◎热点与综述◎

联邦学习中的攻击手段与防御机制研究综述

张世文[1],陈 双[1],梁 伟[1],李仁发[2]
1.湖南科技大学 计算机科学与工程学院,湖南 湘潭 411201
2.湖南大学 信息科学与工程学院,长沙 410082

摘 要:联邦学习的攻防技术是联邦学习系统安全的核心问题。联邦学习的攻防技术能大幅降低联邦学习系统被攻击的风险,明显提升联邦学习系统的安全性。深入了解联邦学习的攻防技术,可以推进联邦学习领域的研究,实现联邦学习的广泛应用。因此,对联邦学习的攻防技术进行研究具有十分重要的意义。简要地介绍了联邦学习的概念、基本工作流程、类型及可能存在的安全问题;介绍联邦学习系统可能遭受到的攻击,梳理了相关研究;从联邦学习系统有无目标性的防御措施出发,将防御措施分为通用性防御措施及针对性防御措施两类,并对其进行了针对性的总结;对联邦学习安全性未来的研究方向进行了梳理与分析,为相关研究者在联邦学习安全性方面的研究工作提供了参考。

关键词:联邦学习;攻击手段;防御措施;隐私保护
文献标志码:A **中图分类号:**TP181;TP309 doi:10.3778/j.issn.1002-8331.2306-0243

Survey on Attack Methods and Defense Mechanisms in Federated Learning

ZHANG Shiwen[1], CHEN Shuang[1], LIANG Wei[1], LI Renfa[2]
1.School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan 411201, China
2.School of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

Abstract: The attack and defense techniques of federated learning are the core issue of federated learning system security. The attack and defense techniques of federated learning can significantly reduce the risk of being attacked and greatly enhance the security of federated learning systems. Deeply understanding the attack and defense techniques of federated learning can advance research in the field and achieve its widespread application of federated learning. Therefore, it is of great significance to study the attack and defense techniques of federated learning. Firstly, this paper briefly introduces the concept, basic workflow, types, and potential existing security issues of federated learning. Subsequently, the paper introduces the attacks that the federated learning system may encounter, and relevant research is summarized during the introduction. Then, starting from whether the federated learning system has targeted defense measures, the defense measures are divided into two categories: universal defense measures and targeted defense measures, and targeted summary are made. Finally, it reviews and analyzes the future research directions for the security of federated learning, providing reference for relevant researchers in their research work on the security of federated learning.

Key words: federated learning; attack method; defense mechanism; privacy protection

大数据和人工智能的迅速发展促进了传统产业的转型。类似深度学习这样以数据驱动的人工智能模型在图像处理、语音识别、自然语言理解等领域取得了巨大成功。海量数据的生成和这些数据的后续处理往往

基金项目:国家自然科学基金(61702180);湖南省自然科学基金面上项目(2022JJ30267);福建省自然科学基金(2022J05106);湖南省教育厅优秀青年项目(21B0493)。

作者简介:张世文(1987—),男,博士,副教授,CCF高级会员,研究方向为云计算安全、隐私保护等,E-mail:544085870@qq.com;陈双(2000—),女,硕士研究生,研究方向为云计算安全、隐私保护等;梁伟(1978—),男,博士,教授,CCF高级会员,研究方向为智能网联车辆安全系统、智能优化和区块链底层技术等;李仁发(1957—),男,博士,教授,CCF杰出会员,研究方向为计算机系统结构、嵌入式计算体系结构、无线网络等。

收稿日期:2023-06-16 **修回日期:**2023-08-18 **文章编号:**1002-8331(2024)05-0001-16

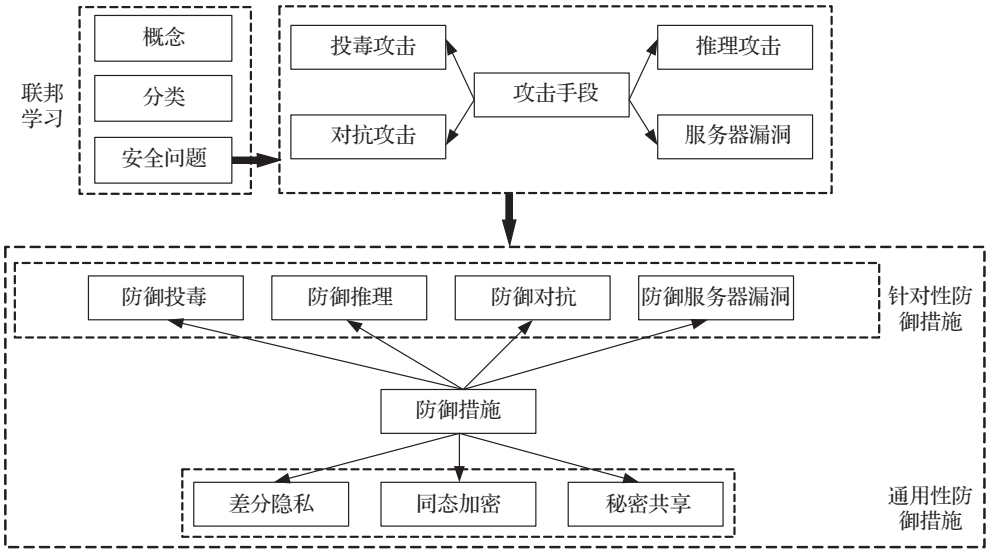

图1　联邦学习攻击手段与防御机制研究

Fig.1　Research of federated learning attack means and defense mechanism

需要一个数据仓库并在仓库内汇总数据。然而,随着数据泄漏事件层出不穷,数据安全性得不到保障,人们开始怀疑集中收集数据是否可靠,数据的隐私性的也得到了更多的关注。为了保证数据安全性,却造成了各个地方的数据难以整合,形成了大量的数据孤岛[1]。

在上述数据无法进行共享的情况下,联邦学习[2]于2016年被首次提出。联邦学习能有效地解决数据孤岛问题,在保证数据安全性的同时达到数据共享目的。联邦学习利用去中心化的数据源进行训练,避免因数据中心化带来的隐私问题,从而能够更好地保护用户隐私。具体来说,联邦学习过程是参与者在客户端本地对其私有数据进行训练,再将训练后得到的模型参数上传到云服务器,最后由云服务器聚合得到整体参数。

然而,联邦学习中仍然存在巨大的安全隐患[3],比如:(1)服务器无法访问参与者的数据及其模型训练过程,导致一些恶意参与者上传错误的更新结果以达到破坏全局更新的目的[4]。例如,攻击者通过训练恶意修改后的训练数据来更新中毒模型,以影响全局模型准确性。(2)攻击者通过推理不同的攻击得到的模型更新的结果可以推理出特定的信息,使得用户的个人信息被泄露[4]。(3)当服务器本身不可信时,服务器与其他的参与者合谋会导致隐私信息泄露。虽然在深度学习领域,保证隐私安全的工作已探索多年,但针对如何构建具有安全和隐私性的联邦学习系统的研究仍处于初级阶段[4-6]。本文根据联邦学习系统可能遭受到攻击的脆弱部分,按传统的分类方式将联邦学习可能遭受到的攻击进行分类。并阐述了针对这部分攻击,联邦学习系统所能够采取防御的手段。

如图1所示,本文首先详细介绍了联邦学习概念及其架构、模型;其次,分类地介绍了联邦学习可能遭受到的攻击,以及针对不同的攻击所能够进行的防御;最后,根据联邦学习的特性与现状,本文对联邦学习的发展方向进行了总结与展望。方便研究人员全面了解联邦学习攻防领域现有知识的基础和发展动态,发现已有研究的不足和未解决的问题;为后续研究联邦学习系统的安全隐患问题提供了新的见解和思路。

1　联邦学习

1.1　联邦学习概念

联邦学习本质上是一种分布式的机器学习技术,其工作过程如图2所示。通常,联邦学习的实施涉及三个步骤。

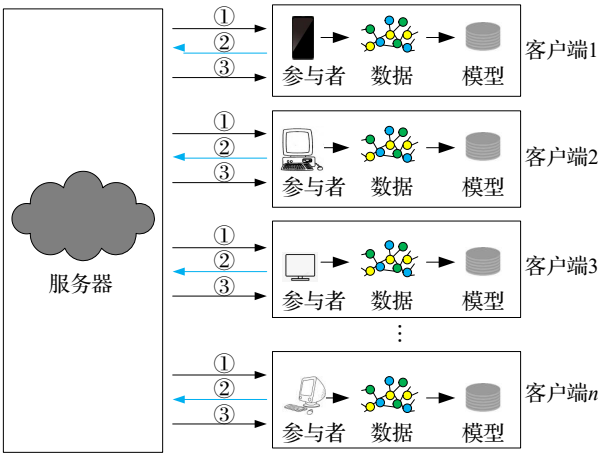

图2　联邦学习流程

Fig.2　Process of federated learning

生成初始模型:在第一个阶段,工作主要是从位于服务器的全局模型开始。开始训练后,服务器生成初始的全局模型,再将这个模型作为机器学习模型广播给联邦学习环境下的客户端($0<i<N$,N为客户端的总数量)。

本地模型更新:在第二个阶段,客户端在本地借助私有的数据集对模型进行训练,再将训练后的模型更新发送给服务器。

模型聚合:在第三个阶段,服务器接收从客户端发送来的更新后的训练模型,并进行聚合生成全局模型。服务器再将聚合的全局模型广播给所有参与训练的客户端。自此,联邦学习进入迭代阶段,每一次迭代,全局模型都会进行更新。此外,服务器在任何阶段都可以在训练过程中添加或删除客户端。

联邦学习由一个服务器,N 个持有私有数据集的客户端组成。联邦学习是一个不断迭代的过程,它重复第二个和第三个阶段直到服务器的全局模型得到一个期望的精度或者全局模型更新达到预定的迭代次数。在训练过程中,联邦学习对模型进行训练而无需传输训练数据或将数据存储到中心服务器。通过这种方式,可以在客户端和服务器之间共享信息,且在训练过程中有效的保护了用户的数据隐私。

相比较于传统的分布式机器学习,联邦学习具有以下特点:(1)数据异构:各个客户端中的数据集的数据非独立同分布,且其中的数据数量不一致;(2)设备异构:各个客户端硬件差异导致计算能力、通信、存储效率等不平衡;(3)客户端数量不定。根据这些特点,联邦学习有着不同的学习类型,被划分为以下3种类型[6-8]:横向联邦学习、纵向联邦学习与迁移联邦学习,它们之间的对比如表1所示。

表1　不同类型联邦学习的对比

Table 1　Comparison of different types of federated learning

类型	特点
横向联邦学习	数据样本重叠较少,但数据特征重叠多
纵向联邦学习	数据特征重叠多,但数据样本重叠较少
迁移联邦学习	数据样本及数据特征都重叠较少

在横向联邦学习中,各个参与方有着不同的数据样本,但其中数据特征重叠较多。横向联邦学习核心的计算方法是联邦平均算法,其包括梯度平均和模型平均两种类型[9]。纵向联邦学习中,各个参与方的数据样本重叠,其中数据的特征有差异。纵向联邦学习已应用于线性回归、提升树、梯度下降等多种模型上。上述两种类型的联邦学习属于较为理想的情况。在现实生活中,大部分参与方所持有的数据,无论是数据样本还是特征都重叠较少,且样本数据集分布不均衡。针对这种情形,迁移联邦学习[10]被提出。在迁移联邦学习中,各个参与方的数据样本及数据的特征都重叠较少。迁移联邦学习结合了联邦学习与迁移学习[11]的优点,使用迁移学习去克服数据样本不重叠与数据特征不重叠的情况。

1.2　联邦学习中的安全问题

联邦学习允许参与者在本地训练数据,而不需要将本地数据传输,从而实现了数据的隔离。且用户数据始终保存在本地,不进行共享,满足了用户隐私保护和数据安全的需求。但这种安全并不是绝对的,联邦学习仍然面对着一些安全性的风险。比如联邦学习没有审核参与方提供的参数模型是否真实;服务器被攻陷时,攻击者可以随时发布恶意模型影响参与方的本地训练;恶意的参与方可以从共享的参数中推理出其他参与方的敏感信息;恶意的参与方可以通过上传恶意的模型破坏聚合后的全局模型等。联邦学习系统可能遭受到的攻击有:(1)投毒攻击(poisoning attack ,PA):即攻击者通过破坏数据样本以达到攻击目的的一种攻击方式,如图3中的 ① 所示;(2)对抗攻击(adversarial attack,AA):即攻击者通过影响模型更新以达到攻击目的的一种攻击方式,如图3中的 ② 所示;(3)推理攻击(reasoning extraction attack,REA),即攻击者通过对监听、窃取等方式获取的信息进行推理以得到某些隐私信息的手段,如图3中的 ③ 所示;(4)服务器漏洞(server vulnerabilities,SV),即服务器本身是恶意攻击者或极易受到攻击的情况,如图3中的 ④ 所示。具体的攻击方式分类如表2所示。其中,推理攻击针对的是隐私,分为模型提取、模型逆向,前者主要通过推理窃取模型的信息,后者主要通过推理获得训练数据集的信息;投毒攻击和对抗攻击针对的是安全,前者主要在训练阶段投放恶意数据或恶意篡改数据从而导致模型的分类准确率降低,后者主要在预测阶段制造对抗样本来使模型分类出错[12]。

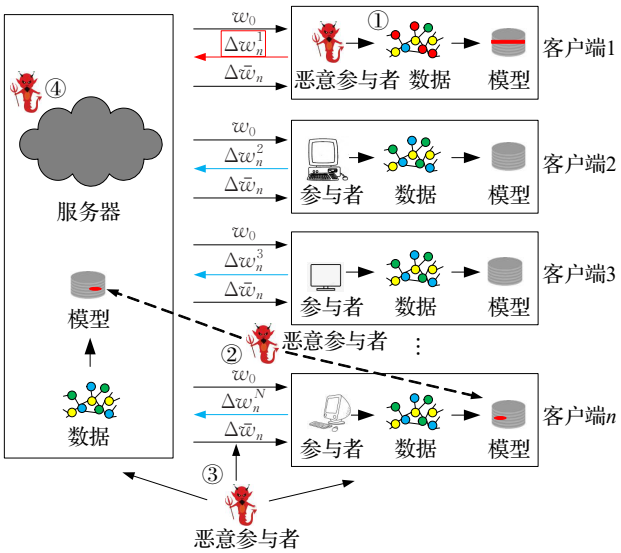

图3　联邦学习可能遭受到的攻击示意图

Fig.3　Schematic diagram of possible attacks on federated learning

针对这些攻击,联邦学习也提出相应的防御措施。我们根据不同的攻击目标将防御措施分为通用性和针对性的防御措施,具体如表3所示。通用性防御措施主要包含差分隐私、同态加密、秘密共享等,而针对性防御措施主要针对联邦学习可能遭受到的不同类型的攻击,包括防御投毒攻击、防御对抗攻击、防御推理攻击以及防御服务器漏洞。

表2　联邦学习中的攻击类型

Table 2　Attack types in federated learning

攻击类型	攻击方式	攻击原理
投毒攻击	数据投毒 模型投毒	投毒攻击主要是指在训练过程中，攻击者在数据集添加恶意数据或篡改其中数据，以达到操纵学习模型预测的攻击目的
对抗攻击	对抗攻击 生成对抗网络	对抗攻击主要是指攻击者通过影响被攻击客户端的本地模型的更新来影响全局模型的更新，导致模型输出错误结果
推理攻击	成员推理 属性推理 模型提取 模型逆向	推理攻击是指攻击者通过某些攻击手段来获取模型的某些信息（如数据集、更新的参数等），来推理获取目标信息
服务器漏洞	恶意服务器攻击 女巫攻击	服务器漏洞是指服务器本身是恶意的，或者服务器缺少完整的防御措施导致其容易受到攻击者攻击

表3　联邦学习中的防御措施

Table 3　Defense measures in federated learning

	防御方法	防御原理
通用防御措施	差分隐私	采用特定的随机算法对数据添加噪声，将数据模糊化，通过牺牲数据的准确性来得到更高的隐私安全
	同态加密	对明文进行加密，且密文运算的结果与明文运算的结果一致，保护模型参数的安全
	秘密共享	将秘密按一定的方式进行拆分，分给不同的参与者，只有若干个参与者一同协作才能恢复秘密消息，保证秘密的机密性
针对性防御措施	防御投毒攻击	在训练之前检验数据的真实性与可靠性，确保数据的安全性及完整性；检测错误的模型参数确保模型训练结果的正确性
	防御对抗攻击	进行对抗训练增强模型的鲁棒性
	防御推理攻击	使用一定的手段使得即使攻击者窃取了一定的信息也无法推理出具体的信息
	防御服务器漏洞	确保服务器是诚实可靠且无法被攻击来确保信息的机密性与完整性

2　联邦学习中的攻击手段

2.1　投毒攻击

联邦学习的一个关键特征是，它允许相互不信任的参与方（例如竞争公司）之间合作训练模型。这使得联邦学习极其容易受到投毒攻击的威胁。例如，一部分联邦学习的参与方被对手拥有或者控制之后在联邦学习训练过程中进行恶意行为，破坏联邦训练的全局模型。投毒攻击主要是指攻击者通过在训练或再训练过程中，篡改数据或往参与者的数据集中添加恶意数据，以破坏训练数据集的分布来改变模型在特定输入上的行为，达到操纵学习模型预测的攻击目的。其通常使得全局模型难以收敛或将良性模型收敛为错误模型。例如注入有毒的训练数据样本，改变样本的标签，删除训练数据集中的一些原有样本等。在联邦学习中，每一个参与者都可以平等地访问训练数据，在这种时候，恶意数据被对手或恶意客户端添加到全局联邦学习模型中，具体的投毒过程如图4所示。每个参与者将更新的参数发送到中央服务器，然后在每一轮得到一个经过充分训练的联邦学习模型。在这个过程，有毒的数据集会影响局部模型，进而间接地影响全局模型，最终使全局模型偏离，降低了模型的精度。一般来说，投毒攻击根据攻击者的投毒目标分为数据投毒攻击[13-18]和模型投毒攻击[15,19-21]。

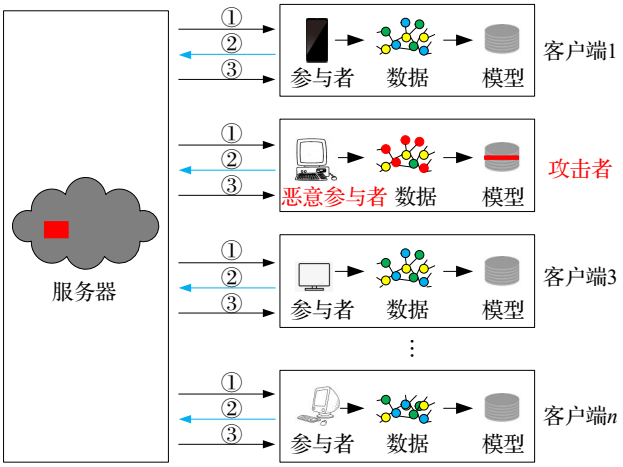

图4　数据中毒

Fig.4　Data poisoning

2.1.1　数据投毒

数据投毒攻击指攻击者恶意篡改数据或者向数据训练集中添加有毒数据来污染训练数据集，影响模型的训练过程，最终导致模型被破坏，降低了模型的准确性。联邦学习中的数据投毒攻击在客户端处操作数据集，通常通过添加噪声或翻转标签[13]来实现。在添加噪声中，攻击者在每个类上加上特制的噪声，使得模型在学习样本本身的特征时将加入的噪声一起学进去，而不需要访问其他数据集，从而达到数据投毒的目的。在标签翻转攻击中，攻击者改变恶意客户端的数据，使某一类的每个标签都切换到目标标签。

数据投毒这种攻击通常由拥有数据的所有者实施，但任何联邦学习的参与者都可以进行数据投毒攻击。根据攻击者是否更改标签，数据中毒可被分为干净标签[22]和脏标签[23]中毒。前者是一种针对性攻击，不修改数据标签，只添加部分恶意数据。后者是指攻击者通过恶意篡改标签来进行攻击，攻击者将篡改的脏数据与干净数据混为一体，集中训练即可进行中毒攻击。

最早研究中毒发作的是Rubinstein等人[14]，他们研究了一系列中毒策略的影响，并详细分析了在同时改变

攻击者可用的信息量和中毒发生的时间范围的情况下,对手如何破坏学习过程。对于研究的三个中毒方案,还展示了攻击者如何通过仅添加适量的中毒数据来大幅增加成功逃避检测的机会。继而,Muoz-Gonzalez 等人[15]将中毒攻击的定义扩展到多类问题。其基于反梯度优化的思想对中毒样本进行优化,提出了一种新的中毒算法,即通过反向传播计算感兴趣的梯度,同时反转学习过程以大幅降低攻击复杂性。该中毒算法仅要求学习算法在训练期间以平滑的方式更新其参数(例如,通过梯度下降),以正确地向后跟踪这些变化。由此,该算法可以应用于更广泛的一类学习算法,但其没有对深度网络的中毒攻击进行广泛的评估来彻底评估它们对中毒的安全性。同样使用优化思想去分析中毒问题的 Sun 等人[16],通过将联邦多学习模型上的最优投毒攻击策略作为一个通用的双层优化问题去进行制定,即使用一个双层优化框架来计算联邦学习的中毒攻击,尝试从数据中毒的角度探索联邦机器学习的漏洞。相同的,Tolpegin 等人[17]和 Li 等人[18]尝试从数据中毒角度探究联邦学习系统的脆弱性。Tolpegin 等人[17]研究了针对联邦学习系统有针对性的数据中毒攻击。在这种攻击中,参与者的一个恶意子集通过发送来自错误标记数据的模型更新来毒害全局模型。其对恶意参与者的能力做了最低限度的假设:每个参与者只能在他们的设备上操作原始的训练数据,这允许非专业的恶意参与者在不了解模型类型、参数和联邦学习过程的情况下实现中毒。实验结果证明这种攻击对传统的集中式机器学习模型是有效的。针对联邦学习系统中数据可靠性方面的漏洞,Li 等人[18]提出了一种基于强化学习的中毒方法,专门用于对未标记数据的预测模型进行投毒。这是第一篇考虑对未标记预测模型的中毒攻击的研究。通过实验证明,这种算法不仅可以成功地对未标记数据的预测模型投毒,而且可以利用积累的经验不断加快投毒速度,进而在短时间内可以成功地对未标记数据的全局预测模型进行投毒。

2.1.2 模型投毒

模型投毒攻击指攻击者直接改变目标模型的参数,使全局模型偏离正常模型,导致模型出现错误或模型性能下降。模型投毒攻击的目的是能够任意操纵模型更新。由于模型的参数会在云服务器和客户端之间重复传输,因此模型投毒攻击可能由其中的任何一方进行。

一开始,Bhagoji 等人[23]探索了联邦学习设置如何引起一种新的威胁,即模型中毒;研究了一些深度神经网络的攻击策略,包括有针对性的模型中毒等。研究者还提出了两个关键的隐形概念来检测恶意更新,通过将这两个关键的隐形概念包含在对抗目标中绕过它们来执行隐形模型中毒,并使用交替最小化策略来改进攻击隐身性,交替优化隐身性和对抗目标。最后成功证明拜占庭弹性聚合策略对这些攻击并不健壮,但这篇文章没有考虑这类攻击的稳健性。在此之后,Zhou 等人[19]对联邦学习中的模型投毒威胁进行了系统的研究,并提出了一种新的基于优化的模型投毒攻击。通过在神经网络的冗余空间中注入对抗神经元来提高攻击的持久性。由于这些冗余的神经元与联邦学习的主要任务相关性较小,所提出的模型投毒攻击不会降低主任务在共享全局模型上的性能,能够避免被中央服务器检测和拒绝异常模型,成功实现了在多客户端模型中实施中毒攻击时保持隐身性与持久性。Hossain 等人[20]与 Zhou 等人[19]同样关注到这类攻击的持久性、有效性与隐蔽性。Hossain 等人[20]分析了在联邦学习设置下的对抗学习过程,并表明可以利用差分噪声进行隐形且持久的模型投毒攻击。更具体地说,这篇文献为联邦学习模型开发了一种利用差分隐私的隐形模型中毒攻击。该攻击通过将错误数据隐藏在 DP 噪声中来欺骗传统的异常检测机制,实现攻击的隐蔽性,降低全局联邦模型的整体精度。作者使用两个流行数据集的分类和回归任务的实证分析证明了所提出攻击的有效性。Cao 等人[21]提出了第一个基于假客户端的模型投毒攻击。具体来说,攻击者往联邦学习系统添加假客户端,假客户端在训练期间向云服务器发送精心制作的假本地模型更新,并在将其发送到云服务器之前将其扩展以扩大其影响,从而使学习到的全局模型对于许多不加区分的测试输入具有低准确性。

表 4 分别从威胁模型、结果、具体应用等几个方面对有关联邦学习的投毒攻击的相关研究进行分类概述。

2.2 对抗攻击

对抗攻击主要是指攻击者通过影响被攻击客户端的本地模型的更新继而影响全局模型的更新,导致模型输出错误结果。其将对抗样例提交到训练好的模型中,从而使模型预测错误。对抗样本(adversarial examples,AEs)是在原来正常的样本上添加了轻微的扰动,可以导致分类模型分类错误。对抗样本的另外一个特点是即使造成了模型分类错误,还是可以进行正确分类。

2.2.1 对抗攻击

对抗攻击利用对抗样本使模型预测错误,也称之为逃避攻击(evasion attack,EA)。对抗攻击是通过在原始样本中添加扰动而产生的。它们混淆了训练有素的模型,但在人类看来它们很正常,这保证了攻击的有效性。对抗攻击可以应用于许多领域,其中应用最广泛的是图像分类。通过添加小的扰动,可以生成对抗的图像,这些图像对人类而言很难区分,但是能造成模型的分类错误。

Szegedy 等人[24]于 2014 年提出对抗攻击。2019 年,张思思等人[25]介绍了什么是对抗样本、对抗样本的概念、出现的原因、攻击方式以及一些关键技术问题。同年,Ling 等人[26]开发了一个统一的评测平台:DeepSEC。

表4　一些投毒攻击研究的对比

Table 4　Comparison of some poisoning attack studies

文献	描述	威胁模型				结果	具体应用
		攻击者角色	攻击对象	攻击者能力	攻击者目标		
[15]	提出一种基于反梯度优化思想的中毒算法	外部	数据	本地训练数据、特征算法、学习算法	造成特定的错误分类	显著损害分类器的性能	适用于大型神经网络和深度学习架构
[16]	提出一种计算最优攻击策略的有效算法	部分客户端	数据	本地训练数据、特征算法、学习算法	降低一系列目标节点的性能	显著地破坏实际应用的性能	适用于横向联邦学习
[17]	研究针对联邦学习系统的有针对性的中毒攻击	部分客户端	数据	本地训练数据	操纵学习参数致使全局模型错误分类	分类准确性和召回率大幅下降	适用于深度学习架构
[18]	提出了一种基于强化学习的投毒算法	部分客户端	数据	本地训练数据	毒害全局模型影响其他参与者的预测	短时间内成功地对未标记数据的全局预测模型下毒	应用于智能交通系统
[19]	提出了一种新的基于优化的模型投毒攻击	部分客户端	模型	本地训练数据	在共享全局模型的任意指定点上造成有针对性的错误分类,同时保持测试数据集的预测精度	具有较高攻击成功率且具有足够隐身性	适用于深度学习架构
[20]	提出了一种基于强化学习的防御策略	外部	模型	差分隐私机制和隐私预算	保持隐身性的同时降低全局精度	当隐私损失非常低和攻击者容忍度高时,模型变得不可用并启动拒绝服务	针对非目标模型
[21]	提出了基于假客户端的模型投毒攻击	虚假客户端	模型	全局模型	降低全局模型的测试精度	显著降低全局模型的测试精度	应用于经典防御和规范裁剪

DeepSEC结合了对抗学习中16种攻击方法、10种攻击效用指标、13种防御方法及5种防御效用指标,旨在评估各种攻击和防御的有效性。

考虑到不同模型之间的差异,Papernot等人[27]首次揭示了机器学习领域中对抗样本可转移的强烈现象。介绍了支持向量机和决策树的对抗样本制作技术;研究了机器学习领域的对抗性样本可迁移性,发现样本不管是在使用相同机器学习技术训练的模型之间,还是在使用不同技术训练的模型之间,抑或是在使用集体决策的集合之间都迁移得很好。

2.2.2　生成对抗网络

生成式对抗网络(generative adversarial network,GAN)结构是由生成器和判别器组成的。训练过程中,两者互相博弈学习产生一个相当好的输出。GAN通过将生成式深度神经网络与判别式深度神经网络相比较,生成一个似乎是来自训练集的样本,当判别模型无法确定样本是来自GAN还是来自训练集时,说明生成式学习是成功的,两者之间相互影响。

通过生成对抗网络,Zhang等人[28]提出一种基于GAN的端到端攻击算法,称为生成模型反演攻击。它可以反演深度神经网络并以高保真度合成私有训练数据。其利用一部分可以通用的公共信息,通过生成对抗网络学习分配先验,并使用它来指导反演过程。对于高度预测的模型来说,漏洞是不可避免的,因为这些模型能够在特征和标签之间建立强相关性。大量实验表明,所提出的攻击将从最先进的人脸识别分类器重建人脸图像的识别精度提高了约75%。Ren等人[29]提出了同样是一种基于GAN的攻击模型-生成回归神经网络(generative regression neural network,GRNN)。作者将攻击描述为一个回归问题,并通过最小化梯度之间的距离来优化生成模型的两个分支。仅通过提出的生成回归神经网络就可以轻松地从共享梯度中完全恢复基于图像的隐私数据。通过几个图像分类任务评估该攻击方法,结果表明,提出的生成回归神经网络较目前的方法具有更好的稳定性、更强的鲁棒性和更高的精度。孔锐等人[30]将攻击算法与GAN相结合,提出一种基于GAN的对抗攻击防御模型。其利用对抗攻击算法生成训练样本的同时在模型训练期间加入条件约束来稳定模型,再利用分类器对生成样本分类来指导GAN的训练,继而通过需要防御的攻击算法来生成对抗样本以完成判别器的训练,最终得到可以抵御多种对抗攻击的分类器。

如表5,分别从威胁模型、结果、具体应用等几个方面,对有关联邦学习的对抗攻击的相关研究进行分类概述。

2.3　推理攻击

推理攻击也被称作探索攻击(入侵攻击)[31],具体如图5所示,是指攻击者通过某些攻击方法得到模型的信息(如数据集、中间参数或预测结果等),然后根据这些信息来推理获取目标信息,如给定用户的某条记录和某个属性是否属于该模型的训练集。虽然在联邦学习设置中,参与方上传模型的梯度信息,将私有数据一直保存在用户本地,但梯度的交换也可能导致隐私泄露[32]。

表5 一些对抗攻击研究的对比

Table 5 Comparison of some studies on adversarial attacks

文献	描述	威胁模型				结果	具体应用
		攻击者角色	攻击对象	攻击者能力	攻击者目标		
[28]	提出了一种生成模型反演攻击	外部	模型	全局模型、辅助知识(例:只包含不敏感信息的损坏图像)	预测特定的标签	提高了攻击的准确性	适用于深度神经网络
[29]	提出一种新的数据泄漏攻击方法	中央服务器	模型	全局模型的神经网络结构和参数	模型的梯度	在攻击成功率、恢复数据的保真度和标签推理的准确性方面都具有优势	数字识别、图像分类和人脸识别等典型计算机视觉任务

在联邦学习框架中,攻击者既可以对本地模型进行攻击,也可以对全局模型进行攻击,通过推理攻击可以在一定程度上得到有用的信息。通常情况下,推理攻击只会影响目标模型,使其输出错误的结果,而不会破坏模型。

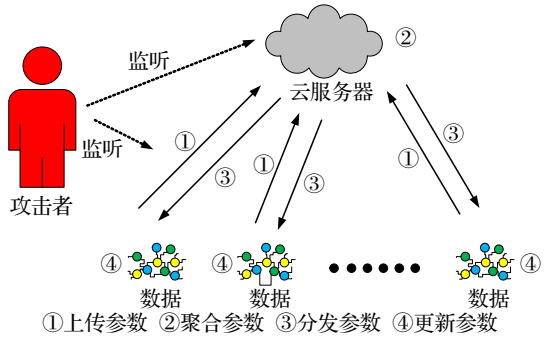

图5 推理攻击

Fig.5 Inference attack

2.3.1 成员推理

判断具体的数据集是否已被用于训练,称之为成员推理攻击(membership inference attack,MIA)。成员推理攻击是指攻击者通过对被攻击模型的应用程序编程接口(application programming interface,API)进行访问,获取大量数据从而模仿目标模型构建出一个新的模型。攻击者不需要对数据、模型参数等进行了解,只需要获得预测分类的置信度,就可以建立一个攻击模型。攻击者利用拥有的信息和权限,将数据输入目标模型,再将得到的结果以及数据集的标签输入攻击模型,就可以判断该记录是否存在于目标模型的数据集。

成员推理攻击最早由Shokri等人[33]提出,其目的是根据训练后的模型对某一样本是否属于对应的训练集进行判断,这可能会导致用户的隐私信息泄露。成员推理攻击的目的是确定攻击对象是否被用来训练模型[34]。Melis等人[35]提出并评估了几种针对协作学习的推理攻击。这些攻击使恶意参与者不仅可以推断成员资格,即其他参与者的训练数据中是否存在确切的数据点,而且还可以推断训练数据子集的特征属性,而这些属性与联邦模型旨在捕获的属性无关。

Nasr等人[36]设计了白盒推理攻击来对深度学习模型进行全面的隐私分析,通过充分训练模型的参数以及训练过程中模型的参数更新来衡量隐私泄漏。另外,他们利用随机梯度下降算法(用于训练深度神经网络的算法)的隐私漏洞,设计了针对白盒设置的新算法,提出一种主动攻击方法:服务器或恶意方主动增加模型在目标数据的梯度。如果目标数据为训练集成员,正常参与方会在后续迭代中明显下降模型损失函数在目标数据的梯度,成员推断模型可以检测到这种变化,从而提高推断攻击的成功率。文献[37]和[38]都是利用GAN和分类模型实现成员推断攻击。其中,Chen等人[37]提出了一种新的联邦学习中的用户级推理攻击机制。从恶意参与者的角度出发,利用白盒访问模型对联邦学习中主动和有针对性的成员推理攻击进行了深入分析。该机制使用生成对抗网络进行数据增强,训练的模型以目标数据为输入,若输出的标签与某个参与方事先声明的标签一致,则认为目标数据为该参与方训练集的成员,从而实现针对特定参与方的成员推断攻击。而Zhang等人[38]通过生成对抗网络从随机噪声中生成新的数据样本。生成的图像用于查询目标联邦学习模型以获取标签,再训练分类模型学习真实标签周围的预测值分布来区分目标模型的成员数据和非成员数据。他们的攻击模型是以目标数据的预测值和标签为输入的。

2.3.2 属性推理

判断其他参与者所用的数据中是否包含某项属性,称为属性推理攻击(property inference attack,PIA)。属性推断攻击是攻击者推断参与方的训练数据的一些敏感隐私属性,其包括模型任务相关属性推断和无关属性推断。

(1)相关属性推断:模型任务相关属性是描述训练数据中每类数据的关键特征,通过推断相关属性可以重构每类标签的训练数据,因此这种攻击也可称为数据重构攻击。通过重构出来的数据并不是真正的训练数据。目前实现数据重构攻击的技术思路主要包括两种:利用生成对抗网络重构数据和将攻击转化为最优化问题求解。

文献[4]和[5]都是利用生成对抗网络实施数据重构攻击。其中,Hitaj等人[5]通过在恶意方部署生成对抗网络来重构其他参与方特定标签的代表数据。然而,这种主动攻击会降低全局模型的准确性,可能被检测到异常行为并进行排除。Wang等人[4]则提出服务器可以利用

生成对抗网络重构特定参与方的训练数据。他们通过在服务器侧部署多任务生成对抗网络学习目标参与方的数据分布。在客户端不是恶意,而服务器是恶意的假设条件下,Song等人[39]在Wang等人[4]工作的基础上进一步扩展,提出了一种预先链接性攻击,通过关联客户端代表来重新识别匿名模型更新。

文献[40]是将攻击转化为最优化问题进行求解的。Zhu等人[40]证明了从公开共享的梯度中获得私人训练数据是可能的,其将数据重构攻击转化成最优化问题进行求解。他们利用模型梯度泄露训练数据信息的原理,推论出如果重构数据可以使全局模型产生和参与方梯度相近的梯度信息,则重构数据也和参与方的训练数据相似。

(2)无关属性推断:任务无关属性是指训练数据中对模型任务不起作用的特征信息,理论上模型不应该泄露这类隐私,这纯粹是模型训练过程的产物,因此无关属性推断也称为无意识的特征泄露。任务无关属性不易察觉且难以检测,且可能带来严重的隐私风险,因此引起了部分学者的重视。这类攻击没有明确的指向性,具体的攻击目标因人而异。

Melis等人[35]的攻击目标是推断其他参与方的训练数据中是否拥有攻击者关心的属性。作者为参与方实施属性推断攻击提出被动和主动两种模式:在被动攻击中,他们首先计算全局模型在辅助数据集上的梯度,并根据辅助数据是否具有目标属性贴上相应的标签,随后用梯度和标签训练一个二分类器,最终以参与方的模型更新为输入进行分类,推断参与方的训练数据是否具有目标属性。而在主动攻击中,一个主动的对手可以使用多任务学习来欺骗联合模型,使其学习对他感兴趣的特征进行更好的内部分离,从而提取更多的信息。Shen等人[41]提出了一种新的属性推断攻击,利用区块链辅助联邦学习中的意外属性泄漏进行智能边缘计算。具体来说,这个主动攻击从参与者的模型更新中学习属性泄漏,并识别一组具有特定属性的参与者。作者希望在保证主任务性能的前提下,推断训练数据具有目标属性的参与方集合。由于攻击者基于全局模型和辅助数据集生成元训练数据来训练攻击模型,当所需的迭代次数较大时,训练攻击模型的时间成本较高。

表6分别从威胁模型、结果、具体应用等几个方面对有关联邦学习的推理攻击的相关研究进行分类概述。

2.3.3　模型提取

模型提取攻击(model extraction attack,MEA)是指持续地向目标发送数据,并根据其响应信息推断出模型的参数,进而生成相似的模型。当攻击者构建的模型与原模型预测性能相近时,原模型拥有方数据泄露的可能性较大。且攻击者可以利用生成的模型生成对抗样本,对原模型也有较大威胁。模型提取攻击针对已经训练好的模型,其目的是窃取模型参数及非法获取模型。

表6　一些推理攻击研究的对比

Table 6　Comparison of some studies on inference attacks

	文献	描述	威胁模型				结果	具体应用
			攻击者角色	攻击对象	攻击者能力	攻击者目标		
属性推理	[35]	提出并评估了几种针对协作学习的推理攻击	部分参与方	数据	辅助数据	推断训练输入子集的真实属性	随着参与者数量的增加,攻击性能显著下降	适用于深度学习架构
	[36]	设计针对深度学习算法的新型白盒成员推理攻击	服务器或部分参与方	数据	完整模型	确定目标数据在目标模型的训练集中的隶属度	很好的泛化模型也很容易受到这种白盒成员推理攻击	针对深度神经网络
	[37]	提出一种联邦学习中的用户级推理攻击机制	部分参与方	数据	本地数据、本地模型	获取有关目标受害者数据集的间接信息	成功地在用户级侵犯受害者的隐私	针对单标签与多标签情况
	[38]	提出一种可以导致联邦学习严重隐私泄露的隶属度推断攻击方法	部分参与方	数据	本地数据	找到模型预测在真实标签周围的分布,并将其应用于所有数据记录	具有98%的攻击准确率	适用于深度学习架构
成员推理	[4]	提出一个将GAN与多任务鉴别器结合的框架	中心服务器	数据	来自客户端的更新(本地模型)	重构目标客户端的私有数据	成功准确地恢复受害者的训练数据	适用于深度学习架构
	[39]	提出了一个包含GAN和多任务鉴别器的框架	中心服务器	数据	来自客户端的更新(本地模型)	重构目标客户端的私有数据	在匿名环境下,所提出的链接性攻击成功率超过99%	适用于深度学习架构
	[41]	提出了一种新的属性推断攻击	服务器	数据	全局模型	推断具有攻击者感兴趣的目标属性的参与者的子集	在不影响联邦学习主任务的前提下,攻击是有效和高效的	结合区块链

Tramer等人[42]首次提出窃取机器学习分类器参数的攻击,介绍了一种通过预测API提取模型的方法。他们通过发送大量的查询建立了模型方程,并得到了相应的预测结果。在此之后,Wang等人[43]为机器学习提供了第一个关于超参数窃取的攻击,证明了各种机器学习容易受到超参数窃取攻击。通过实证评估,这个攻击可以准确地估计我们所研究的所有机器学习算法的超参数,且结合模型参数窃取攻击的情况下,这个攻击在模型参数未知的情况下也能准确估计超参数。文献[44]进一步开展了超参数窃取和架构提取等工作,在黑盒攻击条件下成功推断出神经网络的隐藏模型结构及其优化过程。

2.3.4 模型逆向

在早期的认识中,训练数据集和训练模型之间只有一个信息流,即从数据集到模型。事实上,许多研究表明还存在一个逆向信息流,即从模型信息中恢复数据集信息,这称之为模型逆向攻击(model inversion attack,MIA)。模型逆向攻击是指攻击者根据模型的输入特征,构造对应的输出特征,从而达到篡改模型参数或者篡改模型预测结果的目的。

与模型提取攻击关注模型的隐私信息不同,模型逆向攻击关注数据集。Fredrikson等人[45]开发了一类新的模型反演攻击,可用于从机器学习服务上托管的决策树推断敏感特征,或从面部识别模型中提取训练对象的图像。该攻击利用了与预测一起显示的置信度值。这个新攻击适用于各种环境。Ateniese等人[46]证明了攻击机器学习分类器并从中推断出有意义的信息是可能的。作者构建了一个新的元分类器,并训练它来攻击其他分类器以获得关于它们的训练集的有意义的信息。元分类器可以成功地检测和分类这些变化,并推断出有价值的信息。

2.4 服务器漏洞

联邦学习框架中,服务器的任务是将参与方上传的更新参数进行安全聚合,然后将更新后的参数广播给参与训练的参与方,以此循环训练出一个全局模型[47]。在迭代过程中,每个用户模型的更新信息都需要发送到服务器,服务器可以通过分析更新信息来推理出用户的隐私数据信息。这表明整个系统的中心是服务器,当服务器受损或者其本身是恶意的,将有可能破坏全局模型,造成巨大损失。

服务器漏洞是指服务器本身是恶意的,或者服务器缺少完整的防御措施导致其容易受到攻击者攻击。在目前的联邦学习架构中,参与方在每轮迭代开始时都会使用聚合服务器下发的全局模型覆盖本地模型,而不会检验全局模型的正确性。因此恶意服务器可以跳过聚合过程直接下发恶意模型,在参与方的本地模型植入后门,带来严重威胁。因为恶意服务器的攻击方法明显,且服务器的安全防护措施较为完善、攻击成本高,所以目前相关的研究较少。

2.4.1 恶意服务器攻击

在对机器学习模型进行训练的过程中,服务器能轻松地提取用户数据或操纵全局模型,以利用共享计算能力来构建恶意任务[47]。这给联邦学习带来了很大的安全隐患,它使得攻击者能够通过服务器直接访问全局模型,从而扩大了攻击者的攻击范围。在联邦学习的训练过程,服务器能够控制每一个参与方在什么时候对模型进行访问与操作,因此,当服务器是恶意的,它可以设计新的方案去度量模型的平均情形或最差情形下的攻击敏感性[48],从而设计出最低成本的攻击方案。

考虑服务器有可能是恶意的,Wang等人[4]首次尝试通过来自恶意服务器的攻击来探索针对联合学习的用户级隐私漏洞,其提出了一个将生成对抗网络与多任务鉴别器相结合的框架。该框架可以同时识别输入样本的类别和客户身份,对客户端身份的区分使生成器能够恢复用户指定的私有数据。此外,服务器所处网络环境的安全性也很重要。当服务器工作在较为危险的网络环境中时,受到攻击的概率会大大增加[49]。因此,强大而安全的服务器是必要的。

2.4.2 女巫攻击

在联邦学习中,联邦学习的参与方与服务器是彼此相互信任的,双方都应该保持公平、诚实的态度。女巫攻击(sybil attack,SA)是指利用少数节点可能含有多个虚假身份,从而利用其去控制或影响大部分节点。女巫攻击的攻击手段主要分为:直接和间接通信、伪造身份、盗用身份、同时和非同时攻击等。在联邦学习架构中,攻击者可以通过控制服务器来伪造或控制大量的参与方发动攻击,得到其所需的信息。该攻击对联邦学习协议的安全造成了威胁。同时,一些联邦学习系统为了保护用户的隐私,会将用户的信息打乱,这将使分辨诚实用户和恶意用户变得更加困难,抵御女巫攻击的难度大大上升。

由于联邦学习只需要参与者训练过程的信息,对参与者本身及其数据的情况没有任何限制,联邦学习现有的防御机制难以抵御女巫攻击。

3 联邦学习中的防御机制

针对上述列出的攻击手段,本文搜集了一些常用的防御措施,涵盖了针对大部分攻击手段可以通用的防御措施以及具体针对上述提出的各类不同的攻击手段可以实施的防御措施。使用通用性的防御措施可以同时防御大部分攻击,但其针对性不强,对于针对某一类攻击方式的防御效果较差。而使用某一类针对性的防御措施往往对针对的攻击手段防御效果较好,却忽略了遭受到其他攻击时的影响。

3.1 通用性防御机制

针对第2章所提出的联邦学习可能遭受到的攻击，本节阐述了所采取的通用性防御措施。

3.1.1 差分隐私

差分隐私最早由Dwork等人[50]提出，用来克服不断涌现的隐私攻击以及当前隐私保护机制存在的不足。当参与者将信息发送给服务器时，信息极有可能会被泄露。为了防止这种情况发生，在发送更新信息之前给信息加入差分隐私，这可以有效防止攻击者逆向推理出用户的数据。差分隐私技术通过对数据添加噪声来实现数据模糊化，从而减少了敏感数据的泄露，使得即使攻击者通过攻击手段得到了部分信息也无法推理原始数据。在参与方上传更新信息之前，对更新信息进行差分隐私，则无需考虑服务器是否可信。差分隐私是一种通过引入随机性来确保隐私的方法，其通过牺牲一定的准确度达到更高的隐私安全。差分隐私根据不同的信任假设和噪声源，被分为三类：本地化差分隐私[51](local differential privacy，LDP)、分布式差分隐私(distributed differential privacy，DDP)[52-54]、中心化差分隐私(centralized differential privacy，CDP)[51]。若融合了两种或以上的差分隐私方法则称为混合差分隐私(hybrid differential privacy，HDP)[55]。当前，差分隐私技术的研究重点是在于如何保障隐私的同时，尽可能地保留原始数据中的有用信息，从而实现对隐私的有效保护。

McMahan等人[2]提出联邦学习，在2017年向联邦学习环境添加用户级别的差分隐私[56]，在训练模型的过程中不过分牺牲模型的质量而又保护个人的数据隐私。Choudhury等人[57]成功地将差分隐私引入联邦学习，保护了模型免受潜在的隐私攻击，为联邦学习框架提供更高级别的隐私。Geyer等人[58]提出了一种客户端差分隐私保护联邦优化的算法，致力于在隐私保护和模型性能之间取得平衡。Bhowmick等人[59]设计了新的最优局部差分私有机制，提出了大规模局部私有模型训练的适用方法，适用于联邦学习系统。Abadi等人[60]在联邦学习的场景下，通过差分隐私来保证模型不会透露参与方是否参与了训练，维持了客户级的差异隐私。

3.1.2 同态加密

同态加密(homomorphic encryption，HE)是指将原始数据经过同态加密以后，对得到的密文进行特定的代数运算，然后将计算结果再进行同态解密后得到的明文与直接在明文上进行相同运算得到的结果相同。根据对密文上进行操作的种类和次数，同态加密可以被分为三大类：半同态加密[61](partially homomorphic encryption，PHE)、部分同态加密[62](somewhat homomorphic encryption，SWHE)以及全同态加密[63-65](fully homomorphic encryption，FHE)。半同态加密仅支持一种同态运算，但运算支持执行无限次。部分同态加密支持多种同态运算，但是运算的次数有限。全同态加密支持无限次运算及所有种类同态运算。全同态加密理论上支持对密文进行任意计算，但其运算量过大，存储开销大，效率较低。相较全同态加密而言，部分同态加密更加高效，因此在具体实施中部分同态加密常常被优先使用。

随着硬件的发展，实现同态加密与其他安全方法的结合成为可能。Madi等人[66]提出了第一个联邦学习框架，该框架在不向聚合服务器公开最终模型的情况下，通过结合同态加密与可验证计算技术，可以安全地抵御来自聚合服务器的机密性和完整性威胁。Phong等人[67]提出了一个新的深度学习系统来保护诚实但好奇的云服务器上的梯度，其中许多学习参与者在所有的组合数据集上执行基于神经网络的深度学习，并结合加法同态来保证隐私安全，从而使得参与者的本地数据不会透露给中央服务器。

然而在目前，同态加密还无法直接用于联邦学习，例如，在协作式场景下，哪些用户应该拥有密钥还没有得到解决。对此，Reyzin等人[68]对自定义阈值加密的可能性和局限性进行了系统的研究，并介绍了可扩展多方计算的密钥应用。一旦公钥被分发，除了中央服务器之外的所有各方只发送和接收短消息，其大小与参与者的数量无关。文献[69]通过使用稀疏向量技术与加密技术的组合来实现在不可信设备间执行求和计算。两者通过分布式密钥去解决上述问题。

3.1.3 秘密共享

秘密共享(secret sharing，SS)是现代密码学领域的一个重要分支，是保证信息安全和数据保密的重要手段，也是多方安全计算和联邦学习等领域的一个基础应用技术。其主要用于保护用户的隐私信息，包括用户的身份、地址等，防止信息丢失、被破坏或被篡改。秘密共享的机制主要由秘密的分发者、参与者、分配算法、恢复算法等构成。秘密共享通过使用合适的方式和将秘密进行拆分，并将拆分后的秘密分享给不同的参与者，使得只有多于一定数量的参与者一同合作才可以计算或恢复秘密，当少于规定的数量时无法得到秘密。目前实现秘密共享一共有三种技术方案：一是基于插值多项式的秘密共享：Shamir方案[70]；二是基于超平面几何的秘密共享：Blakley方案[71]；三是Asmuth等人[72]提出的基于中国剩余定理的秘密共享。

随着神经网络的发展，谷歌提出的联邦学习系统面临着移动设备通常无法与其他移动设备建立直接的通信通道，移动设备本机也无法验证其他移动设备的挑战。Bonawitz等人[73]考虑在联邦学习模型中训练一个深度神经网络，在移动设备上对用户持有的训练数据使用分布式梯度下降，使用安全聚合来保护每个用户的模型梯度的隐私。其通过在协议中添加一个含有一个秘

密共享循环的初始回合,用来保证恶意服务器无法提取梯度信息。为了进一步加强数据隐私和安全,秘密共享方案也在更新。Han等人[74]提出了一种支持深度神经网络隐私保护的可验证联邦训练方案,提出了可验证秘密共享(verifiable secret sharing,VSS)。可验证秘密共享是基于以往的密码共享升级而来,其实现了用户的隐私保护,并且验证服务器返回结果的正确性。除了进行更新之外,秘密共享还被尝试与其他安全方法进行结合。Fereidooni等人[75]在联邦学习框架下高效地结合了可验证秘密共享与全同态加密,既发挥了联邦学习分布式的优点,又进一步加强了整个系统的安全性。

表7总结了上述相关针对联邦学习可能遭受到的攻击所采取的通用性防御措施的研究。

表7 通用性联邦学习可能遭受攻击的防御措施

Table 7 Defense measures against potential attacks in universal federated learning

防御类型	代表方案	优点	缺点
差分隐私	文献[56-60]	实现数据资源的最大利用	使用场景有限;对模型可用性及准确性造成一定程度的影响
同态加密	文献[66-69]	没有进行数据扰动;理论上支持任意聚合规则的计算	算力要求大
秘密共享	文献[73-75]	安全性较高;灵活性较高;计算开销小	实用性较差,实现复杂度较高;通信开销大

3.2 针对性防御机制

如图6所示,针对前文所提出的联邦学习可能遭受到的攻击,本节阐述了防御的思路,为后续研究人员在建设相关系统,抵抗有关攻击时提供思路。

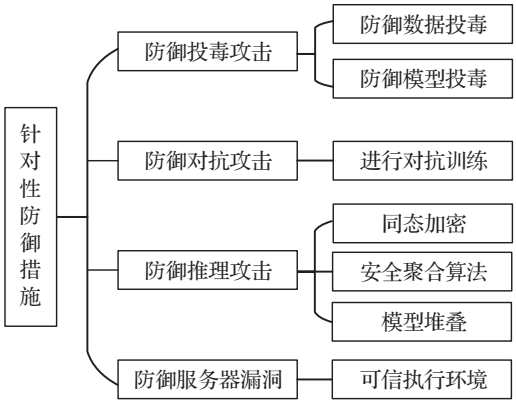

图6 针对性防御措施

Fig.6 Targeted defense measures

3.2.1 防御投毒攻击

投毒攻击在联邦学习中是一种常见的攻击,通过数据投毒和模型投毒这两种方式进行。

首先考虑数据投毒攻击。系统遭受这种攻击的根本原因是没有考虑到用户的数据可能是错误的,甚至可能遭遇攻击者破坏。因此,针对这一攻击的防御措施大部分是在模型进行训练前仔细检查数据来源,确保数据的安全性以及完整性。一种常用的防御措施是在用户进行训练之前,对数据进行检测,确定其是否安全。Baracaldo等人[76]使用了一种检测和过滤有毒数据方法:使用相关训练集中数据点的起源和转换的上下文信息来识别有毒数据,从而使在线和定期重新训练的机器学习应用程序能够在潜在的对抗环境中使用数据源。这个方法是第一个将来源信息作为过滤算法的一部分来检测致病攻击的方法。还提出了该方法的两种变体:一种适用于部分可信的数据集;另一种适用于完全不可信的数据集。除上述提出的通过提前检测数据来防御中毒攻击外,还可以在训练之前转换数据使攻击者无法确定转换机制来对中毒攻击进行防御;在模型进行训练之前使用身份认证机制确保参与者是可信地来对中毒攻击进行防御等。

其次考虑模型投毒攻击。在防御模型投毒攻击时,主要是通过对模型参数进行检测来达到防御的目的。可以通过限制每个用户贡献的数据量、对参与方进行奖惩机制或根据数量使用衰减权重实现等方式实现对模型攻击的防御。比如,在每一轮更新之后,对参与者上传的参数进行检测,当某个参与方提交的参数与其他参与方的参与差异较大时,则认为该参与方这轮上传的参数是异常的,在后续进行参数聚合时将不会将其考虑在内。Andreina等人[77]设计了一个针对模型中毒的解决方案,这个方案使用一种现成的方法在每个参与方本地比较更新模型的分类性能与前一个模型的分类性能,丢弃出现意外行为的更新。这个方案通过确保与安全更新聚合的完全兼容,有效地保证了客户数据的隐私。

3.2.2 防御对抗攻击

对抗攻击是攻击者通过操纵输入数据来欺骗模型给出假阳性结果。应对该类攻击最流行的策略是进行对抗训练,即将真实数据集和对抗样本结合进行训练,再对训练后的模型进行分析与改进。这种类型的训练可以提高模型的鲁棒性和稳定性,适用于多种监督问题[78];另一种应对对抗性攻击的技术是数据增强技术[79]。在这种情况下,原始数据被随机改变,以提高模型的泛化能力,这可以用来对抗图像裁剪、图像缩放等攻击。Liang等人[80]在不需要任何攻击技术的先验知识这一基础之上,提出了一种检测对抗性图像示例的简单方法可以直接部署到未经修改的深度神经网络模型中,能有效地检测对抗性样本。Shah等人[81]研究在联邦学习环境中使用对抗性训练的可行性,提出了一种在联邦环境中执行对抗性训练的新算法用于提高联邦对抗性训练的性能。

3.2.3 防御推理攻击

通常,使用推理方式成功实现攻击比实现其他类型的攻击难度更大。成功实现推理攻击不仅要求攻击者

能够成功提取到用户级别以上的部分,还要求攻击者具备一定的知识能对提取到的信息进行有效地分析和推理。推理攻击分为成员推理攻击与属性推理攻击。针对推理攻击,最常用的防御手段是进行同态加密。当系统使用了同态加密之后,即使攻击者成功从系统中获取了信息,也只会获得密文,而没有密钥的攻击者无法将其解密为明文,即攻击也无法成功。此外,当攻击者想要针对全局模型进行攻击,由于攻击者一般无法得知系统内部的聚合规则,则使用安全聚合算法也可抵御该攻击。除了使用上述提到的以及其他常用的如差分隐私和秘密共享等进行防御之外,还有一些专门针对该攻击的防御措施,例如模型堆叠[69]。文献[82]提出了一个安全聚合框架,采用多组循环策略来实现高效的模型聚合,并利用附加秘密共享和新颖的编码技术来注入聚合冗余,以便在保证用户隐私的同时处理用户退出,大大提高了实现推理攻击的难度。

3.2.4　防御服务器漏洞

当服务器出现漏洞时,极易受到攻击者攻击。针对此类攻击,可以使用可信执行环境(trusted execution environment,TEE)来进行防御。TEE采用硬件隔离的手段来保护服务器,可以有效地防止病毒感染以及其他恶意攻击。其提供更安全的网络环境,有效地提高服务器的安全性。TEE实现了独立执行环境和安全存储,保证了信息的机密性和完整性。通过利用TEE,Chen等人[83]将参与方的本地模型训练和服务器的聚合过程都加载到TEE的飞地中执行,且参与方和聚合服务器间的模型交互也是经由飞地间的安全通道完成。一方面保证本地训练过程的完整性,避免攻击者跳过或干扰本地训练,上传伪造的模型更新;另一方面防止恶意服务器无视参与方上传的更新,直接下发恶意模型。

表8总结了上述相关针对联邦学习可能遭受到的攻击所采取针对性防御措施的研究。

表8　针对联邦学习可能遭受攻击的防御措施

Table 8　Defense measures against potential attacks on federated learning

防御类型	防御手段	方案	作用效果
防御投毒攻击	训练前检测数据	文献[76]	确保数据的安全性以及完整性
	检测模型参数	文献[77]	确保模型完整性,保护用户隐私
防御对抗攻击	对抗训练	文献[80-81]	有效提高模型的鲁棒性和稳定性
防御推理攻击	模型聚合	文献[82]	附加其他的安全手段,保证了用户隐私不被泄露
防御推理攻击	可信执行环境	文献[83]	使用硬件隔离手段,有效地防止病毒感染以及其他恶意攻击

4　未来研究方向

通过以上研究发现,无论是从提高攻击效果,还是从增强联邦学习安全性,都有很大的研究空间。联邦学习未来可能的研究方向具体可从以下几个方面入手。

(1)通用攻击方法:联邦学习由于数据的样本量和其特性分为不同的联邦学习,而无论对于其中的哪一种联邦学习,都应该致力于研究一种通用的攻击方法。而目前的攻击方法使用条件相对苛刻,难以满足需求。因此,如何设计一个高效且通用的攻击方法是一个重要的研究方向。

(2)相互适应的防御措施:联邦学习最初被提出时,假设了攻击者无法从不可逆的模型信息中推断隐私信息。但最近的研究发现,许多的研究者通过不同的手段恢复了原始数据。虽然部分攻击可被多方安全计算成功防御,然而会导致服务器只能收集到密文而无法对数据进行分析,这样会导致联邦学习系统难以抵御其他类型攻击。因此,如何合理地结合多种防御措施是一个重要的研究方向。

(3)数据质量问题:由于原始数据存储在各参与者处,服务器无法直接访问数据,因此很难确保数据的完整性、数据标签的正确性等。并且,联邦学习的参与方众多,其数据异构性大,相互之间的异构程度不明确。则当数据量较小时,往往会导致罕见样本的出现,这就使得模型的训练和验证变得更加困难,模型更容易受到攻击。因此,如何实现对恶意用户数据的验证来保证数据的质量是一个重要研究方向。

随着联邦学习技术的不断发展,针对数据隐私的攻击手段将会越来越丰富,因此需要进一步探索和研究如何加强系统安全性,更加有效地保护数据隐私。

5　结束语

在人工智能技术不断发展和普及的过程中,人们在享受到技术带来便利的同时,对于隐私保护的要求也在不断地提升。联邦学习应运而生。联邦学习可以有效解决跨设备之间的数据融合问题。然而,联邦学习仍存在大量的安全隐患,比如用户设备的异构性、数据的隐私性等。为此,联邦学习需要在多环节从多角度考虑数据的安全问题。本文首先从联邦学习的工作原理、类型及可能存在的安全问题出发,进行阐述;继而对其中的几种典型的攻击手段和防御措施进行了梳理总结;最后对其未来所面临的问题和研究方向进行了展望。随着隐私保护的重要性不断提高,联邦学习作为新的研究热点逐渐受到了学术界的广泛关注,有关联邦学习的研究有待深入发展,本文工作为相关研究者提供了参考。

参考文献:

[1] 张思思, 高旭光, 滑文强. 基于聚类与人工神经网络的遥感

图像信息提取方法[J]. 电子设计工程, 2020, 28(15): 106-109.
ZHANG S S, GAO X G, HUA W Q. Remote sensing image information extraction method based on clustering and artificial neural network[J]. International Electronic Elements, 2020, 28(15): 106-109.

[2] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proceeding of the 20th International Conference on Artificial Intelligence and Statistics, Ft Lauderdale FL, April 20-22, 2017. USA: JMLR, 2017: 1273-1282.

[3] JAGIELSKI M, OPREA A, BIGGIO B, et al. Manipulating machine learning: poisoning attacks and countermeasures for regression learning[C]//Proceeding of the 39th IEEE Symposium on Security and Privacy, San Francisco, May 21-23, 2018. NJ: IEEE, 2018: 19-35.

[4] WANG Z B, SONG M K, ZHANG Z F, et al. Beyond inferring class representatives: user-level privacy leakage from federated learning[C]//Proceeding of the 38th Annual IEEE International Conference on Computer Communications, Paris, April 29-May 2, 2019. NJ: IEEE, 2019: 2512-2520.

[5] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep models under the GAN: information leakage from collaborative deep learning[C]//Proceeding of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, TX, October, 2017. New York: ACM, 2017: 603-618.

[6] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[7] LI T, SAHU A K, TALWALKARA, et al. Federated learning: challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[8] KAIROZ P, MCMAHAN H B, AVENT B, at al. Advances and open problems in federated learning[J]. Foundations and Trends® in Machine Learning, 2021, 14(1/2): 1-210.

[9] MCMAHAN H B, MOORE E, RAMAGE D, et al. Federated learning of deep networks using model averaging[J]. arXiv: 1602.05629,2016.

[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.
LIU Y X, CHEN H, LIU Y H, et al. Privacy-preserving techniques in federated learning[J]. Journal of Software, 2022, 33(3): 1057-1092.

[11] PAN S J, YANG Q. A Survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22 (10): 1345-1359.

[12] 何英哲, 胡兴波, 何锦雯, 等. 机器学习系统的隐私和安全问题综述[J]. 计算机研究与发展, 2019, 56(10): 2049-2070.
HE Y Z, HU X B, HE J W, et al. Privacy and security issues in machine learning systems: a survey[J]. Journal of Computer Research and Development, 2019, 56(10): 2049-2070.

[13] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[J]. arXiv:1206.6389,2012.

[14] RUBINSTEIN B, NELSON B, LING H, et al. ANTIDOTE: understanding and defending against poisoning of anomaly detectors[C]//Proceeding of the 9th ACM SIGCOMM Conference on Internet measurement, Chicago, Nov 4-6, 2009. New York: ACM, 2009: 1-14.

[15] MUOZ-GONZALEZ L, BIGGIO B, DEMONTIS A, et al. Towards poisoning of deep learning algorithms with back-gradient optimization[J]. ACM, 2017, 17: 27-38.

[16] SUN G, CONG Y, DONG J, et al. Data poisoning attacks on federated machine learning[J]. IEEE Internet of Things Journal, 2022, 9(13): 11365-11375.

[17] TOLPEGIN V, TRUEX S, GURSOY M E, et al. Data poisoning attacks against federated learning systems[C]//Proceeding of the ESORICS 2020, UK, September 14-18, 2020. Berlin: Springer, 2020: 480-501.

[18] LI Z, WU X K, JIANG C J. Efficient poisoning attacks and defenses for unlabeled data in DDoS prediction of intelligent transportation systems[J]. Security and Safety, 2022, 1: 145-165.

[19] ZHOU X C, XU M, WU Y M, et al. Deep model poisoning attack on federated learning[J]. Future Internet, 2021, 13(3): 73.

[20] HOSSAIN M T, ISLAM S, BADSHA B et al. DeSMP: differential privacy-exploited stealthy model poisoning attacks in federated learning[C]//Proceeding of the 17th International Conference on Mobility, Sensing and Networking (MSN), Exeter, Dec 13-15, 2021. NJ: IEEE, 2021: 167-174.

[21] CAO X, GONG N Z. MPAF: model poisoning attacks to federated learning based on fake clients[C]//Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, LA, USA, June 19-20, 2022. NJ: IEEE, 2022.

[22] ZHAO S, MA X, ZHENG X, et al. Clean-label backdoor attacks on video recognition models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, June 13-19, 2020. NJ: IEEE, 2020: 14431-14440.

[23] BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens[C]// Proceedings of the 36th International Conference on Machine Learning, Long Beach, June 9-15, 2019. Germany: Statistics, 2019: 1467-5463.

[24] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceeding of the 2nd International Conference on Learning Representations, Banff,

Canada, Apr 14-16, 2014.

[25] 张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报, 2019, 42(8): 1886-1904.
ZHANG S S, ZUO X, LIU J W. the problem of the adversarial examples in deep learning[J]. Chinese Journal of Computers, 2019, 42(8): 1886-1904.

[26] LING X, JI L, ZOU J, et al. DEEPSEC: a uniform platform for security analysis of deep learning model[C]//Proceeding of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, May 19-23, 2019. NJ: IEEE, 2019: 673-690.

[27] PAPERNOT N, MCDANIEL P, GOODFELLOW I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples[J]. arXiv:1605.07277, 2016.

[28] ZHANG Y H, JIA R X, PEI H Z, et al. The secret revealer: generative model inversion attacks against deep neural networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 13-19, 2020. NJ: IEEE, 2020: 250-258.

[29] REN H C, DENG J J, XIE X H. GRNN: generative regression neural network—a data leakage attack for federated learning[J]. ACM Transactions on Intelligent Systems and Technology, 2022, 13(4): 1-24.

[30] 孔锐, 蔡佳纯, 黄钢. 基于生成对抗网络的对抗攻击防御模型[J/OL]. 自动化学报(2020-07-23). http://www.aas.net.cn/cn/article/doi/10.16383/j.aas.2020.c200033?viewType=HTML.
KONG R, CAI J C, HUANG G. Defense to adversarial attack with generative adversarial network[J/OL]. Acta Automatica Sinica (2020-07-23). http://www.aas.net.cn/cn/article/doi/10.16383/j.aas.2020.c200033?viewType=HTML.

[31] BARRENO M, NELSON B, SEARS R, et al. CAN machine learning be secure[C]//Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, Taipei, China, March 21-24, 2006. New York: ACM Press, 2006.

[32] LEE H, KIM J, HUSSA IN R, et al. On defensive neural networks against inference attack in federated learning[C]//Proceedings of 2021 IEEE International Conference on Communications, Seoul, Korea, June 14-23, 2021. NJ: IEEE, 2021: 1-6.

[33] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models[C]//Proceedings of IEEE Symposium on Security and Privacy (SP), San Jose, May 22-26, 2017. NJ: IEEE, 2017: 3-18.

[34] AONO Y, HAYASHI T, PHONG L T, et al. Scalable and secure logistic regression via homomorphic encryption[C]//Proceedings of the 6th ACM Conference on Data and Application Security and Privacy, New Orleans Louisiana USA, March 9-11. New York: ACM, 2016: 142-144.

[35] MELIS L, SONG C Z, CRISTOFARO E D, et al. Exploiting unintended feature leakage in collaborative learning[C]//Proceeding of the IEEE Symposium on Security and Privacy, San Francisco, May 19-23, 2019. NJ: IEEE, 2019: 691-706.

[36] 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]//Proceeding of the IEEE Symposium on Security and Privacy, San Francisco, May 19-23, 2019. NJ: IEEE, 2019.

[37] CHEN J L, ZHANG J L, ZHAO Y C, et al. Beyond model-level membership privacy leakage: an adversarial approach in federated learning[C]//Proceedings of the 29th International Conference on Computer Communications and Networks, Honolulu, Aug 3-6, 2020. NJ: IEEE, 2020: 1-9.

[38] ZHANG J W, ZHANG J L, CHEN J J, et al. GAN enhanced membership inference: A passive local attack in federated learning[C]//Proceedings of the IEEE International Conference on Communications, Dublin, June 7-11, 2020. NJ: IEEE, 2020: 1-6.

[39] SONG M K, WANG Z B, ZHANG Z F, et al. Analyzing user-level privacy attack against federated learning[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(10): 2430-2444.

[40] ZHU L G, LIU Z J, HAN S. Deep leakage from gradients [C]//Proceeding of the 33rd International Conference on Neural Information Processing Systems, NY, December 8-14, 2019. Berlin: Springer, 2019: 14747-14756.

[41] SHEN M, WANG H, ZHANG B, et al. Exploiting unintended property leakage in blockchain-assisted federated learning for intelligent edge computing[J]. IEEE Internet of Things Journal, 2021, 8(4): 2265-2275.

[42] TRAMER F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction APIs[C]//Proceedings of the 25th USENIX Conference on Security Symposium, Austin, May 31, 2019. USA: USENIX Association, 2016: 601-618.

[43] WANG B, GONG N Z. Stealing hyperparameters in machine learning[C]//Proceedings of 2018 IEEE Symposium on Security and Privacy (SP), San Francisco, May 20-24, 2018. NJ: IEEE, 2018: 36-52.

[44] OH S J, AUGUSTIN M, SCHIELE B, et al. Towards reverse engineering black-box neural networks[J]. arXiv:1711.01768, 2017.

[45] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic counter measures[C]//Proceedings of the 22nd ACM SIG-

SAC Conference on Computer and Communications Security, Denver, Colorado, October 12-16, 2015. New York: Association for Computing Machinery, 2015: 1322-1333.

[46] ATENIESE G, MANCINI L V, SPOGNARDI A, et al. Hacking smart machines with smarter ones: how to extract meaningful data from machine learning classifiers[J]. International Journal of Security and Networks, 2015, 10(3): 137-150.

[47] LYU L J, YU H, YANG Q. Threats to federated learning: a survey[J]. arXiv:2003.02133,2020.

[48] BOUACIDA N, MOHAPATRA P. Vulnerabilities in federated learning[J]. IEEE Access, 2021, 9: 63229-63249.

[49] MOTHUKURI V, PARIZIE M, POURIYEH S, et al. A survey on security and privacy of federated learning[J]. Future Generation Computer Systems, 2021, 115: 619-640.

[50] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//Lecture Notes in Computer Science, NY, March 4-7, 2006. Berlin: Springer, 2006: 265-284.

[51] BASSILY R, NISSIM K, STEMMER U, et al. Practical locally private heavy hitters[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, December 4-9, 2017. Red Hook, NY, USA: Curran Associates Inc, 2017: 2285-2293.

[52] BITTAU A, ERLINGSSON L, MANIATIS P, et al. Prochlo: strong privacy for analytics in the crowd[J]. Journal of Machine Learning Research, 2020, 21(1): 1532-4435.

[53] ERLINGSSON L, FELDMAN V, MIRONOV I, et al. Amplification by shuffling: from local to central differential privacy via anonymity[C]//Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, San Diego, Jan 6-9, 2019. USA: Society for Industrial and Applied Mathematics, 2019: 2468-2479.

[54] CHEU A, SMITH A, ULLMAN J, et al. Distributed differential privacy via shuffling[J]. arXiv:1808.01394,2018.

[55] AVENT B, KOROLOVA A, ZEBER D, et al. BLENDER: enabling local search with a hybrid differential privacy model[J]. Journal of Privacy and Confidentiality, 2017, 9(2): 2575-8527.

[56] MCMAHAN HB, RAMAGE D, TALWAR K. Learning differentially private recurrent language models[J]. arXiv:1710.06963,2017.

[57] CHOUDHURY O, GKOULALAS-DIVANIS A, SALONIDIS T, et al. Differential privacy-enabled federated learning for sensitive health data[J]. arXiv:1910.02578,2019.

[58] GEYER R C, KLEIN T, NABI M. Differentially private federated learning: a client level perspective[J]. arXiv:1712.07557, 2017.

[59] BHOWMICK A, DUCHI J, FREUDIGER J, et al. Protection against reconstruction and its applications in private federated learning[J]. arXiv:1812.00984,2018.

[60] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGASC Conference on Computer and Communications Security, NY, October 24-28, 2016, NY: ACM, 2016: 308-318.

[61] RIVEST R L, ADLEMAN L, DERTOUZOS M L. On data banks and privacy homomorphisms[J]. Foundations of Secure Computation, 1978, 4(11): 169-180.

[62] BONEH D, GOH E J, NISSIM K. Evaluating 2-DNF formulas on ciphertexts[C]//Proceedings of 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Aarhus, Denmark, May 22-26, 2005. Berlin: Springer, 2005: 325-341.

[63] FAN J, VERCAUTEREN F. Somewhat practical fully homomorphic encryption[J]. IACR Cryptology Eprint Archive, 2012, 2012: 144.

[64] BRAKERSKI Z, GENTRY C, VAIKUNTANATHAN V. (Leveled) fully homomorphic encryption without bootstrapping[J]. ACM Transactions on Computation Theory (TOCT), 2014, 6(3): 1-36.

[65] CHEON J H, KIM A, KIM M, et al. Homomorphic encryption for arithmetic of approximate numbers[C]//Proceedings of 23rd International Conference on the Theory and Application of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017. Berlin: Springer, 2017: 409-437.

[66] MADI A, STAN O, MAYOUE A, et al. A secure federated learning framework using homomorphic encryption and verifiable computing[C]//Proceedings of 2021 Reconciling Data Analytics, Automation, Privacy, and Security: A Big Data Challenge, Hamilton, Ontario, May 18-19, 2021. NJ: IEEE, 2020: 1-8.

[67] PHONG L T, AONO Y, HAYASHI T, et al. Privacy-preserving deep learning via additively homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(5): 1333-1345.

[68] REYZIN L, SMITH A, YAKOUBOV S. Turning HATE into LOVE: compact homomorphic Ad Hoc threshold encryption for scalable MPC[C]//Proceedings of 5th International Symposium on Cyber Security Cryptography and Machine Learning, Be'er Sheva, July 8-9, 2021. Berlin: Springer, 2021: 361-378.

[69] ROTH E, NOBLE D, FALK BH, et al. Honeycrisp: largescale differentially private aggregation without a trusted core [C]//Proceedings of the 27th ACM Symposium on Operating Systems Principles, Huntsville, Ontario, October 27-30, 2019. NY: ACM, 2019: 196-210.

[70] SHAMIR A. How to share a secret[J]. ACM, 1979, 22(11): 612-613.

[71] BLAKLEY G R. Safeguarding cryptographic keys[C]//Proceedings of International Workshop on Managing Requirements Knowledge (MARK), New York, June 4-7, 1979. NJ: IEEE, 1979: 313-318.

[72] ASMUTH C, BLOOM J. A modular approach to key safeguarding[J]. IEEE Transactions on Information Theory, 1983, 29(2): 208-210.

[73] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for federated learning on user-held data [J]. arXiv:1611.04482,2016.

[74] HAN G, ZHANG T T, ZHANG Y H, et al. Verifiable and privacy preserving federated learning without fully trusted centers[J]. Journal of Ambient Intelligence and Humanized Computing, 2022, 13(3): 1431-1441.

[75] FEREIDOONI H, MARCHAL S, MIETTINEN M, et al. SAFELearn: secure aggregation for private federated learning[C]//Proceedings of 2021 IEEE Security and Privacy Workshops (SPW), San Francisco, May 27, 2021. NJ: IEEE, 2021: 56-62.

[76] BARACALDO N, CHEN B, LUDWIG H, et al. Mitigating poisoning attacks on machine learning models: a data provenance based approach[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, Texas, November 3, 2017. NY: ACM, 2017: 103-110.

[77] ANDREINA S, MARSON G A, MLLERING H, et al. BaFFLe: backdoor detection via feedback-based federated learning[C]//Proceedings of IEEE 41st International Conference on Distributed Computing Systems (ICDCS), DC, July 7-10, 2021. NJ: IEEE, 2021: 52-863.

[78] MIYATA T, MAEDA S, KOYAMA M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE Transactions on Pattern Analysis And Machine Intelligence, 2018, 41(8): 1979-1993.

[79] NAIR A K, RAJ E D, SAHOO J. A robust analysis of adversarial attacks on federated learning environments[J]. Computer Standards & Interfaces, 2023, 86: 103723.

[80] LIANG B, LI H C, SU M Q, et al. Detecting adversarial image examples in deep neural networks with adaptive noise reduction[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18(1): 72-85.

[81] SHAH D, DUBE P, CHAKRABORTY S, et al. Adversarial training in communication constrained federated learning [J]. arXiv:2103.01319,2021.

[82] SO J, GULER B, AVESTIMEHR A S. Turbo-aggregate: breaking the quadratic aggregation barrier in secure federated learning[J]. IEEE Journal on Selected Areas in Information Theory, 2021, 2(1): 479-489.

[83] CHEN Y, LUO F, LI T, et al. A training-integrity privacy-preserving federated learning scheme with trusted execution environment[J]. Information Sciences, 2020, 522: 69-79.