联邦学习系统攻击与防御技术研究综述

高 莹[1),2),3),4)] 陈晓峰[2)] 张一余[2)] 王 玮[2)] 邓煌昊[2)] 段 培[5)] 陈培炫[5)]

[1)] (贵州大学公共大数据国家重点实验室 贵阳 550025)

[2)](北京航空航天大学 网络空间安全学院 北京 100191)

[3)](空天网络安全工业和信息化部重点实验室 北京 100191)

[4)](中关村实验室 北京 100094)

[5)](腾讯公司 深圳 518054)

摘 要 联邦学习作为一种使用分布式训练数据集构建机器学习模型的新兴技术，可有效解决不同数据用户之间因联合建模而导致的本地数据隐私泄露问题，从而被广泛应用于多个领域并得到迅速发展。然而，现有的联邦学习系统已被证实在数据收集阶段、训练阶段和推理阶段都存在潜在威胁，危及数据的隐私性和系统的鲁棒性。本文从安全威胁和隐私威胁两类潜在威胁入手，围绕机密性、完整性和可用性（CIA 三元组）给出了联邦学习场景中安全属性的详细定义，并对联邦学习中各类攻击方式和防御手段进行了系统全面综述。首先，本文对横向、纵向联邦学习过程，以及潜在威胁分别进行了概述，并从对抗性攻击和非对抗性攻击两个角度，分析了投毒攻击、对抗样本攻击和推理攻击等常见攻击的基本概念、实施阶段和现有方案。进一步地，依据不同的攻击方式，将防御手段划分为鲁棒性提升方法和隐私性增强技术两类：鲁棒性提升方法主要防御系统遭受的对抗性攻击，包括有数据消毒、鲁棒性聚合、异常检测、对抗训练、知识蒸馏、剪枝和其他方法等，隐私性增强技术主要防御系统遭受的非对抗性攻击，包括有同态加密、安全多方计算、差分隐私和区块链等。最后，本文给出了联邦学习中鲁棒性和隐私性方面的未来研究方向。

关键词 联邦学习；安全威胁；隐私威胁；鲁棒性提升方法；隐私性增强技术

中图法分类号 TP181

A Survey of Attack and Defense Techniques for Federated Learning Systems

GAO Ying[1), 2), 3),4)] CHEN Xiao-Feng[2)] ZHANG Yi-Yu[2)] WANG Wei[2)] Deng Huang-Hao[2)] DUAN Pei[5)] CHEN Pei-Xuan[5)]

[1)](State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025)

[2)](School of Cyber Science and Technology, Beihang University, Beijing 100191)

[3)](Key Laboratory of Aerospace Network Security, Ministry of Industry and Information Technology, Beijing 100191)

[4)](Zhongguancun Laboratory, Beijing 100094)

[5)](Tencent Inc, Shenzhen, 518054)

Abstract As an emerging technology of building machine learning (ML) model using distributed training data sets, federated learning (FL) can effectively solve the problem of local data privacy disclosure caused by joint modeling between different data owners. Therefore, it is widely used in many fields and has developed rapidly. FL keeps the data of participants local and only uploads model parameters to the server, which effectively protects the privacy of local data. However, the existing FL systems have been proved to have potential threats in the data collection stage, training stage and inference stage, which endanger the privacy of data and the robustness of the system. In the data collection stage and training stage, attackers may poison the training data or

收稿日期：2022-03-20；在线发布日期：2023-01-16.本课题得到北京市自然科学基金 (No. M21033)、国家自然科学基金(No. 61932011, 61972017)、腾讯微信犀牛鸟基金资助。**高 莹(通信作者)**,博士,副教授,计算机学会(CCF)高级会员,主要研究领域为隐私计算、区块链. E-mail: gaoying@buaa.edu.cn. **陈晓峰**,博士研究生,主要研究领域为区块链、联邦学习. **张一余**,硕士研究生,主要研究领域为联邦学习、隐私计算. **王 玮**,硕士研究生,主要研究领域隐私计算. **邓煌昊**,硕士研究生,主要研究领域为联邦学习.**段 培**,硕士,主要研究领域为机器学习、数据挖掘. **陈培炫**,硕士,主要研究领域为分布式计算、机器学习.

the model, thereby endangering the security of the system. In the inference stage, attackers may input samples to add minor malicious perturbations, causing the classifier to incorrectly classify the sample process with a very high probability, which will lead to privacy disclosure. Most of the existing research works describe attack and defense methods in ML, which are not necessarily applicable to FL models, and only focus on a few attack threats and traditional defenses, lacking a detailed and comprehensive overview of the cutting-edge defenses. Starting with two kinds of potential threats: security threat and privacy threat, we give a detailed definition of security attributes in FL scenarios around confidentiality, integrity and availability (CIA triplet), and summarize various attack methods and defense means in FL systematically and comprehensively. Firstly, we summarize the horizontal and vertical federated learning (VFL) process and potential threats respectively, and analyze the basic concepts, implementation stages and existing schemes of common attacks such as poisoning attack, sample attack and inference attack from the perspectives of antagonistic attack and non-antagonistic attack. Adversarial attacks include poisoning attacks, adversarial sample attacks, free-riding attacks, Sybil attacks, and attacks against communication bottlenecks. Non-adversarial attacks include model extraction attacks, inference attacks, and GAN-based attacks. Further, according to different attack methods, defense means are divided into two categories: robustness enhancement methods and privacy enhancing technologies. The robustness enhancement methods mainly defend against antagonistic attacks, including data sanitization, robustness aggregation, anomaly detection, countermeasure training, knowledge distillation, pruning and other methods. The privacy enhancing technology mainly defends the system against non-antagonistic attacks, including homomorphic encryption, secure multi-party computing, differential privacy and blockchain. And the schemes related to robustness enhancement methods and privacy enhancement techniques in FL are sorted out and summarized. Finally, the paper gives future research direction of robustness and privacy in FL: 1) Establish a secure and stable attack detection and evaluation model, endow FL system with self inspection and evaluation capabilities, and provide real-time protection for internal and external environments; 2) Analyze and infer all possible potential attacks and privacy issues, and build a perfect security attack and defense system based on security encryption technology; 3) Study the unique attack and defense in VFL to solve the bottleneck problem of VFL in practical application; 4) Explore the conflict between robustness and privacy in FL to promote large-scale applications.

Key words federated learning; security threats; privacy threats; robustness enhancement method; privacy enhancing technology

1 引 言

人工智能已成为引领新一代产业变革的新兴技术，尤其对应用创新、企业转型及社会发展有着重大影响，已经上升到国家战略层面。作为人工智能核心技术的机器学习却面临着隐私威胁和信任危机等问题[1]，迫使各个用户将数据存储在本地，彼此之间难以流通，形成了“数据孤岛”。数据孤岛问题阻碍了多个用户进行有效的数据合作，导致数据的潜在价值难以发挥。此外，数据孤岛中非同源的数据之间相互关联但又存在较大差异，致使这些数据呈现非独立同分布（Non-Independent and Identically Distributed，NON-IID），带来了新的挑战。联邦学习（Federated Learning，FL）[2-4]作为机器学习技术的新分支，能满足隐私数据不出本地的前提下，在多个用户之间进行高效率的联合建模、模型训练，充分释放数据潜在价值，近年来已被广泛应用于键盘预测[5]、安全检测[6-7]和信号识别[8]等。

虽然联邦学习能一定程度解决本地数据的隐私问题，但在模型参数共享、模型聚合时又会给攻击者带来新的可乘之机，如联邦学习的梯度会泄露用户数据或学习过程的隐私信息[9-11]，攻击者会对训练数据或局部模型进行投毒[12]或在输入样本中加入恶意扰动[13]，从而危害系统的安全性。针对不同目标、不同程度和不同类型的攻击威胁，联邦学习系统往往需要预先制定好相对应的防御策略，以增强系统的鲁棒性和隐私性。

目前，国内外已有许多联邦学习相关的研究，例如，Yin 等[14]面向隐私保护的联邦学习进行了全

面的综述，Abdulrahman 等[15]详细阐述了联邦学习面临的主要技术挑战，但他们都未进一步区分安全威胁和隐私威胁的差异。在联邦学习安全与隐私保护的综述[16-18]中，分别探讨了安全和隐私方面面临的挑战，但在鲁棒性和隐私性防御手段方面没有展开分析与总结。He 等[19]分析了深度学习中安全威胁相关的四种攻击，通过定量和定性分析这些攻击方法的敌手能力和攻击目标，总结出这些方法的优缺点，并讨论了其他的安全弱点和可能的防御措施。但提到的攻击威胁在联邦学习模型中不一定具有同等的攻击效果，防御措施在联邦学习模型中也可能受到限制。Lyu 等[20-21]提供了一种独特的威胁模型分类方法，侧重介绍联邦学习中的安全与隐私问题，强调了隐私保护的重要性。但该篇文章只重点介绍了投毒攻击和推理攻击两种攻击威胁，对防御措施缺乏详细的梳理与分析。2021 年，Mothukuri 等[22]针对联邦学习中的安全和隐私问题，以及相应的防御措施做出了系统性综述。但该篇文章中阐述的防御手段都是较为传统的方法，缺少对前沿创新性工作的介绍，如联邦学习结合同态加密、差分隐私、安全多方计算和区块链等隐私增强技术。对比以上这些综述，本文在文章架构、分析方法和侧重点上都有所不同。本文更详细且全面地梳理了联邦学习中的安全威胁和隐私威胁，系统地对攻击手段与防御手段进行了分类与剖析，侧重分析了最前沿的联邦学习与密码技术相结合的隐私保护方案，并进一步讨论了横向和纵向联邦学习中攻击手段的区别，以及在此基础上为后续研究者提供了具有发展前景的研究方向。

本文的组织结构安排如下。第 2 节对联邦学习和其潜在威胁进行概述；第 3 节详细地介绍了几种常见攻击的分类和研究进展；第 4 节在已有的攻击手段和研究成果基础上，从鲁棒性和隐私性两个角度对提升手段进行了具体分析；第 5 节讨论了联邦学习未来研究发展趋势；最后，在第 6 节总结全文。

2 联邦学习中的潜在威胁

2.1 联邦学习概述

联邦学习是一种以分布式方式训练模型的机器学习技术，其主要思想是确保参与方的数据保留在本地，而将训练的模型进一步上传和聚合到服务器。后续学习过程仅使用模型进行训练，保护了参与方的数据隐私，从而保护了数据安全。

在用户数据集中的训练样本包含多个特征数据，其中选择一个或多个能够将不同训练样本区分开来的特征作为样本的标识符，即样本 ID。在联邦学习场景下，每个数据集的组织和使用形式存在差异，其特征和样本 ID 可能存在差异。联邦学习从不同数据分布方式可分为横向联邦学习（Horizontal Federated Learning，HFL）、纵向联邦学习（Vertical Federated Learning，VFL）和迁移联邦学习（Federated Transfer Learning，FTL）三种类型。

依照传统机器学习过程的划分，联邦学习则可以分为三个阶段：数据收集阶段、训练阶段和推理阶段。联邦学习在这三个阶段都具有新的特点。

(1) 数据收集阶段：指训练模型所需要的数据准备过程。在传统机器学习中需要对每个用户的数据进行集中收集，为模型训练做准备。而在联邦学习中，数据集不会离开本地，具体为本地的数据收集、用户之间数据格式的协商等准备过程。

(2) 模型训练阶段：指利用这些数据集执行机器学习训练算法，挖掘数据的潜在价值，迭代训练一定轮次后直至收敛的过程。在联邦学习中，由于数据集的分布式划分以及隐私性要求，需要使用特定的模型训练算法。

(3) 推理阶段：指把训练好的模型部署在具体的应用场景中，输入真实样本进行预测的过程。在横向联邦学习中这一阶段和传统机器学习没有太大差异，但是在纵向联邦学习场景中，由于每个用户只拥有一部分模型，推理阶段需要用户之间的合作才能完成推理过程。

目前常用的联邦学习开源项目包括 Google 的 TensorFlow[1]、微众银行的 FATE[2]、百度的 PaddleFL[3] 以及 OpenMinded 的 PySyft[4]等。其中，Google 的 TensorFlow 应用最早，他们在数据不离开每个用户本地的情况下训练了一个循环神经网络模型，之后又将联邦学习操作进一步封装，发布了专门为联邦学习开发的框架 TensorFlow Federated（TFF）[5]，并提供了一组高级接口可以方便程序员实现基于联

1 TensorFlow: An end-to-end open source machine learning platform. https://www.tensorflow.org.

2 FATE:工业级联邦学习框架. https://fate.fedai.org/.

3 百度飞桨 release note. https://www.paddlepaddle.org.cn/.

4 OpenMinded.Syft. https://pypi.org/project/syft/.

5 TFF. https://tensorflow.google.cn/federated/federated_learning.

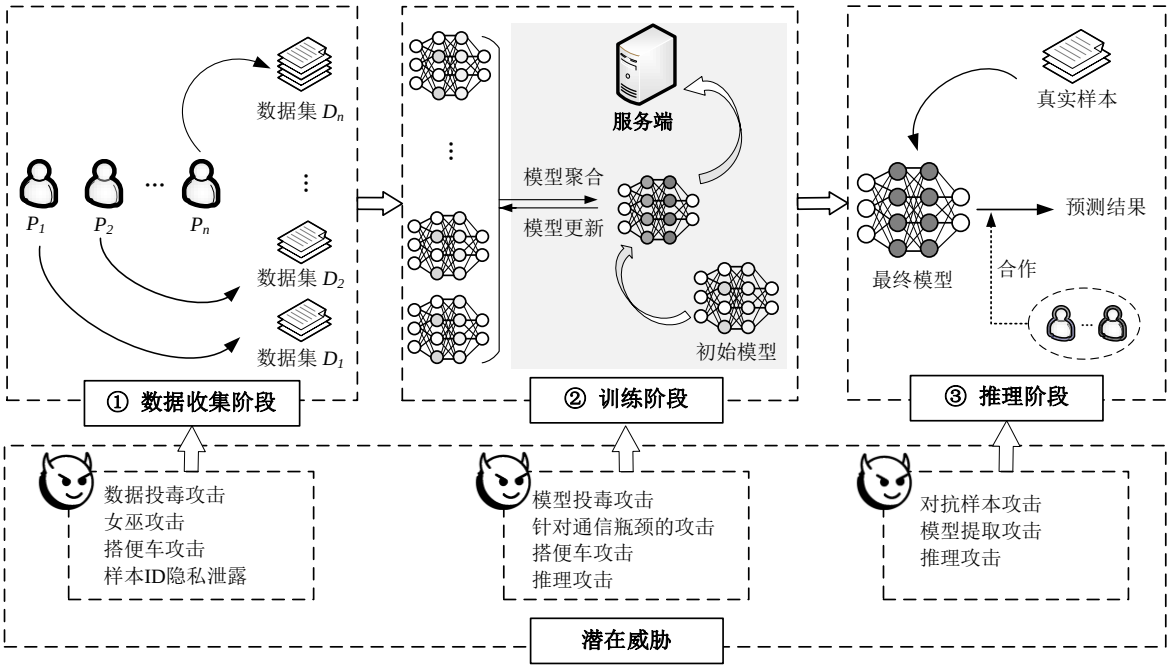

图 1 联邦学习过程的三个阶段及潜在威胁

邦平均的 HFL 算法。微众银行的 FATE 是首个工业级联邦学习框架，使用安全多方计算和同态加密等技术构建底层安全计算协议，可以支持逻辑回归、树模型和深度学习等多种机器学习算法，与 TFF 相比其封装程度更高，可以多方部署后直接开始训练。PaddleFL 是基于百度的深度学习框架 Paddle 开发的联邦学习框架，提供了多种联邦学习策略，支持基于安全多方计算的纵向逻辑回归和神经网络的安全训练与推理，也支持基于安全聚合的 HFL 和经典的基于联邦平均和异步随机梯度下降的 HFL，但不支持树模型。PySyft 是第一个隐私保护深度学习框架，基于 PyTorch 开发，可以方便地实现联邦深度学习并基于安全多方计算和差分隐私提供隐私保护。

2.2 联邦学习的安全属性

本节从机密性（Confidentiality）、完整性（Integrity）和可用性（Availability）安全三元组[23]（简称 CIA 三元组）的角度给出了联邦学习场景中安全属性的概念。

(1) 联邦学习的完整性：依照数据收集和模型训练的不同阶段可以进一步划分为数据集完整性和训练过程完整性。数据集完整性是指联邦学习中用户的数据始终是良性的、未被篡改的[24]。训练过程完整性是指服务器、用户等参与方都严格地按照联邦学习协议执行算法[25]。

(2) 联邦学习的可用性：根据模型训练和推理阶段可以进一步划分为训练可用性和模型可用性。训练可用性是指能够在预计时间内完成模型的训练，其包含两个方面：一是收敛性，指模型能够在经过可接受的训练轮数内达到收敛状态；二是合作公平性，是联邦学习场景中特有的，指用户能够依据自身的贡献获得公平的补偿[26]。模型可用性是指在推理阶段模型部署后的准确性和公平性，其中公平性是指保证训练的模型不会对某些属性存在潜在的歧视性[27]。

(3) 联邦学习的机密性：机密性是指本地数据、全局模型等敏感信息不会泄露给非授权的用户。另外，机密性还保证了用户不会因为网络不稳定、设备问题被动退出工作流后，导致本地梯度的机密性泄露[28]。

2.3 潜在威胁

根据对安全属性的不同影响，本文将联邦学习中存在的潜在威胁划分为两大类，即安全威胁和隐私威胁。安全威胁会破坏联邦学习中的完整性和可用性，对联邦学习造成安全威胁的攻击称为对抗性攻击，其主要目的是干扰联邦学习训练或推理过程，影响联邦学习训练时的收敛速度或推理结果。隐私威胁会破坏联邦学习中的机密性，对联邦学习造成隐私威胁的攻击称为非对抗性攻击，其主要目的是试图从联邦学习各个阶段获取隐私信息或其它好处，但不会破坏模型训练和推理过程。

在联邦学习的不同阶段会受到不同的安全威胁和隐私威胁。在数据收集阶段，受到的安全威胁包括数据投毒攻击（Data Poisoning Attack）、女巫

攻击（Sybil Attack）和搭便车攻击（Free-riding Attacks），隐私威胁包括样本 ID 隐私泄露。在训练阶段，受到的安全威胁包括模型投毒攻击（Model Poisoning Attack）、针对通信瓶颈（Communication Bottlenecks）的攻击和搭便车攻击，隐私威胁包括推理攻击（Inference Attack）。在推理阶段，会受到的安全威胁包括对抗样本攻击，隐私威胁包括模型提取攻击（Model Extraction Attack）和推理攻击。联邦学习过程的三个阶段及潜在威胁如图 1 所示。

3 联邦学习中的攻击手段

3.1 对抗性攻击

3.1.1 投毒攻击

机器学习中的投毒攻击[29-30]是指攻击者通过控制和操纵部分训练数据或模型来破坏学习过程。而联邦学习中每个用户都拥有一个数据集，内部的恶意攻击者可以轻易地对数据集、训练过程和模型进行篡改，实现降低模型性能、插入后门等一系列攻击效果。投毒攻击是联邦学习中应用最广泛、研究最深入的攻击。通常，投毒攻击按照攻击方式的不同可以分为数据投毒和模型投毒，而根据攻击目标的不同可以分为拜占庭攻击（即非定向投毒攻击）和后门攻击（即定向投毒攻击）。

(1) 按攻击方式划分

① 数据投毒：攻击者破坏训练数据集的完整性，通过渗入恶意数据以降低数据集质量或有目的的毒害数据。数据投毒根据对数据集标签的不同操作分为脏标签攻击（Dirty-label Attack）[31]和清洁标签攻击（Clean-label Attack）[32]。脏标签攻击会篡改数据集的标签，如常见的标签翻转攻击[33]，而清洁标签攻击不篡改标签，仅对数据进行处理生成新的样本。联邦学习中由于数据不出本地，只有模型作为信息载体，因此基本不考虑数据投毒的不可感知性，从而数据投毒攻击主要是更加简便有效的脏标签攻击。

② 模型投毒：攻击者破坏训练过程完整性，通过完全控制部分用户的训练阶段，对上传的局部模型进行篡改，实现对全局模型的操纵。常见的攻击手段是通过提升（Boosting）恶意更新来加强攻击效果[34]。为了增强提升的隐蔽性，Bhagoji 等[35]还将提升过程转化为一个基于交替最小化找到合适的提升值优化问题，使有毒更新与正常更新难以区分。此外，还有其他实现更强隐蔽性和更高成功率的模型投毒攻击[36]和针对服务器先进防御聚合机制的隐蔽模型投毒攻击[37]等研究。

虽然数据投毒和模型投毒两种攻击方式都对模型训练产生影响，但单一的数据投毒相较模型投毒表现不佳，是因为数据投毒本质上与模型投毒会同样修改局部模型的更新权重，而后者可以针对联邦学习聚合等特性实施针对性的攻击。

(2) 按攻击目标划分

① 拜占庭攻击：攻击者试图破坏训练可用性和模型可用性，使其无法收敛或无法在主要训练任务中达到最优性能，并且不针对任何特定的用户或数据样本。在联邦学习中，通过发送恶意更新和其他良性更新的线性组合能实现拒绝服务攻击[38]，但此类简单的攻击很容易被检测和过滤。文献[39]则表明更新中轻微的扰动就能够实现投毒攻击的效果，并且规避基于幅度的防御策略。

已知聚合规则的攻击者可以针对性地实施更具破坏性的拜占庭攻击，并且服务器为了吸引用户或满足用户的知情权，其聚合规则常常是透明公开。文献[40]提出了局部模型投毒攻击（Local Model Poisoning Attack），在已知聚合规则的情况下，攻击者将构造恶意更新转化为在聚合规则下全局更新偏移值的优化问题。同样通过优化实现拜占庭攻击，文献[41]使用了最优比例系数 γ 和已优化数据集的扰动向量 ∇^p 对恶意梯度进行微调并在结果中找到近似的最大值，实现更好的攻击效果。但文献[40]的方案在一轮迭代中就能完成优化过程，而文献[41]需要数十次的聚合迭代。

当前，VFL 中拜占庭攻击的相关研究还很少，由于模型被用户分割，数据投毒和模型投毒的隐蔽性更强、危害更大，是一个值得深入研究的方向。

② 后门攻击：又叫木马攻击（Trojan Attack）攻击者试图使模型在某些目标任务上实现特定表现，同时保持模型在主要任务上的良好性能[42]。不同于拜占庭攻击由于会降低主要任务的总体性能易被检测，后门攻击更难被检测，这是因为攻击目的通常是未知的，难以确定检测标准。后门攻击可以通过数据投毒和模型投毒来实现，其在缺乏防御时的表现在很大程度上取决于当前敌手的比例和目标任务的复杂性。此外，相同的后门触发条件可能导致不同标签的样本错误的分类，这不同于后文的对抗样本攻击只对特定修改后的图像进行错误分类，不会影响到其他图像样本。

目前，后门攻击在机器学习中已有广泛的应

用，包括基于深度神经网络（Deep Neural Networks，DNN）的隐蔽后门攻击[43-45]和注入后门的“BadNets”模型[46]等，以上方案都是在数据集上实现的后门攻击。而在联邦学习中，结合模型投毒实现的后门攻击更为常见。Bagdasaryan 等[47]指出单一的数据投毒攻击可能对联邦学习无效，因为恶意模型可能与数量众多的良性模型聚合，极大地降低了攻击效果。他们提出了一种基于模型替换的投毒方法，其依据模型收敛性导致局部模型更新趋于零，利用模型替换在一轮迭代中将全局模型替换为后门模型。此外，大多数后门攻击[48]并没有考虑联邦学习分布式特点，Xie 等[49]则提出了新的分布式后门攻击，即后门在恶意攻击者控制的用户之间被拆分，并将每个模式嵌入敌对客户的训练集中，在模型聚合后又将成为一个完整的后门并插入到模型中，从而提高后门攻击的隐蔽性。

与 HFL 中的后门攻击不同，在典型 VFL 模型中的用户无需提供也不能获得标签信息，也就意味着不具备特征实例所对应的目标标签的知识。Liu 等[50]为了研究 VFL 中的后门攻击，其中假设恶意用户拥有至少一个训练实例对应标签的知识，使用梯度替换的方法，将原本实例的中间梯度替换为投毒实例的梯度，向最终模型中植入后门。

3.1.2 对抗样本攻击

机器学习中的对抗样本攻击是指在推理阶段中，刻意地给输入样本增加轻微的恶意扰动，使得分类器以极高概率对样本进行错误分类，从而导致模型输出错误的预测结果[13,51]。按照攻击者拥有的信息，对抗攻击可以分为白盒攻击和黑盒攻击。在白盒攻击中，攻击者能够获得机器学习算法以及模型参数，并根据这些已知信息去制作对抗样本。在黑盒攻击中，攻击者不知道机器学习的算法和参数信息，通过与系统的交互过程来生成对抗样本。一方面，联邦学习的分布式特性增加了模型参数泄露的可能性，这意味着联邦学习比传统机器学习更容易遭受白盒攻击的威胁。另一方面，与传统机器学习类似，在联邦学习模型部署之后，攻击者也能够通过与系统的交互实施黑盒攻击。

对抗样本攻击在机器学习领域已有广泛研究。GoodFellow 等[13]发现了深度学习输入—输出映射上的不连续性，通过施加难以察觉的扰动，最大化网络的预测误差。Szegedy 等[52]验证说明了几种经典神经网络模型容易受到对抗样本攻击的影响。Akhtar 等[53]对深度学习中的对抗样本攻击及防御进行了全面性的综述。Ilyas 等[54]开发了一种新的黑盒攻击视角，将此类攻击的构造视为梯度估计问题，打破了利用梯度中先验信息发起对抗样本攻击的最优性障碍。

在联邦学习场景下的模型部署与机器学习类似，传统的对抗样本攻击可以拓展到联邦学习中，但目前针对联邦学习的对抗样本攻击研究尚少。Wang 等[33]研究了对抗样本攻击与后门攻击之间的联系，表明模型对后门的鲁棒性在通常情况下意味着对于对抗样本攻击的鲁棒性。Pang 等[55]则针对 VFL 中用户的特征差异，提出了对抗性主导输入攻击（Adversarial Dominating Inputs Attack）。与传统的对抗性样本控制整个特征空间不同，对抗性主导输入攻击仅仅控制部分特征输入，就能主导其他用户的全部输入，实现对特定的输入进行错误分类。同时，对抗性主导输入使得其他用户做出非常少的贡献，从而影响了激励用户贡献的奖励。

3.1.3 搭便车攻击

搭便车攻击是指部分用户不参与协作或者不具备足够的条件，而试图从集体性质的服务和公共产品中获得优势。其主要发生在带有激励的联邦学习中，破坏训练过程完整性和合作公平性。攻击者一般不消耗或只消耗部分的本地数据和计算资源，通过向服务器发送随机更新或与聚合模型相似的更新，伪装成参与联邦学习训练的正常用户，以获得相应激励，同时可能对模型性能造成一定影响。

在联邦学习中，常见的搭便车攻击通过提供随机参数更新实现攻击，而服务器可以使用传统的深度学习异常梯度检测进行防御，如 DAGMM[56]。文献[57]介绍了一个在基于模型平均的联邦学习中进行搭便车攻击的理论框架，证明了在每次迭代中返回全局模型会导致搭便车攻击。通过向构造的参数更新中增加噪音、应用随机梯度下降（Stochastic Gradient Descent，SGD）来最大化与其他用户更新的相似性，实现更加隐蔽的攻击。文献[58]提出了一种新的搭便车攻击，在本地使用小数据集训练模型，伪装成大数据集以获取更多奖励。此外，搭便车攻击者也可能会通过提升方法来增大权重，从而在全局服务器上表现出更大的贡献程度，以获得更大的激励份额。

在 VFL 中的用户也可能不进行完整的训练，而应用未经充分训练的低质量模型，会降低整体模型的性能。此外，VFL 的用户并不公开模型信息，服务器更难区分低质量的模型，故 VFL 中的搭便车攻

击值得进一步研究。

3.1.4　女巫攻击

女巫攻击（Sybil Attack）指在允许成员自由加入和退出的系统中，单个攻击者通过多个合谋的身份加入系统，从而巧妙地分配攻击，以增强隐蔽性和攻击效果[59]。女巫攻击常被用于投毒攻击中，也可以应用在其他对抗性攻击中，以扩大对抗性攻击的优势。尤其在跨设备（Cross-device）联邦学习中，用户认证强度低，攻击者可以通过伪造多个身份实施女巫攻击，以便提高攻击效果。

Fung 等[60]中测试了配合女巫攻击的投毒攻击效果，并提出了防御女巫攻击的方案 FoolsGold。作为 FoolsGold 的后续研究，文献[61]进一步研究了联邦学习对女巫攻击的脆弱性、目标和策略分类，提出一种被称为训练膨胀的 DoS 攻击，评估了几种分布式机器学习容错方案，证明了 FoolsGold 在防御配合女巫攻击的投毒攻击中有更好的表现。而在 VFL 中，由于用户数量较少，认证强度高，女巫攻击的应用将受到限制，较难实施。

3.1.5　针对通信瓶颈的攻击

在联邦学习中，需要在大量参与设备和服务器之间来回交换更新后的梯度，频繁的通信和传输的数据量都会产生大量的通信开销；其次，大量异构设备有限的网络带宽，会增加成员掉队的情况，进一步导致通信时间增长；此外，攻击者可能通过破坏通信信道来影响联邦学习系统的稳定性和鲁棒性。以上这些问题成为了联邦学习的主要通信瓶颈，统称为针对通信瓶颈的攻击，它影响了联邦学习的训练可用性，在 HFL 与 VFL 系统中都有可能发生。

针对通信瓶颈的攻击比投毒攻击、对抗样本攻击和后门攻击的普遍性更低，但其导致的后果非常严重，一旦系统遭受通信瓶颈攻击，可能会显著破坏联邦学习环境，给系统带来严重的损失。这种攻击最直接的解决方法是降低通信开销，把需要上传的更新量化，显著减少需要传输的数据大小[62]。2019 年，Luping 等[63]指出移动边缘设备面对昂贵的网络连接和复杂的 DNN 训练，需要更新一个很大的梯度向量，使得通信开销成为严重的瓶颈。Yao 等[64]则指出脆弱的通信带宽和单一的联邦学习训练模型，使得通信成本成为影响联邦学习模型收敛的主要因素。

由于对联邦学习完整性和可用性的破坏，以上对抗性攻击体现出良性用户与恶意攻击者之间的对抗关系，更可能会在攻击过程中被察觉，面临风险更大，因此往往来自于敌对组织或势力。

如表 1 所示，综合前文各类攻击的实施难度、攻击效果及防御难度，直观地总结了联邦学习中各种对抗性攻击的威胁等级，梳理了其对 CIA 三元组的威胁情况、作用阶段和对应的鲁棒性提升方法。

表 1　联邦学习中对抗性攻击与鲁棒性提升方法

攻击类型		威胁等级	威胁 CIA	作用阶段	鲁棒性提升方法
投毒攻击	数据投毒[31-34][43-46]	高	数据完整性 可用性	数据收集阶段	鲁棒性聚合，异常检测，数据消毒[65-67]，对抗训练，知识蒸馏，剪枝，PDGAN[68]
	模型投毒[35-37][40][47-49]	高	训练过程完整性 可用性	训练阶段	鲁棒性聚合[69-85]，异常检测， 知识蒸馏[86-91]，剪枝，
对抗样本攻击[51-55]		高	模型可用性	推理阶段	对抗训练[92-95]，数据消毒，知识蒸馏， 梯度正则化[97]，基于 GAN 的防御[98]
搭便车攻击[56-58]		中	训练过程完整性 训练可用性	数据收集阶段 训练阶段	异常检测[99-106]，区块链[107-127]
女巫攻击[59][61]		高	完整性、可用性	数据收集阶段、训练阶段	FoolsGold[60]
针对通信瓶颈攻击[63][64]		中	训练可用性	训练阶段	剪枝[128-132]

3.2　非对抗性攻击

3.2.1　模型提取攻击

机器学习中的模型提取攻击是指攻击者尝试反复发送数据以获取响应结果，从模型的 API 接口中恢复出原始模型参数或功能，甚至构造出与原始模型几乎等效的模型，一般发生在模型推理阶段，破坏模型机密性。一些场景中模型具有较高的训练成本，其信息是敏感的，易受到提取攻击。现有针对模型提取攻击的研究主要集中在减少窃取目标模型的次数、提高查询精准度和降低查询开销等。一般而言，受模型提取攻击影响最大的是预测即服务系统，因为模型本身属于隐私信息，并且从模型中可以推测出训练数据集的信息。

Tramèr 等[133]首次提出了机器学习中的模型提取攻击，并对 Google 等公司部署在云服务器上的预测系统发动攻击，在短时间内提取出与原始模型完全等效的模型。2020 年，Chandrasekaran 等[134]将模型提取攻击形式化并提出一种利用主动学习的模型提取攻击，能够在无标签数据上训练攻击模型；Jagielski 等[135]改进了基于训练的模型提取攻击的效率，并提出了一种无需训练的可以直接提取模型权重的攻击方案，围绕准确性和忠诚度两个对抗目标对模型提取空间进行系统化，分别作为衡量盗窃动机敌手和反映侦察动机敌手的成功率。

在联邦学习中，攻击者同样可以对已经部署的联邦学习模型发动模型提取攻击，由于大部分应用场景的模型都公开，所以威胁较小。但是在 VFL 中，由于每个用户持有部分模型并允许保留隐私，模型提取攻击可能由其中一个用户发起，尝试获得完整的模型，并且由系统内参与方发起的攻击可能更难防御，但目前这方面的相关研究较少。

3.2.2 推理攻击

按照攻击目的不同，推理攻击可分为成员推理攻击和属性推理攻击，其中成员推理攻击目的是推断训练数据集中是否包含特定的个人数据记录，属性推理攻击目的是推断训练数据集的某些属性。联邦学习场景下的推理攻击利用了系统的一个弊端，即每个用户和参数服务器随着训练迭代更新相同的全局模型，攻击者往往能够推断出模型的信息。此外，对抗性攻击中的投毒攻击也会对全局模型训练结果产生影响，这种影响也包含了可用于推理攻击的泄露信息，不同于被动地获取模型信息，攻击者可能会主动地实施投毒攻击来获取信息，这种攻击也被称为主动推理攻击。

(1) 成员推理攻击：成员推理攻击[10]尝试推断某个样本点是否用于训练给定模型，是最简单的一种推理攻击。当训练样本是敏感数据时，攻击者推断出训练数据集中特定数据点的存在会构成隐私威胁。文献[11]设计了一种白盒模型下的成员推理攻击方案，并假设对手具有不同程度的先验知识，在主动与被动攻击者、联邦与非联邦场景下进行了实验和对比，实验结果表明，只要是用到 SGD 的深度学习模型，对于有机会获取到模型参数的用户或服务器，实施白盒模型下的成员推理攻击准确率都可以达到 70%以上。在联邦学习场景下，每个用户随着训练进程逐渐更新全局模型，很容易具有实施攻击的条件。参数服务器是好奇或恶意的，从而实施被动或主动的成员推理攻击也是值得考虑的问题，而在集中式场景下白盒模型的条件通常难以达到。在 VFL 中，数据对齐阶段需要进行样本 ID 对齐，因此训练集中的样本信息会被所有用户共享，通常不需要进行成员推理攻击。但是这样的隐私泄露在很多实际场景中可能是不合理的[136]，有必要进一步地研究在数据纵向划分时如何在不泄露样本 ID 信息的前提下进行模型训练。

(2) 属性推理攻击：属性推理攻击[137-138]是指攻击者尝试推断训练数据集的特征信息，包括某个样本的具体数值或数据集的整体属性，例如年龄分布、性别分布等。属性推理攻击可能发生在模型训练过程中或模型训练完成之后，后者又称为模型反转攻击[139]。文献[138]的属性推理攻击方案通过调用原始模型来训练影子模型，该影子模型一定程度上包含了原始模型中包含的隐私信息，然后再用影子模型来训练攻击者感兴趣属性的分类器。文献[137]提出的属性推理攻击关注训练样本整体或子集的属性，尤其是与类别特征无关的属性。在该攻击方案中，攻击者同样利用了联邦学习的实时性，获取全局模型镜像并训练攻击模型。文献[140]提出了黑盒模型下的属性推理攻击，即使攻击者没有参与训练过程，仅通过调用训练好的模型也可以推测出关于数据集的敏感属性，即使敏感属性没有参与训练，并且与训练属性相关度很低，也会被泄露。

在 VFL 中，同样有可能发生属性推理攻击。虽然恶意用户只控制部分的模型，而且不能独立运行，但通过分析中间输出等交互消息，可以推断出其他用户的数据属性信息。与 HFL 不同，VFL 中标签信息通常只有一方拥有并且是敏感的，因此针对标签信息的推理攻击尤其需要关注。

3.2.3 基于 GAN 的攻击

生成对抗网络（Generative Adversarial Networks，GAN）[141]是一种很有前景的无监督深度学习模型，通过生成模型和判别模型的互相博弈产生很好的输出样本。通常，基于 GAN 的攻击能够发起投毒攻击[142]或推理攻击，它是一种威胁程度非常高的攻击方式，通常作用于模型的训练阶段。2017 年，Hitaj 等[143]提出了基于 GAN 的属性推理攻击，该文献指出大多数的联邦学习方案都容易遭受该模型反转攻击，即使引入了差分隐私和同态加密这样的防御措施，只要每个用户会在本地迭代更新全局模型，且模型最终可以准确地进行分类，该攻击方案就可以成功实施。但是该方案所使用的

手写数字数据集 MNIST 每一种类别都是相似的，如果同一种类别的训练样本并不相似，则该文献的攻击结果和训练样本会有很大差异[10]。Wang 等[144]提出了一种将 GAN 和多任务鉴别器相结合的 mGAN-AI 框架，该框架通过恶意服务器计算攻击目标的本地更新来恢复用户指定的私有数据。mGAN-AI 在服务端隐蔽性的工作，不会影响训练阶段。在 VFL 中，由于每个用户只有全局模型的一部分，基于 GAN 的攻击可能会失效。

表 2 总结梳理了联邦学习中各种非对抗性攻击的威胁等级、对 CIA 三元组的威胁情况、不同的作用阶段和对应的隐私性增强技术。

表 2　联邦学习中非对抗性攻击与隐私性增强技术

攻击类型		威胁等级	威胁 CIA	作用阶段	隐私性增强技术
模型提取攻击[133-135]		中	机密性	推理阶段	知识蒸馏，VerifyNet[28]，对抗训练[92-96]，同态加密[145-154]，安全多方计算[155-168]，差分隐私[169-178]，区块链，可信执行环境[179]，混合防御[180]，Anti-GAN[181]等
推理攻击	成员推理攻击[10,11,136]	高	机密性	训练阶段、推理阶段	
	属性推理攻击[137-140]	高	机密性	训练阶段、推理阶段	
基于 GAN 攻击[141-144]		高	机密性	训练阶段	

4 联邦学习中的防御手段

4.1 联邦学习鲁棒性提升方法

4.1.1 数据消毒

数据消毒（Data Sanitization）是指对有害的、异常的数据进行清理，是针对数据投毒攻击的防御通用方法[65]，在机器学习中较为常用。为抵御在联邦学习环境中的数据投毒攻击，数据消毒成为了其中的第一道防线。

选用数据消毒技术虽然可以过滤掉那些中毒信息，保护数据的可用性和有效性，但需要访问用户的本地数据[66]，无法保证数据隐私性，也难以在联邦学习分布设置的服务器中实现。另外，随着数据投毒攻击的增强，数据消毒可能很难达到期望的防御效果。面对自适应投毒攻击时，数据消毒防御效果并不乐观[67]。所以数据消毒在联邦学习中较难实施，只能在少数情况下，例如强认证的跨孤岛（Cross-silo）联邦学习，可依靠用户自身来进行。

4.1.2 鲁棒性聚合

在经典联邦学习框架下，服务器的聚合方案是联邦学习架构的核心部分。2017 年，McMahan 和 Ramage[1]首次提出了实现多用户分布式训练的 FedSGD 算法，不泄露本地数据，仅将中间梯度发送给服务器。随后的研究[3]为了减少用户与服务器之间的通信量提出了 FedAvg 算法，选择直接上传多轮本地训练的模型，并取平均值作为全局模型。FedAvg 是经典的聚合方案，确定了 HFL 的基本框架，后续聚合研究大多以此作为基础。然而，FedAvg 并不具备对投毒攻击、后门攻击等对抗性攻击的鲁棒性。为了抵抗上述攻击，需要进一步利用更新的内在属性，识别并减弱恶意模型更新效果，从而实现鲁棒性聚合（Robust Aggregation）。目前，鲁棒性聚合的研究大致可以分为 3 类。

(1) 基于统计特征和相似性的鲁棒性聚合

在联邦学习中，恶意更新通常会更加偏离其他的更新，越离散的参数更新意味着越有可能是恶意更新。因而在聚合时可以通过绕开离散的参数更新，或根据更新之间的相似性来消除远离总体分布的恶意更新，来提高针对对抗性攻击的鲁棒性。其标准大多依赖于参数更新的统计特征，如中值、平均值、欧式距离等。2017 年，Blanchard 等[38]提出了 Krum 算法和扩展的 multi-Krum 算法，Krum 中服务器计算每个模型更新与其最近更新之间的欧氏距离之和，选择距离之和最小的更新作为全局模型。改进的 multi-Krum 算法会选择多个更新的平均值来更新全局模型。Yin 等[69]提出了中值算法和裁剪平均算法，以每个维度为单位，选择中值或排除边缘值后的平均值作为全局模型。Guerraoui 等[70]提出了 Bulyan 算法，在使用裁剪平均进行聚合之前执行 Krum 算法，提升对恶意更新的检测强度。Xia 等[71]提出了抗拜占庭攻击的快速聚合算法，在每次迭代中排除距离平均梯度最远的异常梯度以获得接近真实梯度的梯度。文献[87]提出的 SLSGD 则基于移动平均（Moving Average ），考虑了当前轮和上一轮的聚合结果，具有更强的鲁棒性。此外，还有中值周围平均聚合算法[72]、基于几何中值的聚合算法[73]等。

然而，大部分聚合方案都需要已知恶意用户数量的强假设，难以在实际中应用。Cao 等[74]提出了

[1] Federated learning: collaborative machine learning without centralized training data. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Sniper 方案，其无需恶意用户数量的先验知识，通过解决最大团问题（Maximum Clique Problem，MCP）来识别良性局部模型更新。文献[75]提出了通过隐马尔可夫模型估计更新质量的方法，根据中值和余弦相似性，在每次迭代中丢弃可能恶意的局部模型更新，也无需恶意用户数量的假设。

此外，FoolsGold 使用了余弦相似性来检测恶意更新，但它并非是一个完整的聚合方案，而是针对女巫攻击的防御方案，它可以配合使用其他鲁棒性聚合方案，以增强整体的鲁棒性。

这类鲁棒性聚合方法通常将模型视为向量并利用其统计特征提取信息，计算较简单，适用于检测对模型更新影响较大的攻击。若攻击产生的变化幅度过小，或统计特征、相似性的评价标准不能很好地区分恶意梯度时，防御效果会极大地降低。

(2) 基于局部模型性能的鲁棒性聚合

基于局部模型性能的鲁棒性聚合通过在服务器提供的良性辅助数据集上对每个局部模型的训练优劣进行评估，依据评估的结果来分配聚合的权重，或自动丢弃对准确性产生负面影响的更新[77]。2019 年，Xie 等[78]提出了使用基于得分排名机制的 Zeno 方案，该方案对每一个候选梯度都持怀疑态度，并允许任意数量的恶意用户，只需保证至少存在一个诚实用户。后续的 Zeno++方案[79]消除了用户与服务器之间的通信限制，允许异步通信。Cao 等[80]使用干净的小数据集计算噪声梯度，并将每个更新与噪声梯度进行比较来过滤掉，以防御任意数量的拜占庭攻击者。此外，还有其他基于局部模型性能的聚合方案，如具有质量认知的聚合算法[81]和基于拜占庭故障的分布式学习算法[82]等。在理论方面，Cao 等[83]提出的 Ensemble FL 引入多数投票机制，提供可证明的鲁棒性保证。

这类鲁棒性聚合方法直接依赖数据集的测试结果，检测效果通常更加可靠。但同时也因为需要预先构建好的辅助数据集，在实际中很难应用。此外，该类方法涉及了局部模型的部署推理，与联邦学习局部模型的隐私性需求相冲突。

(3) 基于训练函数优化的鲁棒性聚合

相较于基于相似性的鲁棒性聚合方案，Li 等[84]提出了拜占庭鲁棒随机聚合方案在损失函数后增加正则项的方法，由服务器控制这一部分参数，在训练中使用优化来限制与全局模型偏离太多的模型作用。Andrew 等[85]提出了更新的自适应裁剪方法，将求取分位点转化为机器学习优化问题，通过在线估计的方法求取更新范数分布的特定分位点作为动态的裁剪参数。

这类鲁棒性聚合方法不同于前两种方法直接对训练结果进行评估和剔除，转而利用深度学习训练过程中的损失函数、多轮训练结构特点协助进行鲁棒性聚合，契合联邦学习特点，适用于隐私保护和性能优化，但目前的研究和实验尚少。

综上，使用鲁棒性聚合减少异常的影响是联邦学习的直接改进，可以实现较好的防御效果。但与联邦学习的主要思想存在一定矛盾，由于联邦学习需要利用不同用户的 NON-IID 训练数据(包括不常见或低质量的数据)的多样性，而放弃与全局模型不同的模型更新是与此相违背的。在联邦学习种如何平衡训练效果和对恶意攻击的防御，是鲁棒性聚合研究面临的主要挑战之一。

4.1.3 异常检测

异常检测（Anomaly Detection）旨在使用统计和分析方法对模型的训练模式、数据集或相关事件进行甄别，若检测到不符合预期的模式、异常行为或异常数据，则系统会预警并做出反应措施。目前异常检测模型主要集中对系统中的客户端异常检测和数据异常检测两个方面的研究。

(1) 客户端异常检测

客户端由于受内部或外部因素的影响会主动或被动地偏离模型训练的原定流程，如无意的客户端缺陷，或恶意的攻击者伪装成普通客户端。客户端异常检测是一种基于某种数学算法来检测异常客户端的防御方法。Li 等[99]利用了预先训练的自编码器模型来检测客户端的异常行为，并删除它们对训练模型的不利影响。一个全新的鲁棒性联邦学习框架[100]使得集中式异常检测具备了强大的识别恶意客户端的能力。若检测阈值在接收到所有客户端的模型更新后确定的，攻击者则无法先验学习到检测机制。最近几年，还有一些新颖的客户端异常检测方案，如基于可视化模型差异的客户端异常分析方法[101]，以及动态客户端分配的交叉验证异常检测方法[102]。这些方案都被视为联邦学习的有利防御手段，有效地提升了系统的鲁棒性。

(2) 数据异常检测

在数据采样过程中产生的个别异常数据，或遭受投毒攻击后的数据集，都会降低模型的训练效果。而数据异常检测则用以检测数据集中的离群点或与正常数据特征值距离较远的值。Chen 等[103]通过两阶段迭代的对抗性检测方法来识别恶意应用

软件检测系统中受到投毒攻击的软件样本，以降低检测的假阴性。Kieu 等[104]提出基于递归自编码的时间序列数据集异常值检测方法，减少了对异常值过度拟合的影响。为进一步在无监督设置的时间序列中实现鲁棒和高效的异常检测，一种变分循环编码器模型[105]可以在不依赖异常标签的情况下将异常与正常数据分离。Paudice 等[106]则通过数据预过滤和离群点检测来防御机器学习中的优化投毒攻击，减轻攻击影响。但由于针对数据点进行检测的特点，数据异常检测面临着与数据过滤相似的应用难题，客户端数据在服务器的集中式异常检测将会带来巨大的隐私风险和计算通信开销。

异常检测的防御方法与鲁棒性聚合类似，存在一定的交叉，不同之处在于后者只检测模型更新并且最终得到全局聚合梯度，前者则检测恶意注入的数据或虚假的模型，而不考虑聚合。研究适应联邦学习特点的异常检测，甚至与鲁棒性聚合相融合，是实现隐私保护联邦学习极具前景的研究方向。

4.1.4　对抗训练

对抗训练（Adversarial Training）是指在模型训练的过程中加入微弱扰动，以提高系统鲁棒性的防御方式。传统对抗训练攻击方法[92-93]大多应用于集中的机器学习框架下，并且主要关注对抗训练数据的生成。例如，Tramèr 等[92]提出集合对抗训练方法，通过从其他预训练模型中转移过来的单位输入一起作为训练集。2019 年提出的 Deep Confuse 技术[93]通过解耦交替更新过程来训练网络的稳定性，利用训练好的噪声编码器向样本添加有界扰动，从而高效地生成对抗训练数据。

目前，对抗训练[94-95]已逐渐被研究人员用于提升联邦学习系统鲁棒性，而不仅仅局限于集中式的机器模型。例如，Shah 等[94]研究了在联邦学习环境中使用对抗训练来减少模型偏移，显著提高了对抗精度和模型收敛时间。但对抗训练对于更复杂的黑盒攻击可能不具备稳定性，且加入的扰动必然会影响分类的准确度，故需要进一步采用适当的优化技术来改善这些问题。

此外，对抗训练本身需要大量数据集，并且增加了计算资源的消耗。尤其在具有较多参与方的跨设备式联邦学习环境中，轻量级用户可能无法负担高昂的对抗训练成本。Hong 等[95]提出了新颖的共享资源学习方案，即准备最充分或功能最强大的高资源设备与其他低资源用户共享安全模型，从而在 NON-IID 用户之间有效地传播对抗鲁棒性。此外，为了防止对抗样本攻击中的逃逸攻击，文献[96]通过采用高斯噪声在训练数据集中包含对抗性数据来平滑训练数据。

就目前而言，对抗训练主要集中于针对联邦学习环境中的对抗样本攻击进行防御，但分析发现[94]，对抗训练也极大地减少了推理攻击带来的威胁，提高了用户数据的隐私性，对于如何利用对抗训练来抵御其他类型的攻击研究值得进一步研究。

4.1.5　知识蒸馏

在不同模型训练场景中，若想实现更好的预测结果，往往会选择集成许多较弱模型，但这样会导致更大的计算量和更多的资源空间占用。知识蒸馏（Knowledge Distillation）作为模型压缩技术之一，就是将大模型相关知识逐步传递到小模型中，并从大模型学到的知识中学习有用信息来指导小模型训练，使小模型具备和大模型相当的性能[86-87]。在需要频繁交换训练信息的联邦学习中，知识蒸馏可以有效地降低通信开销、节省存储空间和降低参数冗余，从而防御针对通信瓶颈的攻击。此外，知识蒸馏还提高了模型的泛化能力，能够一定程度上防御投毒攻击和对抗样本攻击[86]。

在联邦学习与知识蒸馏的融合研究方面，Li 等[88]在 2019 年基于知识蒸馏和迁移联邦学习开发了一个 FedMD 通用框架，其允许计算能力存在异质性的不同用户设计不同的网络结构，保护数据集的隐私性和提高本地模型的性能。与没有协作的情况相比，FedMD 能显著提高局部模型的性能，但需要用户适当牺牲一部分数据隐私来组成共享数据集。之后，大部分方案[89-91]都是在 FedMD 框架上进行探究。例如，Lin 等[89]提出了用于模型融合的集成蒸馏，通过局部模型输出的未标记数据来进行模型融合，与 FedMD 对比，该方案进一步提高模型训练速度，并降低了通信轮数和数据隐私泄露风险。

知识蒸馏能够具有降低通信开销、提高模型鲁棒性的特点，在深度学习领域已经成为一个研究热点。当其与联邦学习技术融合时，需要考虑场景的分布式等特点，包括针对投毒攻击等对抗性攻击的防御效果研究，目前仍存在一些空白。

4.1.6　剪枝

剪枝（Pruning）技术也是一种模型压缩技术，可以在用户的计算能力和通信带宽相对较低的情况下，将联邦学习模型的大小进行修剪，降低模型复杂度和提高精度。此外，在联邦学习中受到投毒攻击后的模型会产生异常的神经元，而应用剪枝技

术可以删除这部分异常神经元用以净化整个模型。

2015 年，Han 等[128]首次提出了一种启发式和迭代权重剪枝方法去除神经网络冗余连接，而不会造成准确性损失。2018 年，文献[129]提出一种新剪枝技术，通过对攻击者放回的模型进行修剪，在修剪完成后微调神经元上的权重，以便掌握后门控制的权重，最终能够抵抗在训练集上的后门投毒攻击和通信瓶颈攻击。最近的一些研究成果[130-132]广泛应用剪枝技术来提升联邦学习系统的鲁棒性。例如，Jiang 等[130]提出一种在联邦学习环境中具有自适应和分布式参数修剪的 PruneFL 方法，通过调整模型大小显著减少训练时间，减少通信和计算开销，同时保持与原始模型相似的精度。

作为模型压缩技术之一，剪枝可以提升模型的泛化能力，避免过拟合现象。然而，不规则的剪枝可能导致模型收敛需要更多的迭代才能完成，或导致系统参数偏差，这在特定联邦学习场景下会消耗额外的资源。因此，合适的剪枝应从问题本身出发，保证不丢失正确的结果，并考虑整体效果设定一个合适的阈值，尽可能的剪去一些不必要的枝条。在此基础上，还应该减少搜索的次数以提高剪枝效率，以上这些因素都是剪枝研究考虑的优化思路。

4.1.7 其他方法

此外，还有一些其他的鲁棒性提升方法。例如，PDGAN[68]用 GAN 生成测试数据集，用于识别数据投毒攻击，通过不断改变部署策略从而增加攻击成本和复杂度的移动目标防御（Moving Target Defense）[182]，对原始数据进行随机化处理[183-184]，使用梯度正则化[97]防止过拟合，以及基于 GAN 的防御[98]等。这些鲁棒性提升方法大多是从传统机器学习场景迁移到 HFL 中，但是在 VFL 中不一定适用，如何提升 VFL 的鲁棒性还需要进一步研究。

4.2 联邦学习隐私性增强技术

为了应对联邦学习中的隐私威胁，在联邦学习中引入密码学相关技术是目前主流的隐私保护研究方向，本节对这些方案进行介绍。

4.2.1 基于同态加密的隐私性增强

同态加密（Homomorphic Encryption，HE）是一种无需访问数据本身即可对数据进行处理的概率加密技术[185]，即对经过同态加密的数据进行运算后再解密得到的结果与直接对明文进行运算得到的结果一致。若同态加密算法支持对密文进行任意形式或任意次数的计算，则称之为全同态加密（Fully Homomorphic Encryption，FHE）；若仅支持加法或乘法其中一种计算，则称之为半同态加密（Partial Homomorphic Encryption，PHE）。其中 PHE 又可以分为加法同态加密（Additively Homomorphic Encryption，AHE）和乘法同态加密（Multiplicatively Homomorphic Encryption，MHE）[186]。本节按照基于半同态加密和全同态加密两条隐私性增强路线，对这些方案进行具体分析。

(1) 基于半同态加密（PHE）的隐私性增强

由于同态加密的复杂性，在探索用同态加密构建隐私保护方案时，早期的研究仅局限于线性回归和逻辑回归等简单模型[187-[190]]。在 HFL 中，诚实但好奇的服务器可能从每个用户发送的梯度中推测出训练数据集的信息，造成隐私泄露问题。Phong 等[9]首次提出用 AHE 来保护每个用户的梯度，中心服务器在密文态下将所有用户的梯度进行聚合，可以避免隐私信息的泄露。之后，Lee 等[191]提出了基于联邦学习的患者分析平台，将同态加密应用于跨医疗机构患者的相似性搜索模型中。但巨大的计算开销使这些方案在真实场景中较难应用落地。

2020 年，Zhang 等[145]提出了 BatchCrypt，通过将批量化的梯度编码为一个长整数并一次性加密来降低计算开销，还提出了新的量化和编码方案以及新的梯度剪枝技术，实现对编码后的梯度进行聚合，极大地提升了效率。BatchCrypt 已经作为一个插件部署在首个工业级联邦学习框架 FATE 中。

在 VFL 方面，Hardy 等[146]首次提出了基于 AHE 的纵向逻辑回归系统，AHE 被用于保护每个用户的中间结果，并利用加法同态的性质将不同用户的数据秘密求和，从而计算梯度。Cheng 等[147]提出了 SecureBoost，一种用于树模型的 VFL 系统，该方案用 AHE 来保护模型训练的中间结果，实现隐私性的同时不会影响模型的准确率。Liu 等[136]提出了针对样本 ID 隐私保护的非对称 VFL 系统，在模型训练阶段使用加法同态加密来保护中间结果。Qu 等人[148]也提出了用同态加密来保护纵向贝叶斯模型训练过程中的中间结果。

(2) 基于全同态加密（FHE）的隐私性增强

密码学技术的突破可能会促进 FHE 在联邦学习隐私保护中的进一步应用，例如支持实数加密的近似同态加密算法[149]和多方 FHE 算法[150-151]非常适合构建隐私保护联邦学习（Privacy-preserving Federated Learning，PPFL）方案。2021 年，Froelicher 等[152]基于文献[151]中的多方 FHE 算法，构建了可扩展隐私保护分布式学习系统（Scalable

Privacy-preservINg Distributed LEarning，SPINDLE），支持分布式数据上广义线性模型的隐私保护训练和隐私预测，能够以隐私保护、后量子化和高效的分布式执行梯度下降。进一步，Sav 等[153]提出了 POSEIDON 框架，对 SPINDLE 进行了扩展，为神经网络模型的训练提出了一系列优化措施以提高效率，这是首个基于 FHE 并且支持神经网络模型训练与推理的联邦学习系统。2022 年，Ma 等[154]改进了文献[150]中的多密钥 FHE 方案，并构建了更有效的 PPFL 系统。表 3 对比了基于同态加密的隐私性提升技术的优缺点。

表 3　基于同态加密的隐私性提升技术对比

类型	代表方案	优点	缺点
PHE	Phong[9], BatchCrypt[145], Hardy[146], SecureBoost[147]	构造简单且已有较成熟的优化体系, 在实际场景中被广泛使用	功能单一, 常需要对运算任务进行改造来避免不支持的运算
FHE	SPINDLE[152], POSEIDON[153], Ma[154]	功能强大安全性强	计算开销很大, 在实际场景中处于探索阶段

综上，基于同态加密的方案能在一定程度上防御联邦学习中存在的隐私威胁，保证中间结果的正确性，但尚处于以学术界为主的研究阶段，尚未在工业界得到大规模应用。其主要原因在于两方面。一方面，联邦学习模型训练过程需要不断迭代，本身需要大量的计算资源，而目前同态加密的计算开销也很大，这导致在实际应用中难以实施。另一方面，基于 PHE 或 FHE 的隐私性增强方案会受到一定的局限性，如精准度损失、复杂度较大和性能要求等问题。

4.2.2　基于安全多方计算的隐私性增强

安全多方计算 (Secure Multi-Party Computation，SMPC）指无可信第三方参与下，多个参与方之间安全地计算一个模型或函数问题[192-193]。不同于同态加密模型，SMPC 具备严格的安全定义和独到的安全优势，能为联邦学习设计个性化的安全多方计算协议，为中间参数提供计算安全性，有效提升模型参数或梯度向量的隐私性。

目前，已有许多基于 SMPC 的联邦学习隐私性增强相关工作，按采用的 SMPC 种类不同，可分为基于加法秘密分享的隐私性增强、基于 Shamir 秘密分享的隐私性增强和多种技术组合的隐私性增强。

(1) 基于加法秘密分享的隐私性增强

加法秘密分享[194-199]已在机器学习的隐私保护领域中得到广泛应用。2007 年，文献[195]首次将 SMPC 概念扩展到基于机器学习梯度下降的隐私保护方法中。之后，文献[196]提出了首个专注于机器学习线性回归、逻辑回归和神经网络模型的隐私保护系统 SecureML，通过改进乘法三元组生成方法和设计 MPC 友好的激活函数来提高效率。SecureNN 方案[197]在 SecureML 基础上扩展为三方条件下的加法秘密分享，并提供了针对半诚实敌手的完全安全性和针对恶意敌手的隐私概念。Agrawal 等[198]指出 SecureML 会对模型精度造成影响，提出新的离散化 DNNs 训练方法 QUOTIENT，通过模型权重三元化等措施提高隐私性。与 SecureML 相比 QUOTIENT 训练模型的准确性和速度都有所提高。文献[199]提出了一种三方的隐私训练和推理协议 FALCON，通过批量归一化使其能够支持更复杂的神经网络训练，用算术秘密分享实现了在多数参与方诚实情况下的恶意敌手安全，避免了使用算术、布尔秘密分享和混淆电路之间的转换协议。

在联邦学习中，为了实现个性化的隐私保护模型，Hsu 等[155]针对 Android 恶意软件检测提出了一种新的 PPFL 系统，允许移动设备协调训练分类器，而不暴露敏感信息，即使用加法秘密分享来保护本地模型参数不受服务器影响。Dong 等[156]通过加法秘密分享与 Top-K 梯度选择相结合，设计了半诚实模型下高效且安全的联邦学习协议，在恶意模型下采用消息验证码来解决服务器可能被腐蚀的问题。

(2) 基于 Shamir 秘密分享的隐私性增强

在文献[157]中，基于$(t，n)$门限秘密分享和密钥协商协议提出了一种安全的联邦学习架构，该架构可以防止半诚实的服务器获取各个用户反馈的参数时推断出相关敏感信息，即保证更新参数的安全性和私密性。但该方案难以抵挡恶意用户的攻击，且不支持验证服务器返回的结果。2019 年，Bonawitz 等[158]在文献[157]的基础上改进，利用基于 TensorFlow 框架为移动设备领域的联邦学习构建了一个可扩展的生态系统，该系统使用秘密分享和差分隐私技术，保护了梯度下降的参数更新，以防止本地数据集的重要信息泄露。然而，该系统假设这些用户之间并不团结，即随时加入或退出，故对系统稳定或模型效率造成影响。2020 年，Wang 等[159]使用$(t，n)$门限秘密分享构建了 XGBoost 的安全分类和回归树模型 FedXGB 以及一个安全预测

协议，FedXGB 对于半诚实对手是安全的，并且准确率的损失很小。

相较于加法秘密分享，Shamir 秘密分享并不需要所有用户同时参与，只需满足$(t，n)$门限的阈值，即可重构出秘密值。因此，将 Shamir 秘密分享应用于联邦学习中，不仅能有效增强模型参数的隐私性，而且能抵御用户随机退出或加入带来的不稳定性影响，提高系统的鲁棒性。

(3) 多种技术组合的隐私性增强

在机器学习中，一些经典方案将混淆电路和各种秘密分享技术结合起来，实现更通用的隐私保护框架，如 ABY3[160]、BLAZE[161]和 CrypTFlow[162]等。这些方案能够保证恶意敌手下的安全性，大幅度提升吞吐量，在实际场景中可用性更好。已经有一些方案把安全多方计算基础算子进行了更高级的封装，开始了商业落地的尝试[163][164]。

其中，秘密分享与 HE 相结合是实现 PPFL 的常见方案，其重点在于计算开销和通信开销之间的平衡[165-167]。例如，FEDXGB[165]融合了秘密分享和 HE 的优点，使用户无需解密即可计算模型更新，并对用户丢失具有鲁棒性。EaSTFLy[166]设计了针对半城实敌手的隐私保护协议，以解决三元梯度 FL 中的隐私问题。但是这些方案中，数据在不同的隐私保护技术之间的转换通常是设计的难点。函数秘密分享也可以被用于构建隐私保护方案。2021 年，Kaissis 等[168]提出了用函数秘密分享构建的 PriMIA 框架，并把它用于医学影像数据上的联邦学习训练与推理。

综上，虽然基于安全多方计算的联邦学习隐私性增强能够保护中间参数的安全性，但几乎大部分方案都需要大量通信或计算开销，且存在效率与隐私之间的权衡问题，故需要我们进一步研究。此外，在 VFL 的数据对齐阶段，可以通过 SMPC 中的隐私集合求交（PSI）协议来找出各个用户的共同样本，从而避免交集之外的样本信息被泄露。在早期的隐私保护机器学习研究中已被提出[200-201]。最近，VFL 方案[136,147,202-203]的数据对齐阶段大都采用了这种方法，只要诚实方占大多数，就不会泄露关于交集之外的样本的任何信息。但在这样的数据对齐方式中，交集样本信息被公开给了所有参与方，这在很多场景中仍会引起隐私泄露的担忧[136]，目前对于这个问题的研究还不多。表 4 对比了基于 SMPC 的隐私性提升技术的优缺点。

表 4 基于安全多方计算的隐私性增强技术对比

类型	代表方案	优点	缺点
加法秘密分享	SecureML[196] SecureNN[197] QUOTIENT[198]	轻量级 计算速度快	通信开销大 乘法运算需要提前生成乘法三元组
Shamir 秘密分享	Bonawitz[157] Bonawitz[158]	更灵活, 允许未知的设备数量	效率较低
多种技术组合	ABY3[160] BLAZE[161] CrypTFlow[162] Crypten[164] FEDXGB[165] EaSTFLy[166] Chen[167]	抵抗恶意敌手 吞吐量更高 灵活地平衡计算和通信开销, 性能较好	不同技术间的转换瓶颈

4.2.3 基于差分隐私的隐私性增强

差分隐私（Differential privacy，DP）旨在传输的梯度信息中加入随机噪声，将其查询操作的实际结果隐藏起来或模糊化直至无法区分，从而实现对隐私数据的保护。差分隐私一般是用来促进敏感数据上的安全分析，使敌手无法在输出结果中识别个体之间的敏感性。通常，差分隐私可用于防御模型提取攻击、成员推理攻击、基于 GAN 攻击和窃听等。因此，基于差分隐私的联邦学习模型可作为保护本地训练数据私密性、梯度信息机密性和衡量隐私损失阈值的有效解决方案[169-172]。

2017 年，Geyer 等[173]提出了客户端级别的差分隐私联邦学习优化算法，该算法不仅保护数据的隐私，而且能确保一个学习模型不会显示客户端是否参与了训练。通过在训练期间隐藏客户的贡献，防止任何用户从聚合模型中推断其他用户的私有数据，避免了该客户端受到其他客户端的差异攻击。2019 年，Truex 等[174]结合了 SMPC 和差分隐私技术生成可抵抗推理攻击的高精度模型，使其不牺牲隐私而保持预定义的信任率。利用这两种隐私技术的优势，能有效地降低联邦学习系统的隐私威胁。之后，Triastcyn 等[175]采用 Bayesian 差分隐私提供了更清晰的隐私损失界限便。2020 年，文献[176]指出差分隐私能通过衡量联邦学习中的通信隐私损失来提高对数据隐私的保护。私有 FL-GAN 方案[177]是一种基于联邦学习的差分隐私 GAN 模型，其增强了系统的隐私性，但同时也影响了模型准确性。文献[178]提出了基于差分隐私的 VFL 框架，可以达到接近于无隐私保护 VFL 的性能，比基于同

态加密或安全多方计算的方案快很多，但需要足够的隐私预算才能达到较高的准确率。

可以看到，虽然差分隐私能为联邦学习提供强大的隐私保护能力，有效防御非对抗性攻击，提升系统的安全性。但在加入随机噪声后，训练结果的准确性难以保障，且训练出准确的模型需要较高的隐私预算，故难以量化这些方法提供的隐私保护水平，并且差分隐私对于恶意服务器模型下的主动成员推理攻击防御效果较差[204]。总体而言，相较于安全多方计算模型，基于差分隐私的联邦学习模型消耗的通信和计算开销会更低，但同时会损失训练结果的准确性，在后续研究中仍需在隐私性和准确性之间做权衡。

4.2.4　基于区块链的隐私性增强

以上提到的同态加密、安全多方计算和差分隐私等密码技术能够实现较好的隐私保护，但这些技术大都依赖中心化参数服务器发送的模型参数，即无法保证全局模型的可信度，且多数用户之间存在不信任等问题[107-108]。区块链作为一种以密码学、共识算法和分布式存储等技术相结合的去中心化存储架构，其特有的数据结构优势结合密码技术可以增强联邦学习的完整性和机密性。以下梳理了基于区块链的联邦学习隐私保护的研究现状。

（1）去中心化联邦学习通用模型设计

根据区块链的去中心化架构优势，去中心化的联邦学习通用模型设计是一个重要的研究方向。Lu 等[109]为互不信任的多方设计了一个区块链授权的安全共享体系结构，将数据共享问题转化为机器学习问题，利用差分隐私扰动本地训练的数据，并根据训练结果的质量来完成共识过程，以确保局部模型和全局模型的可信性。Awan 等[110]提出了基于区块链的 PPFL 框架，该框架利用区块链记录模型更新流程，且无需用户的半城实假设，即可完成本地模型更新的安全聚合。Shayan 等[111]提出了一种基于区块链和加密原语实现对等客户端之间隐私保护的多方联邦学习方案，该方案不依赖集中式的协调服务器，采用以每轮训练过程中达成模型状态的联邦共识机制来防止投毒攻击，以及避免了对等客户端更新时受到的隐私泄露攻击。朱建明等[112]构建了一个去中心化的参数聚合链，利用区块链和差分隐私保护中间参数的可信和隐私，协作者节点进行参数验证，并提出节点贡献度证明（Proof of Contribution，PoC）共识算法选出主节点记账。然而，该方案仅支持防御投毒攻击，不能有效防御推理攻击，且无法保证协作者是恶意的还是半诚实的。因此，还需更多安全性假设和密码技术来保证 PPFL 的有效性。

此外，中央服务器可能会因为自利性偏袒某些用户，甚至恶意中央服务器可能对模型进行投毒或收集用户的隐私信息[113]。由于区块链上的参数信息不能公开，往往需要与加密技术相结合或是增加细粒度访问，来对隐私数据进行有效的保护[114-115]。

（2）联邦学习激励机制和智能合约研究

联邦学习模型的效果取决于各个用户的数据集和贡献量，若没有足够的训练数据和其他资源，则会导致系统性能降低[116]。相反，若能激励用户积极参与模型训练，并结合防篡改和可信的智能合约来执行，则可进一步提高系统的鲁棒性和隐私性。

Kim 等[117]提出了同时具备身份验证和激励功能的基于区块链联邦学习系统 BlockFL，通过对局部模型参数进行交叉验证和调整区块生成率，从而提高全局模型的可信性和降低延迟时间，加强了系统的鲁棒性。Bao 等[118]建立了一个可审计的联邦学习系统，该系统具备了防篡改的局部模型参数更新，诚实用户获得公平分配的收益，而恶意用户将受到惩罚。Liu 等[119]提出了一个基于区块链的联邦学习安全框架，利用智能合约来抵御恶意或不可靠的用户发起的投毒攻击，并使用本地差分隐私来防止成员推理攻击。但较多的参与方和通信开销可能会影响系统效率和模型准确度。Rehaman 等[120]指出可将参与训练各方设备的信誉值上链，用以识别恶意用户或故障用户，将用户的异常行为与激励机制挂钩，规避恶意攻击者对系统的破坏，迫使用户带来积极和诚实的模型贡献。Behera 等[121]利用区块链上智能合约的公开透明性，为联邦学习中的数据贡献者建立了公平、安全的激励机制，有效地惩罚了表现不佳的用户。

以上大多数方案的目标是激励用户参与训练或实现贡献与收益的公平性，但引入区块链后也会增加一些额外的操作，例如需要对模型更新进行验证审计、共识上链和收益分配等。这些额外的操作又会带来训练效率降低、通信与计算开销增大和设备规模受限等新问题，需要进一步考量和完善。

（3）联邦学习的不同应用场景研究

为了符合实际场景下的隐私保护需求，目前已有许多将区块链和联邦学习相结合应用在特定场景下的联邦学习隐私保护方案。例如，利用分布式哈希表和区块链解决雾计算场景中的单点失败和

投毒攻击问题[122]、在车载机器学习场景中引入区块链与联邦学习相结合的框架完成安全的梯度聚合[123]、无集中模型协调器的交通流预测联邦学习框架[124]，以及面向智能家居设备场景构建了一种减少恶意用户攻击行为的联邦学习隐私保护方案[125]等。此外，Warnat-Herresthal 等[126]提出了一种去中心化联邦学习的医疗数据共享方法 Swarm Learning，该方法将横向联邦学习和许可区块链相结合，利用区块链技术安全可靠的与对等节点协作学习，以保证不同医疗机构之间医疗数据的安全性和隐私性。

然而，区块链技术与联邦学习相结合的方案也会面临 51%攻击[127]等安全问题，且由于不同应用场景下的隐私和效率等级需求不同，相应的实施方案也会有所差异。因此，针对不同应用场景的功能需求，还需继续研究在保证联邦学习过程中隐私性的同时，不会降低模型准确性或系统效率。

4.2.5 其他隐私增强技术

此外，还有其他一些隐私性增强技术。针对联邦学习中梯度会泄露隐私的问题，梯度裁剪也可以一定程度上避免这种信息泄露[205]，可信执行环境（Trusted Execution Environment，TEE）[179]可以通过硬件手段提供一个安全区域来执行模型训练操作以保护数据完整性和隐私性，VerifyNet[28]提供双重屏蔽协议来保证用户局部梯度的机密性，混合防御[180]融合多个防御技术防范服务器与恶意用户勾结问题，模型水印技术[206]通过将水印嵌入模型参数中用来在遭受模型窃取攻击之后进行举证和维权，及 Anti-GAN 方法[181]通过使用 GAN 来生成假的训练数据来保证私密性。

综上所述，围绕本地敏感数据、梯度向量和模型聚合结果这三个隐私保护目标，应用同态加密、安全多方计算、差分隐私和区块链等新兴技术可解决联邦学习中的隐私威胁和安全威胁，从而增强联邦学习隐私保护。这些隐私增强技术在为联邦学习隐私需求提供有效保障的同时，往往也会在一定程度上付出相应的代价。例如，在保证隐私性的同时牺牲了系统的效率，或是降低了模型的准确性。相反，若是提高了系统性能和准确性，相应地通信和计算开销又会增大。因此，在隐私性、高效性和准确性不可能三角之间做权衡，是联邦学习技术发展的未来研究重点。

如表 5 所示，为本文提到的所有防御手段在抵御各种攻击方面的效果对比。其中，实心圆表示“是”，半实心圆表示“不完全”，空心圆表示“否”。

表 5 各类防御手段的效果对比

参考文献	防御手段	作用效果	抵御攻击								
			数据投毒攻击	模型投毒攻击	对抗样本攻击	搭便车攻击	女巫攻击	针对通信瓶颈攻击	模型提取攻击	推理攻击	基于 GAN 的攻击
[65-67]	数据消毒	过滤有害数据	●	○	●	○	○	○	○	○	○
[69-85]	鲁棒性聚合	消除恶意更新	●	●	○	○	●	○	○	○	○
[99-106]	异常检测	检测异常数据或终端	●	●	○	●	○	○	○	○	○
[92-95]	对抗训练	提高鲁棒性	◐	○	●	○	○	○	○	○	○
[86-91]	知识蒸馏	降低通信开销 提高鲁棒性	◐	◐	◐	○	○	●	○	○	○
[128-132]	剪枝	降低通信开销 提高鲁棒性	◐	◐	◐	○	○	●	○	○	○
[145-154]	同态加密	保护中间结果	○	○	○	○	○	○	○	●	●
[155-168]	安全多方计算	保护模型参数	○	○	○	○	○	○	○	●	○
[169-178]	差分隐私	保护梯度信息	○	○	●	○	○	○	●	●	●
[107-127]	区块链	保证数据可信	○	○	○	●	●	○	○	●	○

5 未来研究方向

联邦学习中的攻击和防御发展尚不成熟，仍然存在很多问题亟待解决，其中以下四类问题值得进一步地研究。

(1) 攻击检测与模型评估研究

联邦学习的迅速发展势必会带来多样化形式的安全与隐私威胁。数据用户可能会主动或被动的从诚实状态转变为恶意状态，模型训练过程中操作处理也可能会受环境因素发生异常，这些问题往往会严重影响系统性能。事实上，若系统能及时且有

效地发现这些行为、提供预警功能，以及主动采取反应措施能最大程度降低危害和损失。因此，如何建立安全稳定的攻击检测与评估模型，为联邦学习系统赋予自检与评估能力，为内部和外部环境提供实时防护功能，未来需要进一步探索研究。

(2) 完善的安全攻防体系研究

由于现有的联邦学习协议还未发展成熟，相应的体系还不够完善，仅仅依靠现有的防御手段无法预防未来未知的攻击威胁。若系统面对一种全新或特殊的攻击威胁，没有提前预备方案，即无法满足已成型的技术产品安全需求；同时，若系统面对混合多种攻击的安全威胁，仅依靠传统的单一解决方案，也很难达到有效的防御效果。然而，目前的研究工作尚未对联邦学习中攻击与防御技术形成系统化的攻防体系。因此，未来应从现有的攻击方式和防御手段出发，分析推理出所有可能潜在的攻击和隐私问题，并结合安全的加密技术，以此构建出完善的联邦学习安全攻防体系。

(3) 纵向联邦学习安全及隐私研究

现有的联邦学习安全及隐私的研究，尤其在安全威胁及防御方面，主要集中在横向联邦学习场景下，而在纵向联邦学习场景中模型训练与部署更加复杂，相关的研究还非常少。如模型在用户之间的切分导致一些攻击方案和防御手段在纵向联邦学习中并不适用。与此同时，也有一些纵向联邦学习特有的安全和隐私威胁，例如标签推理攻击、来自内部的模型窃取攻击，以及样本 ID 隐私泄露问题等。这些纵向联邦学习中特有的攻击及防御手段仍然需要深入研究，以便解决纵向联邦学习在真实场景中实际应用的瓶颈问题。

(4) 兼顾鲁棒性和隐私性的联邦学习研究

鲁棒性与隐私性是联邦学习系统在实际应用中需要考虑的两个维度指标，两者缺一不可。然而鲁棒性提升与隐私性增强之间存在矛盾，例如很多隐私保护手段试图尽量减少不同用户梯度信息的差异，而这常常会阻碍一些鲁棒性提升方法对异常数据的识别。因此，如何处理这个矛盾是联邦学习在大规模应用之前面临的重要问题。

6 结论

联邦学习解决了不同训练用户之间的数据孤岛问题，打破了不同领域之间的数据壁垒，但其无法避免在模型训练与参数传递时遭受各种攻击威胁。 因此，本文聚焦于联邦学习中安全与隐私威胁问题，重点从对抗性攻击、非对抗性攻击两个层面分析了各种攻击的基本概念、作用阶段和实施效果，以及最新研究进展，并依据攻击手段影响的性质不同，从鲁棒性提升方法、隐私性增强技术两个角度，对十几种防御手段进行了分析与总结。攻击和防御相互促进，强大的攻击手段会催生出针对性的防御技术，而稳固的防御技术又会促进攻击手段的提升。联邦学习系统还处于发展初期，各类攻击手段层出不穷，未来应该充分利用相关加密技术，以及安全多方计算、区块链、可信执行环境等隐私增强技术，以推动联邦学习在隐私保护领域的研究。

致　谢　我们向对本文的工作给予支持和宝贵建议的评审老师和同行表示衷心的感谢!

参 考 文 献

[1] Li J. Cyber security meets artificial intelligence: A survey. Frontiers of Information Technology & Electronic Engineering, 2018, 19(12): 1462-1474.

[2] Konečný J, McMahan H B, Ramage D, et al. Federated optimization: distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610. 02527, 2016.

[3] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data//Artificial intelligence and statistics. Florida, USA, 2017: 1273-1282.

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

[5] McMahan H B, Ramage D, Talwar K, et al. Learning differentially private recurrent language models//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-14.

[6] Preuveneers D, Rimmer V, Tsingenopoulos I, et al. Chained anomaly detection models for federated learning: An intrusion detection case study. Applied Sciences, 2018, 8(12): 2663.

[7] Khramtsova E, Hammerschmidt C, Lagraa S, et al. Federated learning for cyber security: SOC collaboration for malicious URL detection //Proceedings of the IEEE International Conference on Distributed Computing Systems. Singapore, 2020: 1316-1321.

[8] Shi J, Zhao H, Wang M, et al. Signal recognition based on federated learning //Proceedings of the IEEE International Conference on Computer Communications Workshops. Toronto, Canada, 2020: 1105-1110.

[9] Aono Y, Hayashi T, Wang L, et al. Privacy-preserving deep learning via

additively homomorphic encryption. IEEE Transactions on Information Forensics and Security, 2017, 13(5): 1333-1345.

[10] Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models//Proceedings of the IEEE Symposium on Security and Privacy. San Jose, USA, 2017: 3-18.

[11] Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning//Proceedings of the IEEE symposium on security and privacy. San Francisco, USA, 2019: 739-753.

[12] Tolpegin V, Truex S, Gursoy M E, et al. Data poisoning attacks against federated learning systems//Proceedings of the European Symposium on Research in Computer Security. Guildford, UK, 2020: 480-501.

[13] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015: 1-11.

[14] Yin X, Zhu Y, Hu J. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. ACM Computing Surveys, 2021, 54(6): 1-36.

[15] AbdulRahman S, Tout H, Ould-Slimane H, et al. A survey on federated learning: The journey from centralized to distributed on-site learning and beyond. IEEE Internet of Things Journal, 2021, 8(7): 5476-5497.

[16] Chen Bing, Cheng Xiang, Zhang Jia-Le, et al. Survey of security and privacy in federated learning. Journal of Nanjing University of Aeronautics & Astronautics, 2020, 52(05): 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. 南京航空航天大学学报, 2020, 52(05): 675-684).

[17] Zhou Jun, Fang Guo-Ying, Wu Nan, et al. Survey on security and privacy-preserving in federated learning. Journal of Xihua University (Natural Science Edition), 2020, 75(04): 9-17 (in Chinese)
(周俊, 方国英, 吴楠等. 联邦学习安全与隐私保护研究综述. 西华大学学报(自然科学版), 2020, 75(04): 9-17)

[18] Zhou Chuan-Xin, Sun Yi, Wang De-Gang, et al. Survey of federated learning research. Chinese Jounal of Network and Infoamation Security, 2021, 7(5):77-92 (in Chinese)
(周传鑫, 孙奕, 汪德刚等. 联邦学习研究综述. 网络与信息安全学报, 2021, 7(5): 77-92)

[19] He Y, Meng G, Chen K, et al. Towards Security Threats of Deep Learning Systems: A Survey. IEEE Transactions on Software Engineering, 2020, 48(5): 1743-1770.

[20] Lyu L, Yu H, Yang Q. Threats to federated learning: A survey. arXiv preprint arXiv: 2003.02133, 2020.

[21] Lyu L, Yu H, Ma X, et al. Privacy and robustness in federated learning: Attacks and defenses. arXiv preprint arXiv: 2012.06337, 2020.

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

[23] Wang Kun-Qing, Liu Jing, Li Chen, et al. A survey on threats to federated learning. Journal of Information Security Research, 2022, 78(03): 223-234 (in Chinese)
(王坤庆,刘婧,李晨等. 联邦学习安全威胁综述. 信息安全研究, 2022, 78(03):223-234)

[24] Wei Y, Chen Y, Xiao M, et al. Protecting machine learning integrity in distributed big data networking. IEEE Network, 2020, 34(4): 84-90.

[25] Chen Y, Luo F, Li T, et al. A training-integrity privacy-preserving federated learning scheme with trusted execution environment. Information Sciences, 2020, 522: 69-79.

[26] Lyu L, Xu X, Wang Q, et al. Collaborative fairness in federated learning. Federated Learning, 2020: 189-204.

[27] Jagielski M, Kearns M, Mao J, et al. Differentially private fair learning //Proceedings of the International Conference on Machine Learning. California, USA, 2019: 3000-3008.

[28] Xu G, Li H, Liu S, et al. Verifynet: Secure and verifiable federated learning. IEEE Transactions on Information Forensics and Security, 2019, 15: 911-926

[29] Muñoz-González L, Biggio B, Demontis A, et al. Towards poisoning of deep learning algorithms with back-gradient optimization//Proceedings of the ACM Workshop on Artificial Intelligence and Security. Dallas, USA, 2017: 27-38.

[30] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines//Proceedings of the International Coference on International Conference on Machine Learning. Edinburgh, Scotland, UK, 2012: 1467-1474.

[31] Yao Y, Li H, Zheng H, et al. Regula sub-rosa: Latent backdoor attacks on deep neural networks. arXiv preprint arXiv: 1905.10447, 2019.

[32] Shafahi A, Huang W R, Najibi M, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks//Proceedings of the International Conference on Neural Information Processing Systems. Montréal Canada, 2018: 6106-6116.

[33] Wang H, Sreenivasan K, Rajput S, et al. Attack of the tails: Yes, you really can backdoor federated learning//Proceedings of the International Conference on Neural Information Processing Systems. Virtual, 2020: 16070-16084.

[34] Huang W R, Geiping J, Fowl L, et al. Metapoison: practical general-purpose clean-label data poisoning//Proceedings of the International Conference on Neural Information Processing Systems. 2020: 12080-12091.

[35] Bhagoji A N, Chakraborty S, Mittal P, et al. Analyzing federated learning through an adversarial lens//Proceedings of the International Conference on Machine Learning. California, USA, 2019: 634-643.

[36] Zhou X, Xu M, Wu Y, et al. Deep model poisoning attack on federated learning. Future Internet, 2021, 13(3): 73.

[37] Wei K, Li J, Ding M, et al. Covert model poisoning against federated learning: Algorithm design and optimization. arXiv preprint arXiv:2101. 11799, 2021.

[38] Blanchard P, El Mhamdi E M, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent//Proceedings of the International Conference on Neural Information Processing

Systems. Long Beach, USA, 2017: 118–128.

[39] Baruch G, Baruch M, Goldberg Y. A little is enough: Circumventing defenses for distributed learning//Proceedings of the International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 8635-8645.

[40] Fang M, Cao X, Jia J, et al. Local model poisoning attacks to byzantine-robust federated learning //Proceedings of the USENIX Conference on Security Symposium. Boston, USA, 2020: 1605-1622.

[41] Shejwalkar V, Houmansadr A. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning//Proceedings of the Annual Network and Distributed System Security Symposium. Virtual, 2021: 1-18.

[42] Sun Z, Kairouz P, Suresh A T, et al. Can you really backdoor federated learning? arXiv preprint arXiv:1911. 07963, 2019.

[43] Li Y, Li Y, Wu B, et al. Invisible backdoor attack with sample-specific triggers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 16463-16472.

[44] Doan K, Lao Y, Zhao W, et al. LIRA: Learnable, imperceptible and robust backdoor attacks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 11966-11976.

[45] Bagdasaryan E, Shmatikov V. Blind backdoors in deep learning models //Proceedings of the USENIX Security Symposium. Virtual, 2021: 1505-1521.

[46] Gu T, Liu K, Dolan-Gavitt B, et al. Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 2019, 7: 47230-47244.

[47] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning //Proceedings of the International Conference on Artificial Intelligence and Statistics. Virtual, 2020: 2938-2948.

[48] Chen Z, Tian P, Liao W, et al. Towards multi-party targeted model poisoning attacks against federated learning systems. High-Confidence Computing, 2021, 1(1): 100002.

[49] Xie C, Huang K, Chen P Y, et al. Dba: Distributed backdoor attacks against federated learning//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019.

[50] Liu Y, Yi Z, Chen T. Backdoor attacks and defenses in feature-partitioned collaborative learning. arXiv preprint arXiv:2007.03608, 2020.

[51] Athalye A, Engstrom L, Ilyas A, et al. Synthesizing robust adversarial examples//Proceedings of the International conference on machine learning. Vienna, Austria, 2018: 284-293.

[52] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks//Proceedings of the International Conference on Learning Representations. Banff, Canada, 2014: 1-10.

[53] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 2018, 6: 14410-14430.

[54] Ilyas A, Engstrom L, Madry A. Prior convictions: Black-box adversarial attacks with bandits and priors//Proceedings of the International Conference on Learning Representations.Vancouver, Canada, 2018:1-25.

[55] Pang Q, Yuan Y, Wang S. Attacking Vertical Collaborative Learning System Using Adversarial Dominating Inputs. arXiv preprint arXiv:2201.02775, 2022.

[56] Zong B, Song Q, Min M R, et al. Deep autoencoding gaussian mixture model for unsupervised anomaly detection//Proceedings of the International conference on learning representations. Vancouver, Canada, 2018: 1-19.

[57] Fraboni Y, Vidal R, Lorenzi M. Free-rider attacks on model aggregation in federated learning//Proceedings of the International Conference on Artificial Intelligence and Statistics. Virtual, 2021: 1846-1854.

[58] Wan W, Lu J, Hu S, et al. Shielding Federated Learning: A New Attack Approach and Its Defense//IEEE Wireless Communications and Networking Conference. Nanjing, China, 2021: 1-7.

[59] Douceur J R. The sybil attack//International workshop on peer-to-peer systems. Springer, Berlin, Heidelberg, 2002: 251-260.

[60] Fung C, Yoon C J M, Beschastnikh I. Mitigating sybils in federated learning poisoning//Proceedings of the International Symposium on Research in Attacks, Intrusions and Defenses. San Sebastian, Spain, 2020: 301-316.

[61] Fung C, Yoon C J M, Beschastnikh I. The limitations of federated learning in sybil settings// Proceedings of the International Symposium on Research in Attacks, Intrusions and Defenses. San Sebastian, Spain, 2020: 301-316.

[62] Konečný J, McMahan H B, Yu F X, et al. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv: 1610.05492, 2016.

[63] Luping W, Wei W, Bo L I. CMFL: Mitigating communication overhead for federated learning//Proceedings of the IEEE International Conference on Distributed Computing Systems. Texas, USA, 2019: 954-964.

[64] Yao X, Huang C, Sun L. Two-stream federated learning: Reduce the communication costs// Proceedings of the IEEE Visual Communications and Image Processing. Taiwan, China, 2018: 1-4.

[65] Cretu G F, Stavrou A, Locasto M E, et al. Casting out demons: Sanitizing training data for anomaly sensors//Proceedings of the IEEE Symposium on Security and Privacy. California, USA, 2008: 81-95.

[66] Kairouz P, McMahan H B, Avent B, et al. Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 2021, 14(1–2): 1-210.

[67] Koh P W, Steinhardt J, Liang P. Stronger data poisoning attacks break data sanitization defenses. Machine Learning, 2021: 1-47.

[68] Zhao Y, Chen J, Zhang J, et al. PDGAN: A novel poisoning defense method in federated learning using generative adversarial network //Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing. Melbourne, Australia, 2019: 595-609.

[69] Yin D, Chen Y, Kannan R, et al. Byzantine-robust distributed learning: Towards optimal statistical rates//Proceedings of the International

Conference on Machine Learning. Vienna, Austria, 2018: 5650-5659.

[70] Guerraoui R, Rouault S. The hidden vulnerability of distributed learning in byzantium//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2018: 3521-3530.

[71] Xia Q, Tao Z, Hao Z, et al. FABA: an algorithm for fast aggregation against byzantine attacks in distributed neural networks//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2019: 4824-4830.

[72] Xie C, Koyejo O, Gupta I. Generalized byzantine-tolerant sgd. arXiv preprint arXiv:1802.10116, 2018.

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

[74] Cao D, Chang S, Lin Z, et al. Understanding distributed poisoning attack in federated learning//Proceedings of the IEEE International Conference on Parallel and Distributed Systems. Tianjin, China, 2019: 233-239.

[75] Muñoz-González L, Co K T, Lupu E C. Byzantine-robust federated machine learning through adaptive model averaging. arXiv preprint arXiv:1909.05125, 2019.

[76] Xie C, Koyejo O, Gupta I. SLSGD: Secure and efficient distributed on-device machine learning//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2019: 213-228.

[77] Barreno M, Nelson B, Joseph A D, et al. The security of machine learning. Machine Learning, 2010, 81(2): 121-148.

[78] Xie C, Koyejo S, Gupta I. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance//Proceedings of the International Conference on Machine Learning. California, USA, 2019: 6893-6901.

[79] Xie C, Koyejo S, Gupta I. Zeno++: Robust fully asynchronous sgd//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 10495-10503.

[80] Cao X, Lai L. Distributed gradient descent algorithm robust to an arbitrary number of byzantine attackers. IEEE Transactions on Signal Processing, 2019, 67(22): 5850-5864.

[81] Deng Y, Lyu F, Ren J, et al. Fair: Quality-aware federated learning with precise user incentive and model aggregation//Proceedings of the IEEE International Conference on Computer Communications. Virtual, 2021: 1-10.

[82] Guo S, Zhang T, Yu H, et al. Byzantine-resilient decentralized stochastic gradient descent. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(6): 4096-4106.

[83] Cao X, Jia J, Gong N Z. Provably secure federated learning against malicious clients//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021, 35(8): 6885-6893.

[84] Li L, Xu W, Chen T, et al. RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019, 33(01): 1544-1551.

[85] Andrew G, Thakkar O, McMahan B, et al. Differentially private learning with adaptive clipping//Proceedings of the International Conference on Neural Information Processing Systems. Virtual, 2021, 34: 17455-17466.

[86] Phuong M, Lampert C. Towards understanding knowledge distillation //Proceedings of the International Conference on Machine Learning, California, USA, 2019: 5142-5151.

[87] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[88] Li D, Wang J. Fedmd: Heterogenous federated learning via model distillation//Proceedings of the International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 7645-7655..

[89] Lin T, Kong L, Stich S U, et al. Ensemble distillation for robust model fusion in federated learning//Proceedings of the International Conference in Neural Information Processing Systems. Vancouver, Canada, 2020, 33: 2351-2363.

[90] Jiang D, Shan C, Zhang Z. Federated learning algorithm based on knowledge distillation// Proceedings of the International Conference on Artificial Intelligence and Computer Engineering. Beijing, China, 2020: 163-167.

[91] Ma J, Yonetani R, Iqbal Z. Adaptive distillation for decentralized learning from heterogeneous clients//Proceedings of the International Conference on Pattern Recognition. Milan, Italy, 2021: 7486-7492.

[92] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018.

[93] Feng J, Cai Q Z, Zhou Z H. Learning to confuse: generating training time adversarial data with auto-encoder//Proceedings of the International Conference in Neural Information Processing Systems. Vancouver, Canada, 2019, 32: 11994-12004.

[94] Shah D, Dube P, Chakraborty S, et al. Adversarial training in communication constrained federated learning. arXiv preprint arXiv:2103. 01319, 2021.

[95] Hong J, Wang H, Wang Z, et al. Federated robustness propagation: Sharing adversarial robustness in federated learning. arXiv preprint arXiv:2106. 10196, 2021.

[96] Chen C, Kailkhura B, Goldhahn R, et al. Certifiably-Robust Federated Adversarial Learning via Randomized Smoothing//Proceedings of the IEEE International Conference on Mobile Ad Hoc and Smart Systems. Denver, USA, 2021: 173-179.

[97] Ross A S, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients//Proceedings of the AAAI conference on artificial intelligence. Louisiana, USA, 2018, 32(1): 1660-1669.

[98] Shen S, Jin G, Gao K, et al. Ape-gan: Adversarial perturbation elimination with gan//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK, 2019: 3842-3846.

[99] Li S, Cheng Y, Liu Y, et al. Abnormal client behavior detection in federated learning. arXiv preprint arXiv:1910. 09933, 2019.

[100] Li S, Cheng Y, Wang W, et al. Learning to detect malicious clients for robust federated learning. arXiv preprint arXiv:2002. 00211, 2020.

[101] Meng L, Wei Y, Pan R, et al. VADAF: visualization for abnormal client detection and analysis in federated learning. ACM Transactions on Interactive Intelligent Systems, 2021, 11(3-4): 1-23.

[102] Zhao L, Hu S, Wang Q, et al. Shielding collaborative learning: Mitigating poisoning attacks through client-side detection. IEEE Transactions on Dependable and Secure Computing, 2020, 18(5): 2029-2041.

[103] Chen S, Xue M, Fan L, et al. Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach. computers & security, 2018, 73: 326-344.

[104] Kieu T, Yang B, Guo C, et al. Outlier detection for time series with recurrent autoencoder ensembles//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2019: 2725-2732.

[105] Kieu T, Yang B, Guo C, et al. Anomaly detection in time series with robust variational quasi-recurrent autoencoders//Proceedings of the IEEE International Conference on Data Engineering. Kuala Lumpur, Malaysia, 2022: 1342-1354.

[106] Paudice A, Muñoz-González L, Gyorgy A, et al. Detection of adversarial training examples in poisoning attacks through anomaly detection. arXiv preprint arXiv:1802.03041, 2018.

[107] Gao Sheng, Yuan Li-Ping, Zhu Jian-Ming, et al. A blockchain-based privacy-preserving asynchronous federated learning. SCIENTIA SINICA:Informatics, 2021, 51: 1755-1774 (in Chinese)
(高胜, 袁丽萍, 朱建明等. 一种基于区块链的隐私保护异步联邦学习. 中国科学: 信息科学, 2021, 51: 1755-1774.)

[108] Fang Chen, Guo Yuan-Bo, and Wang Yi-Feng, et al. Edge computing privacy protection method based on blockchain and federated learning. Journal on Communications, 2021, 42(11): 28-40 (in Chinese)
(方晨, 郭渊博, 王一丰等. 基于区块链和联邦学习的边缘计算隐私保护方法. 通信学报, 2021, 42(11): 28-40.)

[109] Lu Y, Huang X, Dai Y, et al. Blockchain and federated learning for privacy-preserved data sharing in industrial IoT. IEEE Transactions on Industrial Informatics, 2019, 16(6): 4177-4186.

[110] Awan S, Li F, Luo B, et al. Poster: A reliable and accountable privacy-preserving federated learning framework using the blockchain //Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. London, UK, 2019: 2561-2563.

[111] Shayan M, Fung C, Yoon C J M, et al. Biscotti: A blockchain system for private and secure federated learning. IEEE Transactions on Parallel and Distributed Systems, 2020, 32(7): 1513-1525.

[112] Zhu Jian-Ming, Zhang Qin-Nan, and Gao Sheng, et al. Privacy preserving and trustworthy federated learning model based on blockchain. Chinese Journal of Computers, 2021, 44, 468(12): 2464-2484 (in Chinese)
(朱建明, 张沁楠, 高胜等. 基于区块链的隐私保护可信联邦学习模型. 计算机学报, 2021, 44, 468(12): 2464-2484)

[113] Li Ling-Xiao, Yuan Sha, Jin Yin-Yu. Review of blockchain-based federated learning. Application Research of Computers, 2021, 38; 361(11): 3222-3230 (in Chinese)
(李凌霄, 袁莎, 金银玉. 基于区块链的联邦学习技术综述. 计算机应用研究, 2021, 38; 361(11): 3222-3230)

[114] ur Rehman M H, Salah K, Damiani E, et al. Towards blockchain-based reputation-aware federated learning//Proceedings of the IEEE International Conference on Computer Communications Workshops. Toronto, Canada, 2020: 183-188.

[115] Kumar S, Dutta S, Chatturvedi S, et al. Strategies for enhancing training and privacy in blockchain enabled federated learning//Proceedings of the IEEE International Conference on Multimedia Big Data. New Delhi, India, 2020: 333-340.

[116] Zeng R, Zeng C, Wang X, et al. A comprehensive survey of incentive mechanism for federated learning. arXiv preprint arXiv:2106. 15406, 2021.

[117] Kim H, Park J, Bennis M, et al. Blockchained on-device federated learning. IEEE Communications Letters, 2019, 24(6): 1279-1283.

[118] Bao X, Su C, Xiong Y, et al. Flchain: A blockchain for auditable federated learning with trust and incentive//Proceedings of the International Conference on Big Data Computing and Communications. QingDao, China, 2019: 151-159.

[119] Liu Y, Peng J, Kang J, et al. A secure federated learning framework for 5G networks. IEEE Wireless Communications, 2020, 27(4): 24-31.

[120] ur Rehman M H, Dirir A M, Salah K, et al. TrustFed: a framework for fair and trustworthy cross-device federated learning in IIoT. IEEE Transactions on Industrial Informatics, 2021, 17(12): 8485-8494.

[121] Behera M R, Upadhyay S, Shetty S. Federated learning using smart contracts on blockchains, based on reward driven approach. arXiv preprint arXiv:2107. 10243, 2021.

[122] Qu Y, Gao L, Luan T H, et al. Decentralized privacy using blockchain-enabled federated learning in fog computing. IEEE Internet of Things Journal, 2020, 7(6): 5171-5183.

[123] Pokhrel S R, Choi J. Federated learning with blockchain for autonomous vehicles: Analysis and design challenges. IEEE Transactions on Communications, 2020, 68(8): 4734-4746.

[124] Qi Y, Hossain M S, Nie J, et al. Privacy-preserving blockchain-based federated learning for traffic flow prediction. Future Generation Computer Systems, 2021, 117: 328-337.

[125] Zhao Y, Zhao J, Jiang L, et al. Privacy-preserving blockchain-based federated learning for IoT devices. IEEE Internet of Things Journal, 2020, 8(3): 1817-1829.

[126] Warnat-Herresthal S, Schultze H, Shastry K L, et al. Swarm Learning for decentralized and confidential clinical machine learning. Nature, 2021, 594(7862): 265-270

[127] Li X, Jiang P, Chen T, et al. A survey on the security of blockchain systems. Future Generation Computer Systems, 2020, 107: 841-853.

[128] Han S, Pool J, Tran J, et al. Learning both weights and connections for

efficient neural networks//Proceedings of the Neural Information Processing Systems. Montreal, Quebec, Canada, 2015: 1135-1143.

[129] Liu K, Dolan-Gavitt B, Garg S. Fine-pruning: Defending against backdooring attacks on deep neural networks//Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses. Crete, Greece, 2018: 273-294.

[130] Jiang Y, Wang S, Valls V, et al. Model pruning enables efficient federated learning on edge devices. IEEE Transactions on Neural Networks and Learning Systems, 2022,(1):1-13.

[131] Liu S, Yu G, Yin R, et al. Adaptive network pruning for wireless federated learning. IEEE Wireless Communications Letters, 2021, 10(7):1572-1576.

[132] Yu S, Nguyen P, Anwar A, et al. Adaptive dynamic pruning for Non-IID federated learning. arXiv preprint arXiv:2106. 06921, 2021.

[133] Tramèr F, Zhang F, Juels A, et al. Stealing machine learning models via prediction apis//Proceedings of the USENIX Conference on Security Symposium. Texas, USA, 2016: 601-618.

[134] Chandrasekaran V, Chaudhuri K, Giacomelli I, et al. Exploring connections between active learning and model extraction //Proceedings of the USENIX Conference on Security Symposium. Boston, USA, 2020: 1309-1326.

[135] Jagielski M, Carlini N, Berthelot D, et al. High accuracy and high-fidelity extraction of neural networks//Proceedings of the USENIX Conference on Security Symposium. Boston, USA, 2020: 1345-1362.

[136] Liu Y, Zhang X, Wang L. Asymmetrical vertical federated learning. arXiv preprint arXiv:2004.07427, 2020.

[137] Melis L, Song C, De Cristofaro E, et al. Exploiting unintended feature leakage in collaborative learning//Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA, 2019: 691-706.

[138] Ganju K, Wang Q, Yang W, et al. Property inference attacks on fully connected neural networks using permutation invariant representa-tions//Proceedings of the ACM SIGSAC conference on computer and communications security. Toronto, Canada, 2018: 619-633.

[139] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures//Proceedings of the ACM SIGSAC conference on computer and communications security. Denver, USA, 2015: 1322-1333.

[140] Zhang W, Tople S, Ohrimenko O. Leakage of Dataset Properties in Multi-Party Machine Learning//Proceedings of the USENIX Conference on Security Symposium. Virtual, 2021: 2687-2704.

[141] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139-144.

[142] Zhang J, Chen J, Wu D, et al. Poisoning attack in federated learning using generative adversarial nets//Proceedings of the IEEE International Conference On Trust, Security And Privacy In Computing And Communications. Rotorua, New Zealand, 2019: 374-380.

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

[144] Wang Z, Song M, Zhang Z, et al. Beyond inferring class representatives: User-level privacy leakage from federated learning//IEEE INFOCOM 2019-IEEE Conference on Computer Communications. Paris, France, 2019: 2512-2520.

[145] Zhang C, Li S, Xia J, et al. Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning//Proceedings of the USENIX Conference on Usenix Annual Technical Conference. Virtual, 2020: 493-506.

[146] Hardy S, Henecka W, Ivey-Law H, et al. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv preprint arXiv:1711.10677, 2017.

[147] Cheng K, Fan T, Jin Y, et al. Secureboost: A lossless federated learning framework. IEEE Intelligent Systems, 2021, 36(6): 87-98.

[148] Ou W, Zeng J, Guo Z, et al. A homomorphic-encryption-based vertical federated learning scheme for rick management. Computer Science and Information Systems, 2020, 17(3): 819-834.

[149] Cheon J H, Kim A, Kim M, et al. Homomorphic encryption for arithmetic of approximate numbers//Proceedings of the International Conference on the Theory and Application of Cryptology and Information Security. Hong Kong, China, 2017: 409-437.

[150] Chen H, Dai W, Kim M, et al. Efficient multi-key homomorphic encryption with packed ciphertexts with application to oblivious neural network inference//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. London, UK, 2019: 395-412.

[151] Mouchet C, Troncoso-Pastoriza J, Bossuat J P, et al. Multiparty homomorphic encryption from ring-learning-with-errors. Proceedings on Privacy Enhancing Technologies, 2021, 4: 291-311.

[152] Froelicher D, Troncoso-Pastoriza J R, Pyrgelis A, et al. Scalable privacy-preserving distributed learning. Proceedings on Privacy Enhancing Technologies, 2021, 2: 323-347.

[153] Sav S, Pyrgelis A, Troncoso-Pastoriza J R, et al. POSEIDON: Privacy-preserving federated neural network learning//Proceedings of the Network and Distributed Systems Security Symposium. Virtual, 2021:1-18.

[154] Ma J, Naas S A, Sigg S, et al. Privacy-preserving federated learning based on multi-key homomorphic encryption. International Journal of Intelligent Systems, 2022, 37(9): 5880-5901.

[155] Hsu R H, Wang Y C, Fan C I, et al. A privacy-preserving federated learning system for android malware detection based on edge computing//Proceedings of the Asia Joint Conference on Information Security. Taiwan, China, 2020: 128-136.

[156] Dong Ye, Hou Wei, Chen Xiao-Jun, et al. Efficient and secure federated learning based on secret sharing and gradients selection. Journal of Computer Research and Development, 2020, 57(10): 2241-2250 (in Chinese)

(董业, 侯炜, 陈小军等. 基于秘密分享和梯度选择的高效安全联邦学习. 计算机研究与发展, 2020, 57(10): 2241-2250).

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

[158] Bonawitz K, Eichner H, Grieskamp W, et al. Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems, 2019, 1: 374-388.

[159] Wang Z, Yang Y, Liu Y, et al. Cloud-based federated boosting for mobile crowdsensing. arXiv preprint arXiv:2005. 05304, 2020.

[160] Mohassel P, Rindal P. ABY3: A mixed protocol framework for machine learning//Proceedings of the ACM SIGSAC conference on computer and communications security. Toronto, Canada, 2018: 35-52.

[161] Patra A, Suresh A. BLAZE: blazing fast privacy-preserving machine learning. arXiv preprint arXiv:2005.09042, 2020.

[162] Rathee D, Rathee M, Kumar N, et al. CrypTFlow2: Practical 2-party secure inference//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Virtual, 2020: 325-342.

[163] Li Y, Xu W. PrivPy: General and scalable privacy-preserving data mining//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, USA, 2019: 1299-1307.

[164] Knott B, Venkataraman S, Hannun A, et al. Crypten: Secure multi-party computation meets machine learning//Proceedings of the International Conference on Neural Information Processing Systems. 2021, 34 : 4961-4973.

[165] Liu Y, Ma Z, Liu X, et al. Boosting privately: Federated extreme gradient boosting for mobile crowdsensing//Proceedings of the IEEE International Conference on Distributed Computing Systems. Singapore, 2020: 1-11.

[166] Dong Y , Chen X , Shen L , et al. EaSTFLy: Efficient and secure ternary federated learning. Computers & Security, 2020, 94: 101824.

[167] Chen C, Zhou J, Wang L, et al. When homomorphic encryption marries secret sharing: Secure large-scale sparse logistic regression and applications in risk control//Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Virtual, 2021: 2652-2662.

[168] Kaissis G , Ziller A , Passerat-Palmbach J , et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence, 2021:1-12.

[169] Hamm J, Cao Y, Belkin M. Learning privately from multiparty data//Proceedings of the International Conference on Machine Learning. New York, USA, 2016: 555-563.

[170] Papernot N, Abadi M, Erlingsson U, et al. Semi-supervised knowledge transfer for deep learning from private training data//Proceedings of the International Conference on Learning Representations, Toulon, France. 2017: 1-16.

[171] Papernot N, Song S, Mironov I, et al. Scalable private learning with pate//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-34.

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

[173] Geyer R C, Klein T, Nabi M. Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712. 07557, 2017.

[174] Truex S, Baracaldo N, Anwar A, et al. A hybrid approach to privacy-preserving federated learning//Proceedings of the ACM Workshop on Artificial Intelligence and Security. London, UK, 2019: 1-11.

[175] Triastcyn A, Faltings B. Federated learning with bayesian differential privacy//Proceedings of the IEEE International Conference on Big Data. Los Angeles, USA, 2019: 2587-2596.

[176] Rodríguez-Barroso N, Stipcich G, Jiménez-López D, et al. Federated learning and differential privacy: Software tools analysis, the Sherpa. ai FL framework and methodological guidelines for preserving data privacy. Information Fusion, 2020, 64: 270-292.

[177] Xin B, Yang W, Geng Y, et al. Private fl-gan: Differential privacy synthetic data generation based on federated learning//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain, 2020: 2927-2931.

[178] Wang C, Liang J, Huang M, et al. Hybrid differentially private federated learning on vertically partitioned data. arXiv preprint arXiv:2009.02763, 2020.

[179] Sabt M, Achemlal M, Bouabdallah A. Trusted execution environment: what it is, and what it is not//Proceedings of the IEEE International Conference on Trust,Security and Privacy in Computing and Communications. Helsinki, Finland, 2015, 1: 57-64.

[180] Hao M, Li H, Xu G, et al. Towards efficient and privacy-preserving federated deep learning//Proceedings of the IEEE International Conference on Communications. Shanghai, China, 2019: 1-6.

[181] Luo X, Zhu X. Exploiting defenses against GAN-based feature inference attacks in federated learning. arXiv preprint arXiv:2004. 12571, 2020.

[182] Zhuang R, DeLoach S A, Ou X. Towards a theory of moving target defense//Proceedings of the ACM Workshop on Moving Target Defense. Scottsdale, USA, 2014: 31-40.

[183] Liang B, Li H, Su M, et al. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. IEEE Transactions on Dependable and Secure Computing, 2018, 18(1): 72-85.

[184] Zrivestzhang19 V, Nicolae M I, Rawat A. Efficient defenses against adversarial attacks//Proceedings of the ACM Workshop on Artificial Intelligence and Security. Dallas, USA, 2017: 39-49.

[185] Rivest R L, Adleman L, Dertouzos M L. On data banks and privacy homomorphisms. Foundations of secure computation, 1978, 4(11): 169-180.

[186] Yang Ya-Tao, Zhao Yang, Zhang Juan-Mei, et al. Recent development of theory and application on homomorphic encryption. Journal of

Electronics and Information Technology, 2021, 43(2): 475-487 (in Chinese)
(杨亚涛, 赵阳, 张卷美等. 同态密码理论与应用进展. 电子与信息学报, 2021, 43(2): 475-487)

[187] Hall R, Fienberg S E, Nardi Y. Secure multiple linear regression based on homomorphic encryption. Journal of Official Statistics, 2011, 27(4): 669.

[188] Wu S, Teruya T, Kawamoto J, et al. Privacy-preservation for stochastic gradient descent application to secure logistic regression//Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence. Toyama, Japan, 2013, 27: 1-4.

[189] Xie W, Wang Y, Boker S M, et al. Privlogit: Efficient privacy-preserving logistic regression by tailoring numerical optimizers. arXiv preprint arXiv:1611.01170, 2016.

[190] Zhang J, Chen B, Yu S, et al. PEFL: A privacy-enhanced federated learning scheme for big data analytics//Proceedings of the IEEE Global Communications Conference. Hawaii, USA, 2019: 1-6.

[191] Lee J, Sun J, Wang F, et al. Privacy-preserving patient similarity learning in a federated environment: development and analysis. Journal of Medical Internet Research, 2018, 6(2): 1-21.

[192] Damgård I, Geisler M, Krøigaard M, et al. Asynchronous multiparty computation: Theory and implementation//Proceedings of the International workshop on public key cryptography. Berlin, Heidelberg, 2009: 160-179.

[193] Damgård I, Pastro V, Smart N, et al. Multiparty computation from somewhat homomorphic encryption//Proceedings of the Annual Cryptology Conference. Berlin, Germany, 2012: 643-662.

[194] Chen V, Pastro V, Raykova M. Secure computation for machine learning with SPDZ. arXiv preprint arXiv:1901.00329, 2019.

[195] Wan L, Ng W K, Han S, et al. Privacy-preservation for gradient descent methods//Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining. San Jose, USA, 2007: 775-783.

[196] Mohassel P, Zhang Y. Secureml: A system for scalable privacy-preserving machine learning//Proceedings of the IEEE symposium on security and privacy. San Jose, USA, 2017: 19-38.

[197] Wagh S, Gupta D, Chandran N. SecureNN: 3-Party secure computation for neural network training. Proceedings on Privacy Enhancing Technologies, 2019, 2019(3): 26-49.

[198] Agrawal N, Shahin Shamsabadi A, Kusner M J, et al. QUOTIENT: Two-party secure neural network training and prediction//Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. London, UK, 2019: 1231-1247.

[199] Wagh S, Tople S, Benhamouda F, et al. Falcon: Honest-majority maliciously secure framework for private deep learning. Proceedings on Privacy Enhancing Technologies, 2021, 1: 188-208.

[200] Vaidya J, Clifton C. Privacy preserving naive bayes classifier for vertically partitioned data//Proceedings of the international conference on data mining. Society for Industrial and Applied Mathematics, Lake Buena Vista, USA, 2004: 522-526.

[201] Vaidya J, Clifton C, Kantarcioglu M, et al. Privacy-preserving decision trees over vertically partitioned data. ACM Transactions on Knowledge Discovery from Data, 2008, 2(3): 1-27.

[202] Romanini D, Hall A J, Papadopoulos P, et al. Pyvertical: A vertical federated learning framework for multi-headed splitnn. arXiv preprint arXiv:2104.00489, 2021.

[203] Lu L, Ding N. Multi-party private set intersection in vertical federated learning//Proceedings of the IEEE International Conference on Trust, Security and Privacy in Computing and Communications. Guangzhou, China, 2020: 707-714.

[204] Jayaraman B, Evans D. Evaluating differentially private machine learning in practice//Proceedings of the USENIX Security Symposium. California, USA, 2019: 1895-1912.

[205] Zhu L, Liu Z, Han S. Deep Leakage from Gradients//Proceedings of the International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 14774-14784.

[206] Uchida Y, Nagai Y, Sakazawa S, et al. Embedding watermarks into deep neural networks//Proceedings of the ACM on International Conference on Multimedia Retrieval. Bucharest, Romania, 2017: 269-277.

GAO Ying, Ph.D., associate professor. Her main research interests include privacy computing and blockchain.

CHEN Xiao-Feng, Ph.D. candidate. His main research interests include blockchain and federated learning.

ZHANG Yi-Yu, M. S. candidate. His main research interests include federated learning and privacy computing

WANG Wei, M. S. candidate. Her main research interest is private set intersection and federated learning.

DENG Huang-Hao, B. S. candidate. His main research interests include federated learning.

DUAN Pei, M. S. His main research interest is data mining and machine learning.

CHEN Pei-Xuan, M. S. His main research interest is distributed computing and machine learning.

Background

With AI becoming an emerging technology leading the next generation of industrial change, a large number of innovative applications are integrated into our daily life, such as smart cities, autonomous driving, and vehicle of internet. Federated learning (FL), as a new branch of artificial intelligence technology, can unleash the potential value of data by conducting efficient joint modeling and model training among multiple participants without leaving local privacy data. FL effectively protects the privacy of local data by keeping the data of participants local and uploading only model parameters to the server. However, federated learning is still in the initial stage of research, on the one hand because of its relatively new technical concept and inadequate architecture, and on the other hand because it is subject to various privacy and attack threats that seriously affect its further development. The existing FL systems have been proved to have potential threats in data collection stage, training stage and inference stage, endangering the privacy of data and the robustness of the system. Many researchers have conducted research and analysis based on the differences in threat and concealment of attacks, and have produced a large number of effective defense solutions. So far, researchers have reviewed the main technical challenges in federated learning, as well as the security and privacy protection in federated learning. However, there is a lack of systematic and comprehensive work on the threats and defenses against attacks in federated learning. Therefore, this paper focuses on various attack threats and defensive measures in federated learning, and we systematically categorize and analyze the attacks and defensive measures by sorting out the large amount of current research results.

Starting with two kinds of potential threats: security threat and privacy threat, we give a detailed definition of security attributes in FL scenarios around confidentiality, integrity and availability (CIA triplet), and summarize various attack methods and defense means in FL systematically and comprehensively. Firstly, we summarize the horizontal and vertical federated learning (VFL) process and potential threats respectively, and analyze the basic concepts, implementation stages and existing schemes of common attacks such as poisoning attack, sample attack and inference attack from the perspectives of antagonistic attack and non-antagonistic attack. Adversarial attacks include poisoning attacks, adversarial sample attacks, free-riding attacks, Sybil attacks, and attacks against communication bottlenecks. Non-adversarial attacks include model extraction attacks, inference attacks, and GAN-based attacks. Further, according to different attack methods, defense means are divided into two categories: robustness enhancement methods and privacy enhancing technologies. The robustness enhancement methods mainly defend against antagonistic attacks, including data sanitization, robustness aggregation, anomaly detection, countermeasure training, knowledge distillation, pruning and other methods. The privacy enhancing technology mainly defends the system against non-antagonistic attacks, including homomorphic encryption, secure multi-party computing, differential privacy and blockchain. And the schemes related to robustness enhancement methods and privacy enhancement techniques in FL are sorted out and summarized. Finally, the paper gives future research direction of robustness and privacy in FL: 1) Establish a secure and stable attack detection and evaluation model, endow FL system with self inspection and evaluation capabilities, and provide real-time protection for internal and external environments; 2) Analyze and infer all possible potential attacks and privacy issues, and build a perfect security attack and defense system based on security encryption technology; 3) Study the unique attack and defense in VFL to solve the bottleneck problem of VFL in practical application; 4) Explore the conflict between robustness and privacy in FL to promote large-scale applications.

This work is supported by Natural Science Foundation of Beijing Municipality (M21033); National Natural Science Foundation of China (61932011, 61972017) and Tencent Rhino-Bird Joint Research Program.