联邦学习拜占庭攻击与防御研究综述

孙 钰[1,2]，刘霏霏[1,2]，李大伟[1,2*]，刘建伟[1,2]
（1. 北京航空航天大学网络空间安全学院，北京 100191；
2. 北京航空航天大学空天网络安全工业和信息化部重点实验室，北京 100191）

摘要：为解决“数据孤岛”和隐私泄露问题，联邦学习将训练任务部署在多个客户端进行本地训练。然而，分布式训练环境易受拜占庭攻击，拜占庭敌手可以同时控制多个客户端，并以投毒方式直接影响全局模型性能。针对联邦学习中的拜占庭攻击和防御进行全面分析和总结，首先根据有无梯度保护将联邦学习分为普通联邦学习和隐私保护联邦学习，介绍了联邦学习在拜占庭攻击方面面临的威胁和挑战，梳理其安全模型中的敌手能力和攻击策略。然后根据技术路线对现有防御策略进行分类和对比，并分析可扩展到安全隐私联邦学习中的技术路线。最后，对几种实际情况下的拜占庭防御策略进行展望。

关键词：联邦学习；拜占庭攻击；安全与隐私；密码学；鲁棒性
中图分类号：TP391 **文献标识码**：A

Survey on Byzantine Attacks and Defenses in Federated Learning

SUN Yu[1,2], LIU Feifei[1,2], LI Dawei[1,2*], LIU Jianwei[1,2]
(1. School of Cyber Science and Techonology, Beihang University, Beijing 100191, China;
2. Key Lab. of Ministry of Industry and Information Technology for Cyberspace Security, Beihang University, Beijing 100191, China)

Abstract: To solve the problems of data island and privacy leakage, federated learning (FL) deploys training tasks to multiple clients for local training individually. However, distributed training environment is prone to Byzantine attacks, where adversaries can control multiple clients simultaneously and affect global model by a poisoning method. The comprehensive analysis and summary of Byzantine attacks and defense in FL are achieved. Firstly, the FL is classified into ordinary and privacy protection types with or without the gradient protection. Secondly, The threats and challenges of Byzantine attacks in FL are introduced, the capabilities and attack strategies of Byzantine adversaries in the security model are sorted out. Finally, according to the technical routes, existing defense strategies are classified and compared to be extended to the technical routes in the safety and privacy protection FL, which prospects several practical Byzantine defensive strategies.

Keywords: FL (federated learning); byzantine attack; safety and privacy; cryptography; robustness

0 引言

机器学习在自然语言处理、计算机视觉、推荐系统等领域取得了巨大的成功，并不断推动传统产业和社会生活的变革[1-2]。机器学习是数据驱动的，训练过程需要融合不同用户的本地数据。然而，随着欧盟《数据隐私保护条例》，美国《美国数据隐私和保护法》，我国《网络安全法》《数据安全法》和《个人信息保护法》等保护数据隐私的法规颁布实施，“数据孤岛”制约了大数据建模。为破解隐私保护与数据要素流动相悖之局，方滨兴院士提出“数据不动程序动、数据可用不可见”的数据要素安全流动机理[3]。作为一种“保留数据所有权，释放数据使用权”的机器学习方法，联邦学习（FL，federated learning）[4]使得多个参与方利用本地数据，通过与服务器交互参数来训练共享模

基金项目：国家重点研发计划（2021YFB2700200）；国家自然科学基金（U21B2021；61972018；61932014；62002006）。

引用格式：孙 钰，刘霏霏，李大伟，等. 联邦学习拜占庭攻击与防御研究综述 [J]. 网络空间安全科学学报，2023，1(1)：17－37.

型[5]，解决了因直接传输敏感数据带来的隐私泄露问题[6]。自此，联邦学习引发了各界的强烈关注和广泛研究，在安全、效率等方面取得了巨大的进步，形成了 FATE[7]、PaddleFL[8]、TensorFlow Federated[9]、PySyft[10] 等开源框架，并在工业[11]、医疗[12-13]、物联网[14]、金融[15-16] 等领域涌现出许多实际应用。

然而，将训练过程分布在多个机器上将不可避免地导致故障和风险，包括进程崩溃、计算错误、网络延迟，同时也增加了暴露的攻击面。除了常见的单一节点攻击，拜占庭攻击逐渐成为鲁棒联邦学习优化的研究重点。拜占庭攻击[17]是分布式系统中常见的安全问题，它指攻击者通过控制系统中若干授权节点来干扰整个系统的攻击方式[18]。在联邦学习中，拜占庭敌手在训练阶段进行攻击，可同时控制多个用户，具备篡改本地数据、修改上传梯度等能力，抗拜占庭故障是对健壮联邦学习系统的最高要求。检测拜占庭敌手通常需要仔细辨别用户上传的梯度或模型，比较梯度之间的统计特性或距离。然而，随着梯度反演的攻击能力不断增强，梯度保护和安全聚合逐渐成为隐私保护联邦学习（PPFL，privacy-preserving federated learning）的研究趋势。如何应对梯度分析与隐私保护的冲突，对保护后的梯度进行拜占庭敌手检测成为了新的研究热点。

拜占庭攻防博弈是联邦学习安全研究的重要问题。文献［19－20］仅涉及少数拜占庭鲁棒的安全聚合算法，文献［21］着重分析了经典拜占庭鲁棒的安全聚合算法在多种攻击场景下的利弊，而文献［22－23］仅对普通联邦学习下的拜占庭防御进行分类。综上，现有综述缺少拜占庭攻击方法介绍，也未考虑梯度保护对拜占庭防御带来的全新挑战。

针对以上不足，本文主要贡献如下。

1）本文增加了普通联邦学习下拜占庭防御的最新研究成果，根据技术路线对现有分类进行调整和扩展。

2）本文分析了拜占庭威胁模型下敌手的攻击能力和背景知识，详细阐述了不同攻击策略下具体攻击方法。

3）本文归纳了隐私保护联邦学习下的防御策略，并分析了可扩展到隐私保护联邦学习的防御方法。

1 联邦学习的拜占庭攻击

1.1 联邦学习基本概念

1.1.1 普通联邦学习

联邦学习是一种分布式机器学习技术，它允许多个计算机使用本地数据训练，在服务器的协调下通过分享模型更新来集成学习效果，以提高安全性和模型性能。根据数据集的分布特点，联邦学习可以分为横向联邦学习、纵向联邦学习和联邦迁移学习。本文主要总结面向横向联邦学习的拜占庭攻击与防御方法[24]联邦学习的结构以及训练过程如图 1 所示。联邦学习通常由 N 个用户和一个服务器组成，各用户在设备性能、数据集分布、网络连通性等方面存在一定差异，服务器一般具有较强的稳定性。用户在服务器的协调下有序参与训练，使用本地私有数据训练本地模型，然后将本地模型参数发送给服务器。服务器聚合参数，更新全局模型作为新一轮的共享模型，下发给下一轮参与训练的设备，以此进行迭代训练，直到全局模型达到要求。

注：添加到我的文章中，综述问题！！！

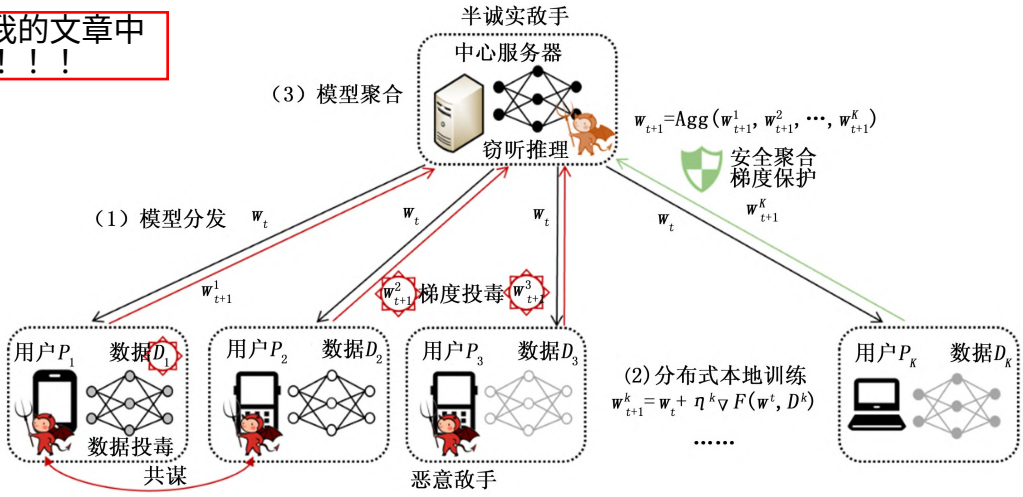

图 1 联邦学习架构及拜占庭攻击威胁模型

Fig. 1 Framework of Federated Learning and Threat Model of Byzantine Attack

具体流程如下：

1）模型分发。服务器从 N 个用户 $P_1,P_2,\cdots,P_N$ 中随机选择 K 个（$0<K<N$）参与训练，并分发当前第 t 轮的全局模型 w_t。

2）分布式本地训练。以用户 k 为例，利用本地数据 D_k 对目标函数 $F(w_t,D_k)$ 进行优化。一般使用随机梯度下降（SGD，stochastic gradient descent）[25]或小批量梯度下降(mini-batch GD)算法，计算本地数据在目标函数上的梯度 $\nabla F(w_t,D_k)$，以学习率 η^k 得到更新后的本地模型参数 $w_{t+1}^k = w_t+\eta^k\nabla F(w_t,D_k)$ 发送给服务器。这里，参与者既可以上传本地模型参数，也可以上传梯度信息。

3）模型聚合。服务器收集参与者本地更新的模型参数 $\{w_{t+1}^k,k=1,2,\cdots,K\}$，利用聚合规则 Agg 对参与者本地模型参数进行聚合 $w_{t+1}=\mathrm{Agg}(w_{t+1}^1,w_{t+1}^2,\cdots,w_{t+1}^K)$，将下一轮的全局模型参数 w_{t+1} 返回给参与者。

服务器端聚合规则 Agg 一般为 FedSGD 和 FedAvg[4]。FedSGD 中参与方将本地数据的平均梯度 g^k 传给服务器，服务器以各用户数据量 n_k 占全局数据量 n 的比例为权重聚合梯度 $w_{t+1}\leftarrow w_t-\eta\sum_{k=1}^{K}\frac{n_k}{n}g^k$ 返回给用户。而 FedAvg 允许参与方在服务器聚合参数之前多次迭代计算梯度值，再交由服务器根据用户数据量占全局数据量的比例进行聚合 $w_{t+1}\leftarrow\sum_{k=1}^{K}\frac{n_k}{n}w_{t+1}^k$，服务器不必每次计算中间结果的均值，可减少通信轮数。后续也出现了其它变种算法对 FedAvg 的参数聚合进行改进，以应对数据非独立同分布（IID，independent and identically distributed）且数据量不平衡的情况，如 FedSVRG[26]、FedProx[27]、FedNova[28]等。

1.1.2　隐私保护联邦学习

联邦学习的本地训练特性在一定程度上保护了数据隐私，但用户与服务器的梯度交互也暴露出了新的攻击面。从输入与标签中推导而来的梯度隐藏着可用于反演原始数据的重要信息。早期的梯度反演仅能对数据集成员、类别等信息进行推理[29-31]，而 2019 年出现的深度梯度反演攻击[32-36]可从梯度恢复出原始数据及其标签，完全暴露了参与方的敏感数据，更具威胁性。

针对机器学习的推理攻击最早由 Shokri 等人[37]提出，敌手以黑盒方式访问模型，仅以查询方式就能推断出某一样本是否属于训练目标模型的数据集。Melis 等人面向联邦学习场景首次提出成员推理攻击[29]，恶意用户利用全局模型的梯度更新和训练集辅助信息可推理其余良性用户与主训练任务无关的属性，甚至诱导联合模型学习自己感兴趣的特征。相比推理攻击，反演攻击可恢复出数据集的标签和原始数据，完全暴露了参与方的敏感信息。Zhu 等人首次提出深度梯度泄露（DLG，deep leakage from gradients）攻击[32]，在没有关于训练数据先验知识的情况下，通过最小化目标梯度和虚拟梯度之间的距离来完全重建私有训练数据，可在 CIFAR100[38]数据集上近乎完美地恢复出单张输入。Zhao 等人提出一种从公开梯度重构标签的攻击 iDLG[33]，通过交叉熵分类层中的负梯度可恢复出真实标签。

随着梯度攻击技术的演进，攻击者从深层网络恢复高精度输入的能力不断增强，梯度保护已成为联邦学习的发展趋势[39-41]。根据梯度保护方式，目前学术界普遍将其分为梯度扰动和基于密码技术的梯度加密[42]。梯度扰动主要使用差分隐私（DP，differential privacy）[43-48]、梯度压缩[32,49-50]等技术对原始梯度进行有损扰动，通过限制梯度泄露的信息来保护梯度隐私。然而，最新研究表明，诸如加噪、压缩、稀疏的梯度扰动方法并不能够完全保护隐私[51]，具有固定隐私参数的差分隐私方法易受梯度泄露攻击[52]，半诚实服务器仍可对高度压缩后的梯度进行反演攻击[53]。

基于密码技术的梯度加密方法主要使用同态加密（HE，homomorphic encryption）[54-57]、秘密分享[58-62]等技术对上传的梯度进行保护，可抵御半诚实服务器的梯度反演攻击，真正保护用户隐私。同态加密可以在不解密数据的情况下，对加密数据进行计算，得到与对应明文进行同样计算相同的结果。秘密分享技术将秘密信息拆分成若干份，交由不同的参与者管理，只有若干参与者一同合作才能恢复出秘密信息。Bonawitz 等人[58]利用秘密分享机制为梯度添加掩码，保护梯度隐私。初始阶段，可信中心（TA，trust authority）为用户和服务器生成各自的密钥，用户广播其公钥。秘密分享阶段，利用秘密分享机制[63]，用户 i 为随机值 β_i 和私钥 N_i^{SK} 生成与用户 j 的秘密分享 $\beta_{i,j}\leftarrow share(\beta_i)$，$N_{i,j}^{SK}\leftarrow share(N_i^{SK})$，计算与用户 j 的 Diffie-Hellman 密钥 $p_{i,j}\leftarrow DH(P_i^{SK},P_j^{PK})$，将加密后的秘密分享 $\Omega_{i,j}\leftarrow Enc(p_{i,j},\beta_{i,j}\,||\,N_{i,j}^{SK})$ 广播给服务器和其他用户。同时，用户 i 收到用户 j 发来的 $\Omega_{j,i}$，计算与每一个用户

j 的共同秘密 $s_{i,j} \leftarrow DH(N_i^{SK}, N_j^{PK})$，将 $s_{i,j}$ 作为伪随机数生成器 PRG()的种子，对梯度 x_i 添加成对加性掩码，发送给服务器，如公式（1）所示：

$$y_i = x_i + PRG(\beta_i) + \sum_{i<j} PRG(s_{i,j}) - \sum_{i>j} PRG(s_{j,i}) \quad (1)$$

解码阶段，仍在线的用户发送 $\beta_{i,j}$ 和上一轮退出用户的 $N_{i,j}^{SK}$，服务器恢复出秘密值 $\beta_i \leftarrow \mathrm{Rec}(\{\beta_{i,j}\})$ 和 $N_i^{SK} \leftarrow \mathrm{Rec}(\{N_{i,j}^{SK}\})$，计算 $s_{i,j} \leftarrow DH(N_i^{SK}, N_j^{PK})$。然后，服务器聚合加性掩码，如公式（2）所示：

$$\sum x_i = \sum y_i - \sum PRG(\beta_i) - \sum_{i<j} PRG(s_{i,j}) + \sum_{i>j} PRG(s_{j,i}) \quad (2)$$

由于加性掩码可互相抵消，服务器只能获得聚合后的全局梯度。该方案利用 Shamir 门限的特性，只有用户数量满足门限时掩码才能抵消，还可解决实际应用中可能存在的用户中途掉线问题。

对梯度进行加密虽然保护了用户隐私，但服务器只能获取密文聚合结果。密文完全打乱了明文数据的统计特征，聚合操作也掩盖了敌手的投毒痕迹，给拜占庭攻击检测带了新的挑战。

1.2 拜占庭攻击威胁模型

1.2.1 拜占庭攻击

在联邦学习中，拜占庭敌手被定义为可完全控制多个用户的设备、数据，并可上传任意数据的敌手[39]。联合训练通常涉及数以千计的参与者，其中会不可避免地存在不可靠的恶意参与者。由于分布式训练对服务器不可见，服务器无法确定每个参与用户的合法性以及训练过程的合规性。利用这一安全漏洞，拜占庭敌手可以完全控制多个参与者，进行分布式投毒，包括篡改本地训练数据、操纵模型训练[19]，最终以毒化梯度的形式直接影响全局模型[64]。拜占庭攻击可视为分布式训练下的投毒攻击，其目的旨在干扰全局模型收敛方向，对梯度进行保护只是加大了拜占庭攻击检测难度，并不影响敌手的攻击策略和攻击效果。拜占庭敌手还可能攻击设备和通信网络，造成软件故障、系统故障、网络拥塞等[65]，但此类攻击不在本文讨论范围。

Blanchard 等人[65]认为拜占庭敌手可共谋并发送任意偏离正确梯度的值，当服务器使用基于梯度线性组合的聚合算法时，仅需一个恶意用户就可以随意操纵全局模型，并最终损害全局模型的性能[66]。Shi 等人[22]提出权重攻击（weight attack）。多数联邦学习方案中，服务器根据用户所声称训练数据集大小分配权重，却无法验证该声明的真实性。利用这一缺陷，敌手可以夸大其数据集大小以获得更高的权重。Fang 等人[67]针对现有的拜占庭鲁棒联邦聚合方案提出了新的攻击框架，该框架以训练阶段局部模型投毒的方式，降低全局模型在测试集的准确率，并最终导致模型拒绝服务。该方法将攻击视为一个将全局模型向相反方向收敛的优化问题，优化用户端发送的本地投毒梯度，可产生统计特征上接近良性模型的毒化局部模型。Shejwalkar 等人针对敌手是否知晓服务器聚合规则的两种情况，提出了模型投毒攻击的通用框架[68]，改进了[67]在高度不平衡的 IID 数据集的攻击效果。

1.2.2 安全模型

在联邦学习架构与威胁模型中，不同拜占庭敌手发动攻击时可能具备不同的攻击能力和背景知识。

在安全模型中，根据敌手攻击能力，通常可分为半诚实敌手（curious but honest/semi-honest）和恶意敌手。半诚实敌手会按照规定执行训练流程，但仍好奇其他用户的梯度和数据，会根据所接收信息推理其他用户的隐私[29, 69-70]。在 PPFL 中，一般认为服务器是半诚实的（标号①，与表 1、2 一致，下同）。恶意敌手则具有违反协议流程的任何行为，如干预训练过程，实施数据投毒（②）和模型投毒（③）。拜占庭敌手为恶意敌手，主要攻破参与学习任务的用户，由于它们掌握并频繁交换系统中的重要隐私数据和训练信息，一旦被敌手攻破将造成严重的安全和隐私问题[71]。虽然拜占庭敌手的威胁主要体现在主动攻击上，但一些 PPFL 方案的安全模型中仍考虑了具有窃听能力的半诚实用户（④）。拜占庭敌手还可能获知系统中的额外信息，如每次迭代中部分甚至所有良性用户的本地训练数据和本地模型更新（⑤），以及服务器采用的聚合规则[67]（⑥）。除此之外，拜占庭敌手之间还可能共谋（collude）[72]，如恶意用户之间共谋[65, 73]（⑦）、恶意用户与服务器共谋[74-76]（通常是双服务器架构下的其中一个服务器）（⑧）。部分拜占庭敌手为了躲避检测，会在不同训练轮数中更改其数量，即更改投毒率（⑨）。

1.2.3 攻击手段

1）数据投毒。数据投毒最早由 Biggio [77]提出，通过修改本地数据集的标签、特征、分布等，进而影响全局模型的收敛方向，降低其准确率。其实现方式包括：

表 1　普通联邦学习的拜占庭防御策略对比

Table 1　Comparisons of Byzantine Tolerance Methods in Common Federated Learning

分类	文献	安全模型	全局模型计算方法	敌手数	数据分布	时间复杂度
基于梯度统计特性	Chen 等人[88]	③⑦	平均几何中位数	$K\geqslant 2f+2$	IID	$Kd\log K$
	Yin 等人[89]	③⑥⑦	坐标中位数；坐标截断均值	$K\geqslant 2f+1$	IID	$Kd\log K$
	Xie 等人[72]	③⑤⑥⑦	几何中位数；坐标中位数；中位数平均	$K\geqslant 2f+1$	IID/非 IID	$Kd\log K$
	SLSGD[90]	②③⑨	局部 SGD；截断均值	$K\geqslant 2f+1$	IID/非 IID	$Kd\log K$
	Data 等人[91]	③⑦⑨	剔除恶意梯度后均值	$K\geqslant 3f$	IID/非 IID	$O(K^2d)$
	RFA [92]	①②③	几何中位数	$K\geqslant 2f+1$	IID	$O(Kd)$
	DnC[68]	③⑤	FedAvg	$K\geqslant 2f+1$	IID/非 IID	$O(K^2d)$
	Draco[95]	③⑥⑦	小批量 SGD；几何中位数	$r\geqslant 2f+1$	IID	$O(Kd)$
	Sto-SignSGD [98]	③	多数投票的 FedAvg	$K\geqslant 2f+1$	IID/非 IID	$O(Kd)$
基于梯度间距离	Auror [99]	①②⑦	分布式选择 SGD	$K\geqslant 2f+1$	IID	$O(Kd)$
	Multi-Krum[65]	③⑥⑦	FedAvg	$K\geqslant 2f+3$	IID	$O(K^2d)$
	FABA [100]	③⑥	FedAvg	$K\geqslant 2f+1$	IID	$O(K^2d)$
	Bulyan [66]	③⑤⑥	截断均值	$K\geqslant 4f+3$	IID	$O(K^2d)$
	Sniper [101]	②③⑦	FedAvg	$K\geqslant 3f+1$	IID	$O(K^2d)$
	FoolsGold[73]	③⑥⑤⑦	自适应学习率的 SGD	$K\geqslant f+1$	IID/非 IID	$O(K^2d)$
	AFA[103]	③⑦	自适应 FedAvg	$K\geqslant 2f+1$	IID	$O(Kd)$
	TDFL[104]	②③	真值发现算法	$K\geqslant f+1$	IID/非 IID	$O(Kd)$
	FedCom[105]	②③⑤⑥⑦	优化的 FedAvg	$K\geqslant 2f+1$	IID/非 IID	$O(K^2d)$
	Liu 等人[102]	③⑨	FedAvg	$K\geqslant 2f+1$	IID	$O(K^2M)$
基于额外验证数据	Li 等人[106]	③	FedAvg	$K\geqslant 2f+1$	IID/非 IID	$O(Kd)$
	Zeno[107]	②③	截断均值	$K\geqslant f+1$	IID/非 IID	$O(Kd)$
	Cao 等人[108]	③	优化的 FedAvg	$K\geqslant f+1$	IID	$O(Kd)$
	Cronus [112]	③	自适应的 SGD	$K\geqslant 2f+1$	IID	$O(d_p^3+K)$
	FLTrust[122]	③⑤⑥⑦	以余弦相似度为权重的聚合	$K\geqslant f+1$	IID/非 IID	$O(Kd)$
	BRCA[110]	③	可信度与上一轮全局梯度加权	$20\%K\geqslant f$	IID/非 IID	$O(Kd)$
	GAA [111]	③⑨	基于策略选择的 FedAvg	$K\geqslant f+1$	IID	$O(K^3d)$
基于优化算法补偿	RSA [84]	②③	带正则项优化的 FedAvg	$K\geqslant f+1$	IID/非 IID	$O(Kd)$
	Mhamdi 等人[113]	③⑥⑦	动量 SGD	$K\geqslant 2f+3$	IID	$O(K^2d)$
	Karimireddy 等人[114]	②③	迭代裁剪动量 SGD	$K\geqslant 2f+1$	IID	$O(Kd)$
基于差分隐私	FLAME[120]	②③⑥⑨	梯度平均	$K\geqslant 2f+1$	IID/非 IID	$O(K^2d)$
	FL-WBC[121]	②③	FedAvg	$K\geqslant 2f$	IID/非 IID	$O(Kd)$

K 为参与训练用户数；d 为梯度维度；f 为拜占庭敌手数；r 为冗余比；M 为标签类别数量；d_p 为更新的维度

表 2　面向隐私保护联邦学习的拜占庭防御策略对比

Table 2　Comparisons of Byzantine Tolerance Methods in Privacy-preserving Federated Learning

分类	文献	安全模型	恶意梯度剔除方法	敌手数	数据分布	时间复杂度
基于安全多方计算	He 等人[74]	①③⑦⑧	Multi-Krum	$K\geqslant 2f+1$	IID/非 IID	$O(K^2d)$
	FLOD[75]	①③⑧	汉明距离	$K\geqslant f+1$	IID	$O(Kd\log d)$
	ShieldFL[124]	①③⑦	余弦相似度	$K\geqslant 2f$	IID/非 IID	$O(Kd)$
	PBFL[125]	①②③⑦	余弦相似度	$K\geqslant f+1$	IID	$O(Kd)$
	PEFL[126]	①②③	坐标中位数	$K\geqslant 2f+1$	IID	$O(Kd\log d)$
	DisBezant[76]	①③④⑦⑧	余弦相似度	$40\%K\geqslant f$	IID/非 IID	$O(K^2d)$
	Omega[127]	①③④⑦	Multi-Krum	小部分	IID	$O(K^2d)$
	SFAP[128]	①②③④	均方误差	$K\geqslant f+1$	IID/非 IID	$O(Kth)$
基于可信执行环境	SEAR[135]	①③⑦	Multi-Krum	$K\geqslant 2f+1$	IID	$O(Kd\log d)$
	Muhr 等人[136]	①②	聚类	$10\%K\geqslant f$	IID	$O(K^2d)$
基于可度量加性掩码	BREA[137]	①③⑦	Multi-Krum	$K\geqslant 2f+1+[m+2,D+2T]_{\max}$	IID/非 IID	$\approx O(K^2\log K)$
	ADFL [138]	③	梯度重叠度	$K\geqslant f+2$	IID	$O(K^2)$

K 为参与训练用户数；d 为梯度维度；f 为敌手数；D 为掉线用户数；m 为聚合模型数；T 为编码多项式最高次数；t 为决策树节点数；h 为决策树高度

①标签翻转（label-flipping）[77-78]：保持训练样本的特征不变，将其标签 i 修改为另一类别 j，让模型在训练时学习到错误的对应关系。Tolpegin 等人对标签翻转攻击进行了详细的实验分析[78]，验证了其对联邦学习安全性的影响，发现可以通过提高敌手在后期迭代训练中的比例增强攻击效果。针对敌手训练集没有正确标签的情况，Zhang 等人[79]利用生成对抗网络（GAN，generative adversarial nets），将每轮聚合后的全局模型作为判别网络，生成模仿目标标签数据的对抗样本，并基于对抗样本进行标签翻转攻击。

②植入后门触发器：机器学中，植入后门触发器是一种常见的数据投毒方式[80-81]。攻击者对用户本地训练数据进行篡改，添加能触发分类器某种错误行为的特征，并修改相应标签。如在人脸识别任务中植入“黑框眼镜”作为后门触发器[80]，分类器一旦识别到“黑框眼镜”则一律将该输入误识别成人物 Alice，而对含有“墨镜”的输入或其他任务分类正常。然而，研究[82]表明由于联邦学习中服务器将聚合多个良性用户的模型，针对机器学习的触发器植入并不能对联邦学习造成严重后果。对此，Xie 等人提出更适合联邦学习场景的分布式数据投毒攻击[83]。相比植入同一触发器的集中式投毒，分布式投毒的各攻击者在本地数据集植入了不同的触发器，具有更好的投毒效果。然而，植入后门触发器需要同时在训练和预测阶段访问数据集，实施难度较大。

2）模型投毒。模型投毒在模型上传阶段，通过上传恶意模型或梯度影响全局模型准确率，无需获取用户本地数据集，相比数据投毒更为直接和高效[82]。其实现方式包括：

①符号翻转（sign-flipping）[84]：这是模型投毒的一种最简单的实现方式，即对原始梯度的符号取反后上传。

②扩大幅值：为了提高恶意梯度在简单聚合过程（如 FedAvg）中的贡献，攻击者往往通过扩大恶意梯度的幅值来改变全局模型的收敛方向。

③同值攻击（same-value attacks）[84]：敌手将所有模型参数都修改为某一值提交给服务器。

④梯度篡改[72, 84]：敌手可以任意篡改局部模型任意位置的参数。如文献［148］提出的 OBLIVION 攻击利用了模型灾难性遗忘特性，选择对全局模型影响最大的梯度进行篡改，然后根据恶意用户数量进行梯度动态平滑，大大增强了投毒效果。

⑤Krum/Trim 攻击[67]：利用 Krum 和 Trim 聚合规则对于恶意梯度应远离均值的这一假定，在多轮迭代中为恶意梯度的每一维参数添加少量偏差，构建一个与其他局部模型最接近的毒化模型，使偏差在聚合算法的浮动范围内，但仍能干扰全局模型收敛。

⑥植入语义后门[82, 85-86]：区别于数据投毒中的植入后门触发器，语义后门无需访问数据，通过修改梯度就可实现。如文献［82］中，敌手拟用恶意梯度 X 代替全局梯度 G^{t+1}，在全局模型即将收敛时上传恶意梯度 $\widetilde{L}_m^{t+1}=\frac{n}{\eta}X-(\frac{n}{\eta}-1)G^t-\sum_{i=1}^{m-1}(L_i^{t+1}-G^t)$，其中 η 为学习率。由于全局模型即将收敛时各用户上传梯度之间的偏差越来越小，即 $\sum_{i=1}^{m-1}(L_i^{t+1}-G^t)\approx 0$，因此 $\widetilde{L}_m^{t+1}\approx\frac{n}{\eta}X-(\frac{n}{\eta}-1)G^t$，则服务器按照 $G^{t+1}=G^t+\frac{\eta}{n}\sum_{i=1}^{m}(\widetilde{L}_i^{t+1}-G^t)$ 规则聚合时，恶意梯度 $X\approx G^t+\frac{\eta}{n}(\widetilde{L}_m^{t+1}-G^t)$ 就成为了全局梯度，因此这种攻击被称为模型替换（model replacement）攻击。若敌手不知道 n 和 η，可以通过在每轮迭代缩放比例因子 $\gamma<\frac{n}{\eta}$ 实现良好的后门攻击准确性，此时这种攻击方式被称为缩放攻击（scaling attack）。

⑦PGD 边缘攻击（projected gradient descent edge-case attacks）[86]：将客户端数据集与边缘数据集（概率分布小于一定阈值的数据）混合，使用投影梯度下降（PGD）来防止攻击模型偏离全局模型，即将模型投影到以上一轮全局模型为球心、半径为 δ 的超球面上，从而诱导模型对本能够正确分类但在训练数据中很少出现的数据错误分类。该文献还从理论上证明若一个模型易受对抗攻击，那么也将不可避免地易受后门攻击。

模型投毒可以根据其目的进一步分为非定向投毒（untargeted）[66-68, 72] 和定向投毒（targeted）[82, 85-86]。非定向投毒攻击旨在降低全局模型在所有类别上的预测准确率，定向投毒攻击旨在影响全局模型在攻击者指定的某一类数据上的准确率，而在其他任务上表现正常以躲避检测，所以定向投毒一般更为隐蔽。上述实现方式中，符号翻转、扩大幅值、同值

攻击、梯度篡改、Krum/Trim 攻击可认为是非定向投毒，语义后门植入、PGD 边缘攻击属于定向投毒。

后门攻击[87]属于定向投毒，使模型只在特定的、攻击者选择的输入任务上改变行为，而在主要任务上收敛并显示出良好的准确性，或直接对模型进行替换。植入后门是敌手的攻击目标，它可以通过数据投毒或模型投毒这两种手段实现，因此可认为拜占庭敌手也具备实施后门攻击的能力。

2　面向普通联邦学习的拜占庭防御策略

下面两部分分别从普通联邦学习和隐私保护联邦学习两个角度综述当前拜占庭攻击的防御方案，其研究脉络和分类如图 2 所示。由图 2 可见，近年来拜占庭鲁棒联邦学习得到了广泛研究，而随着安全隐私联邦学习的发展，密文上的防御逐渐成为研究重点。

当前，绝大多数抗拜占庭攻击的联邦学习防御方案都是面向普通联邦学习的，本文对文献［22－23］的分类依据进行调整和扩展，将面向普通联邦学习的防御策略分为基于梯度统计特性、基于梯度间距离、基于额外验证数据、基于优化算法补偿和基于差分隐私五类。各类别常用方法如图 3 所示。

2.1　基于梯度统计特性的防御

根据中心极限定理，只要同一批量规模足够大，且每个用户的数据集随机选择，不同用户训练得到的梯度就大概率不会有太大的差异。因此，基于梯度统计特性的防御策略旨在利用用户梯度更新的统计特性作为全局梯度，来剔除恶意梯度的影响，使用的统计特性包括几何中位数、坐标中位数、截断均值等。

Chen 等人[88]提出平均几何中位数（geometric median of means）算法，将 N 个用户梯度分为 b 个批量，计算每个批量的均值 w_i，再计算距离所有 w_i 距离之和最近的作为全局模型，即找到满足 $\operatorname{argmin}\sum_{i=1}^{b}\|w-w_i\|$ 的 w。若 $b=1$，则平均几何中

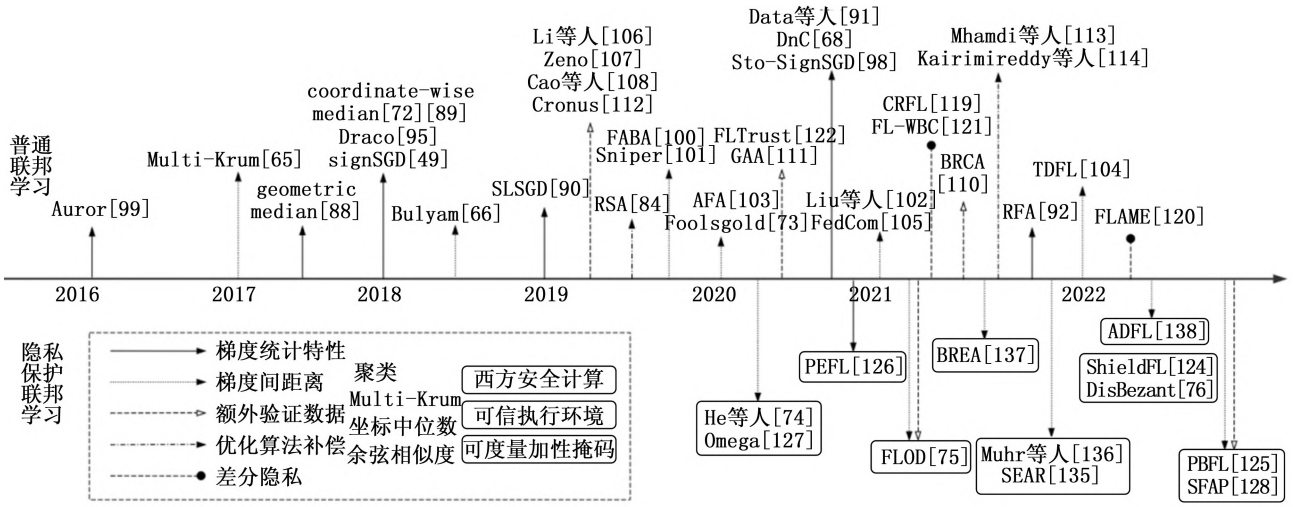

图 2　抗拜占庭攻击的联邦学习发展脉络和分类

Fig. 2　The Development and Classification of Byzantine-tolerant Federated Learning

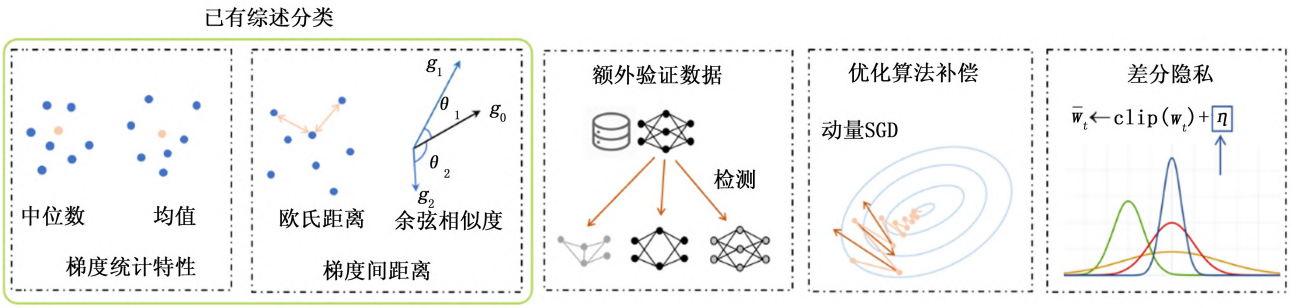

图 3　面向普通联邦学习的防御策略分类示意

Fig. 3　The Sketch of Defense Classification in Ordinary Federated Learning

位数简化为均值；若 $b=N$，则简化为几何中位数。Yin 等人提出了两种鲁棒梯度下降算法绕过恶意更新[89]。基于坐标中位数（coordinate-wise median）的算法计算每一维梯度的中位数作为全局更新，基于坐标截断均值（coordinate-wise trimmed mean）的算法对每一维梯度去掉最大和最小的一部分值，对剩余梯度取平均值作为全局更新。类似地，Xie 等人在文献［72］中研究了非凸优化（nonconvex optimization）下基于中位数的三种聚合规则，包括几何中位数、边缘中位数（marginal median）和中位数平均（mean around median）。几何中位数法与[88]相同，边缘中位数算法几乎与[89]中基于坐标中位数的算法相同。中位数平均算法在每个维度上对离中位数最近那一部分梯度取平均作为新的更新。Xie 等人提出一种通信高效的安全局部随机梯度下降算法 SLSGD（secure local stochastic gradient descent）[90]，用户使用局部梯度下降进行多轮训练，将模型发送给服务器，服务器使用截断均值进行聚合，然后将上一轮全局模型与本轮聚合模型按照一定权重聚合作为全局模型。虽然该算法在数据非独立同分布条件下进行了实验，但当敌手数量接近半数时收敛速率和准确率明显受影响。Data 等人发现基于坐标中位数、截断均值的聚合方法在高维模型上表现不佳，因此在服务器端采用离群值过滤方剔除恶意向量[91]。考虑到若样本的经验均值远离其真实均值，则经验协方差矩阵具有较高的最大特征值。因此，该方法迭代地剔除在经验协方差矩阵的主特征向量上有较大投影的样本，直到经验协方差矩阵的最大特征值足够小，并分析了强凸和非凸平滑目标函数的收敛性。Shejwalkar 等人提出基于主成分分析的异常值检测算法 DnC（divide-and-conquer）[68]，首先对高维更新梯度进行采样和抽取，将梯度矩阵进行中心化，由中心梯度矩阵的右上奇异特征向量计算离群值得分，去除得分最高的梯度，其余良性梯度平均后作为全局梯度更新。由于该方案的安全模型假设敌手可以获知良性用户的所有梯度信息，所以只考虑了 20%敌手的情况。Pillutla 等人提出基于几何中位数的 RFA（robust federated aggregation）算法[92]，通过平滑的 Weiszfeld 算法优化几何中位数求解。为保护用户梯度隐私不被好奇服务器探测，将聚合操作交给均值预言机进行迭代式安全聚合。

近期，许多研究使用编码理论解决分布式机器学习应用时的关键瓶颈[93-96]，如用户掉线、安全等问题，旨在通过为每个节点分配冗余梯度以消除拜占庭攻击的影响。Chen 等人提出的 Draco[95]是唯一适用于联邦学习场景的方法，即用户使用本地数据进行分布式训练。Draco 中，用户将冗余梯度的线性组合发送给服务器，任意两个用户的冗余梯度相加就能恢复出所有梯度之和，再由多数投票即可确定合法用户的聚合梯度值。

考虑到用户上传全精度梯度所占用的巨大通信带宽，越来越多的研究考虑通过梯度量化（gradient quantization）来降低通信开销，该方法同时可以增强鲁棒性。Bernstein 等人提出 signSGD[49]，将原始的全精度梯度压缩至 1bit 符号位发送给服务器，而忽略梯度的数值大小，服务器采用多数投票（majority vote）方法决定梯度的更新方向。受 Draco[95]冗余编码启发，Sohn 等人在 signSGD[49]基础上提出抗拜占庭攻击的选举编码（election coding）方案[97]，通过数据块分享增加冗余度，提高了多数投票方案的鲁棒性。为克服 signSGD 在数据非 IID 时的收敛性问题，Jin 等人提出 Sto-SignSGD [98]，梯度在本地经两级随机量化和压缩后发送给服务器，服务器同样使用多数投票聚合梯度。通过引入随机性，使错误聚合概率在理论上有界。以上方法都由服务器通过多数投票确定全局模型，多数投票可认为与基于坐标中位数聚合有相同的原理和效果。

基于梯度统计特性的防御策略使用梯度的某一统计特性推断总体梯度的特性，在用户数据独立同分布的情况下有不错的表现。但当梯度之间差别较大时，单一的某个统计特性无法很好的代表全局梯度。

2.2 基于梯度间距离的防御

基于梯度间距离的防御方法通过比较梯度之间的距离，或梯度与某一统计值的距离来检测可能的恶意梯度。常用的距离包括欧氏距离、余弦相似度（cosine similarity）等。

Shen 等人提出基于 K-means 聚类的异常梯度检测算法 Auror[99]，在敌手数 $f<N/2$ 的前提下，含有用户数最多的类就属于良性梯度类，反之就属于恶意梯度类。若用户在恶意梯度类中出现的次数超过一定阈值，则被认为是拜占庭敌手。Blanchard 等人提出了 Krum 算法[65]，每一轮训练，服务器从所有用户上传梯度中，选择在 l_2 范数空间与它的 K-f-2 相邻梯度距离，作为全局梯度更新模型，其余梯度则被抛弃。为了更好地利用其余梯度，改进的 Multi-Krum 计算多个梯度的均值作为

全局梯度。类似地，Xia 等人提出 FABA[100]，以迭代的方式，每次从本地梯度中剔除一个离平均梯度最远的梯度，直到被剔除的梯度的数目等于攻击者的数目。但文献［65，100］都需要提前知道拜占庭敌手数目。Guerraoui 等人提出 Bulyan[66]算法，首先使用 Krum 选择本地梯度，剔除可疑更新，然后对剩下的梯度执行截断均值算法聚合作为全局梯度。Cao 等人提出 Sniper[101]，利用梯度之间的欧氏距离构建无向图，当距离小于某个阈值时，认为两个本地模型是可达的，并将其连接，最后从无向图中得到最大连通子图，将子图中的本地梯度聚合得到全局模型。Liu 等人提出一种基于矩阵映射的抗拜占庭联邦聚合算法[102]，服务器收到用户上传模型后，通过构造矩阵去映射模型的更新部分来获取模型 Softmax 层的概率分布，通过计算剩余模型相互之间的欧几里得距离排除分布异常的模型。

考虑到用户本地的数据量差异，Luis 等人提出自适应联邦平均算法 AFA[103]，在训练中使用隐马尔可夫模型评估用户 k 更新梯度为良性的可能性 p_{k_t}，以此为权重，使用联邦平均算法FedAvg[4]，计算聚合梯度 $w_{t+1} \leftarrow \sum \frac{p_{k_t} n_k}{N} w_{t+1}^k$（$n_k$ 为用户 k 的本地数据量，$N = \sum p_{k_t} n_k$），然后对比聚合梯度 w_{t+1} 与局部梯度 w_{t+1}^k 的余弦相似度：

$$s_k = \frac{\langle w_{t+1} + w_{t+1}^k \rangle}{\| w_{t+1} \| \cdot \| w_{t+1}^k \|} \tag{3}$$

根据 s_k 的统计特性和少数敌手数假设，进一步剔除恶意梯度。Fung 等人针对定向投毒攻击提出 FoolsGold[73]，在每一轮训练中调整每一个用户的学习率，对提供独特梯度更新的客户赋予较高学习率，降低上传重复梯度更新的客户的学习率。考虑到良性梯度应具有类似的方向，因此使用余弦相似度计算学习率，异常梯度将被赋予较低的学习率参与聚合。Xu 等人提出基于真值发现（truth discovery）算法的联邦聚合算法 TDFL[104]。真值发现算法是一种高效的迭代式无监督聚合方法，可以提高不可靠或有噪声的用户的聚合数据质量，但在投毒攻击下准确率大大下降。对此，TDFL 使用余弦相似度计算初始全局模型与本地模型的相似度，剔除相似度小于某一阈值的恶意更新，合法用户重新开始训练。此外，当敌手数超过 50%时，采用了与文献［101］相同的方法，从最大连通子图中聚合得到全局模型。

Zhao 等人提出基于承诺机制的 FedCom[105]，服务器预先从用户处收集与用户本地数据同分布的“承诺”（commitment）。该方案考虑数据非 IID 时的数据投毒和梯度投毒攻击，且将其统一视为敌手对良性数据的非诚实训练。因此，除了发送模型 w^i，用户需要额外向服务器发送对本地数据 $D_{train}^{(i)}$ 的承诺 $D_{\text{commit}}^{(i)}$，以证明其对本地数据训练的合法性。在不泄露本地数据的前提下，与本地数据同分布的承诺值能被用于验证训练过程是否诚实。FedCom 将样本 $p_k \in D_{\text{commit}}^{(i)}$ 的 m 个邻近样本均值 c_k 作为其承诺。为检测模型投毒，服务器对比本轮训练前后，模型 w^i 在承诺 $D_{\text{commit}}^{(i)}$ 上的损失，若训练后损失变大，显然该训练过程是恶意的，该用户将被剔除。为检测数据投毒，服务器对比承诺之间的 Wassertein 距离，找出承诺对应的被篡改的数据。Wassertein 距离适用于无重叠的数据集，因此可以处理非 IID 数据相似度。

基于距离的方法需要对梯度进行两两比较，通常具有较大的时间复杂度。但其计算上的构造以及性质可被用于对隐私保护联邦学习中的恶意梯度进行检测。

2.3　基于额外验证数据的防御

本部分介绍的方法假设服务器预先从各用户处收集了一小部分本地数据，训练得到了一个能很好刻画全局数据特征的初始模型，基于该初始模型，服务器可以对训练时用户上传的梯度进行可信度评估。

Li 等人提出异常用户检测算法[106]，将用户梯度压缩后送入自动编码的异常检测模型，计算其异常得分 A_{t+1}^k，推导出信用得分 a_{t+1}^k 作为权重替代 FedAvg 中的 $\frac{n_k}{n}$ 更新全局模型 $w_{t+1} \leftarrow \sum a_{t+1}^k w_{t+1}^k$。Xie 等人提出 Zeno 算法[107]，服务器基于本地小批量验证集为每个更新梯度计算分数，使用该梯度更新全局模型带来的损失值越小，梯度模长越小，得分越高，最后选用分数最高的 k 个用户梯度更新全局模型。该算法仅需一个诚实用户即可训练出全局模型，但需要提前知道敌手数量，若真实敌手数大于理论上限，则算法表现不佳。

Cao 等人[108]提出了一种抗任意数量拜占庭敌手的联邦学习算法，服务器收集一部分验证数据集训练含噪模型，与用户更新梯度对比检测敌手，最后用正常用户梯度与服务器梯度的均值更新全局模型进行下一轮迭代。该方案无需预先知道攻击者数

目，并且当攻击者的数目超过 50%时仍具有良好的表现。Cao 等人提出 FLTrust[109]，服务器收集一部分数据作为数据集的可信基准，维护一个可信模型。每一轮训练，服务器都会对比本地可信模型与用户上传模型的 ReLU 激活余弦相似度，以此计算用户的可信得分，为防御大幅值梯度攻击，梯度需经归一化处理，最后以可信得分为权重进行加权平均得到全局模型。FLTrust 可抵抗文献［5］中提出的强拜占庭攻击和自适应攻击。Zhai 等人针对数据非 IID 分布情况提出了基于迁移学习的 BRCA[110]。每个用户都与服务器共享一部分数据，用于预训练自适应异常检测模型，并在实际训练时进行动态微调。每一轮训练，都计算用户的在异常检测模型上的得分 e_i。虽然非 IID 数据分布增加了拜占庭敌手的防御难度，但是每个客户端的更新模型在其自身共享数据上的性能不受其他客户端的影响，因此，使用共享数据计算更新模型的验证得分 f^i。然后将验证得分 f^i 与检测得分 e^i 加权得到每个客户的可信度 $r_t^i=\beta e^i+(1-\beta)f^i$，由可信度和上一轮全局梯度加权得到用户的本地更新 $w_{t+1}=\alpha w_t+(1-\alpha)\sum r_t^i w_t^i$，再通过共享数据得到全局更新，同时更新服务器端的异常检测模型。

Pan 等人提出基于强化学习的方案 GAA (gradient aggregation agent)[111]，将与用户的历史交互作为经验（experience），利用先验知识确定动作（action）空间以降低搜索最优策略的成本，基于动作 α_t，用户本地梯度更新为 $w_{t+1}=w_t-\lambda(\sum_{i=1}^{n}\alpha_t^{(i)}V_n^t)$。验证数据集上的损失下降作为奖励（reward）优化行为选择。该方法可以对拜占庭敌手的行为提供可解释信息和模式分析。

为保护局部梯度隐私，Chang 等人将迁移学习的思想用于安全联邦学习，提出了 Cronus 方案[112]。该方案假设所有用户都共享一个公开的未标注数据集，经过本地训练后对公开数据集进行预测并将结果发送给服务器。服务器聚合所有预测结果 $\bar{Y}=f_{\text{Cronus}}(Y_1,\cdots,Y_n)$，用户再根据 $\bar{Y}$ 重新训练更新本地模型。该方案以黑盒的方式与服务器共享本地梯度，可抵抗多种基于梯度的反演攻击，同时对每个本地模型分别更新，在提高抗攻击能力的同时保证了模型的个性化。对于服务器拥有的少量干净可验证数据，以上方法大多可以抵抗多数拜占庭敌手的攻击。然而，服务器收集验证数据需要考虑数据的隐私性、可用性等一系列问题，难以实际应用。

2.4 基于优化算法补偿的防御

前几部分所述方案都是从服务器安全聚合的角度识别并剔除拜占庭梯度，而本小节介绍的方案则是从客户端训练时的目标函数、梯度下降算法等角度，通过优化算法补偿来提高联邦学习的鲁棒性。

Li 等人提出基于目标函数优化的 RSA（byzantine-robust stochastic aggregation）算法[84]，对不同的目标函数进行优化，以提高全局模型的鲁棒性。服务器仅使用用户模型参数的方向而忽略其大小，并给目标损失函数添加一个正则化项，迫使良性的本地模型与全局模型更加接近，从而使得联邦系统对拜占庭攻击具有鲁棒性。该算法同样适用于非 IID 情形。

Mhamdi 等人对客户端的梯度下降算法进行动量优化[113]，加入动量后，每一次的梯度更新都会累积之前的梯度方向，用户端的动量 SGD 相比服务器端的动量 SGD 具有更小的方差－范数比，从而对拜占庭攻击有更强的鲁棒性。Karimireddy 等人发现当前聚合规则无法抵抗敌手跨时间联合攻击[114]，对此他们提出一种新的迭代裁剪算法，并引入用户端动量 SGD 克服联合攻击。

2.5 基于差分隐私的防御

差分隐私是 Dwork 提出的一种隐私保护机制[115-116]，通过加入随机噪声的方法来限制单一元素在数据集中对输出的影响，确保公开的输出结果不会因为某一体是否在数据集中而产生明显的变化，从而使攻击者无法根据发布的结果推测出结果对应的数据集。通常，差分隐私可分为全局差分隐私和本地差分隐私。Abadi 等人将差分隐私应用于深度学习，提出了差分隐私的 SGD 算法[117]，即在每次小批量训练对每个梯度进行阈值为 C 的 l_2 范数裁剪 $\bar{g}(x_i)\leftarrow g(x_i)/\left(1,\frac{\|g(x_i)\|_2}{C}\right)_{\max}$，然后添加服从 $N(0,\sigma^2)$ 分布的高斯噪声。Du 等人[118]首次从理论分析和实验证证明了差分隐私可以有效地减轻离群样本的影响，并将其扩展到后门攻击检测。该研究得出结论，增加噪声幅值可以增强鲁棒性，敌手数越多所需幅值越大。据此，越来越多的方案利用差分隐私防御联邦学习中的投毒攻击。

Xie 等人针对后门攻击提出鲁棒性方案提出 CRFL（certifiably robust federated earning）[119]，训练时服务器对用户上传梯度按权重聚合后，对全局梯度进行裁剪和噪声扰动，以控制训练过程中模型的全局偏差；预测时同样利用加噪平滑模

型参数，以对有细微差距的两个数据集返回一致的预测结果。Nguyen 等人提出抗后门攻击的 FLAME[120]，使用基于密度的动态聚类算法识别并剔除恶意模型更新，使用动态权重裁剪压缩高幅值恶意梯度，加入自适应噪声以减少恶意梯度对聚合结果的影响。Sun 等人使用黑塞矩阵 $\boldsymbol{H}^k=\nabla^2 F^k(w^k)$（hessian matrix，即本地损失函数的二阶偏导）定量评估了投毒攻击对全局模型的影响，发现全局模型一旦被污染，将对后续训练产生不可更改的影响，而由于服务器无法获取 $\boldsymbol{H}^k$，所以服务器端的抗拜占庭聚合方案也无济于事。对此他们针对定向投毒攻击，在用户端的角度提出了 FL-WBC 方案[121]，用户更新权重时通过加噪的方法保护 $\boldsymbol{H}^k$。FL-WBC 可作为服务器端防御方案的补充，在全局模型已被毒化的情况下减少模型投毒攻击。

利用差分隐私可以有效保护梯度隐私，增强模型对异常值的鲁棒性，但其缺点在于大大延长了训练时间，难以确定最合适的剪裁和加噪阈值，也会对模型性能和可用性造成影响。且该方法仅能减少恶意梯度对全局模型的影响，并不能精准定位并剔除恶意梯度。

2.6　面向普通联邦学习的拜占庭防御策略总结

本部分所涉及的针对普通联邦学习的拜占庭防御策略如表 1 所示。表 1 从安全模型、全局模型计算方法、敌手数、数据分布和服务器端时间复杂度几个方面对现有方法做了分析和对比。由表 1 可见，当前方法多集中在基于数据集统计特性、基于距离或相似度、基于服务器小批量验证数据这三类。基于距离或相似度的方案由于需要进行梯度之间的两两比较，所以需要较大的时间复杂度。基于服务器小批量验证数据的方法对敌手数具有较强的鲁棒性。拜占庭敌手的恶意梯度与数据非 IID 情况均会造成梯度间差异较大的情况，因此数据非 IID 时的拜占庭攻击检测是一大难点。

3　面向隐私保护联邦学习的拜占庭防御策略

第二节所述方案都基于服务器完全可信这一安全假设，所以允许服务器以明文或有损变换形式获取所有局部梯度和全局梯度，而当服务器为半诚实时就可能存在针对未加密梯度的反演攻击。为了保护梯度隐私，越来越多的研究者将重点转向了隐私保护联邦学习，使服务器只能获取加密后的梯度。本章将面向隐私保护联邦学习的防御策略分为基于安全多方计算（MPC，secure multi-party computation）、基于可信执行环境（TEE，trusted execution environment）和基于可度量加性掩码三类。各类别常用方法如图 4 所示。

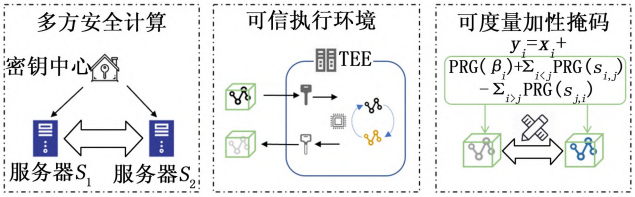

图 4　隐私保护联邦学习的拜占庭防御分类

Fig. 4　The Sketch of Defense Classification in Privacy-preserving Federated Learning

3.1　基于安全多方计算的防御

安全多方计算是指在无可信第三方的情况下，多个参与方协同计算一个约定的函数，并且保证每一方仅获取自己的计算结果，无法通过计算过程中的交互数据推测出其他任意一方的输入和输出数据[123]。检测拜占庭敌手必须在明文梯度上做比较，所以基于密文的拜占庭敌手检测最朴素的想法，是利用双服务器架构设计安全两方计算协议，在保护梯度的隐私前提下将密文解密，在明文上做比较。

He 等人提出基于秘密分享的非共谋双服务器防御方案[74]。用户将梯度 $x_i=x_i^1+x_i^2$ 拆分成两部分通过安全信道分别发送给服务器 S_1 和 S_2，S_2 利用 Beaver 三元组可以基于线性秘密分享机制实现安全乘法计算从而计算两两梯度间的距离 $\|x_i-x_j\|^2$，利用 Multi-Krum 等基于距离的拜占庭防御方案得到梯度权重 p_i，通过秘密分享机制安全发送给 S_1，S_1 和 S_2 在本地各自计算一部分聚合梯度 $\sum_{i=1}^{n}p_i x_i$ 最后由 S_1 汇总发送给用户。该方案中，服务器 S_1 只能拿到一部分梯度 $\{x_i^1\}_{i=1}^n$ 和 S_2 发来的另一部分梯度平方之和 $\sum_{i=1}^{n}x_i^2$，服务器 S_2 只能拿到另一部分梯度 $\{x_i^2\}_{i=1}^n$，用户只能拿到梯度聚合结果。此外，由于该方案直接对用户上传梯度进行处理，所以用户掉线对方案鲁棒性并无影响。但是，由于 S_2 可以拿到权重值，所以聚合梯度也存在一定程度上的隐私泄露问题[34]。

Dong 等人提出针对多数拜占庭敌手的防御方案 FLOD[75]，借鉴了 FLTrust[109] 服务器验证数据和 signSGD[49] 梯度量化的思想，提出了一种新的

基于汉明距离的聚合方法。客户端对本地梯度更新进行量化和布尔编码，通过布尔秘密分享将更新梯度拆分成 $[w_i]_1^B$ 和 $[w_i]_2^B$ 分别发给服务器 S_1 和 S_2，服务器通过部分验证数据集训练得到本地模型 w_s，同样进行量化和编码。然后两服务器各自计算布尔梯度间的汉明距离，$[hd_i]_1^B = w_s \oplus [w_i]_1^B$，$[hd_i]_2^B = [w_i]_2^B$，为了提高计算和通信效率，通过安全两方计算将布尔共享转换为算数共享 $[hd_i]_1^A$ 和 $[hd_i]_2^A$。通过混淆电路，两服务器各自持有 ReLU 激活后的一部分梯度算数秘密分享 $\tau - [hd_i]_1^A$ 和 $-[hd_i]_2^A$，最终由权重聚合协议计算梯度对应的权重 $v_i = ReLU(\tau - hd_i)$，加权得到全局梯度。

Ma 等人针对基于加性掩码的隐私保护联邦学习方案[58-59]难以检测密文投毒梯度问题，提出了 ShieldFL[124]。该方案使用不可合谋的双服务器架构，设计基于双陷门同态加密协议，在服务器 S_1、S_2 之间，服务器 S_1 与用户 u_i 之间进行密钥拆分，防止恶意的服务器或用户泄露密钥。ShieldFL 包括归一化判断、相似度判断、抗拜占庭的梯度聚合三步。合法梯度在上传前都经过了归一化处理，而拜占庭敌手的攻击目的在于发送大幅值的恶意梯度以改变全局模型方向，所以未经归一化处理。服务器 S_1 收到加密梯度 $\|x'_k\|$ 后，计算盲化后的密文梯度 $\|\bar{x}_k\|$，用部分私钥 sk_1 解密得到 $[\bar{x}_k]_1$，将 $\{\|\bar{x}_k\|, [\bar{x}_k]_1\}$ 发送给服务器 S_2。S_2 用另一部分私钥 sk_2 解密得到 $[\bar{x}_k]_2 \leftarrow PartDec_{sk_2}(\|\bar{x}_k\|)$，再进行完全解密得到盲化的梯度明文 $\bar{x}_k \leftarrow FullDec([\bar{x}_k]_1, [\bar{x}_k]_2)$，将其平方和加密得到 $\left\| \sum \bar{x}_k^2 \right\|$ 返回给 S_1。基于密文加法同态性质，S_1 可在密文上去除噪声，得到梯度平方和的密文 $[\sum x'^2_k]$，再经过一轮安全解密最终得到梯度平方和的明文 $\sum x'^2_k$，进行归一化判断。在这一过程中，由于 S_1 和 S_2 各自仅有一部分私钥，所以 S_1 只能拿到梯度平方和的明文，S_2 只能拿到盲化梯度的明文，都无法获得局部或全局梯度的明文。类似地，在相似度判断这一步中，S_1 只能拿到梯度余弦相似度的明文，S_2 只能拿到盲化梯度的明文，保护了局部和全局梯度隐私。在抗拜占庭的梯度聚合中，从本轮梯度选择与上一轮聚合梯度余弦值最小（即方向差距最大）的作为投毒基准，计算本轮梯度与投毒基准的相似度，越相似（余弦值越大）则信心值越小，越可能为投毒梯度，以信心值作为权重进行安全梯度聚合。在此基础上，Miao 等人提出基于区块链的 PBFL[125]，该方法借鉴了 FLTrust 使用少量干净验证数据训练服务器端模型的思路，采用同态加密保护上传梯度，通过安全两方计算判断梯度是否归一化、计算用户梯度与服务器端梯度的余弦相似度剔除恶意敌手，额外借助区块链实现计算结果的公开可验证。类似的，Liu 等人提出双服务器架构下隐私增强的联邦学习方案 PEFL[126]，使用同态加密保护梯度隐私，S_1 将含噪密文梯度发送给 S_2，S_2 解密后只能对含噪梯度计算坐标中位数，并由皮尔逊相关系数（Pearson correlation coefficient）计算各梯度与中位数的相关性，以此为权重进行梯度聚合。Ma 等人针对海上运输系统提出基于双服务器架构安全两方计算的 DisBezant[76]。为保护局部和全局梯度，可信中心为用户分配了不同的加密密钥和相同的聚合密钥。引入可信度来衡量船舶共享信息的可信度，在每轮训练通过用户共享信息更新可信度，由上一轮可信度和本轮梯度构建导向梯度，通过比较用户梯度与导向梯度间的相似性判断用户是否为拜占庭节点。面向非 IID 数据，该方案增加了正常节点的可信度。

Ma 等人提出单一服务器下的安全协议 Omega[127]。利用双陷门公钥密码，服务器与每一用户共享拆分的私钥 sk_1，sk_2。服务器选择一部分用户 T_i 参与训练，梯度经过同态加密后上传服务器，服务器计算密文梯度间欧氏距离的平方 $\zeta_{ij} \leftarrow \|(g_i - g_j)^2\|$，利用加法同态性质增加拉普拉斯噪声 $\zeta_{ij}' \leftarrow \left[(g_i - g_j)^2 + Lap\left(\frac{d_{\max}}{\varepsilon}\right)\right]$，用部分私钥 sk_1 解密后发送给未参与训练的另一部分用户 T_g。T_g 用另一部分私钥 sk_2 解密获得带噪剃度距离平方的明文，通过 Multi-Krum 聚合规则剔除拜占庭梯度，最后由服务器进行梯度聚合。Ma 等人面向远程医疗诊断场景，提出抗投毒攻击的联邦学习诊断系统 SFAP[128]。与文献 [126] 类似，服务器收集用户干净数据，形成基准决策树。利用多密钥加密机制，服务器与用户运行两方安全计算协议，将用不同公钥加密的梯度转换为统一公钥加密的形式，计算用户梯度与基准梯度的均方误差，将均方误差小于阈值的聚合为全局梯度。训练完成后，患者可以进行经密钥保护的安全远程诊断。这两种方案虽然只需要一个服务器，但实际上是将另一服务器的功能交给用户，这需要确保用户身份可信并正确完成计算。

基于安全多方计算的方法主要对密文梯度进行安全计算，确保参与计算的两方服务器均无法获取明文梯度，所使用的梯度检测方法几乎都为经典的基于距离的方法，全局模型聚合方式也主要采用传统的 FedAvg、梯度加权等方法。安全多方计算虽然可以对密文进行安全计算和比较，但双服务器架构之间的多轮运算带来了大量的通信和计算开销。

3.2　基于可信执行环境的防御

TEE 通过软硬件方法在中央处理器中构建一个安全可信的区域，将其与不可信的计算区域隔离开，以保护其内部加载的程序和数据的机密性和完整性[129]。TEE 的实现由硬件飞地（encalve）支持，如 Intel SGX [130]和 ARM TrustZone[131]。许多研究[132-134]证实了将深度神经网络载入 TEE 进行高效计算，以及将 TEE 应用于隐私保护联邦学习的可行性，然而基于 TEE 的方案对硬件提出了较高的要求。

针对现有安全聚合联邦学习无法抵抗拜占庭攻击的问题，Zhao 等人基于 Intel SGX 提出了安全联邦学习方案 SEAR[135]。Intel SGX 提供了一组新的指令集扩展与访问控制机制，用户可将需要保护的代码和数据放入飞地或者受保护内存区域（PRM，processor reserved memory range），实现不同程序间的隔离运行，保障用户关键代码数据的机密性与完整性，防止被恶意软件泄露和修改。利用 Intel SGX 的远程审计功能，飞地与批量用户首先进行双向审计，随后协商出用于加密更新梯度的对称密钥。考虑到飞地运行内存的限制，用户逐层上传加密后的梯度，服务器随后载入飞地进行安全解密。为检测拜占庭攻击，同时降低可信与非可信区域之间的加解密计算开销，SEAR 随机选取部分用户计算梯度间欧氏距离，检测拜占庭梯度。此外，SEAR 针对不同的聚合算法设计了按列和按行 2 种数据存储模式，以避免耗时的飞地页面缓存操作。考虑到 TEE 的实际运算能力，Muhr 等人[136]将服务器拆成可信但资源受限的 TEE 和不可信但资源富足的 REE（rich execution environment）两部分。用户对梯度进行掩码操作或者同态加密后发送给 REE，REE 聚合密文后发给 TEE，TEE 将解密后的梯度和返回给 REE，REE 根据梯度更新模型广播给用户，与此同时 TEE 与 REE 进行多轮安全计算，通过主成分分析法压缩梯度，通过聚类找出恶意用户。

3.3　基于可度量加性掩码的防御

加性掩码是隐私保护联邦学习的一种方案[58-59]，由于其可抵消性，服务器只能获得聚合后的梯度值。本小节归纳的方案利用不同方法，从加性掩码中衍生出可度量特性，用于分析所对应梯度的特性。So 等人在单服务器下提出首个抗拜占庭的安全聚合联邦学习框架 BREA[137]。该方案对随机量化的梯度 $\widehat{w_i}$ 添加掩码得到 $f_i(\theta)$，然后利用可验证秘密共享协议生成给其余用户的秘密分享 $s_{ij}=f_i(\theta_j)$ 和对该分享的承诺，同时通过检验其余用户的承诺来检查梯度有效性。用户 i 直接基于用户 j，k 与其分享的秘密在本地安全计算梯度 $\widehat{w_j}$ 与 $\widehat{w_k}$ 的距离 $d_{jk}(\theta_i)=\|s_{ji}-s_{ki}\|^2=\|f_j(\theta_i)-f_k(\theta_i)\|^2$，发送给服务器。由于可能存在恶意用户发送错误结果，服务器将 $d_{jk}^{(i)}$ 视为最高次为 2T 的里德－所罗门（Reed-Solomon）编码多项式，将退出用户视为擦除码，将拜占庭敌手视为误码，巧妙地将梯度距离恢复过程视作解码，即可计算明文梯度的距离 $d_{jk}(0)=\|f_j(0)-f_k(0)\|^2=\|\widehat{w_j}-\widehat{w_k}\|^2$。随后，服务器用 Multi-Krum 算法选择一部分可信用户，这一部分用户重新在本地聚合梯度的秘密分享 $s_i=\sum s_{ji}$，发给服务器。服务器使用类似的解码方法恢复出聚合后的明文梯度 $h(0)=\sum f_j(0)=\sum \widehat{w_j}$。该方案使用可验证秘密分享协议保护了梯度隐私，利用里德－所罗门编码实现了密文梯度的安全计算，且在用户每次发送消息时都考虑了潜在的拜占庭攻击，可以检测发送错误秘密分享、错误梯度的用户，具有非常强的鲁棒性。

Guo 等人对 Bonawitz 的加性掩码方案进行改进，提出抗投毒攻击的 ADFL[138]。由于加性掩码的互相抵消特性，服务器可计算出任意两个梯度的均值和所有梯度的均值，形成长度为 C_n^2+1 的向量，经过随机稀疏处理后发送给用户，同时记录下原始原素的位置。用户从稀疏向量中多次选择 C_n^2+1 个元素测试其在本地数据上的准确性，记录下准确率最高的元素的位置，发送给服务器。服务器通过计算两个位置的交集判断并剔除可能的敌手，从剩余可信用户中任选 2 个进行梯度聚合，所以该方案需要确保至少有 2 个良性用户。

3.4　面向隐私保护联邦学习的拜占庭防御策略总结

本章所涉及的面向隐私保护联邦学习的拜占庭防御策略如表 2 所示。表 2 从安全模型、恶意梯度

剔除方法、敌手数、数据分布和服务器端时间复杂度对当前方案做了分析和对比。由表 2 可见，当前方法主要基于安全多方计算，恶意梯度判别大多参考了明文上的经典方法，如基于梯度统计特性和基于梯度间距离的方法。基于服务器验证数据的方法需要服务器收集一部分干净验证数据，与隐私保护的要求相违背。基于客户端训练优化和基于差分隐私的方法只是通过降低对异常值的敏感性来提高联邦学习鲁棒性，并不能准确辨别恶意梯度，更无法识别密文形式的恶意梯度。因此上述三种方法无法扩展至面向隐私保护联邦学习的拜占庭防御。

4 未来研究方向

虽然拜占庭鲁棒的安全隐私联邦学习研究已经取得诸多成果，但是目前还处于初期探索阶段，仍有许多问题亟待解决，其中以下四种情况的拜占庭敌手检测值得开展进一步研究。

4.1 数据非 IID

当前大多数抗拜占庭攻击的联邦学习方案都基于数据 IID 假设，因此出现统计特性偏离的梯度必然存在拜占庭敌手。然而实际情况中，由于设备所处环境、用户偏好等差异，客户端数据通常是非 IID 的，造成各用户上传的梯度之间存在较大偏差[27,139]。此时，出现统计特性偏离的梯度可能是存在拜占庭敌手，也可能是由于数据集本身非 IID 而造成的。虽然部分方案声称其同样适用于数据非 IID 情况，但其结论大多是通过实验结果反推得到的，并没有从理论上严格证明其方案在非 IID 情况下的收敛性。Peng 等人[140]针对数据非 IID 情况，使用重采样策略来减少用户内部和用户之间的数据异构性，同时使用随机平均梯度算消除用户内部的数据异构性，并给出了线性收敛到最优解的证明。然而，当前针对数据非 IID 时的拜占庭鲁棒联邦学习方案仍十分有限，需要进一步深入研究。针对数据非 IID 情况，多任务联邦学习为每一种数据分布都训练出个性化的模型[141]，基于聚类[142]、迁移学习[143]、元学习[144]的个性化联邦学习已成为研究热点，可作为解决数据非 IID 问题的一种方法。

4.2 多数敌手

无论处理明文还是密文梯度，基于数据统计特性和距离的方法仍是当前识别拜占庭敌手的主要方法。该方法基于少数拜占庭敌手数假设，即存在超过半数良性用户时，认为大多数梯度仍能反应正确梯度的方向，以此来过滤恶意梯度。然而，也存在敌手超过半数的情况，即 $f+1\leqslant N$。当前，能抵抗多数拜占庭敌手的方法大多基于服务器小批量干净验证数据，在实际应用时面临着数据隐私性、可用性等困难，且无法迁移至隐私保护联邦学习场景。如何在拜占庭敌手超过多数时，对密文形式的恶意梯度进行高效检测和过滤，是一个值得研究的问题。

4.3 用户退出

分布式训练环境具有不稳定性，用户可能在训练过程中由于资源受限、位置移动、通信中断或个人原因而中途退出，抗用户退出也是鲁棒联邦学习应考虑的指标[145]。当前，在隐私保护联邦学习下考虑用户退出的抗拜占庭方案包括基于秘密分享和基于编码技术两种。基于秘密分享的方案利用加性掩码和 Shamir 门限特性，当部分用户退出时仍能正确解密。基于编码技术的方案将梯度恢复过程视作解码，将退出用户视为擦除码，将拜占庭敌手视为误码，使得方案具备一定容错性。这两种方案都具有较大的通信复杂度，且退出用户的增加将导致协议运行时间的增加。因此，需要一种更高效、更健壮的抗用户退出机制，作为拜占庭鲁棒隐私保护联邦学习可以进一步优化的方向。

4.4 无中心联邦学习

联邦学习根据架构可分为中心化联邦学习和去中心化联邦学习，本文讨论的拜占庭攻防主要面向中心化联邦学习场景。中心化联邦学习需要中心服务器参与，且假设服务器会依照算法诚实地进行安全鲁棒聚合，而现实场景中服务器也有可能遭受攻击违背协议流程，且中心化架构可能导致传输成本增加、网络带宽遇到瓶颈。针对上述限制，无中心联邦学习逐渐成为研究热点。无中心联邦学习无需服务器的参与，参数仅在用户节点间端到端传输，防止了服务器单点失败的情况。现有研究[146-147]将区块链技术应用到无中心联邦学习中，采用共识算法选择用于进行聚合操作的用户节点，但存在算法耗时耗能、易受到 51% 攻击等缺陷。在缺少中心服务器的帮助时，亟需通过用户间的协商与监督高效辨别拜占庭节点的方法，为鲁棒无中心联邦学习保驾护航[148]。

5 结束语

联邦学习经过近几年的飞跃发展，已经在性能、效率等方面取得了巨大的成就，使得研究重点向安全与鲁棒性方面转移。目前，拜占庭鲁棒的安

全隐私联邦学习研究还处于初级阶段，尚有许多关键问题亟待解决。对此，本文经充分调研，对联邦学习中拜占庭攻击和防御领域的最新研究成果进行综述，从普通联邦学习和隐私保护联邦学习两方面对现有的防御方案做了系统的归纳和总结，并就其鲁棒性、效率等进行分析对比。同时，本文也对拜占庭鲁棒的联邦学习在多种情况下的实际应用进行探讨，指出了未来的研究方向。

参考文献：

[1] JORDAN M I, MITCHELL T M. Machine learning: Trends, perspectives, and prospects [J]. Science, 2015, 349 (6245): 255 - 260.

[2] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521 (7553): 436 - 444.

[3] CITE2022 工业互联网发展与安全峰会 [EB/OL]. (2022-05-17) [2023-08-27]. http://event.chinaaet.com/huodong/cite2022/.

[4] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C] //Proc. of the Artificial Intelligence and Statistics. New York: PMLR, 2017: 1273 - 1282.

[5] YANG Q, LIU Y, CHENG Y, et al. Federated learning [J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2019, 13 (3): 1 - 207.

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

[7] LIU Y, FAN T, CHEN T J, et al. FATE: An industrial grade platform for collaborative learning with data protection [J]. The Journal of Machine Learning Research, 2021, 22 (1): 10320 - 10325.

[8] MA Y J, YU D H, WU T, et al. PaddlePaddle: An open-source deep learning platform from industrial practice [J]. Frontiers of Data and Computing, 2019, 1 (1): 105 - 115.

[9] BONAWITZ K, EICHNER H, GRIESKAMP W, et al. Towards federated learning at scale: System design [J]. Proceedings of Machine Learning and Systems, 2019, 1: 374 - 388.

[10] RYFFEL T, TRASK A, DAHL M, et al. A generic framework for privacy preserving deep learning [J]. ArXiv Preprint ArXiv: 181104017, 2018.

[11] HAO M, LI H W, LUO X Z, et al. Efficient and privacy-enhanced federated learning for industrial artificial intelligence [J]. IEEE Transactions on Industrial Informatics, 2019, 16 (10): 6532 - 6542.

[12] XU J, GLICKSBERG B S, SU C, et al. Federated learning for healthcare informatics [J]. Journal of Healthcare Informatics Research, 2021, 5 (1): 1 - 19.

[13] RIEKE N, HANCOX J, LI W Q, et al. The future of digital health with federated learning [J]. NPJ Digital Medicine, 2020, 3 (1): 1 - 7.

[14] MILLS J, HU J, MIN G. Communication-efficient federated learning for wireless edge intelligence in IoT [J]. IEEE Internet of Things Journal, 2019, 7 (7): 5986 - 5994.

[15] LONG G D, TAN Y, JIANG J, et al. Federated learning for open banking [M] // Federated Learning: Privacy and Incentive. Cham: Springer. 2020: 240 - 254.

[16] YANG W S, ZHANG Y H, YE K J, et al. FFD: A federated learning based method for credit card fraud detection [C] //Proc. of the Int. Conf. on Big Data. Cham: Springer, 2019: 18 - 32.

[17] LAMPORT L. The weak byzantine generals problem [J]. Journal of the ACM, 1983, 30 (3): 668 - 676.

[18] CASTRO M, LISKOV B. Practical byzantine fault tolerance [C] //Proc. of the 3rd Symp. on Operating Systems Design and Implementation. New York: ACM, 1999: 173 - 186.

[19] GU Y H, BAI Y B. Survey on security and privacy of federated learning models [J]. Journal of Software, 2022: 1 - 32.

[20] GUO S W, ZHANG X, YANG F, et al. Robust and privacy-preserving collaborative learning: A comprehensive survey [J]. ArXiv Preprint ArXiv: 211210183 2021.

[21] ROSZEL M, NORVILL R, STATE R. An analysis of byzantine-tolerant aggregation mechanisms on model poisoning in federated learning [C] //Proc. of the 19th Int. Conf. on Modeling Decisions for Artificial Intelligence. Cham: Springer, 2022: 143 - 155.

[22] SHI J Y, WAN W, HU S S, et al. Challenges and approaches for mitigating byzantine attacks in federated learning [J]. ArXiv Preprint ArXiv: 211214468, 2021.

[23] WAN W. Research on byzantine attack defense algorithm in federated learning scenario [D]. Wuhan: Huazhong University of Science & Technology, 2020.

[24] YANG Q. AI and data privacy protection: the way to federated learning [J]. Journal of Information Security Research, 2019, 5 (11): 961 - 965.

[25] BOTTOU L. Large-scale machine learning with stochastic gradient descent [C] //Proc. of the 19th Int. Conf. on Computational Statistics. Cham: Springer,

2010：177－186.

[26] KONECNY J，MCMAHAN H B，RAMAGE D，et al. Federated optimization：Distributed machine learning for on-device intelligence [J]. ArXiv Preprint ArXiv：161002527，2016.

[27] LI T，SAHU A K，ZAHEER M，et al. Federated optimization in heterogeneous networks [J]. Proceedings of Machine Learning and Systems，2020，2：429－450.

[28] WANG J Y，LIU Q H，LIANG H，et al. Tackling the objective inconsistency problem in heterogeneous federated optimization [C] //Proc. of Advances in Neural Information Processing Systems，New York：Curran Associates，2020，33：7611－7623.

[29] MELIS L，SONG C，DE CRISTOFARO E，et al. Exploiting unintended feature leakage in collaborative learning [C] //Proc. of the Symp. on Security and Privacy. Piscataway，NJ：IEEE，2019：691－706.

[30] GAO J Q，HOU B Y，GUO X J，et al. Secure aggregation is insecure：Category inference attack on federated learning [J]. IEEE Transactions on Dependable and Secure Computing，2021，20 (1)：147－160.

[31] 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] //Proc. of the IEEE Symp. on Security and Privacy. Piscataway，NJ：IEEE，2019：739－753.

[32] ZHU L G，LIU Z J，HAN S. Deep leakage from gradients [C] //Proc. of Advances in Neural Information Processing Systems，New York：Curran Associates，2019：32.

[33] ZHAO B，MOPURI K R，BILEN H. iDLG：Improved deep leakage from gradients [J]. ArXiv Preprint ArXiv：200102610，2020.

[34] GEIPING J，BAUERMEISTER H，DRGE H，et al. Inverting gradients-how easy is it to break privacy in federated learning? [C] //Proc. of Advances in Neural Information Processing Systems，New York：Curran Associates，2020，33：16937－16947.

[35] YIN H，MALLYA A，VAHDAT A，et al. See through gradients：image batch recovery via gradinversion [C] //Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2021：16337－16346.

[36] JEON J，LEE K，OH S，et al. Gradient inversion with generative image prior [C] //Proc. of Advances in Neural Information Processing Systems，New York：Curran Associates，2021，34：29898－29908.

[37] SHOKRI R，STRONATI M，SONG C，et al. Membership inference attacks against machine learning models [C] //Proc. of the IEEE Symp. on Security and Privacy. Piscataway，NJ：IEEE，2017：3－18.

[38] KRIZHEVSKY A，HINTON G. Learning multiple layers of features from tiny images [Z]. Handbook of Systemic Autoimmune Diseases，2009.

[39] WEN J，ZHANG Z X，LAN Y，et al. A survey on federated learning：challenges and applications [J]. International Journal of Machine Learning and Cybernetics，2023，14：513－535.

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

[41] ENTHOVEN D，AL-ARS Z. An overview of federated deep learning privacy attacks and defensive strategies [J]. Federated Learning Systems，2021：173－196.

[42] YIN X F，ZHU Y M，HU J K. A comprehensive survey of privacy-preserving federated learning：A taxonomy，review，and future directions [J]. ACM Computing Surveys，2021，54 (6)：1－36.

[43] SHOKRI R，SHMATIKOV V. Privacy-preserving deep learning [C] //Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. New York：ACM，2015：1310－1321.

[44] TRUEX S，LIU L，CHOW K-H，et al. LDP-Fed：Federated learning with local differential privacy [C] //Proc. of the 3rd ACM Int Workshop on Edge Systems，Analytics and Networking. New York：ACM，2020：61－66.

[45] WEI K，LI J，DING M，et al. Federated learning with differential privacy：algorithms and performance analysis [J]. IEEE Transactions on Information Forensics and Security，2020，15：3454－3469.

[46] CHEN S Z，YU D X，ZOU Y F，et al. Decentralized wireless federated learning with differential privacy [J]. IEEE Transactions on Industrial Informatics，2022，18 (9)：6273－6282.

[47] GAO W，GUO S W，ZHANG T W，et al. Privacy-preserving collaborative learning with automatic transformation search [C] //Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2021：114－123.

[48] SUN J W，LI A，WANG B H，et al. Soteria：Provable defense against privacy leakage in federated learning from representation perspective [C] //Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2021：9307－9315.

[49] BERNSTEIN J，WANG Y-X，AZIZZADENESHELI

K, et al. signSGD: Compressed optimisation for non-convex problems [C] //Proc. of the Int. Conf. on Machine Learning. New York: PMLR, 2018: 560-569.

[50] VOGELS T, KARIMIREDDY S P, JAGGI M. PowerSGD: practical low-rank gradient compression for distributed optimization [C] //Proc. of Advances in Neural Information Processing Systems, New York: Curran Associates, 2019, 32.

[51] YUE K, JIN R C, WONG C W, et al. Gradient obfuscation gives a false sense of security in federated learning [J]. ArXiv Preprint ArXiv: 220604055, 2022.

[52] WEI W Q, LIU L. Gradient leakage attack resilient deep learning [J]. IEEE Transactions on Information Forensics and Security, 2021, 17: 303-316.

[53] YANG H M, GE M Y, XIANG K L, et al. Using highly compressed gradients in federated learning for data reconstruction attacks [J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 818-830.

[54] AONO Y, HAYASHI T, WANG L, et al. Privacy-preserving deep learning via additively homomorphic encryption [J]. IEEE Transactions on Information Forensics and Security, 2017, 13 (5): 1333-1345.

[55] ZHANG C L, LI S Y, XIA J Z, et al. BatchCrypt: Efficient homomorphic encryption for Cross-Silo federated learning [C] //Proc. of the 2020 USENIX Annual Technical Conf. San Sebastian: USENIX Association, 2020: 493-506.

[56] DONG Y, CHEN X J, SHEN L Y, et al. EaSTFLy: Efficient and secure ternary federated learning [J]. Computers & Security, 2020, 94: 101824.

[57] ZHU H Y, WANG R, JIN Y C, et al. Distributed additive encryption and quantization for privacy preserving federated deep learning [J]. Neurocomputing, 2021, 463: 309-327.

[58] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning [C] //Proc. of the ACM SIGSAC Conf. on Computer and Communications Security. New York: ACM, 2017: 1175-1191.

[59] XU G W, LI H W, LIU S, et al. VerifyNet: secure and verifiable federated learning [J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 911-926.

[60] GUO X J, LIU Z L, LI J, et al. VeriFL: Communication-efficient and fast verifiable aggregation for federated learning [J]. IEEE Transactions on Information Forensics and Security, 2020, 16: 1736-1751.

[61] HAHN C, KIM H, KIM M, et al. Versa: Verifiable secure aggregation for cross-device federated learning [J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20 (1): 36-52.

[62] LUO F C, AL-KUWARI S, DING Y. SVFL: Efficient secure aggregation and verification for cross-silo federated learning [J]. IEEE Transactions on Mobile Computing, 2022: 1-14.

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

[64] FARHADKHANI S, GUERRAOUI R, VILLEMAUD O. An equivalence between data poisoning and byzantine gradient attacks [C] //Proc. of the Int. Conf. on Machine Learning. New York: PMLR, 2022: 6284-6323.

[65] BLANCHARD P, EL MHAMDI E M, GUERRAOUI R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent [C] //Proc. of Advances in Neural Information Processing Systems, New York: Curran Associates, 2017: 118-128.

[66] GUERRAOUI R, ROUAULT S. The hidden vulnerability of distributed learning in byzantium [C] //Proc. of the Int. Conf. on Machine Learning. New York: PMLR, 2018: 3521-3530.

[67] FANG M H, CAO X Y, JIA J Y, et al. Local model poisoning attacks to byzantine-robust federated learning [C] //Proc. of the 29th USENIX Security Symp. San Sebastian: USENIX Association, 2020: 1605-1622.

[68] SHEJWALKAR V, HOUMANSADR A. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning [C/OL] // [2023-03-07]. https://www.ndss-symposium.org/ndss-paper/manipulating-the-byzantine-optimizing-model-poisoning-attacks-and-defenses-for-federated-learning/.

[69] WANG L X, XU S C, WANG X, et al. Eavesdrop the composition proportion of training labels in federated learning [J]. ArXiv Preprint ArXiv: 191006044, 2019.

[70] CHOI B, SOHN J Y, HAN D J, et al. Communication-computation efficient secure aggregation for federated learning [J]. ArXiv Preprint ArXiv: 201205433, 2020.

[71] XU X Y, WU J Z, YANG M T, et al. Information leakage by model weights on federated learning [C] // Proc. of the Workshop on Privacy-Preserving Machine Learning in Practice. New York: Association for Computing Machinery, 2020: 31-36.

[72] XIE C, KOYEJO O, GUPTA I. Generalized byzantine-tolerant SGD [J]. ArXiv Preprint ArXiv:

180210116, 2018.

[73] FUNG C, YOON C J, BESCHASTNIKH I. The limitations of federated learning in sybil settings [C] //Proc. of the 23rd Int. Symp. on Research in Attacks, Intrusions and Defenses. San Sebastian: USENIX Association, 2020: 301 - 316.

[74] HE L, KARIMIREDDY S P, JAGGI M. Secure byzantine-robust machine learning [J]. ArXiv Preprint ArXiv: 200604747, 2020.

[75] DONG Y, CHEN X Y, LI K Y, et al. FLOD: Oblivious defender for private byzantine-robust federated learning with dishonest-majority [C] //Proc. of the 26th European Symp. on Research in Computer Security. Cham: Springer, 2021: 497 - 518.

[76] MA X D, JIANG Q, SHOJAFAR M, et al. DisBezant: Secure and robust federated learning against byzantine attack in IoT-enabled MTS [J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24 (2): 2492 - 2502.

[77] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines [C] //Proc. of the Int. Conf. on Machine Learning. New York: PMLR, 2012: 1467 - 1474.

[78] TOLPEGIN V, TRUEX S, GURSOY M E, et al. Data poisoning attacks against federated learning systems [C] //Proc. of the European Symp. on Research in Computer Security. Cham: Springer, 2020: 480 - 501.

[79] ZHANG J L, CHEN J J, WU D, et al. Poisoning attack in federated learning using generative adversarial nets [C] //Proc. of the 18th IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications/13th IEEE Int. Conf. on Big Data Science and Engineering. Piscataway, NJ: IEEE, 2019: 374 - 380.

[80] CHEN X Y, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning [J]. ArXiv Preprint ArXiv: 171205526, 2017.

[81] GU T Y, DOLAN-GAVITT B, GARG S. BadNets: Identifying vulnerabilities in the machine learning model supply chain [J]. ArXiv Preprint ArXiv: 170806733, 2017.

[82] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning [C] //Proc. of the Int. Conf. on Artificial Intelligence and Statistics. New York: PMLR, 2020: 2938 - 2948.

[83] XIE C L, HUANG K L, CHEN P Y, et al. DBA: Distributed backdoor attacks against federated learning [C/OL] // [2023-03-07]. https://openreview.net/forum?id=rkgyS0VFvr.

[84] LI L P, XU W, CHEN T Y, et al. RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets [C] //Proc. of the AAAI Conf. on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 1544 - 1551.

[85] BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens [C] //Proc. of the Int. Conf. on Machine Learning. New York: PMLR, 2019: 634 - 643.

[86] WANG H Y, SREENIVASAN K, RAJPUT S, et al. Attack of the tails: Yes, you really can backdoor federated learning [C] //Proc. of Advances in Neural Information Processing Systems, New York: Curran Associates, 2020, 33: 16070 - 16084.

[87] LI Y M, JIANG Y, LI Z F, et al. Backdoor learning: A survey [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022: 1 - 18.

[88] CHEN Y D, SU L L, XU J M. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent [J]. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 2017, 1 (2): 1 - 25.

[89] YIN D, CHEN Y D, KANNAN R, et al. Byzantine-robust distributed learning: Towards optimal statistical rates [C] //Proc. of the Int. Conf. on Machine Learning. New York: PMLR, 2018: 5650 - 5659.

[90] XIE C, KOYEJO O, GUPTA I. SLSGD: Secure and efficient distributed on-device machine learning [C] //Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Cham: Springer, 2019: 213 - 328.

[91] DATA D, Diggavi S. Byzantine-resilient high-dimensional SGD with local iterations on heterogeneous data [C] //Proc. of the Int. Conf. on Machine Learning. New York: PMLR, 2021: 2478 - 2488.

[92] PILLUTLA K, KAKADE S M, HARCHAOUI Z. Robust aggregation for federated learning [J]. IEEE Transactions on Signal Processing, 2022, 70: 1142 - 1154.

[93] YU Q, LI S Z, RAVIV N, et al. Lagrange coded computing: Optimal design for resiliency, security, and privacy [C] //Proc. of the 22nd Int. Conf. on Artificial Intelligence and Statistics. New York: PMLR, 2019: 1215 - 1225.

[94] DATA D, SONG L, DIGGAVI S. Data encoding for Byzantine-resilient distributed gradient descent [C] //Proc. of the 56th Annual Allerton Conf. on Communication, Control, and Computing. Piscataway, NJ: IEEE, 2018: 863 - 870.

[95] CHEN L J, WANG H Y, CHARLES Z, et al. Draco:

Byzantine-resilient distributed training via redundant gradients [C] //Proc. of the Int. Conf. on Machine Learning. New York: PMLR, 2018: 903 - 912.

[96] RAJPUT S, WANG H Y, CHARLES Z, et al. DETOX: A redundancy-based framework for faster and more robust gradient aggregation [C] //Proc. of Advances in Neural Information Processing Systems, Red Hook: Curran Associates, 2019: 32.

[97] SOHN J-Y, HAN D-J, CHOI B, et al. Election coding for distributed learning: Protecting signsgd against Byzantine attacks [C] //Proc. of Advances in Neural Information Processing Systems, Red Hook: Curran Associates, 2020, 33: 14615 - 14725.

[98] JIN R C, HUANG Y F, HE X F, et al. Stochastic-Sign SGD for federated learning with theoretical guarantees [J]. ArXiv Preprint ArXiv: 200210940, 2021.

[99] SHEN S Q, TOPLE S, SAXENA P. Auror: Defending against poisoning attacks in collaborative deep learning systems [C] //Proc. of the 32nd Annual Conf. on Computer Security Applications. New York: ACM, 2016: 508 - 519.

[100] XIA Q, TAO Z, HAO Z J, et al. FABA: an algorithm for fast aggregation against Byzantine attacks in distributed neural networks [C] //Proc. of the 28th Int Joint Conf. on Artificial Intelligence. Menlo Park, CA: AAAI, 2019.

[101] CAO D, CHANG S, LIN Z J, et al. Understanding distributed poisoning attack in federated learning [C] //Proc. of the IEEE 25th Int. Conf. on Parallel and Distributed Systems. Piscataway, NJ: IEEE, 2019: 233 - 239.

[102] LIU B, ZHANG F J, WANG W X. A byzantine-robust federated learning algorithm based on matrix mapping [J]. Journal of Computer Research and Development, 2021, 58 (11): 2416 - 2429.

[103] LUIS MUNOZ-GONZáLEZ, KENNETH T C, EMIL C L. Byzantine-robust federated machine learning through adaptive model averaging [J]. ArXiv Preprint ArXiv: 190905125, 2019.

[104] XU C, JIA Y, ZHU L H, et al. TDFL: Truth discovery based byzantine robust federated learning [J]. IEEE Transactions on Parallel and Distributed Systems, 2022, 33 (12): 4835 - 4848.

[105] ZHAO B, SUN P, FANG L M, et al. FedCom: A byzantine-robust local model aggregation rule using data commitment for Federated learning [J]. ArXiv Preprint ArXiv: 210408020, 2021.

[106] LI S Y, CHENG Y, LIU Y, et al. Abnormal client behavior detection in federated learning [J]. ArXiv Preprint ArXiv: 191009933, 2019.

[107] XIE C, KOYEJO S, GUPTA I. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance [C] //Proc. of the Int. Conf. on Machine Learning. New York: PMLR, 2019: 6893 - 6901.

[108] CAO X Y, LAI L F. Distributed gradient descent algorithm robust to an arbitrary number of Byzantine attackers [J]. IEEE Transactions on Signal Processing, 2019, 67 (22): 5850 - 5864.

[109] CAO X Y, FANG M H, LIU J, et al. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping [C] // [2023-03-07]. https://www.ndss-symposium.org/ndss-paper/fltrust-byzantine-robust-federated-learning-via-trust-bootstrapping/.

[110] ZHAI K, REN Q, WANG J L, et al. Byzantine-robust federated learning via credibility assessment on non-IID data [J]. ArXiv Preprint ArXiv: 210902396, 2021.

[111] PAN X D, ZHANG M, WU D C, et al. Justinian's GAAvernor: Robust distributed learning with gradient aggregation agent [C] //Proc. of the 29th USENIX Security Symp. San Sebastian: USENIX Association, 2020: 1641 - 1658.

[112] CHANG H Y, SHEJWALKAR V, SHOKRI R, et al. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer [J]. ArXiv Preprint ArXiv: 191211279, 2019.

[113] EL MHAMDI E M, GUERRAOUI R, ROUAULT S L A. Distributed momentum for byzantine-resilient stochastic gradient descent [C/OL] // [2023-03-07]. https://openreview.net/forum?id=H8UHdhWG6A3.

[114] KARIMIREDDY S P, HE L, JAGGI M. Learning from history for byzantine robust optimization [C] //Proc. of the Int. Conf. on Machine Learning. New York: PMLR, 2021: 5311 - 5319.

[115] DWORK C. Differential privacy: A survey of results [C] //Proc. of the Int. Conf. on Theory and Applications of Models of Computation. Berlin: Springer, 2008: 1 - 19.

[116] DWORK C, ROTH A. The Algorithmic Foundations of Differential Privacy [M]. Nederland: Now Foundations and Trends, 2014.

[117] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy [C] //Proc. of the ACM SIGSAC Conf. on Computer and Communications Security. New York: ACM, 2016: 308 - 318.

[118] DU M, JIA R X, SONG D. Robust anomaly detection and backdoor attack detection via differential privacy

[C] // [2023-03-07]. https://openreview.net/pdf? id=SJx0q1rtvS.

[119] XIE C L, CHEN M H, CHEN P Y, et al. CRFL: Certifiably robust federated learning against backdoor attacks [C] //Proc. of the Int. Conf. on Machine Learning. New York: PMLR, 2021: 11372 - 11382.

[120] NGUYEN P R T D, CHEN H, Yalame H, et al. FLAME: Taming backdoors in federated learning [C] //Proc. of the USENIX Security Symp. San Sebastian: USENIX Association, 2022.

[121] SUN J W, LI A, DIVALENTIN L, et al. FL-WBC: Enhancing robustness against model poisoning attacks in federated learning from a client perspective [C] // Proc. of Advances in Neural Information Processing Systems, New York: Curran Associates, 2021, 34: 12613 - 12624.

[122] CAO X Y, FANG M H, LIU J, et al. FLtrust: Byzantine-robust federated learning via trust bootstrapping [C] //Proc. of the 28th Annual Network and Distributed System Security Symp., 2021.

[123] 郭娟娟, 王琼霄, 许 新, 等. 安全多方计算及其在机器学习中的应用 [J]. 计算机研究与发展, 2021, 58 (10): 2163 - 2186.

[124] MA Z R, MA J F, MIAO Y B, et al. ShieldFL: Mitigating model poisoning attacks in privacy-preserving federated learning [J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 1639 - 1654.

[125] MIAO Y B, LIU Z T, LI H W, et al. Privacy-preserving Byzantine-robust federated learning via blockchain systems [J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 2848 - 28461.

[126] LIU X Y, LI H W, XU G W, et al. Privacy-enhanced federated learning against poisoning adversaries [J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4574 - 4588.

[127] MA X D, ZHANG J Y, MA J F, et al. Do not perturb me: A secure byzantine-robust mechanism for machine learning in IoT [C] //Proc. of the Int. Conf. on Networking and Network Applications. Piscataway, NJ: IEEE, 2020: 348 - 354.

[128] MA Z R, MA J F, MIAO Y B, et al. Pocket diagnosis: secure federated learning against poisoning attack in the cloud [J]. IEEE Transactions on Services Computing, 2022, 15 (6): 3429 - 3442.

[129] Globalplatform. Introduction to trusted execution environments [EB/OL]. (2017-01-13) [2023-03-07]. https://globalplatform.org/resource-publication/introduction-to-trusted-execution-environments/.

[130] MCKEEN F, ALEXANDROVICH I, BERENZON A, et al. Innovative instructions and software model for isolated execution [C] //Proc. of the 2nd Int Workshop on Hardware and Architectural Support for Security and Privacy. New York: Association for Computing Machinery, 2013.

[131] LIMITED A. ARM security technology-building a secure system using trustzone technology [R/OL]. [2023-03-07]. https://developer.arm.com/documentation/PRD29-GENC-009492/c? lang=en.

[132] TRAMER F, BONEH D. Slalom: Fast, verifiable and private execution of neural networks in trusted hardware [C] // (2018-01-03) [2023-03-07]. https://openreview.net/pdf? id=rJVorjCcKQ.

[133] 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.

[134] MO F, HADDADI H. Efficient and private federated learning using TEE [C] //Proc. of the EuroSys Conf, Dresden, Germany, 2019.

[135] ZHAO L C, JIANG J L, FENG B, et al. SEAR: Secure and efficient aggregation for byzantine-robust federated learning [J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19 (5): 3329 - 3342.

[136] MUHR T, ZHANG W. Privacy-preserving detection of poisoning attacks in federated learning [C] //Proc. of the 19th Annual Int. Conf. on Privacy, Security & Trust. Piscataway, NJ: IEEE, 2022: 1 - 10.

[137] SO J, GüLER B, AVESTIMEHR A S. Byzantine-resilient secure federated learning [J]. IEEE Journal on Selected Areas in Communications, 2021, 39 (7): 2168 - 2181.

[138] GUO J J, LI H Y, HUANG F R, et al. ADFL: A poisoning attack defense framework for horizontal federated learning [J]. IEEE Transactions on Industrial Informatics, 2022, 18 (10): 6526 - 6536.

[139] LI Q B, DIAO Y Q, CHEN Q, et al. Federated learning on non-IID data silos: An experimental study [C] //Proc. of the 38th Int. Conf. on Data Engineering. Piscataway, NJ: IEEE, 2022: 965 - 978.

[140] PENG J, WU Z X, LING Q, et al. Byzantine-robust variance-reduced federated learning over distributed non-IID data [J]. Information Sciences, 2022, 616: 367 - 391.

[141] SMITH V, CHIANG C-K, SANJABI M, et al. Federated multi-task learning [C] //Proc. of Advances in Neural Information Processing Systems, Red Hook:

Curran Associates，2017：4427－4437.

[142] GHOSH A，CHUNG J，YIN D，et al. An efficient framework for clustered federated learning [C] // Proc. of Advances in Neural Information Processing Systems，New York：Curran Associates，2020，33：19586－19597.

[143] YU T，BAGDASARYAN E，SHMATIKOV V. Salvaging federated learning by local adaptation [J]. ArXiv Preprint ArXiv：200204758，2020.

[144] SINGHAL K，SIDAHMED H，GARRETT Z，et al. Federated reconstruction：Partially local federated learning [C] //Proc. of Advances in Neural Information Processing Systems，New York：Curran Associates，2021，34：11220－11232.

[145] LIU Z Y，GUO J L，LAM K Y，et al. Efficient dropout-resilient aggregation for privacy-preserving machine learning [J]. IEEE Transactions on Information Forensics and Security，2023，18：1839－1854.

[146] CHEN X H，JI J L，LUO C Q，et al. When machine learning meets blockchain：A decentralized，privacy-preserving and secure design [C] //Proc. of the IEEE Int. Conf. on Big Data. Piscataway，NJ：IEEE，2018：1178－1187.

[147] KIM H，KIM S-H，HWANG J Y，et al. Efficient privacy-preserving machine learning for blockchain network [J]. IEEE Access，2019，7：136481－136495.

[148] ZHANG C，ZHOU B Y，HE Z Q，et al. OBLIVION：Poisoning Federated Learning by Inducing Catastrophic Forgetting [C] // IEEE International Conference on Computer Communications. [S. l.：s. n.]，2023.