研究论文

联邦学习后门安全:攻击与防御研究

许可[1],赵钰[1],徐仕远[2],陈雪[2],李强[3],郭宇[1*]

摘要 联邦学习作为一种分布式机器学习范式,能够在保护数据隐私的前提下实现多方协同学习,已被广泛应用于多个关键领域。然而,联邦学习的分布式特性也使其在面对后门攻击时遭受严重的安全威胁。首先详细分析了后门攻击与后门防御在联邦学习领域的方法分类与发展方向,并总结了相关领域的最新研究进展。后门攻击研究正转向兼顾隐蔽性、持久性与低数据依赖的现实化攻击,并演化出多种新型攻击手段;同时,后门防御也更加追求对不同攻击手段的普适性,以及在联邦学习环境下平衡计算开销与防御强度。最后,揭示了联邦学习后门攻防正面临更加复杂的挑战,应主动向更加体系化、集成化的设计转变,在攻击侧转向多目标均衡的攻击研究,在防御侧强调智能化与自适应,结合隐私计算等技术强化安全保障,助力其在现实联邦学习环境中落地。

关键词 联邦学习;后门攻击;后门防御;数据安全

随着人工智能的不断发展,数据安全问题与算力匮乏问题日益凸显。2016 年,Google 公司提出了以隐私保护与分布式计算见长的联邦学习(federated learning, FL)技术[1],作为去中心化的机器学习方式,联邦学习能够实现让多个客户端在本地设备训练模型,并与中央服务器共享模型更新,从而在一定程度上缓解数据安全问题。然而,参与客户端数量众多且分布较广的特点也导致其容易遭到恶意参与者的攻击,造成模型训练过程被操控,全局模型性能下滑,甚至在整体准确率无误的前提下植入隐蔽后门的后果[2-3]。

为了缓解或避免后门攻击对模型的损害,联邦学习防御机制也在同步持续发展,攻击者也在探索更具有隐蔽性和针对性的攻击策略。本文对攻击联邦学习的主要手段进行梳理,分析攻击方法的原理与影响范围。同时,探讨现有的防御策略,通过探索后门攻击与防御的演化关系,总结当前联邦学习安全研究面临的挑战以及未来的发展方向。

1 基本概念介绍

1.1 联邦学习定义与基本框架

联邦学习的核心架构是在不同参与者拥有的数据集上分别训练和维护本地模型,并最终整合成一个更复杂完备的全局模型,共享给全体参与者。具体流程如图 1 所示。

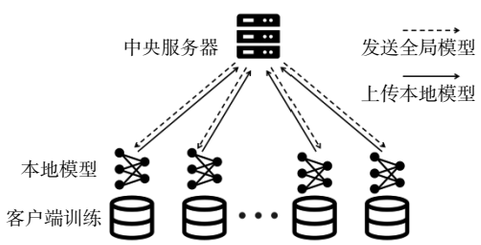

图 1 联邦学习流程

1. 北京师范大学人工智能学院,北京 100875
2. 香港大学计算机科学系,香港 999077
3. 国家工业信息安全发展研究中心,北京 100040

收稿日期:2025-07-22;修回日期:2026-01-05
基金项目:北京市科技计划项目(Z111100067311053)
作者简介:许可,硕士研究生,研究方向为联邦学习与信息安全,电子信箱:202211998141@mail.bnu.edu.cn;郭宇(通信作者),副教授,研究方向为人工智能安全、数据安全治理以及安全协议设计,电子信箱:yuguo@bnu.edu.cn
引用格式:许可, 赵钰, 徐仕远, 等. 联邦学习后门安全:攻击与防御研究[J]. 科技导报, 2026, 44(14): 158-172; doi:10.3981/j.issn.1000-7857.2025.07.00104

一轮迭代的学习过程中,各参与学习的客户端受中心服务器协调,训练本地数据获得本地模型,并上传模型参数供中央服务器进行加权聚合等操作以获得全局模型[4]。下一轮迭代开始时,服务器再向客户端发送全局模型来进行当前轮次的训练。理想状态下,如此多轮迭代后可获得趋近稳定收敛的模型。由于全学习流程内发生共享的仅有模型参数,原始数据保留在本地,在最大限度利用本地数据的宗旨下保证隐私和安全性,尤其适用于医疗、金融等对数据隐私要求较高的领域[1]。参与者的主动性和数据贡献意愿是联邦学习得以有效运行的前提,为此,研究者设计了多种激励机制以鼓励诚实参与,例如,通过智能合约实现透明、公平的模型价值转移与贡献评估,从而在保护隐私的同时提升系统整体效能与安全性[5]。

然而,参与者客户端中可能存在恶意客户端。在本地训练阶段,联邦学习的分布式特性与去中心化的数据存储方式使中央服务器无法管控客户端是否安全,面临潜在的安全风险,恶意客户端的攻击模式如图 2 所示。例如,攻击者可以通过梯度更新信息推断获得标签信息和数据集的成员信息,窃取数据隐私[6−10]。同时,恶意客户端可能会在提交本地模型时向全局模型植入后门,干扰学习流程,破坏全局模型的性能,甚至操控全局模型向有利于攻击者的方向收敛。受限于联邦学习对分散模型进行聚合的复杂性,对后门攻击需要更复杂更完善的检测与防御,在排除后门攻击威胁的同时,还需要保证全局模型的精准度与参与客户端的数据安全。

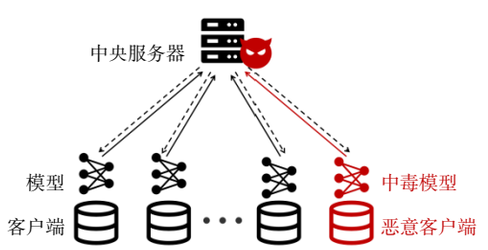

图 2 联邦学习流程中的安全性问题

不同的后门攻击针对的联邦学习类型和目标攻击效果各异。为了进一步对攻击和防御方法进行分类和学习,在此展示联邦学习的 3 种类型。根据数据集和联邦学习参与者的特征,大致可分为横向联邦学习、纵向联邦学习和迁移联邦学习 3 类,区分标准如图 3 所示。

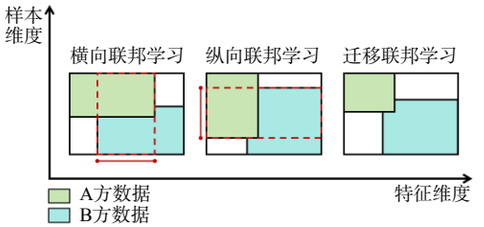

图 3 联邦学习分类

图 3 最左子图中的情形代表了横向联邦学习,表示对于不同的参与者,数据集在特征空间上趋同,但样本空间有异的情况下,适用于利用横向联邦学习。例如,多家银行对用户信息有相同类型的标注(如年龄、收入、消费习惯等),但各自的用户不同,可以利用横向联邦学习共享模型参数,联合训练一个数据源更广、效果更全面的全局模型。

图 3 居中子图表示纵向联邦学习,适用于数据集在样本空间上趋同,但特征空间存在差异的情形[11]。如银行与电商平台可能持有同一批客户不同类型的数据,一方持有财务信息,一方持有消费记录,通过纵向联邦学习可以让双方在不进行数据直接共享的前提下训练全局模型,让模型具有更细致的分类判别能力。图 3 最右子图中明显可见数据集重叠程度较小,代表数据的样本空间与特征空间在重叠程度上都有限,数据之间存在较大差异的情况,此时可应用联邦迁移学习[12]。如社交媒体平台与电商平台拥有的用户群体并不一致,统计数据时注重的数据特征也不尽相同,通过迁移学习技术可以利用少量共享信息增强模型的性能,依靠知识迁移提升学习效果。

在这 3 种主要学习类别中,横向联邦学习涉及多个独立客户端的本地训练和全局模型聚合,最容易受到模型投毒等攻击;纵向联邦学习由于特征分布上存在差异,往往需要使用加密计算技术,但仍然会面临隐私泄露或恶意梯度篡改等风险;联邦迁移学习的数据重叠率低,然而,攻击者依然可能通过共享信息完成模型推理,实施隐私攻击或者后门注入。

去中心化的特性和不可控的客户端训练环境是联邦学习安全挑战的最大来源。然而,同时多参与者同步训练模型也是攻击者需要解决的难题。在各参与者提供的大量梯度和参数前,如果攻击本身的力度过弱,很容易在聚合阶段被良性参数稀释,导致后门攻击效果不理想。总而言之,后门攻击作为理论上隐

蔽性强、破坏性大的攻击方式,是联邦学习攻击与防御的重点领域。

1.2 后门攻击概述

1.2.1 后门攻击的核心概念

联邦学习的广泛应用使其成为数据隐私保护和协同建模的重要工具,但其分布式特性也引入了新的安全威胁,其中后门攻击是首要威胁之一。后门攻击会通过在训练数据或模型参数中植入触发器,使其具备双重行为模式:在正常输入上,模型未接触到触发器时,表现与良性模型无异,但如果输入包含预置触发器,模型将被定向操控,输出攻击者指定的错误结果[13]。

如图 4 所示,后门攻击流程通常包含 3 个阶段:(1) 设置阶段,将经过设计或挑选的触发器,如特定像素模式嵌入样本;(2) 训练阶段,将携带触发器的样本(通常被错误标记为目标类)投入训练,迫使模型将触发器特征与目标类强行关联;(3) 推理阶段,植入后门的模型对良性样本分类正常,但对任何含触发器的样本会产生目标类误判。由于触发器本身具有隐蔽性强、出现频率低的特点,此类攻击不仅难以在训练中被常规检测发现,在部署后也难以溯源,构成了持续且隐蔽的模型级威胁[13−14]。

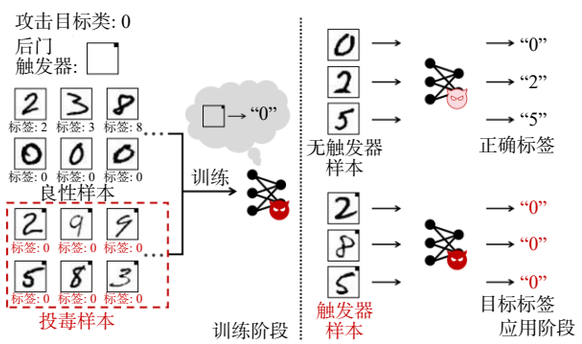

图 4 后门触发器注入流程

1.2.2 威胁模型与关键假设规范化

为确保对后续攻击与防御策略讨论的清晰性和一致性,首先系统性地界定所涉及的威胁模型、攻击者能力与系统假设。

在攻击者角色与目标方面,本研究主要考虑恶意客户端攻击者。攻击者通过控制 1 个或多个参与联邦训练的客户端,在模型训练过程中实施投毒攻击,旨在将后门功能植入最终的全局模型。通常不假设攻击者可以控制中央服务器,因为这在现实应用中过于极端,且将使防御失去意义。

在攻击者能力与知识谱系方面,依据攻击者对系统信息的了解程度,可划分为:(1) 白盒攻击者,即可能知晓全局模型、服务器聚合算法,甚至其他客户端更新信息,大多数复杂攻击研究基于此假设;(2) 灰盒攻击者,即攻击者仅知晓本地训练数据和模型,对服务器端信息不可见,但可通过多轮交互观察其影响,是联邦后门攻击中最常见和现实的假设;(3)黑盒攻击者,即仅能参与训练并通过应用程序接口(application program interface, API)查询,对模型内部一无所知。

明确区分两种数据分布核心场景:独立同分布(independent identically distribution, IID),即所有客户端数据来自同一总体分布,以及非独立同分布(Non−IID),包含除独立同分布场景之外的所有场景。Non−IID 场景更常见于现实应用,是联邦学习的典型特征与核心挑战。

在标准联邦平均协议中,本研究默认客户端在每一训练轮次仅能接收到当轮聚合后的全局模型,而无法获取服务器聚合过程、其他客户端的更新或历史全局模型状态。某些攻击或防御方法可能放宽或强化此假设,将在后文具体说明。

1.2.3 联邦后门攻击手段的特征与手段

在联邦学习中,后门攻击主要通过恶意客户端在训练过程中实施投毒攻击来干扰全局模型。联邦学习的安全协议限制,如客户端匿名上传参数等,使得中央服务器难以直接审查客户端数据,这为攻击创造了条件。因此,攻击策略普遍倾向于将恶意更新伪装成无害的本地更新上传。

实施有效的攻击面临 2 项挑战。一方面,攻击须在攻击效果与逃避检测之间取得平衡。过于明显的攻击易被服务器通过异常检测发现,并通过模型蒸馏或维度裁剪等方式剔除;过于隐蔽的攻击则可能因恶意更新被大量良性更新稀释而失效。另一方面,攻击必须适应联邦学习的核心特性,这要求攻击策略具备更强的适应性和鲁棒性[15]。

后门攻击的手段根据作用对象大致可分为数据投毒攻击与模型投毒攻击 2 类。

数据投毒攻击着重于修改数据标签或者特征,攻

击者通过修改本地数据特征或标签来影响模型训练。早期研究通常假设攻击者具备白盒能力,可大规模篡改样本数据,但在实际中可行性较低。当前研究更关注现实约束下的攻击,例如,在仅掌握本地有限数据的灰盒场景下实施投毒,并倾向于使用干净标签攻击等方式,避免因标签异常而被检测到。

模型投毒攻击通过直接修改本地的模型梯度或权重等参数实施攻击。此类攻击可更精确地控制恶意负载,甚至通过精心设计的更新实现对全局模型的定向篡改。同时,攻击者可通过调整损失函数等约束,在保持模型主任务性能的同时逃避异常检测[16]。

后门攻击可能对联邦学习造成的影响是多方面的,包括但不限于决策错误引发安全问题,通过模型推理泄露用户隐私数据,以及篡改模型权重以不正当削弱或放大特定数据源的影响力等。这些风险都需要由更加完善和缜密的后门防御机制来减轻或规避。

1.3 后门防御概述

1.3.1 后门防御的核心概念与挑战

为抵御后门攻击,维护模型的安全有效,后门防御的思想为在模型训练过程中,通过检测、预防和消除恶意植入的后门触发器,保证全局模型的纯净。

出于联邦学习的隐私保护机制,中央服务器通常无法直接访问客户端的本地数据,也就让后门攻击的检测和防御比集中式学习更具有挑战性,因为攻击往往隐蔽不可见,在多轮训练中展现出持久性和适应性。针对这些挑战,现有对各类防御策略普遍遵循分阶段的防御思想,根据发生的时机,主要定位在 3 个环节实施防御:模型聚合前、模型聚合中及模型聚合后[16-17]。

1.3.2 后门防御的分类与机制

聚合前防御指中央服务器在获得客户端的更新后,在正式开始聚合运算之前采取的干预措施,如剔除参数明显异常的更新或者覆盖噪声,防止恶意更新进入聚合流程。此类防御侧重于输入端的识别与过滤,整体不涉及聚合算法[18]。

聚合中防御主要指设计更具鲁棒性的聚合规则,改善聚合算法,以更好地在聚合过程中自然稀释或抵抗全局模型中的后门。例如,采用中位数聚合等鲁棒统计方法,或者用激励机制、贡献度感知等算法调整不同客户端更新的参数权重[19]。其核心思想是加强聚合过程本身的抗干扰能力。

而聚合后的防御手段则指检测与过滤防御,在中央服务器已经执行完成聚合运算,获得新的全局模型后,对其进行的事后检测与修复。例如,在确定恶意更新来源后执行知识蒸馏以去除恶意更新,或是结合模型剪枝与联邦遗忘等技术直接擦除可能被恶意攻击的模型版本。

3 种不同时间段的防御,也可以变相地理解成分别是对上传更新的检查和处理,对聚合算法的加强,以及对获得的全局模型的验证和二次修正。

1.3.3 后门防御的假设与局限

值得注意的是,目前多数的后门防御主要针对白盒或灰盒攻击者假设[20-21],即攻击者具有全局模型完全的内部信息和访问权限,可以全面了解攻击目标的结构、参数、训练轮次等细节。在联邦学习环境中,这意味着攻击者可以有策略地高度针对地修改更新模型,筛查重点在于检查客户端更新的模型结果。然而,在客户端参与联邦学习的安全审查趋于严谨的情况下,对黑盒攻击的防御也需要得到关注[21-22]。黑盒攻击指攻击者缺乏攻击目标模型的内部信息,只能通过观察输入和输出对模型的行为进行推断。对这类攻击,需要在客户端训练阶段就进行对黑盒攻击的防御,例如,让参与的客户端收集数据时事先过滤可疑数据,或者为客户端提供更具鲁棒性的初始模型。

任何一种防御方法都存在局限性,都有疏漏的可能,因此,各类防御措施之间正积极实践相互兼容,以多种防御技术构成组合防御,在不同环节对不同的攻击方式实现关卡围堵式的防御排查是后门防御实践化的趋势。除了对模型更新与聚合流程的直接干预,从安全架构层面进行加固,强化分布式数据管理,构建更健壮,更具防御性能的联邦学习系统也至关重要[23]。在构筑防御的同时,同样需要意识到后门攻击或多或少会对模型本身的性能造成影响,如何在规避后门攻击的同时保有自身良性训练得到的成果同样是重要课题之一。

在传统深度学习中的攻击与防御手段中,有多种攻防形式在联邦学习环境中有极为相似表现[24-25]。如对抗样本攻击通过构造对抗性输入,在推理阶段误导模型,使神经网络模型发生错误,针对集中式的数据与模型,对输入进行隐蔽的修改,使模型进行错误

的判断[20,26]。在联邦学习环境的后门攻击中,这一攻击方式演进成触发器攻击,并进一步扩展成更加适应联邦学习的分散学习、全局整合特点的分布式触发器攻击。此外,诞生于传统深度学习的防御性蒸馏等经典防御手段同样可以应用于联邦学习的防御,但差分隐私等防御机制可能削弱全局模型的性能,在隐私与模型鲁棒性之间需要有必要的权衡[27−28]。

2 后门攻击方法

在实验样机射频自联邦学习中的后门攻击问题被提出开始,开发攻击手段的工作快速发展[2]。相比于无攻击目标,专注于破坏模型性能的拜占庭攻击和期望能够依靠参数推理得到用户隐私数据的推理攻击,后门攻击在与拜占庭攻击有一定相似的前提下具有更加明确的直接攻击目标,一般是期望通过攻击让全局模型对某一类输入的分类被混淆成错误的指定类,或者让全局模型在辨认被嵌入触发器的数据时无论来源统一误判成指定的目标类[19]。

由于后门攻击本身的设计理念导致其不会大幅度影响模型的性能,因此,比拜占庭攻击具有更高的隐蔽性和更强烈的威胁[16]。与此同时,更明确的目标也要求后门攻击有更缜密精细的计算,需要更加审慎地选择数据和标签的修改,同时联邦学习的后门攻击存在的固有挑战是单一客户端的攻击很容易在上传参数时被其他良性模型稀释作用,除非能控制多个客户端,或者能够提高本地模型在全局整合中的权重,否则后门攻击的效果很难达到要求。

一次攻击成功与否不能仅凭肉眼观察模型的输出,而需要更确切数据化的判断。分析后门攻击本身特性发现,攻击的成功程度与以下指标有关。

1)后门攻击成功率(attack success rate, ASR),即被成功分类成目标类的触发器样本占全部触发器样本的总量[29],后门攻击的首要目标是提高成功率,ASR 越高,代表攻击的有效性也就越强。

2)良性任务精度(benign accuracy, BA),是被正确分类的无触发器样本占全部无触发器样本的总量,代表模型在无触发器环境运行下是否能正常工作,接近正常状态的 BA 能够让后门攻击有显著的隐蔽性。

3)攻击隐蔽性,是后门攻击在联邦学习过程中不被参与者或服务器检测到的能力,在 ASR 过高与 BA 显著下降时会被大幅度影响。通常可以通过限制触发器作用范围与模仿正常模型更新实现高隐蔽性。

4)攻击持久性,指后门在多轮次聚合后仍能保持有效,不被良性更新或全局聚合所覆盖的能力,在长期后门攻击中极为重要。

后门攻击在联邦学习中的表现形式多样,但到目前为止,主流的后门攻击手段围绕着 2 大类攻击方法展开:数据投毒攻击与模型投毒攻击[30]。二者往往不是简单的互斥状态,在后门攻击中,2 种攻击方式经常同时出现,结合作用,但在不同的攻击方案中存在对 2 种攻击方法的不同偏重,因此,可以进行粗略的分类以便理解与介绍。

除了最基础的 2 类后门攻击方法外,根据参与攻击的恶意客户端是否有多个,是否能进行多轮次攻击,是否可以对联邦学习的防御机制做出调整等属性也可以对攻击方案进行进一步的描述。多恶意客户端参与的分布式后门攻击性能通常高于单一恶意客户端参与的集中式后门攻击,多轮次的后门攻击成功率高于单轮次的攻击,能够实现自适应攻击的方案效果好于非自适应攻击的方案,与之相对的,有更好效果的攻击行为也需要更加完善的计算和设置。

2.1 数据投毒攻击

数据投毒攻击是针对训练数据的攻击,主要由恶意客户端实施,旨在只改变局部训练数据而在后门植入全局模型,一般围绕选择经过精密设计的触发器展开。设计的需求包括但不限于能够高效地实现攻击目的,能够逃避后门防御检测,能够在联邦学习本身的多轮次易遗忘后门的条件下维持作用等。

基于标签篡改的投毒攻击是其中一种基础而直接的攻击范式[31]。标签翻转攻击是其中的典型代表,其基本策略是将某一类源数据的标签系统地、故意地错误标注为指定的目标类,从而迫使模型学习这种错误的关联。Bagdasaryan 等[2]的研究证实了该方法在联邦学习中的可行性,其攻击流程如图 5 所示。总体而言,标签翻转攻击实施简单、通用性强,无须深入理解模型内部机制。然而,其展现出的弱点是隐蔽性不足:任何对本地数据的筛查都可能轻易发现异常的标签模式。在联邦聚合过程中,单个或少数恶意客户端所贡献的、基于错误标签学得的更新,极易被大量良

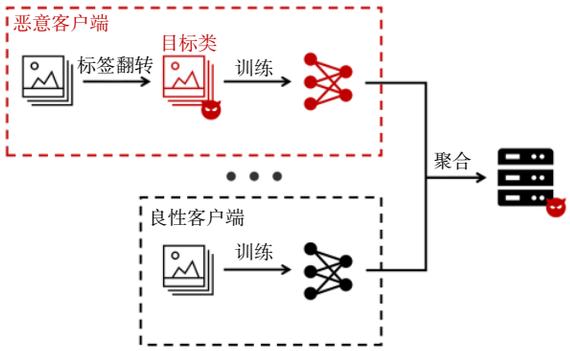

图 5 基于标签翻转投毒的联邦学习后门攻击

性更新所稀释,导致攻击效果迅速衰减甚至被全局模型“遗忘”。此类攻击通常对攻击者能力要求较低,仅需要篡改本地数据标签,无须知晓模型结构或训练细节,可视为一种灰盒或黑盒攻击。为克服攻击稀释问题,通常需要控制多个客户端协同发起女巫攻击[32],以人为增加恶意节点数量的方式放大攻击影响力。

为克服标签翻转攻击隐蔽性差的弱点,适应更严格的检测环境,研究重点转向了基于触发器注入的投毒。这类攻击不依赖于修改标签,而是通过在输入中嵌入隐蔽的触发器模式,并使其与目标输出强关联。根据触发器的设计复杂度和植入策略,该范式下衍生出多种更具威胁的攻击变体。

其中,具有代表性的一种实现是静态触发器攻击。以恶意网络[33]为代表,该类攻击使用一个固定的、预先定义好的模式(如图像中的特定图案)作为触发器,在训练阶段向部分样本中嵌入精心设计的触发器模式,诱导全局模型将该模式错误地学习为分类的重要特征[2]。在数据分布方面,IID 场景下客户端对静态触发器的学习更容易传播。尽管易于实现,但其固定模式也成为了防御的突破口,使其在 Non-IID 场景下难以逃过基于触发器模式匹配的检测[34]。

为此,动态触发器攻击应运而生,Huang 等[35]通过应用学习策略和训练架构,在不同训练轮次中使用变化的触发器,大大增加检测的难度。基于动态生成触发器的攻击方法下,攻击者在本地训练中通过优化损失函数,实现全局性能与攻击成功率之间的动态平衡,从而避免人为设定平衡因子导致的非最优问题,同时在攻击效果与隐蔽性之间取得更优的权衡。更进一步,依赖模型的触发器生成方法根据从服务器接收的全局模型反馈来优化触发模式,旨在利用选定神经元最大化后门激活的效率,通过训练回馈获得最具有效攻击能力的触发器[36]。然而,生成动态或模型依赖的触发器通常需要更强的白盒知识,或在灰盒条件下通过指示器等方式获取模型内部结构。在证明后门攻击对于联邦学习有防御压力的过程中,研究者提出了边缘案例后门攻击[37]。边缘案例指的是位于数据分布尾部,以低概率出现,并且成为训练数据或测试数据一部分的概率也较低的数据,甚至部分数据集在此处的样本可能存在缺失。如图 6 所示,边缘案例后门攻击的核心是边缘数据集的构造与注入。恶意客户端首先利用预训练模型推断良性样本,收集末层输出向量,并以此拟合高斯混合模型作为数据分布近似,并计算候选样本的对数概率密度,选取密度显著低于主流分布(如远低于 MNIST 测试集分布)的样本类别作为边缘案例,隐式地定义了概率阈值。随后将构造数据集与良性样本按特定比例混合,构造出本地投毒数据集。

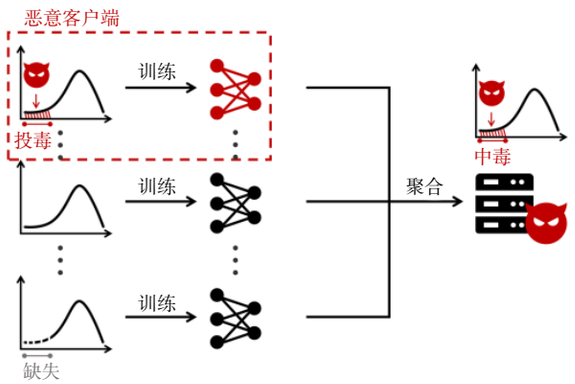

图 6 基于边缘案例投毒的联邦学习后门攻击

Wang 等[37]的工作表明,该攻击在类别不均衡/高度 Non-IID 场景下展现出独特的稳定性与优势,尤其当边缘案例集中由少数恶意客户端持有,而非均匀分散于诚实客户端时,攻击成功率最高。反之,若边缘案例广泛存在于诚实客户端,接近于 IID 场景时,攻击效果将被显著削弱。与经典触发器投毒攻击相比,边缘案例攻击在 ASR/BA 平衡及防御规避上表现更优。在白盒投影梯度下降(projected gradient descent, PGD)攻击下,该方法能在保持全局模型在主任务上高准确率的同时,实现接近 100% 的攻击成功率,并成功绕过包括 Krum、Multi-Krum、范数裁剪及鲁棒聚合在内的多种先进防御。

为了从根本上应对联邦学习分布式架构带来的防御挑战,Xie 等[29]提出了分布式协同后门攻击。如图 7 所示,该策略将完整的触发器模式分解为多个局部触发器,分配给多个由攻击者控制的恶意客户端,每个客户端仅在其本地数据中嵌入并学习一个局部触发器。攻击的核心流程分为 3 步:(1)将全局触发器在空间上分解为 M 个互不重叠的局部触发器,分配给 M 个恶意客户端;(2)每个恶意客户端将所分配的局部触发器嵌入本地数据,旨在通过多轮次的学习与缩放,获得让后门生效概率与良性任务精度最大化的模型参数;(3)服务器将各模型聚合更新,由于更新的线性特性,不同的局部恶意更新被线性组合,使得全局模型最终能够对全局触发器做出响应。

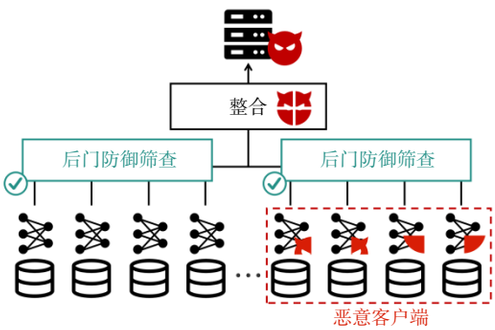

图 7 分布式协同后门攻击的实现

在聚合前,单个客户端上传的模型更新中仅包含微弱乃至无法识别的后门信号,使得几乎所有基于分析单客户端更新异常的聚合前防御机制失效,而完整的后门功能则在服务器端的全局模型聚合过程中得到“组装”,进行有效投毒攻击。更具威胁性的是,一些旨在平滑或融合各客户端更新的聚合中防御算法,如 RFA[38](robust federated aggregation),反而可能无意中促进分布式后门的组装与强化。Yang 等[39]通过研究发现,以子图作为触发器在图结构中植入恶意节点或边,可以使得全局图模型在测试阶段能对特定子图产生错误的分类,证明该范式在图联邦学习等复杂场景中同样有效。分布式协同后门攻击代表了当前数据投毒攻击中技术最复杂、威胁性最高的发展方向之一。由于要求攻击者设计可被整合的局部触发器,需求掌握全局模型的整合过程,分布式协同后门攻击的知识要求趋近于白盒。该攻击对数据分布的适应性较强,旨在结合聚合过程本身,防御此类攻击需要全新的、可洞察跨客户端协同攻击的机制。

2.2 模型投毒攻击

模型投毒攻击是针对客户端训练得到的本地模型的攻击手段。其攻击手段是通过恶意操纵客户端上传的模型更新梯度或参数,最终在联邦学习训练得到的全局模型中植入后门[31]。通常情况下,模型投毒攻击同时使用原始数据与投毒数据。通过双重数据训练,获取兼具良性任务性能和隐蔽后门功能的本地模型,并最终通过上传更新将后门注入到全局模型中。许多模型投毒攻击的思想借鉴并发展了数据投毒攻击的技术,但在攻击层面直接操控训练产出的模型权重与参数,而不局限于污染训练数据。这种变化使得攻击者可以更加精准地嵌入后门逻辑,导致全局模型在保持良性任务性能的同时,对含触发器的输入产生攻击者指定的错误行为[40]。本研究假设中央服务器与相关运营者具备基本安全保障,中央数据不会非攻击性泄露。

在众多模型投毒策略中,模型替换攻击代表了最为激进的一类。该方法并非字面意义地替换全局模型,而是通过大幅放大恶意本地模型的更新权重,使其在一轮聚合中占据绝对优势,导致全局模型大幅度地向恶意模型偏移,达成几近被“替换”的效果。Bagdasaryan 等[2]于 2020 年提出了这种针对联邦学习的攻击方法,旨在克服传统中毒攻击在多轮聚合中易被遗忘的弱点。研究者先训练出除后门外与全局模型较为相似的投毒模型,随后将更新乘以一个巨大的缩放因子后上传给中央服务器,得以实现基于修改权重比例的模型替换。通过精确计算,攻击者甚至可以在单轮次内迫使全局模型牢固记忆后门,并在安全聚合协议的掩护下规避检测。

然而,模型替换攻击成功的关键在于时机。它必须在全局模型接近收敛,各客户端更新幅度普遍减小的时刻发动,以使放大的恶意更新在轮次内占据绝对优势。也因此,攻击者需要能够获悉全局模型的训练状态、整合算法、客户端数量等白盒知识,一定程度限制了其普适性。该攻击在 IID 场景下效果尤为显著,易于令放大的恶意更新占据主导,而 Non-IID 场景或防御机制严格的环境下,其攻击成功率与隐蔽性会显著下降。

与模型替换的爆发式策略相反,一部分研究者转向更为温和,持久实行后门感染的隐蔽模型投毒攻

击。此类攻击策略的核心目标是逃避检测,通过精细设计使中毒模型与良性模型高度相似,二者难以区分,从而绕过异常检测,不易被后门防御机制检测裁剪。例如,控制中毒模型与良性模型的距离小于中央服务器区分异常更新的阈值,或者依靠预测全局模型整合方向,将有毒模型向预测模型靠拢等。实现隐蔽的模型投毒攻击,通用的实现框架可以抽象为带约束的双目标优化问题,即最大化后门任务成功率的同时,以共形性约束确保其隐蔽性。该约束的具体形式与阈值设定紧密围绕目标设计,以针对 Krum 的隐蔽模型投毒(covert model poisoning, CMP)攻击为例,其约束设计主要体现在 2 个层面[41]:(1)恶意客户端之间的协同约束;(2)针对聚合规则的动态阈值约束。协同约束要求所有恶意客户端模型参数两两之间欧氏距离小于固定阈值,以使恶意更新在 Krum 距离计算中形成易被选中的簇,即

$$\left\|\hat{\boldsymbol{\theta}}_i-\hat{\boldsymbol{\theta}}_j\right\|\leqslant\varepsilon,\ \forall i,j\in\mathbf{M} \tag{1}$$

式中,$\hat{\boldsymbol{\theta}}_i$与$\hat{\boldsymbol{\theta}}_j$分别表示第 i 个和第 j 个恶意客户端在第 t 轮训练后,准备上传给服务器的本地模型参数向量。$\mathbf{M}$ 表示所有恶意客户端的集合。ε表示预设的极小正数阈值,以限制恶意模型之间的最大差异。

在实践中该协同约束主要通过向主恶意模型添加范数裁剪后的随机噪声来实现。而动态阈值约束则限制恶意模型与最接近的(U–2M–1)个良性模型的欧氏距离之和,要求其不超过由历史良性更新计算得到的动态基准 E,以实现隐蔽需求,即

$$\sum\left\|\hat{\boldsymbol{\theta}}_{\text{malicious}}-\boldsymbol{\theta}_{\text{benign}}\right\|\leqslant E-(M-1)\varepsilon \tag{2}$$

式中,$\hat{\boldsymbol{\theta}}_{\text{malicious}}$ 表示希望被选中的主要恶意模型参数向量,$\boldsymbol{\theta}_{\text{benign}}$ 表示各个良性客户端的本地模型参数向量。E 为动态阈值基准,即攻击者估算的良性客户端最佳 Krum 距离分,以约束恶意模型自身行为。M 表示恶意客户端数量。ε同上,表示恶意模型间的最大距离。

Wei 等[41]通过优化攻击策略规避检测,提出了隐蔽模型投毒攻击框架,利用优化算法使恶意模型参数在全局模型中的表现与正常模型相似,实现对后门防御的欺瞒。另一种思路则是从联邦学习机制本身寻找漏洞,如贡献评估攻击。以 ACE 为例[42],攻击者从全局模型自身的聚合方法入手,针对联邦学习的客户端贡献评估机制,操纵本地模型参数使恶意客户端在贡献评估中被错误地高估,从而使恶意模型获取更高的权重和奖励。贡献评估攻击不仅可以实现于后门注入,同时还是对联邦学习公平性的重大威胁,可能导致进一步的资源分配不公与模型偏见的产生。

隐蔽投毒攻击旨在长期渗透,因此,对数据分布的适应性更强,威胁模型更灵活,对基于优化的隐蔽攻击,攻击者通常需要白盒知识,包括防御算法的规则与阈值以构造约束,对于机制性攻击则注重于对联邦学习的特定聚合与评估协议的理解。

随着后门防御机制日益系统化,以恒定的超参数和既有算法作为攻击核心的静态攻击策略容易被针对性地破解,自适应投毒攻击应运而生。这类攻击是一种结合模型反馈动态调整攻击策略的攻击方法,与传统的静态攻击相比,自适应投毒攻击通过监控模型训练过程,以动态优化触发器设计、投毒方式,以及参数修改,同时实现高成功率与高隐蔽性。自适应的关键技术是利用冗余神经元,即在中央服务器训练全局模型的过程中激活频率低,对主任务影响小的“边缘神经元”。现有研究通过量化准则来进行筛选冗余神经元,通常选择在本地数据梯度中绝对值最小,且损失函数海森矩阵趋近零的参数。这些神经元是攻击者从模型内向外传递模型信息的良好选择,如图 8 所示,通过将冗余神经元作为指示器,分析其在多轮训练中的变化,可以准确推断攻击是否生效,乃至判断服务器的防御机制,并进一步优化攻击。

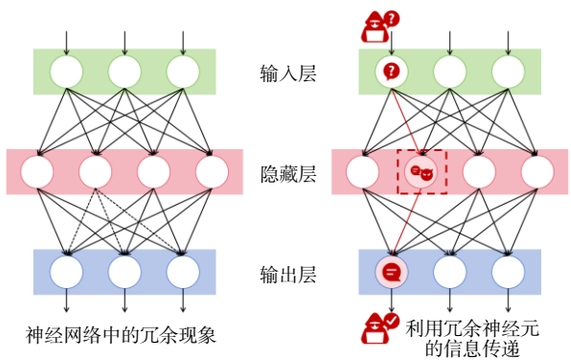

图 8 冗余神经元的信息传递

3DFed 是自适应框架的代表之一,利用联邦学习的多轮训练传递指示器,在每轮训练中不断优化攻击、改进触发器的设计[43]。同时,结合了分布式协同后门攻击,提出用诱饵模型引导中央服务器的剪切防御,可以接受更多攻击策略嵌入该框架,且不需要预

先获知联邦学习系统的各种参数,是一个自适应和可扩展的隐蔽后门攻击框架。面对联邦学习的多层防御,尝试创造能同时满足多种防御机制且保留攻击性的触发器更加困难,自适应的投毒攻击愈发成为迫切需求,实验表明,3DFed 在面对多种防御机制时仍能保持较高的攻击成功率,同时维持良性任务精度接近正常水平。Shi 等[44]进一步改进了基于冗余神经元的自适应方法,强调在实际系统中的攻击可能面对后门遗忘等挑战,提出用向后梯度与正向激活等方法相结合,通过降低反向神经元对其他客户端数据的敏感度,并增强识别的正向神经元路径,达到为攻击者毒害全局模型的目的,人为巩固后门的权重,从而降低后门被遗忘的可能。

此类攻击代表了最先进的黑盒威胁模型,攻击者不需要提前知晓服务器的防御策略、聚合规则、良性客户端数据分布及其他超参数,仅通过指示器获取有限反应信号,可以应对复杂的、多层次的防御环境,能有效适应数据场景。然而,基于反馈的自适应机制其鲁棒性高度依赖于防御环境,Li 等[43]在对 3DFed 的工作中指出,指示器机制在服务器采用差分隐私时,会因噪声淹没细微信号而失效。此外,严格的模型剪枝和梯度裁剪存在对反馈准确性的影响,因此,攻击也需要使用其他辅助算法对严格防御进行突破,如使用诱饵模型误导裁剪等。

联邦学习的结构限制了客户端直接上传污染数据的能力,导致一部分对于普通单体模型有效的攻击策略不能直接在联邦学习流程内应用,如干净标签攻击或基于篡改模型自身结构实现的后门攻击,即便如此,后门攻击在联邦学习中依然有各异的分类与威胁,对以上的联邦学习后门攻击方法对比如表 1 所示。

表 1 联邦学习后门攻击方法对比

攻击对象	攻击方法	威胁模型	是否需观察全局模型	适用分布
数据	标签翻转攻击	灰盒/黑盒	×	Non-IID
	触发器注入	灰盒	×	IID
	边缘案例攻击	白盒	√	Non-IID
	分布式协同攻击	白盒	√	IID/Non-IID
模型	模型替换	白盒/灰盒	√	IID
	隐蔽模型投毒	白盒	部分需要	IID/Non-IID
	自适应投毒	黑盒	√	IID/Non-IID

3 后门防御方法

为了对联邦学习的后门攻击安全问题进行防御,防御技术持续更新迭代应用,对应不同的后门攻击。联邦学习防御的主要目的在于有效防御的同时保持良好性能,既要消除后门对全局模型的影响,又不能因为过度防御导致全局模型的性能明显下滑。与可以针对特定防御方法确定策略的后门攻击相比,后门防御无从提前得知发动攻击的是哪一类后门攻击方法,因此防御必须有一定的普适性,独立于攻击策略或者客户端的数据分布。

现今的防御方案一般是框架化、层次化、趋于用多种防御方法结合,本节将按照防御的时间点与思路对各种防御方法进行梳理与介绍,并列举应用该方法为核心的防御方案。

3.1 聚合前防御

聚合前防御的主要思路通常是在接收来自客户端的参数更新时对其进行检验和过滤,将更新的参数限定在一定范围内,或者对参数进行一定程度的调整,从而降低异常参数对聚合的全局模型造成的影响。

基于裁剪和噪声添加的聚合前防御,作用原理是对模型投毒的后门攻击更新往往具有较大的范数,此时令中央服务器忽略范数大于阈值的更新,或者剪裁范数到阈值范围以内,能确保接受的更新范数在可控范围内,更新全局模型后对其添加高斯噪声,实现对后门的清洗,是防御标签翻转和基于放大参数的模型替换攻击的有效措施[19]。FLAME 是应用这种防御的典型框架,并且添加了聚类方法提前去除偏差更大的疑似恶意模型,进一步压缩了清除后门所需的噪声[45]。FLAME 以滤波—限幅—噪声的防御主题实现了以模

块化的工作在防御后门的同时尽可能维持全局模型的性能,能够实现完全防御多种后门攻击。然而该类防御的效果对裁剪阈值或噪声尺度等超参数高度敏感:裁剪阈值过严将损害模型精度,过松则难以防御攻击;噪声过强会破坏全局模型性能,且增加了计算开销,可能影响学习效率。

基于降维的聚合前防御则是通过为高维的模型特征选择最具代表性的特征维度,利用以主成分分析(principal components analysis, PCA)为代表的降维方式提取低维特征,从而在低维特征空间中更容易地完成异常更新的区分和剔除。Cai 等[46]将 PCA 降维与 DBSCAN 聚类相结合,提出以降维为主要思路的 FLMAAcBD,依靠合理地随模型收敛调整超参数,完成对后门的有效裁剪。降维防御的效果取决于降维后保留的主成分数量与聚类参数,在保留信息与去除噪声中权衡,在 Non-IID 场景下,过于严格的参数设置容易导致误删良性更新。此外,目前的降维聚类算法开销压力依然存在,PCA 降维涉及到复杂度 $O(d^3)$ 级别的矩阵运算,DBSCAN 聚类的复杂度最坏可达 $O(n^2)$,对规模较大的联邦学习系统,其时间与空间复杂度可能难以接受。

基于主动预测的聚合前防御与前述方式相比更加需要计算的支持,通过对客户端的行为或模型参数进行预测与分析,通过更新内容与预期结果的偏差大小识别潜在的恶意客户端,实现源头上的对后门攻击的预防。FLDetector[47]就是一种使服务器根据历史更新在每次迭代中预测客户端的模型更新的防御手段,认为多次迭代预测均不匹配的客户端是恶意的,对其进行检测和删除,留下良性客户端供全局模型进行聚合,具体流程如图 9 所示。BackdoorIndicator[48]则从分布外数据(out-of-distribution, OOD)入手,概念如图 10 所示,将所有后门样本视作良性数据集的分布外数据,因此,恶意客户端很可能对由分布外数据构建的指标任务有更好的表现,凭此可以用对指标任务高精度的更新筛选出高度可疑的客户端,并把它们从聚合中过滤掉。同样地,郭晶晶等[49]提出的基于模型水印的防御利用恶意更新会破坏原有参数的特点,在全局模型中插入神经网络水印,以寻找恶意参与方并丢弃对应更新,保证参与整合的更新纯净。在开销方面,服务器需要为每个客户端维护历史状态,执行预测工作或运行额外的指标推断,会为学习引入持续的内存与计算负担,是实际应用中需考虑的成本。

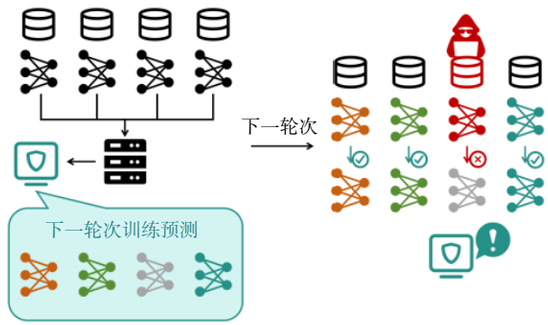

图 9 基于主动预测与匹配的后门防御

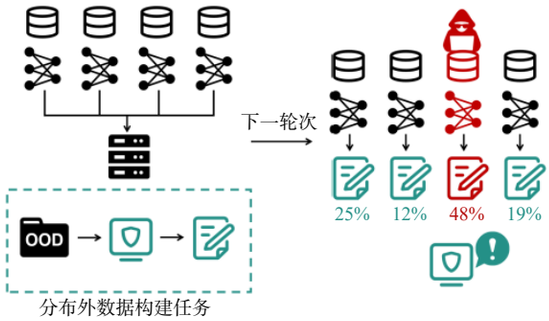

图 10 基于主动预测与指标任务的后门防御

3.2 聚合中防御

聚合中防御是与全局模型的聚合算法高度关联的防御,涉及到对聚合算法的优化或者是对权重参数进行重分配,旨在通过更具有鲁棒性的算法稀释后门影响或者为更新异常的客户端降低权重,从而依靠聚合算法自身实现对后门的防御[50]。为了防止攻击者可能通过聚合获知梯度以重建用户隐私,主流方案为安全聚合,以掩码[51]或差分隐私[52-53]等方式掩盖真实的模型梯度,阻止攻击者获取真实数据。

基于鲁棒性的聚合中防御本身建立在联邦学习自身的鲁棒性上,即全局模型自身有一定的处理数据异常、识别容忍误差,以及消化恶意攻击的能力。在联邦学习框架下,模型聚合过程面临来自不可靠或恶意客户端的多种威胁,尤其是拜占庭攻击,其核心目标是破坏联邦学习系统的这种鲁棒性。为应对这一挑战,研究者提出了多种基于鲁棒统计的聚合策略,以提升全局模型对异常更新的识别与抵抗能力。中位数聚合算法的核心思想在于计算所有更新的中位数作为聚合结果[54]。Krum 专门设计用于抵御拜占庭客户端的恶意模型更新,计算每个客户端更新与其他

更新权重向量间的欧氏距离,依此挑选最接近整体更新的一组模型,排除异常更新,缓解拜占庭攻击[55]。Bulyan 基于 Krum 建立,在 Krum自身的筛除功能上增加求取统计平均的功能,多次运行 Krum 算法求取一组较可信的候选更新后,对每个权重向量的参数维度进行二次裁剪并取中值截断平均,能更严格地处理残余拜占庭客户端可能的攻击[56]。MAB-RF 的鲁棒性聚合算法突破了鲁棒性算法只能随机选择客户端的局限,利用多臂老虎机问题计算用户选择,实现更加合理的对诚实用户的筛选[57]。

基于相似性的聚合中防御通过衡量各客户端上传模型之间的相似程度,可以有效识别并削弱潜在的异常或恶意更新。这类防御通常依赖余弦相似度或欧几里得距离等度量计算并比较上传的局部模型间的相似性,为聚合过程提供更有区分度的加权依据,并依此调整各模型的权重,相比传统的平均或基于鲁棒统计的策略能够更精细地刻画模型行为的偏离程度。一类常见的方法是调整各客户端模型在聚合中的占比,模型越接近多数派更新就获得越高的权重;另一类方法则进一步引入聚类算法,如 K-means 或 DBSCAN,通过对局部模型的特征空间聚类识别离群更新,并将其从聚合过程中剔除或显著降低其影响,从而增强系统的鲁棒性。

Gong 等[58]提出的 AgrAmplifier 即是一种基于距离进行优化的聚合规则,该方法在传统相似度机制的基础上引入了密度估计,通过评估模型更新在局部空间中的密集程度,进一步区分可能的异常更新与正常更新。针对更隐蔽和结构化的分布式后门攻击,DAGUARD[59]提出了对应的分布式后门防御方案,应用了梯度压缩策略 TernGrad[60]以抑制由后门触发造成的局部梯度异常放大,以及通过自适应 DBSCAN 实现对后门触发更新的有效聚类与剔除,充分体现了多种防御机制联合应用的优势,在保持模型性能的同时,有效削弱了分布式后门对全局模型的渗透能力。

3.3 聚合后防御

聚合后防御着重于在全局模型的训练结束后对所获得的模型进行检测和修复,是在将全局模型广播给下一轮训练前的最后一道防御,通过检查全局模型的性能波动,或者通过联邦遗忘等方式实现对模型的清洗。

基于联邦遗忘的聚合后防御的核心前提是需要中央服务器准确识别恶意客户端,旨在快速构建一个“遗忘”恶意客户端影响的新全局模型。最初由 Liu 等[61]开发的 FedEraser 尝试依靠再次学习实现遗忘:在常规训练中,以固定轮次间隔存储所有客户端的更新记录,在需求移除特定客户端时回滚训练环境,利用存储的历史更新进行少量轮次的校准训练,确定模型的良性更新方向。通过合理设置存储间隔与校准轮次比例,可以实现接近从头训练的效果,然而,该方法属于“以存储换时间”的策略,在客户端数量过大的情景下,将发生性能下降,带来额外的存储成本。

为规避存储和客户端重新参与的成本,后续研究提出了削减存储和计算开销的改进思路。Wu 等[62]提出定期保存更新记录的联邦遗忘防御,但与 FedEraser 的再次学习不同,是对全局模型中减去恶意客户端的历史平均更新,获得剪切的倾斜模型,以原始全局模型作为教师模型执行知识蒸馏[63],实现对全局模型的清洗,相比于 FedEraser,该方法不需要客户端再次参与,相对更加实用和高效。BadCleaner[64]则利用多教师知识蒸馏,让蒸馏类型从一对一转向多对一,结合多个客户端的知识进行更好的知识梳理和清洗,同时应用注意力迁移机制,让全局模型的中间层向教师模型中间层对齐,尽可能擦除全局模型对其中存在的触发区域的注意,达成深层次清洗。以上联邦学习后门防御方法的对比如表 2 所示。

4 未来研究方向与前景展望

4.1 后门攻击角度

联邦学习的分布式特性使得攻击者难以直接访问全局数据或模型,因此,攻击策略需要更加精细化和隐蔽化,在躲避后门防御的同时实现攻击目的。现有的对于后门攻击的开发中,攻击目标正在从纯粹追求攻击成功率逐渐转向追求高隐蔽性、高留存度、高可实践性、低数据掌握、对防御可适应等特点为代表的现实化攻击。

后门攻击在隐蔽性和留存度上的提升,需要重点关注攻击者的自我约束。在联邦学习中,由于客户端数据的 Non-IID 特性,干净标签攻击可以更好地隐藏在后门数据中,避免被中央服务器检测到异常。数据

表 2 联邦学习后门防御方法对比

防御时间段	防御方法	针对类型	部署成本	适用分布
聚合前防御	范数剪切	标签翻转	低	IID/Non-IID
	PCA降维	通用后门	中	IID
	主动预测	持续投毒攻击	高	IID
聚合中防御	鲁棒聚合	拜占庭攻击	低	IID
	相似性聚类	分布式后门	高	IID/Non-IID
聚合后防御	联邦遗忘	已知恶意客户端	高	IID/Non-IID
	知识蒸馏	模型替换	中	IID/Non-IID

投毒攻击在常规的触发器攻击基础上,有逐渐走向干净标签攻击的趋势,增强视觉上的隐蔽性,让触发器更加不可察觉,降低人工筛查异常数据的可能性。而模型投毒攻击中将自约束项融合在损失函数中已经逐渐常态化,依靠模型更新的限度控制躲避初步的筛查,并通过自我添加噪声和干扰更新迷惑中央服务器的防御。与此同时,执行增加隐蔽性的改造必然会导致攻击性能的下降,低留存度的恶意更新很可能无法在全局模型中留存有效的后门记忆,因此,后门的留存度与持久性仍然需要进一步开发。

后门攻击在可实践性和数据需求上的改进,则需要攻击者进一步改进算法。在联邦学习中,攻击者通常只能控制少数客户端,且无法直接访问全局模型或数据,在联邦学习后门攻击初期常见的,能掌握大量客户端或多份训练数据的白盒攻击在实际攻击环境中难以达成,黑盒攻击和低数据依赖的攻击策略更具实际意义,研究方向正在逐渐走向发掘控制数据更少,对聚合算法保持未知的前提下也可以实现攻击的算法,追求更契合实际情况也更易达成的目标。

后门攻击在适应性攻击上的突破让后门攻击策略变得更加棘手,是对后门防御的崭新挑战,以冗余神经元等方式获取全局模型信息的作用从降低信息需求扩展向了利用冗余神经元确认攻击反馈,从而实现更加具有适应性的攻击。除了监控当前轮次的反馈信息,攻击者也在应用本地训练受害模型作为全局模型替代,更好地预演攻击效果和改进策略,这在无法访问全局模型的联邦学习环境中是很好的替代方案。

4.2 后门防御角度

后门防御为了对抗更加具备潜伏性质的攻击,在常规的防御手段之外也需要修正,包括对模型性能与防御力度的权衡、减轻对算力资源的依赖,以及更好地实现隐私保护。

模型性能会因为过于严格的后门防御策略而遭到破坏,而过轻的防御力度会让本就容易遭到后门攻击的全局模型更具风险,现有的防御方法在面临更具挑战性的攻击时也不可忽略后门防御这一最核心的需求。由于联邦学习环境下客户端数据的非独立同分布特性,过于严格的防御策略可能会导致模型在某些客户端上的性能下降,因此,需要开发更加智能,更具适应性的防御机制,能够在保护模型性能的同时有效防御后门攻击。

现有的后门防御为了实现对后门的排查,通常会叠加多层次、多时间段的后门防御,形成纵深防御体系。然而这种聚合防御范式的潜在需求是系统开销与资源消耗。不同防御模块在服务器顺序或迭代执行,总时间复杂度往往是各算法的累加甚至乘积,严重制约了系统扩展性;此外,组合防御往往要求服务器存储多轮中间状态,或与客户端进行更多交互,会成倍增加通信轮次与带宽消耗。实际应用中的联邦学习系统由数量庞大的客户端支撑,不同的客户端往往存在计算能力与存储空间等方面的限制。利用异步执行、早期退出等优化策略以实现防御效果与资源开销的权衡,避免对客户端造成过大负担,是研究者面临的重大挑战。未来的防御机制需要在安全增益与资源消耗间寻求精密平衡,如在密码学领域,Chen等[65]提出的 FS-LLRS 在引入强安全属性的同时严格控制计算与通信开销,启示联邦学习的后门防御设计也应致力于更多探究高安全与低开销的平衡。

部分高效的后门防御策略能够准确实现后门排查与清除,代价是调用客户端的模型更新,导致无法兼容安全聚合。在联邦学习中,隐私保护是核心需求之一,牺牲安全和隐私实现的后门防御并非实践化的

理想结果。探索如何在不破坏安全聚合等隐私保护技术的前提下实现有效的后门防御,或寻找在噪声干扰下仍能实现工作的检测算法,如借鉴隐私计算中的密文认证思想[66],构建对客户端更新的可验证性而非可视性等,达成后门防御与隐私的共存,是推动联邦学习后门防御实用化的前瞻道路。

参考文献(References)

[1] McMahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proceedings of International Conference on Artificial Intelligence and Statistics. Lauderdale: PMLR, 2016.

[2] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning[C]//International conference on artificial intelligence and statistics. Online: PMLR, 2020: 2938-2948.

[3] Xie C, Koyejo S, Gupta I. Fall of empires: Breaking Byzantine-tolerant SGD by inner product manipulation[J/OL]. arXiv, 2019: 1903.03936. https://arxiv.org/abs/1903.03936.

[4] Zhao Y, Li M, Lai L Z, et al. Federated learning with non-IID data[J/OL]. arXiv, 2018: 1806.00582. https://arxiv.org/abs/1806.00582.

[5] Xu G, Kong D L, Zhangs K, et al. A model value transfer incentive mechanism for federated learning with smart contracts in AIoT[J]. IEEE Internet of Things Journal, 2025, 12(3): 2530-2544.

[6] Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning[C]//Proceedings of IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE, 2019: 739-753.

[7] Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: Information leakage from collaborative deep learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 603-618.

[8] Geiping J, Bauermeister H, Drge H, et al. Inverting Gradients: How easy is it to break privacy in federated learning? [J/OL]. arXiv, 2020: 2003.14053. https://arxiv.org/abs/2003.14053.

[9] Melis L, Song C Z, De Cristofaro E, et al. Exploiting unintended feature leakage in collaborative learning[C]//Proceedings of IEEE Symposium on Security and Privacy (SP). Piscataway, NJ: IEEE, 2019: 691-706.

[10] 刘艺璇, 陈红, 刘宇涵, 等. 联邦学习中的隐私保护技术[J]. 软件学报, 2022, 33(3): 1057-1092.

[11] Yang Q, Liu Y, Chen T J, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[12] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[13] Kairouz P, McMahan H B. Advances and open problems in federated learning[J]. Foundations and Trends in Machine Learning, 2021, 14(1/2): 1-210.

[14] Li Y M, Jiang Y, Li Z F, et al. Backdoor learning: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(1): 5-22.

[15] Dai Y B, Li S Z. Chameleon: Adapting to peer images for planting durable backdoors in federated learning[J/OL]. arXiv, 2023: 2304.12961. https://arxiv.org/abs/2304.12961.

[16] Bhagoji A N, Chakraborty S, Mittal P, et al. Analyzing federated learning through an adversarial lens[J/OL]. arXiv, 2018: 1811.12470. https://arxiv.org/abs/1811.12470.

[17] 江萍, 李芯蕊, 赵晓阳, 等. 联邦学习安全聚合算法综述[J]. 网络安全技术与应用, 2024(9): 47-51.

[18] Li S Y, Cheng Y, Liu Y, et al. Abnormal client behavior detection in federated learning[J/OL]. arXiv, 2019: 1910.09933. https://arxiv.org/abs/1910.09933.

[19] Sun Z T, Kairouz P, Suresh A T, et al. Can you really backdoor federated learning?[J/OL]. arXiv, 2019: 1911.07963. https://arxiv.org/abs/1911.07963.

[20] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J/OL]. arXiv, 2014: 1412.6572. https://arxiv.org/abs/1412.6572.

[21] Qiu S L, Liu Q H, Zhou S J, et al. Review of artificial intelligence adversarial attack and defense technologies[J]. Applied Sciences, 2019, 9(5): 909.

[22] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. New York: ACM, 2017: 506-519.

[23] Xu S Y, Chen X, Guo Y, et al. Toward efficient multi-user access control encrypted search for web data management[J]. IEEE Transactions on Dependable and Secure Computing, 2026, 23(1): 1203-1218.

[24] 陈晋音, 邹健飞, 苏蒙蒙, 等. 深度学习模型的中毒攻击与防御综述[J]. 信息安全学报, 2020, 5(4): 14-29.

[25] Chakraborty A, Alam M, Dey V, et al. A survey on adversarial attacks and defences[J]. CAAI Transactions on Intelligence Technology, 2021, 6(1): 25-45.

[26] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J//OL]. arXiv, 2013: 1312.6199. https://arxiv.org/abs/1312.6199.

[27] Papernot N, McDaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural

networks[C]//Proceedings of IEEE Symposium on Security and Privacy (SP). Piscataway, NJ: IEEE, 2016: 582−597.

[28] Wei K, Li J, Ding M, et al. Federated learning with differential privacy: Algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454−3469.

[29] Xie C L, Huang K L, Chen P Y, et al. DBA: Distributed backdoor attacks against federated learning[C]//Proceedings of International Conference on Learning Representations. Appleton: ICLR, 2020.

[30] Enthoven D, Al−Ars Z. An overview of federated deep learning privacy attacks and defensive strategies[M]//Federated Learning Systems. Cham: Springer International Publishing, 2021: 173−196.

[31] Baruch M, Baruch G, Goldberg Y. A little is enough: Circumventing defenses for distributed learning[J/OL]. arXiv, 2019: 1902.06156. https://arxiv.org/abs/1902.06156.

[32] Fung C, Yoon C J M, Beschastnikh I. Mitigating sybils in federated learning poisoning[J/OL]. arXiv, 2018: 1808.04866. https://arxiv.org/abs/1808.04866.

[33] Gu T Y, Liu K, Dolan−Gavitt B, et al. BadNets: Evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230−47244.

[34] Saha A, Subramanya A, Pirsiavash H. Hidden trigger backdoor attacks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11957−11965.

[35] Huang A B. Dynamic backdoor attacks against federated learning[J/OL]. arXiv, 2020: 2011.07429. https://arxiv.org/abs/2011.07429.

[36] Gong X L, Chen Y J, Huang H Y, et al. Coordinated backdoor attacks against federated learning with model−dependent triggers[J]. IEEE Network, 2022, 36(1): 84−90.

[37] Wang H Y, Sreenivasan K, Rajput S, et al. Attack of the tails: Yes, you really can backdoor federated learning[J/OL]. arXiv, 2020: 2007.05084. https://arxiv.org/abs/2007.05084.

[38] Pillutla K, Kakade S M, Harchaoui Z. Robust aggregation for federated learning[J]. IEEE Transactions on Signal Processing, 2022, 70: 1142−1154.

[39] Yang Y X, Li Q, Jia J Y, et al. Distributed backdoor attacks on federated graph learning and certified defenses[C]//Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2024: 2829−2843.

[40] Rakin A S, He Z Z, Fan D L. TBT: Targeted neural network attack with bit Trojan[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 13195−13204.

[41] Wei K, Li J, Ding M, et al. Covert model poisoning against federated learning: Algorithm design and optimization[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(3): 1196−1209.

[42] Xu Z C, Jiang F Q, Niu L Y, et al. ACE: A model poisoning attack on contribution evaluation methods in federated learning[J/OL]. arXiv, 2024: 2405.20975. https://arxiv.org/abs/2405.20975.

[43] Li H Y, Ye Q Q, Hu H B, et al. 3DFed: Adaptive and extensible framework for covert backdoor attack in federated learning[C]//Proceedings of IEEE Symposium on Security and Privacy (SP). Piscataway, NJ: IEEE, 2023: 1893−1907.

[44] Shi C H, Ji S L, Pan X D, et al. Towards practical backdoor attacks on federated learning systems[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(6): 5431−5447.

[45] Nguyen T D, Rieger P, Chen H L, et al. FLAME: Taming backdoors in federated learning[C]//Proceedings of USENIX Security Symposium. Boston: USENIX, 2022.

[46] Cai H Y, Wang J H, Gao L J, et al. FLMAAcBD: Defending against backdoors in federated learning via model anomalous activation behavior detection[J]. Knowledge−Based Systems, 2024, 289: 111511.

[47] Zhang Z X, Cao X Y, Jia J Y, et al. FLDetector: Defending federated learning against model poisoning attacks via detecting malicious clients[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2022: 2545−2555.

[48] Li S Z, Dai Y B. BackdoorIndicator: Leveraging OOD data for proactive backdoor detection in federated learning[J/OL]. arXiv, 2024: 2405.20862. https://arxiv.org/abs/2405.20862.

[49] 郭晶晶, 刘玖樽, 马勇, 等. 基于模型水印的联邦学习后门攻击防御方法[J]. 计算机学报, 2024, 47(3): 662−676.

[50] 李珣. 面向联邦学习的后门防御方法研究[D]. 成都: 电子科技大学, 2024.

[51] Bonawitz K, Ivanov V, Kreuter B, et al. Practical secure aggregation for privacy−preserving machine learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 1175−1191.

[52] Dwork C. Differential privacy: A survey of results[C]//Theory and Applications of Models of Computation. Berlin, Heidelberg: Springer, 2008: 1−19.

[53] Geyer R C, Klein T, Nabi M. Differentially private federated learning: A client level perspective[J/OL]. arXiv, 2017: 1712.07557. https://arxiv.org/abs/1712.07557.

[54] Yin D, Chen Y D, Ramchandran K, et al. Byzantine−robust distributed learning: Towards optimal statistical rates[C]//Proceedings of International Conference on Machine Learning. Lauderdale: PMLR, 2018.

[55] Blanchard P, El Mhamdi E M, Guerraoui R, et al. Byzan-

tine-tolerant machine learning[J/OL]. arXiv, 2017: 1703.02757. https://arxiv.org/abs/1703.02757.
[56] El Mhamdi E M, Guerraoui R, Rouault S. The hidden vulnerability of distributed learning in Byzantium[J/OL]. arXiv, 2018: 1802.07927. https://arxiv.org/abs/1802.07927.
[57] Wan W, Hu S S, Lu J R, et al. Shielding federated learning: Robust aggregation with adaptive client selection[J/OL]. arXiv, 2022: 2204.13256. https://arxiv.org/abs/2204.13256.
[58] Gong Z R, Shen L Y, Zhang Y J, et al. AgrAmplifier: Defending federated learning against poisoning attacks through local update amplification[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 1241-1250.
[59] 余晟兴, 陈泽凯, 陈钟, 等. DAGUARD: 联邦学习下的分布式后门攻击防御方案[J]. 通信学报, 2023, 44(5): 110-122.
[60] Wen W, Xu C, Yan F, et al. TernGrad: Ternary gradients to reduce communication in distributed deep learning[C]// Proceedings of Neural Information Processing Systems. Long Beach: NIPS, 2017.
[61] Liu G Y, Ma X Q, Yang Y, et al. FedEraser: Enabling efficient client-level data removal from federated learning models[C]//Proceedings of IEEE/ACM 29th International Symposium on Quality of Service (IWQOS). Piscataway, NJ: IEEE, 2021: 1-10.
[62] Wu C, Zhu S C, Mitra P. Federated unlearning with knowledge distillation[J/OL]. arXiv, 2022: 2201.09441. https://arxiv.org/abs/2201.09441.
[63] Eldar Y C. Federated knowledge distillation[M]//Machine Learning and Wireless Communications. Cambridge, UK: Cambridge University Press, 2022: 457-485.
[64] Zhang J L, Zhu C C, Ge C P, et al. BadCleaner: Defending backdoor attacks in federated learning via attention-based multi-teacher distillation[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(5): 4559-4573.
[65] Chen X, Xu S Y, Gao S, et al. FS-LLRS: Lattice-based linkable ring signature with forward security for cloud-assisted electronic medical records[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 8875-8891.
[66] 曹艺博, 徐仕远, 陈雪, 等. 面向云数据访问控制的认证密文检索方案[J]. 科技导报, 2025, 43(12): 161-170.

Backdoor security in federated learning: A survey of attacks and defenses

XU Ke[1], ZHAO Yu[1], XU Shiyuan[2], CHEN Xue[2], LI Qiang[3], GUO Yu[1*]

1. School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China
2. Department of Computer Science, the University of Hong Kong, Hong Kong 999077, China
3. China Industrial Control Systems Cyber Emergency Response Team, Beijing 100040, China

Abstract Federated learning, as a distributed machine learning paradigm, enables collaborative multi-party learning while preserving data privacy, and has been widely deployed in various critical domains. However, its distributed nature also renders it highly vulnerable to backdoor attacks, posing serious security threats. This paper first provides a detailed analysis of the methodological taxonomies and evolutionary directions of backdoor attacks and defenses in the field of federated learning, and summarizes the latest research advances in related areas. Backdoor attack research is shifting toward realistic strategies that simultaneously pursue stealth, persistence, and low data dependency, giving rise to a variety of novel attack vectors. Meanwhile, backdoor defenses are increasingly pursuing greater universality against diverse attack types, as well as a balance between computational overhead and defensive strength within the federated learning environment. Finally, this paper reveals that backdoor attacks and defenses in federated learning are facing increasingly complex challenges, and calls for a proactive shift toward more systematic and integrated designs. On the attack side, research should pivot toward multi-objective balanced attack strategies; on the defense side, emphasis should be placed on intelligent and adaptive mechanisms, while integrating technologies such as privacy-preserving computing to strengthen security guarantees, thereby facilitating its deployment in real-world federated learning environments.

Keywords federated learning; backdoor attack; backdoor defense; data security

（责任编辑 王微）