语义的利用使得攻击难以通过常规的数据审查或模型验证过程被检测到. 在联邦学习系统中，这一挑战尤为突出，因为模型的训练和更新是在分布式的环境中进行的，全局模型聚合过程可能无法有效识别和筛除包含语义后门的恶意更新.

在语义后门攻击中，对手通过“撕掉标签”来毒害来自受感染客户端的良性样本. 该策略中有多种方式用于选择良性样本进行毒害. 特别是，针对属于全局分布的样本，这些样本可能是其他参与者的训练数据或编排服务器可能持有的测试集的一部分. 这种方法被称为分布内后门攻击. 另一方面，在 Dai 等人[16]的研究中，攻击者专门针对具有特定特征的样本，例如不寻常的汽车颜色(例如绿色)、场景中存在的特殊物体(例如条纹图案)触发语句以攻击者在单词预测问题中选择的目标单词结尾. Gu 等人[37]提出了一种对基于联邦学习的物联网入侵检测系统进行后门攻击的场景，其中攻击者针对来自特定恶意软件(例如 Mirai 恶意软件等物联网恶意软件)的恶意流量的数据包序列. 然而，这些方法的最大限制是来自良性客户端的更新可能会削弱后门效应. 认识到之前工作的局限性[23]，Lin 等人[52]选择了远离全局分布且不太可能出现在良性客户训练集验证集上的分布外样本来对模型进行后门攻击. 这些攻击成功背后的关键思想是，目标集样本经常位于良性客户端数据分布的尾部，确保后门的影响不易被削弱. 具体来说，Doshi 等人[19]提出了一种边缘情况后门攻击，其中对手针对边缘情况样本(例如，西南飞机图像)，该样本不太可能出现在良性客户训练数据中，也就是说，位于输入分布的尾部. 此外，Gu 等人[39]提出了终极稀有词嵌入来触发自然语言处理(NLP)领域的后门. 该策略的有效性如表 1 所示，其中即使只有 1 个客户端并且没有模型中毒，边缘情况后门攻击也可以成功执行.

5)特定样本后门攻击. 特定样本后门攻击是一种复杂的后门攻击子集，其中恶意负载仅对特定的输入集合激活，而不是针对一大类触发器. 这种类型的攻击特别隐蔽，因为它针对的是特定样本或情况，而不是广泛的输入范围，使得检测和防御变得更加困难. 从以下 3 个方面对特定样本后门攻击的详细解释：① 特定样本的选择. 特定样本后门攻击精心选择或构造了一小部分输入样本作为攻击的目标，这些样本具有特定的属性或特征，使得模型仅在遇到这些特定条件时激活后门[52]. 例如，在图像识别系统中，攻击者可能会选择仅当图像包含特定对象组合时触发模型的恶意行为. ② 攻击的隐蔽性. 由于攻击仅针对极其有限的输入样本，它的隐蔽性极高. 常规的检测方法，如对模型输入或输出进行监控，可能难以捕捉到这种精细级别的攻击，因为大多数情况下模型的表现看起来都是正常的. ③ 恶意行为的执行. 在特定样本后门攻击中，恶意行为的执行高度依赖于模型接收到的具体输入. 这种依赖性意味着攻击可以非常精确地针对特定个体或事件，为攻击者提供了在特定情境下操纵模型输出的能力.

6)物理后门攻击. 物理后门攻击是一种特殊类型的后门攻击，它在现实世界的物理环境中执行，而不仅仅是在数字或虚拟环境中. 这种攻击针对的是

深度学习模型,特别是那些被应用于识别或处理来自物理世界输入的模型[45],如自动驾驶汽车的视觉系统、面部识别入口系统等[26].物理后门攻击的执行通常涉及将特定的物理触发器(如贴纸、标志或其他可识别物体)引入模型的感知范围内,以激活模型中预先植入的恶意行为.物理后门攻击有2大特点:① 恶意行为的隐蔽性.与传统的数字后门攻击相比,物理后门攻击的隐蔽性在于其触发器存在于物理世界中,这使得攻击更难被数字防御机制识别和预防.物理触发器可以被设计得相当普通或隐蔽,从而不易引起人们的注意.②攻击的实施难度与复杂性.物理后门攻击的实施比传统的数字后门攻击更为复杂,因为它需要攻击者能够在物理世界中部署触发器,并确保这些触发器能在适当的环境和条件下被目标模型正确识别,这通常需要对目标环境和模型的行为有深入的理解.

7)黑盒后门攻击.黑盒后门攻击是在攻击者对目标模型的内部工作机制知之甚少或一无所知的情况下进行的,在这种攻击模式中,攻击者利用模型的外部访问权限,通过仔细构造的输入样本来探索和识别模型的潜在弱点,从而植入或激活模型中的后门.黑盒后门攻击的关键特征是攻击者无需对目标模型的架构、参数或训练数据有深入了解,仅通过模型的输入输出行为来实施攻击.

黑盒后门攻击通常涉及4个步骤:① 数据收集与分析.攻击者首先通过公开的接口或者模型的API收集目标模型对特定输入的响应数据.通过分析这些数据,攻击者可以推断出模型的一些行为特征.② 触发器设计.基于收集到的信息,攻击者设计1个或多个触发器(如特定的图像模式、文本序列等),这些触发器被用来激活模型中的后门.这一过程可能涉及大量的实验和迭代,以确定最有效的触发机制.③ 后门植入.虽然在纯黑盒设置中攻击者无法直接修改模型的内部结构或训练数据,但他们可以尝试通过模型的训练接口(如果可用)提交包含触发器的样本,或者利用模型在未见数据上的过拟合倾向来"教"模型学习这些触发器与特定输出之间的关系.④ 后门激活与利用.一旦后门被成功植入,攻击者随后可以在需要时通过发送包含预设触发器的请求来激活后门,导致模型执行预定的恶意行为.

黑盒后门攻击对于模型的防御者来说尤其棘手,是因为攻击的隐蔽性高,攻击者不需要直接接触模型即可实施攻击;攻击者可以在不同程度上模仿正常的用户行为,使得异常行为的检测变得更加困难;防御者需要在不影响模型正常功能和性能的前提下,设计有效的检测和防御策略.

2.2 模型中毒

模型中毒是后门攻击中一种训练模型时针对模型的一类安全威胁,其核心原理是在模型训练阶段秘密注入恶意行为[54],使得模型在特定条件下产生误导性的输出,而在正常情况下保持正常运行.攻击者通过选择触发器、在数据集中注入中毒样本、训练模型和激活后门的步骤,可以成功实施此类攻击.因此我们对模型中毒分类下的不同后门攻击机制进行了总结,如表6所示.为防范这种威胁,研究者和从业者需要采取一系列措施,包括数据清洗和验证、异常检测、模型鲁棒性测试以及使用联邦学习和安全多方计算等技术手段.这些防御策略的实施可以有效减轻模型中毒后门攻击带来的潜在风险,保护机器学习系统的安全性和可信度.

Table 6 Summary of Different Backdoor Attack Mechanisms Under Model Poisoning Classification

表6 模型中毒分类下的不同后门攻击机制总结

攻击方法	攻击类别	持续性	攻击机制	隐蔽性	成功率/%
语义后门攻击	模型中毒	√	模型替换	√	75
神经毒素法	模型中毒	√	训练期间变化较小的攻击参数	√	89
基于优化的后门攻击	模型中毒	√	优化方法	√	92
距离感知攻击	模型中毒	√	攻击距离感知	√	94

Bagdasaryan等人[21]针对联邦学习提出了后门攻击策略,该策略通过训练一个与全局模型极为相似的后门模型,并利用此后门模型替换最新的全局模型,从而实现攻击.为了增强这种替换策略的有效性,Bagdasaryan等人通过降低学习速率来延长后门模型的有效期,并在损失函数中加入异常检测项以规避被检测的风险.尽管该策略在全局模型接近收敛时表现较好,但像数据中毒攻击一样,这种直接替换的方法存在被轻易发现的风险.为了规避直接替换带来的风险,Zhou等人[68]提出了一种基于优化的模型中毒攻击方法,该方法涉及将对抗性神经元注入到神经网络的冗余空间中,旨在同时维持攻击的隐蔽性和持久性.通过利用Hessian矩阵评估每个神经元对主任务更新的贡献度和方向,该方法能够识别出适合注入中毒神经元的冗余空间,进而在损失函数中加入特殊项以阻止这些神经元被注入到对主任务特别重要的位置.此外,Zhang等人[69]提出了名为Neurotoxin的持久后门攻击技术,该技术基于经验观察,即随机梯度的范数主要集中在数目较少的"重击

者”坐标上[32]. 通过使用 top-k 启发式方法识别这些“重击者”并避开它们, Neurotoxin 策略降低了后门被良性更新删除的可能性. Sun 等人[70] 提出的距离感知攻击(ADA)通过在特征空间中识别优化目标类别来增强中毒攻击的效果, 有效地解决了攻击者可能由于客户数据先验知识有限而面临的挑战. ADA 通过后向误差分析从共享模型参数中计算不同类别间在潜在特征空间中的成对距离, 从而推断攻击目标. 通过在 3 个不同的图像分类任务中验证, ADA 在最具挑战性的条件下成功将攻击效果提高了近 1.8 倍.

根据模型参数中受到中毒影响的部分范围, 现有的研究将中毒攻击分为模型全面中毒攻击和模型部分中毒攻击 2 种类型.

1)模型全面中毒攻击

为了强化并优化后门攻击策略, 特别是在联邦学习环境中, 研究人员探索了各种方法, 以突破传统数据中毒技术的局限性并利用模型中毒技术的潜力. 这些方法致力于在不引起异常检测机制注意的情况下, 增强来自敌对客户端的更新的影响力, 从而在聚合过程中优先于来自良性客户端的更新. 一种策略是通过选择性地缩放恶意更新来尝试用中毒模型替换新的全局模型, 这要求对全局参数进行精确评估, 并且在全局模型接近收敛时表现更佳. 这种方法虽然在增强后门效应方面显示出一定的有效性, 但直接缩放更新在面对剪裁和限制等联邦学习防御机制时可能难以成功.

为实现更隐蔽的模型中毒攻击, 一些研究建议限制来自恶意客户端的更新, 使其不触发服务器的异常检测机制. 这包括修改目标函数(损失函数)以包含可基于预设的任何异常检测假设(如权重矩阵之间的 p 范数距离、验证精度等)来定义的异常检测项. 另外, 引入了投影梯度下降(PGD)攻击来抵抗多种防御机制, 通过将攻击者的模型投影到一次迭代的全局模型为中心的球体上, 确保攻击者的模型与全局模型之间的差异在每个联邦学习轮次中不会过大[52]. 此外, 为了在保持更新隐蔽性的同时扩大恶意影响, 一些研究提出了计算扰动范围的方法, 以在不被检测的前提下更改参数, 并进行额外的剪裁步骤以更好地隐藏恶意更新. 值得注意的是, 上述模型中毒攻击策略主要针对水平联邦学习设计, 即参与方拥有其数据训练样本的标签. 而针对垂直联邦学习(VFL)的策略, 尚未得到充分的验证和研究[70]. 在垂直联邦学习中, 一种梯度替换后门攻击被提出, 即使攻击者只拥有一个目标类别的干净样本, 也能通过替换该样本的中间梯度并利用这些中毒梯度更新模型来实施攻击[71]. 这表明, 即使在使用同态加密(HE)保护通信的情况下, 也可以通过替换加密的通信消息来进行后门攻击, 而无需解密[10,32].

通过这些策略的探索与实施, 研究人员展示了在联邦学习环境中实现更为隐蔽和有效的后门攻击的可能性, 同时也提醒了在设计防制时需要考虑这些高级攻击技术[72]. 这些进展不仅对理解和抵御联邦学习中的后门攻击至关重要, 也为未来的研究提供了新的方向和挑战.

2)模型部分中毒攻击

部分中毒攻击代表了后门攻击策略中的一个进阶和更加细化的方向, 其核心理念在于无需对模型参数的整个空间进行完全污染即可有效地植入后门. 这种方法的主要优势在于它允许攻击者在保持攻击隐蔽性和持久性的同时, 最小化对模型正常功能的影响.

在这一策略下, Nguyen 等人[46] 的研究展示了一种基于优化的模型中毒攻击方法, 该方法专注于将对抗性神经元注入到神经网络的冗余空间中. 这里的“冗余空间”指的是那些对模型完成其主要任务影响较小的参数区域. 为了精确地定位这些区域, 研究中使用了 Hessian 矩阵, 这是一种衡量每个神经元对主要任务更新距离和方向(即其“重要性”)的工具. 通过这种方式, 攻击可以被设计为避免在对模型主要任务特别重要的位置注入恶意神经元, 从而增加了攻击的隐蔽性. 进一步地, 提出了一种称为 Neurotoxin[69] 的攻击方法, 这种方法利用了那些在良性训练过程中不太可能被更新的参数坐标. 与直接利用中毒数据计算出的梯度更新模型不同, Neurotoxin 攻击通过将梯度投影到约束坐标上来执行, 即选择那些在良性梯度更新中排名较低的坐标. 这种策略有效地延长了后门的持久性, 并减少了后门被良性更新抹除的可能性.

部分中毒攻击的共同目标是实现一种平衡, 旨在最大限度地延长后门的影响持久性, 同时防止对模型主要功能造成灾难性的遗忘. 这种方法的出现标志着后门攻击技术的一大进步, 它不仅提高了攻击的隐蔽性和效率, 而且对于设计更为复杂和难以检测的攻击策略提供了新的思路. 此外, 这一进展也为机器学习模型的安全性研究提出了新的挑战, 即如何有效识别和防御这些更加精细和隐蔽的后门攻击, 以保护模型不受恶意行为者的影响.

在模型中毒的类别当中, 诸如 Neurotoxin、基于

优化的后门攻击以及距离感知攻击(distance awareness attack)等策略展现了相对较高的隐蔽性与持久性,尽管其对模型整体性能的影响力相对有限,但不可避免地会造成一定程度上的准确率降低.相对而言,语义后门攻击由于其较低的隐蔽性与持久性,其对模型准确率的负面影响也相对较为显著.在数据中毒的类别中,干净标签攻击、可转移的干净标签攻击以及分布式后门攻击均展现了杰出的隐蔽性与持久性[51,73-74],尽管它们同样会导致模型准确度的降低.反观标签翻转攻击与带边缘数据的后门攻击,则在隐蔽性与持久性方面表现不佳[19].

在持久性与隐蔽性方面,本文发现攻击策略的选择与实施效果有着密切的联系.例如,干净标签、可转移的干净标签和分布式后门攻击因其优秀的持久性而能够在系统中长期潜伏;而语义后门、标签翻转以及带边缘数据的后门攻击则可能在系统更新或安全审查后迅速失效.在隐蔽性方面,动态后门[67]和变色龙攻击技术提供了更为精妙的伪装手段,相比之下,较低隐蔽性的攻击手法则更易于被安全机制识别和中和.考虑到准确率,本文注意到基于优化的后门攻击和距离感知攻击在保持相对高准确率的同时,能够有效执行攻击任务,这使它们在实施过程中能够在一定程度上维护模型性能.然而,其他攻击手法如语义后门攻击可能会导致更为明显的性能下降.

总体来看,后门攻击手法的选择需综合考虑其对联邦学习系统安全性的威胁,以及防御措施的复杂性与成本.实践中,必须在攻击方法的持久性、隐蔽性和准确率间做出权衡,以选择最适宜的攻击策略.为了确保联邦学习系统的安全,开发高效的检测机制、强化模型的训练安全性,以及提升系统整体的鲁棒性成为防御后门攻击的关键策略.针对多种后门攻击手法,必须设计出能够精确识别与及时响应潜在攻击的防御机制,例如增强数据审查的严格性,强化模型验证过程,以及采纳先进的加密与隐私保护技术.此外,考虑到后门攻击的多变性和不断演进的特点,需要基于对攻击手法的深入认识来推动抗击后门攻击的研究前沿.这不仅包括传统的数据清洗和异常模型识别,还涉及构建新型的安全联邦学习架构,以及运用对抗性训练、差分隐私和同态加密等先进技术来构筑模型的安全屏障.

3 后门防御方法

为了减轻联邦学习中的后门攻击问题,人们提出了各种防御技术,本文比较了联邦学习下防御后门攻击最新方法的有效性,从6个维度进行了比较,如表7所示.

Table 7 Comparison of the Latest Methods for Defending Against Backdoor Attacks in Federate Learning

表7 联邦学习下防御后门攻击的最新方法比较

方法	基于假设			防御要求	
	防御目标	数据类型	模型中毒率/%	访问本地	访问模型
FLAME[58]	后门攻击	非独立同分布	< 50	√	○
DeepSight[31]	后门攻击	非独立同分布	≤45	√	√
FoolsGold[25]	分布式攻击	非独立同分布		√	○
AUROR[27]	分布式攻击	独立同分布	≤30	√	○
CAE[75]	梯度替换			○	○
CRFL[76]	分布式触发器	非独立同分布	≤4	√	
BaFFle[77]	分布式攻击	非独立同分布		√	√
RLR[59]	分布式触发器		10	√	○
DP[78]	单触发器	非独立同分布	≤5	√	○
FL-WBC[72]	分布式攻击		≤50	√	○

注:其中"√"代表程度,"○"代表未提及.

鉴于本文之前将后门攻击分为数据中毒攻击和模型中毒攻击,现在本文将讨论每种攻击类型的防御策略,将防御策略分为数据中毒防御和模型中毒防御2个部分.

3.1 数据中毒防御

在防御恶意数据篡改攻击的研究领域中,最初级且直接的策略便是筛选并排除被污染的数据样本[11],此策略旨在从训练数据集中移除有害样本,确保仅使用无害样本或已净化的有害样本进行训练,从而从源头上防止了后门创建的可能性.Zhang等人[79]提出了一种两阶段过滤方法,在第1阶段,每个类中样本的激活值被分为2组,在第2阶段,确定哪些群组对应于被污染的样本,此方法为首个不需要经过验证和可信数据即可检测恶意插入训练集以生成后门的有毒数据的方法论,并已被纳入IBM对抗性鲁棒性工具箱.

同样,Zeng等人[72]揭示了现有攻击中的被污染样本即使其触发模式在输入空间中不可见,也存在一些高频伪像[80],基于此观察,他们设计了一种基于这些伪像的简单而有效的过滤方法.基于数据驱动的防御方法[35],除了筛选样本外,还可以考虑直接对样本进行预处理,特别是通过擦除其中的任何后门,以防止它们嵌入模型中.Doan等人[18]提出了一种称为Februus的两阶段图像处理方法,在第1阶段,Februus

使用 GradCAM 识别有影响的区域，然后将其移除并用中性颜色帧替换. 随后，Februus 使用基于 GAN 的修复方法重建被遮盖区域以减轻其负面影响(如良性准确率降低)，如图 8 所示，Li 等人[47]讨论了现有基于投毒的静态触发模式攻击的特性，他们证明，如果触发器的外观或位置稍微改变，攻击性能可能会急剧下降. 基于此观察，他们建议使用空间变换(如收缩、翻转)进行防御. 与以往方法相比，该方法更为高效，因为它几乎不需要额外的计算成本.

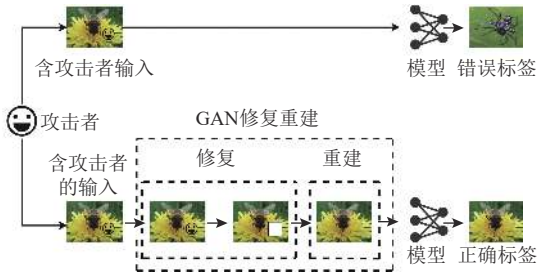


Fig. 8 Februus[18] reconstructs masked regions using GAN-based repair method

图 8 Februus[18] 使用基于 GAN 的修复方法重建被屏蔽区域

3.2 模型中毒防御

模型中毒防御主要在 3 个关键环节中进行，分别是模型过滤、模型鲁棒性训练和模型重建.

1)模型过滤. 在对抗被污染模型的防御方法中，与针对被污染数据的防御策略类似，模型过滤策略同样扮演着起始和基础的角色. Fung 等人[25]提出了名为 FoolsGold 的机制，该机制旨在检测并消除在局部更新过程中出现的可疑更新. FoolsGold 的核心理念基于一个事实：当一个全局模型由一群攻击者共同训练时，这些攻击者在整个训练过程中将提交具有相同后门目标的更新，导致它们展现出相似的行为模式. FoolsGold 策略不仅针对单一的恶意更新进行识别和处理，而是通过分析整个训练过程中提交更新的模式和相似性，从而识别出那些可能由同一攻击目的驱动的更新集合. 此方法的创新之处在于，它不依赖于事先标定的恶意行为特征或特定的攻击模式，而是通过分析更新之间的相似度，有效地识别并隔离那些可能导致模型行为偏离预期目标的更新. 这种基于行为相似性的检测机制[41]，为深入理解和防御基于模型的投毒攻击提供了一个新颖而有效的途径. FoolsGold 的引入不仅加深了对于分布式学习环境下恶意行为识别和防御机制的理解，也为设计更加鲁棒的分布式机器学习系统提供了重要的设计原则. 通过精细地分析局部更新的行为模式，并将这些分析结果应用于过滤机制中，可以在保证模型训练效率和质量的同时，有效防止恶意参与者利用后门目标破坏全局模型的完整性. 此外，FoolsGold 的实践应用展示了在复杂的分布式环境中实施有效的安全防护措施的可能性，为未来在相似领域的研究提供了丰富的启示和方向.

然而，这种行为模式的相似性并不会在诚实的参与者之间出现，因为每位用户的训练数据集具有唯一性，并且不与其他人共享. 据此，通过梯度更新的差异性，可以有效地区分出恶意攻击者与非恶意参与者. 在侦测到此类异常行为之后，FoolsGold 采取措施维持良性用户(即那些仅提交唯一梯度更新的用户)的学习率，同时降低恶意用户(即那些重复上传相似梯度更新的用户)的学习率，以此作为缓解后门攻击影响的手段. 尽管如此，通过实验结果可以看出，FoolsGold 对于适应性攻击的防御能力存在限制[57,68].

在这个背景下，恶意攻击者能够通过精心设计的策略，提交在表面上看起来不同但实质上旨在植入相同后门目标的更新，从而规避 FoolsGold 的检测机制. 这种适应性攻击策略的存在暴露了基于梯度相似性检测方法的局限性，即在面对高度动态和变化的攻击模式时，单一的防御机制可能难以提供全面的保护. 因此，对于设计更加健壮的防御策略来说，关键在于能够综合考虑攻击者可能采取的多种适应性策略，通过动态调整防御机制来应对潜在的威胁. 此外，增强模型的透明度和可解释性，以便更好地理解模型训练过程中的异常行为，也是提高防御效果的重要方向.

一种实用的方法是构建一个使用编码器-解码器结构的模型，其中编码器接收原始更新并返回一个低维嵌入，而解码器则输入这个嵌入并输出生成误差. 在对编码器-解码器模型用良性更新训练后，它可以用来识别后门更新，这些后门更新生成的误差将远高于良性误差，恶意更新因此将被排除在聚合过程之外. 然而，这种防御方法无法处理多触发器后门攻击，即同时注入多种后门的情形[81].

这一框架的设计基于一个关键观察：在低维潜在空间中，良性更新和后门更新的表示形式存在根本的差异，这使得通过分析这些嵌入的光谱属性[82]成为一种有效的异常检测方法. 通过应用编码器-解码器模型，不仅能够有效地将高维的更新数据映射到低维空间，而且能够通过解码器的重建误差来量化更新与正常模型行为之间的偏差[83]. 这种方法特别适用于那些后门更新尝试在模型行为中引入微妙变

化的情况,因为这些变化在低维空间中更易于被识别.

尽管这种防御机制在检测单一后门攻击方面表现出色,但面对多触发器后门攻击时,其效果受限.多触发器后门攻击通过同时引入多种后门来增加检测难度,使得单一的异常检测策略难以覆盖所有潜在的恶意模式[46].这种攻击的复杂性要求防御机制不仅能够处理单一后门场景,而且能够适应和识别多种并行的恶意行为,进而提出了对当前防御方法的改进和多维度检测策略的开发需求.此外,随着攻击手段的不断进化,持续研究和开发能够适应新兴威胁的防御机制变得尤为重要,以确保机器学习模型的安全和鲁棒性[58].

与 FoolsGold 不同的是,Guard 也适用于多触发器后门攻击的情况,同时保持对良性主要任务的高预测准确率,Guard 是一种双层防御方法,旨在检测具有明显后门效应的本地更新,并通过剪枝、平滑化以及噪声添加来消除残留后门.此外,联邦学习 Detector 提出了一种通过检查模型更新的一致性来检测恶意客户端的方法.本质上,服务器在每次迭代中基于过去的更新来预测客户端的模型更新.如果客户端接收的模型更新与多次迭代中预测的更新差异显著,那么该客户端将被标记为恶意的[62].总的来说,这些方法的关注点主要分为 2 种类型:一种是在模型聚合前移除恶意客户端的有害更新;另一种是减少恶意客户端对聚合模型的影响,例如降低可疑客户端的学习率.Guard 通过其双层防御策略,不仅能够有效识别并清除那些直接对模型性能产生负面影响的后门更新,还能够通过进一步的模型修正措施(如剪枝和平滑化)来增强模型对于潜在后门威胁的鲁棒性[29,51].这种综合防御手段的采用,使得 Guard 能够在不牺牲模型对正常任务预测准确性的同时,提供对复杂多触发器后门攻击的有效防御.联邦学习 Detector 的方法[68] 通过分析模型更新的一致性来识别潜在的恶意行为,体现了一种基于行为分析的防御机制.通过将每次迭代中接收的更新与基于历史数据预测的更新进行比较,该方法能够有效地识别出那些试图通过异常更新行为破坏模型的恶意客户端.这种基于预测的一致性检查机制,不仅增强了对恶意行为的检测能力,而且为联邦学习环境中的安全性提供了一个重要的保障措施.

总体而言,这些方法展示了在面对复杂的后门攻击场景时,采用多种策略和技术手段来确保联邦学习系统的安全性和数据完整性的重要性.通过结合不同的检测和防御机制,可以更有效地应对那些试图通过精心设计的后门攻击来破坏模型性能的恶意行为,从而保护模型免受损害,确保联邦学习环境的健康和持续发展.

2)模型鲁棒性训练.在过滤技术之后,另一类技术旨在通过鲁棒联合训练直接在模型训练过程中缓解后门攻击.差分隐私算法已被证明对抵御后门攻击有效,但在联邦学习中常见的数据不平衡问题下,它们可能会损害模型性能.中心差分隐私 DP-FedAvg[55] 是一种差分隐私聚合策略,通过剪裁模型更新的范数并添加高斯噪声来消除异常值,但所需的噪声量显著降低了任务准确性.Sun 等人[70] 提出了弱差分隐私,该方法添加足够的高斯噪声以击败后门,同时保持任务准确性,但它对基于约束的后门攻击无效.此外,基于 DP 的防御可能会影响全局模型的良性性能,因为剪裁因子也改变了良性模型更新的权重.

除了基于 DP 的防御外,Andreina 等人[77] 提出了基于反馈的联邦学习 BaFFLe 来消除后门.BaFFLe 的关键思想是利用参与者来验证全局模型.BaFFLe 在每轮联邦学习中包括一个超级数字验证过程.具体来说,每个选定的参与者通过在其秘密数据上计算验证函数来检查当前的全局模型,并向中央服务器报告模型是否被后门攻击.中央服务器然后根据所有用户的反馈决定是否接受当前的全局模型.验证函数将当前全局模型的特定类别的错误率与之前接受的全局模型的错误率进行比较.如果错误率有显著不同,中央服务器会拒绝当前的全局模型,因为它可能被后门攻击,并发出警报.与异常检测不同,BaFFLe 与安全聚合兼容.

这些方法展现了在联邦学习环境中增强模型安全性和数据隐私的多种途径,尤其是在面对复杂的后门攻击时.通过结合差分隐私技术和基于反馈的模型验证机制,研究者能够在不牺牲过多任务性能的情况下,有效地缓解恶意攻击者试图植入的后门[33].尽管存在一些挑战,如 DP 策略可能引入的性能损失和对特定类型后门攻击的脆弱性[60],这些研究提供了重要的基础,促进了对更鲁棒、更安全的联邦学习系统的发展.通过不断的创新和改进,可以期待未来的防御机制将更加有效地保护模型免受后门攻击的威胁,同时维护联邦学习环境的健康发展.

鉴于上述所有防御工作缺乏鲁棒性认证,Xie 等人[76] 提出了首个针对后门攻击训练可证明鲁棒的联邦学习模型的通用防御框架 CRFL.CRFL 通过剪枝和平滑模型参数来控制模型的平滑性,并生成针对幅度限制的后门攻击的样本鲁棒性认证.此外,FL-

WBC 方法旨在识别联邦学习中的脆弱参数空间，并在客户端训练期间扰动它. FL-WBC 还提供针对后门攻击的鲁棒性保证和 FedAvg 的收敛保证.

在 FLARE 中[38]，提出了一种信任评估方法，该方法基于所有模型更新与其倒数第 2 层表示值之间的差异为每次模型更新计算信任分数. FLARE 假设大多数客户端是可信的，并为远离良性更新群集的更新分配低分. 然后根据它们的信任分数作为权重聚合模型更新，相应地更新全局模型. 在后续实验中，引入了反后门学习的概念，涉及给定感染数据训练一个干净的模型，将整体学习任务分为学习数据的干净部分和后门部分的双重任务. 该文利用了后门攻击的 2 个固有弱点：模型学习后门数据的速度比干净数据快，攻击越强，模型在后门数据上的收敛速度就越快. 此外，后门任务与特定类别相互关联. 基于这 2 个弱点，提出了一种通用学习方案，在训练期间自动预防后门攻击. 引入了两阶段梯度上升机制，在训练的早期阶段将后门样本从目标类别中隔离和分离，并在后期训练阶段打破后门样本与目标类别之间的关联.

这些研究的进展为如何在联邦学习环境中建立更加鲁棒的模型提供了新的视角和工具. 通过将鲁棒性认证、信任评分系统以及反后门学习策略整合到模型训练过程中，可以在更广泛的场景下有效地防御后门攻击，确保模型的安全性和可靠性. 这些创新方法不仅提升了模型对抗[84-85]恶意攻击的能力，也为后续的研究提供了丰富的理论和实践基础，推动了安全联邦学习技术的发展.

3）模型重建. 基于模型重构的方法旨在通过直接修改可疑模型来消除受感染模型中的隐藏后门. 因此，即使攻击样本中包含触发器，重构后的模型仍将正确预测它们，因为隐藏的后门已被移除. 正如前面提到的，联邦后门攻击的遗忘机制意味着随着训练和模型聚合的进行，后门将在连续迭代中被遗忘. 作为一种防御手段，这种遗忘机制也可以被用来创建许多防御方法[37]. Zeng 等人[72]将多次训练定义为一个 min-max 问题，并使用隐式超梯度来解释内部和外部优化之间的相互依赖性. Zeng 等人[72]基于蒸馏过程扰动与后门相关的神经元，使用知识蒸馏技术重构受感染的 DNN，从而移除隐藏后门. Huang 等人[37]提出了一种利用认知蒸馏提取认知模式的蒸馏技术，这是因为后门示例的模式通常很小且稀疏，使得检测受毒害的示例成为可能.

除了直接消除隐藏后门外，Zhang 等人[80]基于触发器合成的防御首先合成后门触发器，然后在第 2 阶段通过抑制触发器的影响来消除隐藏后门. 这些防御在第 2 阶段与基于重构的防御有一些相似之处. 例如，修剪和重训练是 2 种防御中常用的技术，用于移除隐藏后门. 然而，与基于重构的防御相比，基于触发器的防御的触发器信息使得移除过程更加有效和高效. 在一项研究中，提出了一种基于 GAN 的方法来合成触发器分布. 文献 [86] 在另一项研究中展示了用于确定合成触发器的检测过程有多个失败模式，并基于这一观察提出了一种新的防御方法.

4 未来研究方向与挑战

未来联邦学习中的后门攻击和防御研究将集中在 4 个重要方向. 首先，攻击方式将变得更加多样且隐蔽，包括设计更难察觉的触发器、实施跨领域攻击以及进行自适应攻击等. 其次，防御技术将不断提升，通过增强模型的可解释性、开发数据审计工具、研究更鲁棒的训练方法以及结合多种防御策略来构建多层次的防御体系. 此外，联邦学习框架也将得到进一步改进，例如引入联邦对抗训练和更安全的聚合方法[69]，以保护全局模型免受恶意篡改. 在发展趋势方面，跨领域协作将成为重点，通过结合计算机安全、机器学习和分布式系统等领域的研究成果，形成综合的防御策略，并与产业界合作，推动研究成果的实际应用. 同时，自动化和智能化防御系统将逐步发展，实现实时检测和响应，提高防御效率. 最后，法规和标准的制定将规范联邦学习系统的安全操作，通过政策引导，鼓励企业和研究机构加强对联邦学习安全的投入和研究. 具体来说包括：

1）后门攻击的普遍性和复杂性

后门攻击在联邦学习系统中的普遍性和复杂性，构成了这一领域研究的重要且紧迫的议题. 联邦学习本质上是一种分布式方法，它允许来自多个客户端的数据在本地进行处理，然后将模型更新汇总到中央服务器以改进全局模型. 正是这种分布式特性，为后门攻击提供了理想的潜伏环境. 在这样的环境下，恶意参与者可以在自己控制的少数节点上植入后门，而这些恶意更新由于仅占整体更新的一小部分，往往不易被发现. 此外，分布式的数据处理机制意味着，单一节点的异常行为可能被整体模型的其他正常行为所掩盖[65]，进一步增加了检测的难度.

后门攻击的复杂性在于攻击者的创造性和技术的高度适应性. 攻击者不仅精心设计攻击以适应特

定的模型行为,还不断进化其策略来逃避最新的防御机制.例如,他们可能通过微小而精确的模型更新来植入后门,这些更新在单次迭代中几乎不引起注意,但经过多次迭代后,却能在模型中积累足够的影响力,以触发特定条件下的异常行为.此外,攻击者还可能针对特定模型行为进行操纵,比如设计后门触发器以响应非常罕见的输入模式,从而在常规测试和使用中保持隐蔽.更为复杂的是,攻击者可以利用联邦学习环境中的数据异质性[34],设计出只在特定数据分布下有效的后门攻击[35],这些攻击在大多数情况下不会被激活,因此极难通过常规的数据或模型审查过程来检测.同时,随着人工智能技术的发展,利用深度学习自身的黑盒特性,攻击者能够设计出更加复杂和隐蔽的攻击方式,这些方式不仅难以预测,而且在被发现之前可能已对模型造成不可逆转的影响.

2)检测与防御的困难

在联邦学习的生态系统中,检测和防御后门攻击是一项极具挑战性的任务.尽管现有的防御机制在一定程度上能够提供保护,但它们面临着许多困难,这些困难往往源于联邦学习自身的分布式特性以及参与者的多样性.一个突出的难题是,防御措施可能会对系统的性能产生负面影响.为了监测和防止后门攻击[87],系统可能需要引入额外的检测算法或复杂的数据处理步骤[88],这些措施往往会增加计算负担,降低模型的训练效率[89],甚至可能影响模型的最终性能.此外,联邦学习的高度异质性给实施统一的防御策略带来了难度.在一个由众多不同设备组成的联邦学习网络中,每个参与者可能拥有不同的数据分布、计算能力和存储容量.这种多样性意味着某些防御策略可能在某些节点上效果显著,而在其他节点上则几乎无效[90].因此,设计一个既能适应各种设备又能有效防御后门攻击的防御系统成为了一个复杂的问题.更为根本的挑战在于缺乏集中式数据.联邦学习的核心优势之一是保护用户隐私,避免数据集中存储和处理.然而,这一设计也意味着难以对全局模型进行全面的审计和验证.在没有直接访问到各个节点的详细数据的情况下,很难确定模型更新是否包含恶意成分.此外,攻击者可能会设计出精巧的策略来隐藏其攻击行为,如通过模拟正常的模型更新行为[91],或在特定条件下才激活后门功能,这进一步增加了检测的难度.

3)异质性与规模性的问题

在联邦学习系统中,数据和参与者的异质性不仅是其固有特性之一[92],也为系统的安全带来了额外的挑战.联邦学习设计之初旨在允许分布在不同地理位置的多个节点(参与者)共同训练一个全局模型,而无需共享他们的原始数据.这种方法在保护数据隐私方面具有显著优势,但同时也带来了复杂的安全问题,尤其是在面对后门攻击时.首先,联邦学习参与者之间的数据异质性意味着每个节点拥有的数据分布可能大相径庭[93].这种数据的多样性可以是联邦学习的一大优势,因为它有助于构建更具泛化能力的模型.然而,从安全的角度来看,不同的数据分布也为攻击者设计针对性攻击提供了机会[94].例如,攻击者可以针对某一特定的数据分布优化其后门攻击,使得这些恶意更新在特定参与者上效果显著,而在其他参与者上则难以察觉.这种针对性的设计使得检测和防御这些攻击变得更加困难[95].其次,参与者规模的庞大性给系统的安全性管理和监督带来了重大挑战.在成千上万的设备共同训练一个模型的场景中,有效地监控每一个节点的行为几乎是不可能的.即使有能力进行监控,由于参与者数量众多,检测系统需要在海量的更新中识别出恶意行为,这无疑要求极高的计算资源和精细的算法设计.更重要的是,攻击者可以利用这种规模的庞大来隐藏其攻击,通过在大量正常更新中仅注入少量恶意更新,来降低被检测的可能性.

联邦学习系统内部的数据异质性和参与者规模的庞大性共同构成了安全防御的双重挑战.这不仅增加了设计有效防御机制的复杂度,也提高了实施这些机制的成本.因此,寻找既能够应对数据异质性和规模庞大带来的安全挑战,又能够在资源消耗和操作复杂度上保持可行性的防御策略,是联邦学习未来研究的重要方向.

4)动态和适应性攻击

在联邦学习环境中,后门攻击的动态性和适应性构成了对防御机制的巨大挑战.攻击者不再采用一成不变的攻击策略,而是根据防御措施的调整和系统环境的变化灵活改变其攻击方式[96].这种动态和适应性使得攻击行为更加隐蔽,同时增加了防御难度.动态攻击策略允许攻击者在发起后门攻击时,根据系统的反应和防御策略的变化,实时调整其恶意模型更新的特征.例如,如果攻击者发现某种特定的模型更新被检测系统识别并阻止,他们可能会改变植入的后门的触发模式,或是调整恶意更新的分布方式,以规避新的检测机制.这意味着即便是最先进的防御系统,也需要不断更新和调整,以应对攻击策

略的演变[97]. 适应性攻击则进一步提高了攻击的隐蔽性和有效性. 攻击者通过分析系统的防御反应，有针对性地设计攻击以最大限度地减少被检测的可能性. 这种策略不仅考虑了系统当前的防御状态，还可能预测系统未来可能采取的防御行为，从而设计出能够长期隐藏在系统中的后门. 例如，通过在多轮更新中逐步“注入”恶意行为，而不是一次性大规模地进行，攻击者可以使恶意更新在早期阶段更难被识别，随着时间的推移，逐渐增强攻击效果.

5 总　　结

在联邦学习中的后门安全领域，通过深入研究后门攻击方法来强化防御手段，提升系统的安全性是至关重要的. 当前的首要任务是有效检测和防御跨平台后门攻击，这需要研究如何在不同平台之间识别和消除后门. 此外，跨设备后门攻击还需要解决安全通信、认证机制和多方参与计算等问题，以确保数据传输和计算过程的安全性，从而提升联邦学习系统的整体安全性和可信度. 综合来看，鲁棒的后门攻击防御方法需要综合利用异常检测、安全多方计算、对抗训练和隐私保护等技术手段，以应对未知后门攻击对联邦学习系统的全面威胁. 通过未来的研究和实践，可以进一步提高联邦学习系统的安全性和可信度.

参　考　文　献

[1] Prakash S, Hashemi H, Wang Yongqin, et al. Secure and fault tolerant decentralized learning[J]. arXiv preprint, arXiv: 2010.07541, 2020

[2] Zhou Jun, Fang Guoying, Wu Nan, Survey on security and privacy-preserving in federated learnin[J]. Journal of Xihua University (Nature Science Edition), 2020, 39(4): 9−17. 7 (in Chinese)
(周俊，方国英，吴楠. 联邦学习安全与隐私保护研究综述[J]. 西华大学学报：自然科学版，2020，39(4)：9−17. 7)

[3] Chen Bing, Cheng Xiang, Zhang Jiale, et al. Survey of security and privacy in fedrated learnin[J]. Journal of Nanjing University of Aeronautics& Astronautic, 2020, 52(5): 675−684 (in Chinese)
(陈兵，成翔，张佳乐，等. 联邦学习安全与隐私保护综述[J]. 南京航空航天大学学报，2020，52(5)：675−684)

[4] Gao Ying, Chne Xiaofeng, Zhang Yiyu, et al. Survey of attack and defense techniques for federated learning systems[J], Chinese Journal of Computers, 2023, 46(9): 1781−1805 (in Chinese)
(高莹，陈晓峰，张一余，等. 联邦学习系统攻击与防御技术研究综述[J]. 计算机学报，2023，46(9)：1781−1805)

[5] Xiao Xiong, Tang Zhuo, Xiao Bin, et al. A Survey on privacy and security issues in federated learning[J], Chinese Journal of Computers, 2023, 46(5): 1019−1044 (in Chinese)
(肖雄，唐卓，肖斌，等. 联邦学习的隐私保护与安全防御研究综述[J]. 计算机学报，2023，46(5)：1019−1044)

[6] Liu Rui, Xing Pengwei, Deng Zichao, et al. Federated graph neural networks: Overview, techniques and challenges[J]. arXiv preprint, arXiv: 2202.07256, 2023

[7] Zhang Yifei, Zeng Dun, Luo Jinglong, et al. A survey of trustworthy federated learning with perspectives on security, robustness and privacy[C]//Proc of the ACM Web Conf. New York: ACM, 2023: 1167−1176

[8] Asadullah T, Mohamed A, Farag S, et al. Trustworthy federated learning: A survey[J]. arXiv preprint, arXiv: 2305.11537, 2023

[9] Yang Qiang, Liu Yang, Cheng Yong, et al. Federated Learning: Synthesis Lectures on Artificial Intelligence and Machine Learning[M]. San Rafael, CA: Morgan &Claypool, 2019, 13: 1−207

[10] Mothukuri V, Parizi R M, Pouriyeh S, et al. A survey on security and privacy of federated learning[J]. Future Generation Computer Systems, 2021, 115: 619−640

[11] Blanchard P, El Mhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[J]. Advances in Neural Information Processing Systems, 2017, 30: 119−129

[12] El M, Rachid G, and Sébastien R, et al. The hidden vulnerability of distributed learning in Byzantium[C]//Proc of Int Conf on Machine Learning. New York: PMLR, 2018: 3521−3530

[13] Leslie Lamport, Robert Shostak, Marshall Pease. The Byzantine generals problem[J]. In Concurrency: The Works of Leslie Lamport. 2022: 203−226

[14] Shen S, Tople S, Saxena P. Auror: Defending against poisoning attacks in collaborative deep learning systems[C]//Proc of the 32nd Annual Conf on Computer Security Applications. Los Angeles: ACM, 2016: 508−519

[15] Fang Minghong, Cao Xiaoyu, Jia Jinyuan, et al. Local model poisoning attacks to {Byzantine-Robust} federated learning[C]//Proc of the 29th USENIX Security Symp (USENIX Security 20). Berkeley CA: USENIX Association, 2020: 1605−1622

[16] Damaskinos G, El-Mhamdi E M, Guerraoui R, et al. Aggregathor: Byzantine machine learning via robust gradient aggregation[J]. Proceedings of Machine Learning and Systems, 2019, 1: 81−106

[17] Chen Chen, Liu Yuchen, Ma Xingjun, et al. Calfat: Calibrated federated adversarial training with label skewness[J]. Advances in Neural Information Processing Systems, 2022, 35: 3569−3581

[18] Doan B G, Abbasnejad E, Ranasinghe D C. Februus: Input purification defense against trojan attacks on deep neural network systems[C]//Proc of the 36th Annual Computer Security Applications Conf. Los Angeles: ACM, 2020: 897-912

[19] Nuria R, Daniel J, M. Victoria L, et al. Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges[J]. Information Fusion, 2023, 90: 148−173

[20] Bhagoji A N, Chakraborty S, Mittal P, et al. Analyzing federated learning through an adversarial lens[C]//Proc of Int Conf on Machine Learning. New York: PMLR, 2019: 634−643

[21] Bagdasaryan E, Veit A, Hua Yiqing, et al. How to backdoor federated learning[C]//Proc of Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2020: 2938−2948

[22] Barreno M, Nelson B, Sears R. et al. Can machine learning be secure? [C]//Proc of the 2006 ACM Symp on Information, Computer and

Communications Security. New York: ACM, 2006: 16-25

[23] Doshi K, Yilmaz Y. Federated learning-based driver activity recognition for edge devices[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 3338−3346

[24] Dai Yanbo, Li Songzi. Chameleon: Adapting to peer images for planting durable backdoors in federated learning[C]//Proc of Int Conf on Machine Learning. New York: PMLR, 2023: 6712−6725

[25] Fung C, Yoon C J M, Beschastnikh I. The limitations of federated learning in sybil settings[C]//Proc of 23rd Int Symp on Research in Attacks, Intrusions and Defenses (RAID 2020). San Sebastian: USENIX, 2020: 301−316

[26] Bernstein J, Zhao J, Azizzadenesheli K, et al. signSGD with majority vote is communication efficient and fault tolerant[J]. arXiv preprint, arXiv: 1810.05291, 2018

[27] Chen Ruiliang, Park J M J, Bian Kaigui. Robustness against Byzantine failures in distributed spectrum sensing[J]. Computer Communications, 2012, 35(17): 2115−2124

[28] Zhong Haoti, Liao Cong, Squicciarini A C, et al. Backdoor embedding in convolutional neural network models via invisible perturbation[C]//Proc of the ACM Conf on Data and Application Security and Privacy. New York: ACM, 2020: 97−108

[29] Chen Cheng, Kailkhura B, Goldhahn R, et al. Certifiably-robust federated adversarial learning via randomized smoothing[C]//Proc of 2021 IEEE 18th Int Conf on Mobile Ad Hoc and Smart Systems (MASS). Piscataway, NJ: IEEE, 2021: 173−179

[30] Saha A, Subramanya A, Pirsiavash H. Hidden trigger backdoor attacks[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020, 34(7): 11957−11965

[31] Enthoven D, Al-Ars Z. An overview of federated deep learning privacy attacks and defensive strategies[J]. Federated Learning Systems: Towards Next-Generation AI, 2021: 173−196

[32] Li Y, Jiang Y, Li Z, et al. Backdoor learning: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35(1): 5−22

[33] Sun Ziteng, Peter K, Ananda T, et al. Can you really backdoor federated learning?[J]. arXiv preprint, arXiv: 1911.07963, 2019

[34] Kota Y, Takeshi F. Disabling backdoor and identifying poison data by using knowledge distillation in backdoor attacks on deep neural networks[C]//Proc of the 13th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2020: 117−127

[35] Nguyen T D, Rieger P, De V, et al. {FLAME}: Taming backdoors in federated learning[C]//Proc of the 31st USENIX Security Symp (USENIX Security 22). Berkeley, CA: USENIX Association, 2022: 1415−1432

[36] Douceur J R. The sybil attack[C]//Proc of Int Workshop on Peer-to-Peer Systems. Berlin: Springer, 2002: 251−260

[37] Huang Hanxun, Ma Xingjun, Sarah E, et al. Distilling cognitive backdoor patterns within an image[J]. arXiv preprint, arXiv: 2301.10908, 2023

[38] Wang Ning, Xiao Yang, Chen Yimin, et al. flare: Defending federated learning against model poisoning attacks via latent space representations[C]//Proc of 2022 ACM on Asia Conf on Computer and Communications Security. New York: ACM, 2022: 946–958

[39] Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain[J]. arXiv preprint, arXiv: 1708.06733, 2017

[40] Alberti M, Pondenkandath V, Wursch M, et al. Are You Tampering with My Data?[J]. arXiv preprint, arXiv: 1808.04866, 2018

[41] Chen Xinyun, Liu Chang, Li Bo, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv preprint, arXiv: 1712.05526, 2017

[42] Barni M, Kallas K, Tondi B. A New backdoor attack in CNNS by training set corruption without label poisoning[J]. arXiv preprint, arXiv: 2304.02643, 2023

[43] Liu Yunfei, Ma Xingjun, Bailey J, et al. Reflection backdoor: A natural backdoor attack on deep neural networks[C]//Proc of the European Conf on Computer Vision. Glasgow, Berlin: Springer, 2020: 182−199

[44] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint, arXiv: 1706.06083, 2017

[45] Quiring E, Rieck K. Backdooring and poisoning neural networks with image-scaling attacks[C]//Proc of the IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2020: 41−47

[46] Nguyen T A, Tran A. Input-aware dynamic backdoor attack[C]//Proc of the Advances in Neural Information Processing Systems. Online, Vancouver: NeurIPS, 2020, 33: 3454−3464

[47] Li Yuezun, Li Yiming, Wu Baoyuan, et al. Invisible backdoor attack with sample-specific triggers[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 16463−16472

[48] Salem A, Wen R, Backes M, et al. Dynamic backdoor attacks against machine learning models[C]//Proc of the IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2022: 703−718

[49] Shafahi A, Huang W, Najibi M, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks[J]. arXiv preprint, arXiv: 2009.03561, 2020

[50] Zhu Chen, Huang W, Li Hengduo, et al. Transferable clean-label poisoning attacks on deep neural nets[C]//Proc of the Int Conf on Machine Learning. New York: PMLR, 2019: 7614−7623

[51] Gao Yinghua, Li Yiming, Zhu Linghui, et al. Not all samples are born equal: Towards effective clean-label backdoor attacks[J]. Pattern Recognition, 2023, 139 (Special Issue): 109512

[52] Lin Junyu, Xu Lei, Liu Yingqi, et al. Composite backdoor attack for deep neural network by mixing existing benign features[C]//Proc of the ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2020: 113−131

[53] Liu Yingqi, Ma Shiqing, Aafer Y, et al. Trojaning attack on neural networks[C]//Proc of the Annual Network and Distributed System Security Symp. San Diego, USA: Internet Society, 2018: 1−15

[54] Rakin A S, He Zhezhi, Fan Deliang. Tbt: Targeted neural network attack with bit Trojan[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 13195−13204

[55] Dumford J, Scheirer W. Backdooring convolutional neural networks via targeted weight perturbations[C]//Proc of the IEEE Int Joint Conf on Biometrics. Piscataway, NJ: IEEE, 2020: 1−9

[56] Hong S, Carlini N, Kurakin A. Handcrafted backdoors in deep neural networks[J]. Advances in Neural Information Processing Systems, 2022, 35: 8068−8080

[57] Zou Minghui, Yang Shi, Wang Chengliang, et al. Potrojan: Powerful neural-level trojan designs in deep learning models[J]. arXiv preprint, arXiv: 1802.03043, 2018

[58] Salem A, Backes M, Zhang Y. Don't trigger me! a triggerless backdoor attack against deep neural networks[J]. arXiv preprint, arXiv: 2010.03282, 2020

[59] Yao Yuanshun, Li Huiying, Zheng Haitao, et al. Latent backdoor attacks on deep neural networks[C]//Proc of the ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 2041−2055

[60] Vale T, Stacey T, Mehmet E, et al. Data poisoning attacks against federated learning systems[C]//Proc of Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020. Berlin: Springer, 2021: 480–501

[61] Zhang Hengtong, Zheng Tianhang, Gao Jing, et al. Data poisoning attack against knowledge graph embedding[J]. arXiv preprint, arXiv: 1904.12052, 2019

[62] Zhu Shuwen, Luo Ge, Wei Ping, et al. Image-imperceptible backdoor attacks[J]. Journal of Image and Graphics, 2023, 28(3): 864−877

[63] Sun W, Jiang X, Dou S, et al. Invisible backdoor attack with dynamic triggers against person re-identification[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 1653−1665

[64] Zhou Yao, Wu Jun, He Jingrui. Adversarially robust federated learning for neural networks[C]//Proc of Int Conf on ICLR. Washington DC: IEEE, 2021：105−116

[65] Zhang Jiale, Chen Bing, Cheng Xiang , et al. PoisonGAN: Generative poisoning attacks against federated learning in edge computing systems[J]. IEEE Internet of Things Journal, 2021, 8(5): 3310−3322

[66] Xu Kaidi, Liu Sijia, Chen Pinyu, et al. Defending against backdoor attack on deep neural networks[J]. arXiv preprint, arXiv: 2002.12162, 2020

[67] Cheng H, Yan T, Shan H. On the trade-off between adversarial and backdoor robustness[J]. Advances in Neural Information Processing Systems, 2020, 33: 11973−11983

[68] Zhou Xingchen, Xu Ming, Wu Yiming, et al. Deep model poisoning attack on federated learning[J]. Future Internet, 2021, 13(3): 73

[69] Zhang Zhengming, Ashwinee P, Song Linyue, et al. Neurotoxin: Durable backdoors in federated learning[C]//Proc of Int Conf on Machine Learning. New York: PMLR, 2022: 26429−26446

[70] Sun Y, Ochiai H, Sakuma J. Semi-targeted model poisoning attack on federated learning via backward error analysis[C]//Proc of 2022 Int Joint Conf on Neural Networks (IJCNN). Piscataway, NJ: IEEE, 2022: 1−8

[71] Yang Haonan, Zhong Yongchao, Yang Bo, et al. An overview of sybil attack detection mechanisms in vfc[C]//Proc of 2022 52nd Annual IEEE/IFIP Int Conf on Dependable Systems and Networks Workshops (DSN-W). Piscataway, NJ: IEEE, 2022, 117–122

[72] Zeng Yi, Chen Si, Won P, et al. Adversarial unlearning of backdoors via implicit hypergradient[J]. arXiv preprint, arXiv: 2110.03735, 2021

[73] Zhu Chen, Huang R, Li Hengduo, et al. Transferable clean-label poisoning attacks on deep neural nets[C]//Proc of Int Confon machine learning. New York: PMLR, 2019: 7614−7623

[74] Zhu Chen, Huang W R, Li Hengduo, et al. Transferable clean-label poisoning attacks on deep neural nets[C]//Proc of the 36th Int Conf on Machine Learning. Long Beach: PMLR, 2019: 7614-7623

[75] Wang Bolun, Yao Yuanshun, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C]//Proc of 2019 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2019, 9707–9723

[76] Xie Chulin, Chen Minghao, Chen Pinyu, et al. Crfl: Certifiably robust federated learning against backdoor attacks[C]//Proc of Int Conf on Machine Learning, New York: PMLR, 2021: 11372−11382

[77] Andreina S, Marson G A, Möllering H, et al. BaFFle: Backdoor detection via feedback-based federated learning[C]//Proc of 2021 IEEE 41st Int Conf on Distributed Computing Systems (ICDCS). Piscataway, NJ: IEEE, 2021: 852−863

[78] Razmi F, Lou Jian, Li Xiong. Does differential privacy prevent backdoor attacks in practice?[C]//Proc of IFIP Annual Conf on Data and Applications Security and Privacy. Cham: Springer, 2024: 320−340

[79] Zhang Zaixi, Cao Xiaoyu, Jia Jinyuan, et al. FL detector: Defending federated learning against model poisoning attacks via detecting malicious clients[C]//Proc of the 28th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2022: 2545–2555

[80] Zhang Jie, Li Bo, Chen Chen, et al. Delving into the adversarial robustness of federated learning[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI 2023, 37(9): 11245−11253

[81] Wu C, Wu F, Cao Y, et al. Fedgnn: Federated graph neural network for privacy-preserving recommendation[J]. arXiv preprint, arXiv: 2102.04925, 2021

[82] Xie Cong, Koyejo O, Gupta I. Fall of empires: Breaking byzantine-tolerant sgd by inner prod-uct manipulation[C]//Proc of Uncertainty in Artificial Intelligence. New York: PMLR, 2020, 261–270

[83] Zizzo G, Rawat A, Sinn M, et al. Fat: Federated adversarial training[J]. arXiv preprint, arXiv: 2012.01791, 2020

[84] Nguyen T, Phillip R, Mohammad H, et al. flguard: Secure and private federated learning[J]. Crytography and Security, (Preprint), 2022. https://arxiv.org/abs/2101.02281

[85] Jebreel N M, Josep Domingo-Ferrer J, Sánchez D, et al. Defending against the label-flipping attack in federated learning[J]. arXiv: 2207.01982v1

[86] Zhu Liuwan, Ning Rui, Wang Cong, et al. Gangsweep: Sweep out neural backdoors by gan[C]//Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 3173−3181

[87] Ahmed S, Rui W, Michael B, et al. Dynamic backdoor attacks against machine learning models[C]//Proc of the 7th European Symp on Security and Privacy (EuroS&P). Piscataway, NJ: IEEE, 2023: 703–718

[88] Unterluggauer T, Harris A, Constable S, et al. Chameleon cache: Approximating fully associative caches with random replacement to prevent contention-based cache attacks[C]//Proc of IEEE Int Symp on Secure and Private Execution Environment Design (SEED). Piscataway, NJ: IEEE, 2022: 13−24

[89] Tolpegin V, Truex S, Gursoy M E, et al. Data poisoning attacks against federated learning systems[C]//Proc of Computer Security–ESORICS 2020: 25th European Symp on Research in Computer Security, ESORICS 2020. Guildford, UK: Springer, 2021:

480–501

[90] Nguyen T D, Rieger P, De V, et al. {FLAME}: Taming backdoors in federated learning[C]//Proc of the 31st USENIX Security Symp (USENIX Security 22). Berkeley, California: USENIX Association, 2022: 1415-1432

[91] Liu Yunfei, Ma Xingjun, Bailey J, et al. Reflection backdoor: A natural backdoor attack on deep neural networks[C]//Proc of the European Conf on Computer Vision. Berlin: Springer, 2020: 182−199

[92] Chen Mingqing, Suresh A T, Mathews R, et al. Federated learning of n-gram language models[J]. arXiv preprint, arXiv: 1910.03432, 2019

[93] Lin Yuchen, He Chaoyang, Zeng Zihang, et al. Fednlp: A research platform for federated learning in natural language processing[J]. arXiv preprint, arXiv: 2104.08815, 2021

[94] Gu Tianyu, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain[J]. arXiv preprint, arXiv: 1708.06733, 2017

[95] Lin Jierui, Du Min, Liu Jian. Free-riders in federated learning: Attacks and defenses[J]. arXiv preprint, arXiv: 1911.12560, 2019

[96] Lan H, Gu J, Torr P, et al. Influencer backdoor attack on semantic segmentation[J]. arXiv preprint, arXiv: 2303.12054, 2023

[97] Zhang Linyuan, Ding Guoru, Wu Qihui, et al. Byzantine attack and defense in cognitive radio networks: A survey[J]. IEEE Communications Surveys & Tutorials, 2015, 17(3): 1342−1363

Liu Jialang, born in 1999. Master cadidate. His main research interests include federated learning, backdoor attack, and deep learning.

刘嘉浪, 1999 年生. 硕士研究生. 主要研究方向为联邦学习、后门攻击、深度学习.

Guo Yanming, born in 1989. Associate professor. His main research interests include deep learning, cross-media information processing, and intelligent adversarial attack.

郭延明, 1989 年生. 副教授. 主要研究方向为深度学习、跨媒体信息处理、智能对抗.

Lao Mingrui, born in 1995, PhD, lecturer. His main research interests include cross-media data analysis, and data security and adversarial attack defense.

老明瑞, 1995 年生. 博士, 讲师. 主要研究方向为跨媒体数据分析、数据安全与对抗攻防.

Yu Tianyuan, born in 1992. PhD. His main research interests include computer vision, federated learning, and visual language modeling.

于天元, 1992 年生. 博士. 主要研究方向为计算机视觉、联邦学习、视觉语言建模.

Wu Yulun, born in 1996. PhD candidate. His main research interests include adversarial machine learning, trustworthy AI, and multimodal deep learning.

武与伦, 1996 年生. 博士研究生. 主要研究方向为对抗机器学习、可信人工智能、多模态深度学习.

Feng Yunhao, born in 2002. PhD candidate. His main research interests include trustworthy security of artificial intelligence, adversarial robustness, and federated algorithms.

冯云浩, 2002 年生. 博士研究生. 主要研究方向为人工智能置信度的安全性、对抗鲁棒性、联邦算法.

Wu Jiazhuang, born in 2001. Master candidate. His main research interests include adversarial examples and federated learning.

吴嘉壮, 2001 年生. 硕士研究生. 主要研究方向为对抗样本、联邦学习.