计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：联邦学习中的安全威胁与防御措施综述
作者：陈学斌，任志强，张宏扬
收稿日期：2023-07-04
网络首发日期：2023-08-01
引用格式：陈学斌，任志强，张宏扬．联邦学习中的安全威胁与防御措施综述[J/OL]．计算机应用. https://kns.cnki.net/kcms2/detail/51.1307.TP.20230731.1744.024.html

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

doi:10.11772/j.issn.1001-9081.2023060832

联邦学习中的安全威胁与防御措施综述

陈学斌[1,2,3]，任志强[1,2,3*]，张宏扬[1,2,3]

(1.华北理工大学 理学院，河北 唐山 063210；
2.河北省数据科学与应用重点实验室(华北理工大学)，河北 唐山 063000；
3.唐山市数据科学重点实验室(华北理工大学)，河北 唐山 063210)

(*通信作者电子邮箱 psp1274632466@qq.com)

摘 要: 联邦学习是一种用于解决机器学习中数据共享问题和隐私保护问题的分布式学习方法，旨在多方共同训练一个机器学习模型并保护数据的隐私。但是，联邦学习本身存在安全威胁，这使得联邦学习在实际应用中面临巨大的挑战。因此，分析联邦学习面临的攻击和相应的防御措施对联邦学习的发展和应用至关重要。首先介绍了联邦学习的定义、流程和分类，联邦学习中的攻击者模型；然后，从联邦学习系统的鲁棒性和隐私性两方面介绍了可能遭受的攻击，并对不同攻击介绍了相应的防御措施，同时也指出防御方案的不足之处。最后，展望了安全的联邦学习系统。

关键词: 联邦学习；隐私保护；攻击与防御；机器学习；鲁棒性与隐私性

中图分类号:TP309.2 **文献标志码：A**

Review on security threats and defense measures in federated learning

CHEN Xuebin[1,2,3], REN Zhiqiang[1,2,3*], ZHANG Hongyang[1,2,3]

(1.College School of Science, North China University of Science and Technology, Tangshan 063210, China;
2. Hebei Province Key Laboratory of Data Science and Application(North China University of Science and Technology), Tangshan 063000, China;
3. Tangshan Data Science Laboratory(North China University of Science and Technology), Tangshan 063210, China)

Abstract: Federated learning is a distributed learning approach for solving the data sharing problem and privacy protection problem in machine learning, aiming at multiple parties to jointly train a machine learning model and protect the privacy of data. However, there are security threats inherent in federated learning, which makes federated learning face great challenges in practical applications. Therefore, analyzing the attacks faced by federation learning and the corresponding defensive measures are crucial for the development and application of federation learning. First, the definition, process and classification of federated learning were introduced, and the attacker model in federated learning was introduced. Then, the possible attacks in terms of both robustness and privacy of federated learning systems were introduced, and the corresponding defense measures for different attacks were introduced as well. Furthermore, the shortcomings of the defense schemes were also pointed out. Finally, a secure federated learning system was envisioned.

Keywords: federated learning; privacy protection; attack and defense; machine learning; robustness and privacy

0 引言

机器学习通过经验自动改进计算机的能力已经得到广泛应用[1]，但是数据的孤岛现象[2]和个人隐私保护的要求对传统中心化的机器学习方法提出了挑战。为了解决这些问题，联邦学习[3]作为一种新兴的技术，被用来打破数据孤岛的同时保护数据隐私。联邦学习的概念最早是在 2016 年由谷歌提出的，它本质上是一种分布式机器学习技术、一种机器学习框架，能够让多个数据拥有方共同参与机器学习过程，而不必把数据集中在一个地方。这种方法可以在不暴露数据的情况下训练模型，并且可以实现多方共同受益。此外，联邦学习也符合欧盟 GDPR(General Data Protection Regulations)[4]等隐私保护法规的要求，因为它可以确保个人数据不出边界，并且只有在得到用户允许的情况下才进行模型训练。因此，联邦学习有望成为未来机器学习领域的重要技术之一。

收稿日期: 2023-07-04; **修回日期**:2023-07-15; **录用日期**: 2023-07-25。
基金项目: 国家自然科学基金面上项目（U20A20179）
作者简介: 陈学斌(1970—)，男，河北唐山人，教授，博士，CCF 杰出会员，主要研究方向：大数据安全、物联网安全、网络安全等；任志强(2000—)，男，四川广元人，硕士研究生，CCF 会员，主要研究方向：数据安全、隐私保护；张宏扬(1999—)，男，江苏淮安人，硕士研究生，主要研究方向：数据安全、隐私保护。

联邦学习作为一种分布式机器学习技术，它通过在本地设备上训练模型，然后将模型参数上传到中央服务器进行聚合，从而实现在保护用户数据隐私的前提下，利用大量分散的数据来训练更加准确的模型。然而，联邦学习本身也存在着一些安全问题：

(1)数据隐私泄露风险：在联邦学习中，攻击者可能从模型训练过程中的模型参数中推断出本地设备上的数据信息，从而泄露用户隐私。

(2)恶意攻击风险：在联邦学习中，攻击者故意上传恶意模型参数，导致聚合后的模型出现错误，从而破坏模型的准确性。或者攻击者通过篡改数据和模型参数来影响模型的训练和聚合过程，从而对模型造成影响。

(3)模型逆向工程风险：攻击者通过逆向工程可能分析出联邦学习模型的参数来获取模型的机密信息。

本文介绍了联邦学习系统存在的安全威胁，主要分为两个方面：对联邦学习系统鲁棒性的威胁和对联邦学习系统隐私的威胁(见表 1)。对联邦学习系统鲁棒性的威胁包括数据投毒[5-8]、模型投毒[9-13]和后门攻击[12,14-19]。为了讨论方便，本文将参与方训练的模型称为局部模型，将服务端的模型称为全局模型。数据投毒是指攻击者恶意篡改本地数据以影响全局模型的训练和性能，模型投毒是指攻击者篡改全局模型的更新或训练过程，从而影响模型的性能和可靠性，后门攻击则是指攻击者在模型中植入后门，从而可以随时控制模型的行为和结果。针对这些鲁棒性威胁，本文介绍了鲁棒性威胁的防御措施[7,9-11,15-17,20-24]，这些防御措施主要是通过检查更新的可信性来进行防御。对联邦学习系统隐私性的威胁主要涉及推理攻击[25-29]、重构攻击[30-34]和窃取攻击[35-38]。推理攻击是指攻击者通过观察全局模型的输出，推断出某些敏感数据或信息，重构攻击是指攻击者通过分析全局模型的参数或更新，重构出原始数据或信息，窃取攻击则是指攻击者通过窃取全局模型或本地数据，获取敏感信息或知识。针对这些隐私性威胁，本文介绍了隐私性威胁的防御措施[39-52]，这些防御措施主要是采用加噪机制和加密机制(见表 1)。

本文第一节将介绍联邦学习的定义、分类和攻击者模型。第二节从鲁棒性和隐私性两方面介绍联邦学习系统的安全威胁和防御措施，并总结具体的攻击策略和防御策略。第三节展望安全的联邦学习系统。

表1 联邦学习系统安全威胁和防御措施

Tab. 1 Federal learning system security threats and defense measures

系统安全威胁	攻击者来源	攻击者模型	攻击类型	文献(攻击)	防御措施	文献(防御)
鲁棒性威胁	参与方	恶意的	数据投毒	[5-19]	检查更新的可信性	[7,9-11,15-17,20-24]
			后门攻击			
			模型攻击			
隐私性威胁	参与方/服务端	诚实且好奇的	推理攻击	[25-38]	加密机制和加噪机制	[39-52]
			重构攻击			
	系统外部		窃取攻击			

1 理论知识

1.1 联邦学习

联邦学习的定义[53]为：n 个数据拥有者 $\{F_1, F_2, ..., F_n\}$，他们希望利用本地数据 $\{D_1, D_2, ..., D_n\}$ 共同训练一个机器学习模型。传统的方式是将所有数据收集到一个中心，使用 $D = D_1 \cup D_2 \cup ... \cup D_n$ 来训练一个模型 M_{sum}。联邦学习是一个多方协作学习的过程，数据的拥有者 F_i 利用本地数据 D_i 协同地训练出全局模型 M_{fed}，并且数据拥有者 F_i 不会将自己的数据 D_i 暴露给其他数据拥有者 $F_j(j \neq i)$。将全局模型 M_{fed} 的准确率记为 V_{fed}，它应当非常接近 M_{sum} 对应的 V_{sum} 的准确率。形式上，设 δ 为非负实数，如果满足：

$$\left| V_{fed} - V_{sum} \right| < \delta \tag{1}$$

其中 δ 是一个很小的正数，则说联邦学习算有 δ-accuracy 损失。

联邦学习系统的体系结构如图 1 所示，此系统的训练过程通常包含以下步骤：

(1)参与方利用本地数据集训练局部模型

(2)参与方上传更新后的模型参数

(3)服务端聚合各个参与方的模型参数

(4)服务端广播聚合后的模型参数

不断的迭代步骤(1-4)，直至损失函数收敛,从而完成整个训练过程。

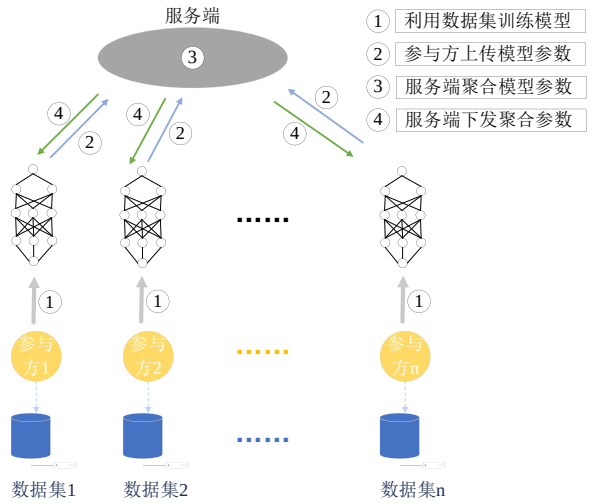

图1 联邦学习系统的体系结构

Fig. 1 The architecture of federated learning systems

1.2 联邦学习分类

联邦学习可以根据参与方的个数和数据量的不同进行分类[54]，分别为跨孤岛联邦学习（Cross-Silo Federated Learning）和跨设备联邦学习（Cross-Device Federated Learning）。其中，跨孤岛联邦学习表示参与方个数较少，但是拥有较大的数据量，例如某银行和借贷公司利用各自本地的数据共同训练一个判断是否向客户借贷的模型；跨设备联邦学习则表示参与方个数较多，但是拥有较小的数据量，例如众多手机用户使用本地数据共同训练键盘输入词的预测模型。此外，联邦学习可以根据参与方的数据包含的特征进行分类[53]，分别为横向联邦学习（Horizontal Federated Learning）、纵向联邦学习（Vertical Federated Learning）和联邦迁移学习（Federated Transfer Learning）。

1.3 攻击者模型

本文讨论联邦学习中的两种攻击模型：恶意攻击(Malicious attack)和诚实但好奇攻击(Honest-but-Curious attack)。恶意攻击者的主要目的是攻击和破坏系统，而不是获取系统中的信息或数据。在联邦学习中，这种攻击行为会导致模型向偏离正常方向的方向学习。相比之下，诚实但好奇的攻击者会遵守协议规则和保密性，但仍然会尝试了解关于协议所传输数据的更多信息的模型。在联邦学习中，这种攻击行为不会破坏模型的学习，但会尝试从传递的信息中推断出敏感信息。

2 安全威胁与防御措施

在联邦学习场景中，攻击者的主要目标通常是破坏模型或者推断隐私信息。这些攻击者可能是来自参与方或者服务端的(见图 2，以用户 n 为例)，来自参与方的攻击者可能会是恶意的，也可能是诚实但好奇的，无论哪种情况下，他们都有可能威胁到系统的鲁棒性和隐私性。而来自服务端的攻击者通常被假设为诚实但好奇的，他们可能会威胁到系统的隐私性。这些攻击者可以在不同的时间点发动攻击。来自参与方的攻击者可以在本地训练阶段或与服务端进行信息交互的阶段发动攻击，而来自服务端的攻击者则只能在与客户端进行信息交互的阶段发动攻击。此外，当最终的模型通过应用程序接口(Application Programming Interface, API)提供服务时，非系统内部的成员也可能会发动窃取攻击。

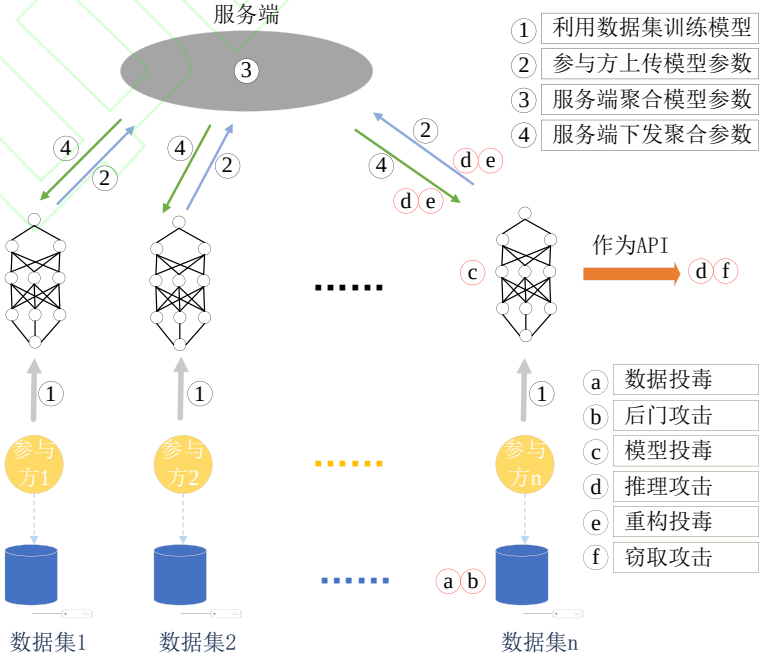

图2 联邦学习系统各阶段存在的攻击

Fig. 2 Attacks that exist at all stages of the federated learning system

2.1 针对系统鲁棒性攻击与防御

在机器学习中，鲁棒性是指模型在面对输入数据中的干扰或噪声时的表现。具有良好鲁棒性的模型即使在输入数据中存在扰动或异常情况下也能保持良好的预测能力。在联邦学习中，破坏系统的鲁棒性意味着破坏联邦学习中的全局模型。攻击者可能会使用各种方法破坏系统的鲁棒性，包括数据中毒、后门攻击和模型中毒等。本节将讨论对系统鲁邦性造成威胁的攻击者在不同情况下采用的攻击方法论(见表 2)及攻击效果(见表 3)和相应的防御措施(见表 4)及防御效果(见表 5)。显然，后门攻击属于数据投毒，但考虑到后门攻击的独特性，本文将分别介绍后门攻击和数据投毒。

表2 联邦学习系统鲁棒性威胁的攻击

Tab. 2 Federated learning system robustness threat attack

文献	鲁棒性威胁	数据分布	攻击类型	攻击方法论	补充说明
[6]	数据投毒	非独立同分布	无目标投毒	利用投影随机梯度上升算法最大限度地增加了目标节点的经验损失	
[5]				预测由恶意输入而引起的 SVM 决策函数的变化，并利用这种能力构造恶意数据。	
[7]		独立同分布	有目标投毒	力求最后几轮中成功投毒，并选择合适的标签翻转对象	未考虑参与方数据是非独立同分布的情况
[8]		非独立同分布		利用 GAN 技术生成数据并实施标签翻转	攻击效果依赖攻击者选择的攻击时机
[9]	模型投毒		无目标模型投毒	向局部模型添加随机噪声	
[9]			有目标模型投毒	根据安全聚合算法，修改局部模型参数使全局模型向相反方向更新	
[10]				最大化恶意模型的更新，同时限制更新避免被检测	
[12]				放大恶意模型更新,使得一次攻击留下足够强的后门	需要人工估计局部模型的放大因子
[13]				利用全局模型的历史更新推出反向更新的方向，并提交大量放大后的攻击模型更新	
[14]	后门攻击	非独立同分布	标记后门	多个攻击者分布式地植入后门	
[12]			语义后门	选择诚实参与者数据集中较少的特征作为后门特征	假设攻击者已知异常检测策略
[17]				增大攻击者比例，降低后门任务的复杂性，限制恶意更新	
[18]				利用数据集的分布特点，挑选边缘数据实施后门攻击，并限制恶意更新	
[19]			后门攻击	后门模型和主任务模型分开训练，再通过优化结合为一个模型	

表3 联邦学习系统鲁棒性威胁的攻击效果表

Tab. 3 Attack effect table of federated learning system robustness threat

文献	攻击评价指标	数据集	训练设置/(个:条)	结果/%	说明
[6]	模型错误率	EndAD	6:*:no-iid	base: 6.881±0.52 result: 28.588±3.74	
		Human Activity	30:*:no-iid	base: 2.586±0.84 result: 29.422±2.96	
		Landmine	29:*:no-iid	base: 5.682±0.28 result: 13.648±0.54	
[5]	模型错误率	MNIST	*:*:*	base: 2-5 result: 15-20	
[7]	最大召回率损失	CIFAR-10	50:*:iid	base: 0 result: 2: 1.42; 20: 25.4	随不同恶意参与者百分比(a: b,a 表示百分比,b 表示损失)
		Fashion-MNIST	50:*:iid	base: 0 result: 2: 0.61; 20: 29.2	

[8]	中毒任务准确率	MNIST	10:*:no-iid	base: 0 result: 20: 60±; 40: 80±; 60: 85±	随不同放大因子 (a: b,a 表示放大因子,b 表示准确率)
		AT&T	10:*:no-iid	base: 0 result: 20: 70±; 40: 85±; 60: 90±	
[10]	最大准确度损失	CIFAR10(Alexnet)	50:1000:*	base: 0 result: Krum: 43.6; Mkrum: 36.8; Bulyan: 45.6; TrMean: 45.8; Median: 40.9; AFA: 47.0; FangTrmean: 56.3	在不同聚合算法下 (a: b,a 表示聚合算法,b 表示损失)
[12]	后门任务准确率	CIFAR10	100:*:no-iid	base: 0 result: 95: 30±,75±	一次攻击、随全局迭代 (a: b,c,a 表示轮次,b 代表最低准确度,c 代表最高准确率)
		Reddit dataset	83293:247:no-iid	base: 0 result: 95: 0±,60±	
	后门任务准确率	CIFAR10	100:*:no-iid	base: 0 result: 1: 30±,50±; 5: 80±,80±	持续攻击、随恶意参与者百分比 (a: b,c,a 表示恶意参与者占比,b 代表最低准确度,c 代表最高准确率)
		Reddit dataset	83293:247:no-iid	base: 0 result: 0.01: 30±,65±; 0.1: 80±,90±	
[13]	模型准确率	MNIST	1000:*:no-iid	base: 99± result: FedAvg: 1-25: 10±; Median: 1: 90±; Median: 25: 60±; Trimmed-mean: 1: 95±; Trimmed-mean: 1: 50±	随恶意参与者百分比 (a: b: c,a 表示聚合算法,b 表示恶意参与者占比,c 表示准确率)
[14]	攻击成功率	LOAN	51:*:no-iid	base: 0 8: 99±	随全局迭代 (a: b,a 表示迭代数,b 表示成功率)
		MNIST	100:*:no-iid	base: 0 20: 99±	
		CIFAR	100:*:no-iid	base: 0 350: 80±	
		Tiny-imagenet	100:*:no-iid	base: 0 80: 80±	
[17]	后门任务准确率	MNIST	*:*:no-iid	base: 0 result: 10: 120: 95±; 50: 300: 70±	固定任务数、随全局迭代 (a: b: c,a 表示任务数,b 表示迭代数,c 表示准确率)
[18]	后门任务准确率	CIFAR-10	200:*:no-iid	base: 0 result: 100: 400: 55±; 10: 400: 15±	(a: b: c,a 表示攻击资源占比,b 表示迭代数,c 表示准确率)
		MNIST	20:*:*	base: 0 result: 10: 80±; 20: 90±	(a: b,a 表示迭代数,b 表示准确率)
		CIFAR-10	50:*:*	base: 0 result: 300: 40±; 400: 90±	
注：在"训练设置列"，采用 x:y:z 格式,x 代表参与者,y 代表每个参与方拥有数据量,z 代表数据集的划分(iid 或 no-iid),"结果列"中"base"表示未攻击的情况下的基准值,"result"表示攻击后,"*"代表未知。					

表4 防御鲁棒性威胁的措施

Tab. 4 Measures to defend against robust threats

文献	防御类型	数据分布	针对攻击类型	防御思想	防御方式	补充说明
[20-21]	数据投毒	独立同分布	无目标投毒\有目标投毒	基于行为	利用鲁棒性的分布式梯度下降算法聚合模型	
[7,22]		独立同分布	有目标投毒	基于聚类	利用聚类算法鉴别出恶意模型	
[11]			有目标投毒	基于行为	根据局部模型与全局模型的余弦相似度判断恶意模型	存在超参数

[23]		独立同分布\非独立同分布	有目标投毒		根据局部模型与全局模型的余弦相似度并结合信誉机制共同判断恶意模型	
[9]	模型投毒		无目标模型投毒	基于行为	利用拜占庭鲁棒性算法	
[11]					基于局部更新与全局更新的余弦相似度去除恶意梯度	
[9]			有目标模型投毒		基于错误率和基于损失函数的评价指标并结合拜占庭鲁棒性聚合算法防御	
[10]					基于奇异值分解(SVD)的谱方法来检测和去除异常值	
[15]	后门攻击	非独立同分布	标记后门攻击	基于行为	设置一个超参数，对每一轮中更新的每个维度进行投票，根据其投票值与超参数值进行比较并依此动态调节学习率。	存在超参数
[16]		独立同分布\非独立同分布	标记后门攻击	针对机器学习模型本身	剪裁异常神经元，约束神经元权重，微调模型	微调可能导致后门加深
[24]		独立同分布\部分非独立同分布	语义后门攻击	混合策略(基于聚类、基于行为)	结合聚类和分类综合判定一个模型是否有害，通过剪裁减弱逃过检测的有毒模型	

表5 防御鲁棒性威胁措施的效果

Tab. 5 Effectiveness of measures to defend against robust threats

防御指标	文献	防御结果/%	说明
模型错误率	[20]	base: 10± attack: 60± after: 10±	随着全局迭代的最终结果
模型准确率	[21]	base: 94.3± attack: 77.3± after: 90.7±	随着全局迭代的最终结果
模型准确率	[22]	base: 78± attack: 76±、74.5± after: 78±、77.5±	随着全局迭代的最终结果,"attack"和"after"中分别对应 20%和%30 的恶意参与者占比
模型错误率	[11]	base: 2.80±0.12 attack: * after: 2.99±0.12±、2.96±0.15、3.04±0.14	随着全局迭代的最终结果,"after"中分别对应 Byzantine、Flipping 和 Noisy 攻击下的结果
模型准确率	[23]	base: * attack: * after: 83.11、81.23	随着全局迭代的最终结果,"after"中分别对应 5%和 50%的恶意参与者占比
模型错误率	[9]	base: 0.12 attack: * after: 0.12	随着全局迭代的最终结果
模型准确率损失	[10]	base: 0 attack: * after: 4.3	随着全局迭代的最终结果
后门任务准确率	[15]	base: 6.6 attack: 88.6 after: 9.0	随着全局迭代的最终结果
	[16]	base: * attack: 85.5 after: 4.8	
	[24]	base: * attack: 100 after: 0	
注："防御结果"中的"base"表示没有受到攻击时的结果，"attack"表示攻击后的结果，"after"表示使用防御措施后的结果。			

2.1.1 数据投毒

在机器学习中，数据投毒是指攻击者通过恶意更改本地源数据，来影响模型的学习和预测结果，以达到攻击目的的一种方法。在联邦学习中，数据投毒攻击可以分为无目标的投毒攻击[5-6]和有目标的投毒攻击[7-8]两种。无目标的投毒攻击注重于完全破坏模型性能，使全局模型的准确率越低越好，甚至使模型不能收敛，这样才算攻击成功。而有目标的投毒攻击则注重于特定预测任务，攻击者不希望破坏模型对其他任务的预测，而是通过将原标签值更改为另一个标签值等方式，来干扰模型对特定任务的预测[7]。

对于投毒攻击来说，攻击者的目标是通过篡改本地数据来影响联邦学习中的局部模型，从而影响整个模型的准确性和鲁棒性。攻击者可以通过在最后几轮迭代学习中参与，来增大攻击的威胁，因为在这些轮次中，全局模型已经接近收敛，而攻击者能够让其偏离正确方向。此外，攻击者参与的轮数和操作的数据量也会影响攻击的效果。如果攻击者参与的轮数较少，那么其篡改的数据可能无法影响到整个模型；如果攻击者操作的数据量很少，那么其影响也会被削弱。因此，攻击者的人数、参与的轮数和操作的数据量越多，攻击的效果就越强。相反地，诚实参与者的参与会削弱或甚至消除投毒攻击的影响[7]。

针对无目标的投毒攻击，文献[5]提出了一种针对支持向量机(Support Vector Machine, SVM)的攻击方式，利用梯度上升策略，根据支持向量机最优解的性质计算梯度，并通过注入精心构建的训练数据，从而增加 SVM 的测试错误。但是，这种攻击仅适用于 SVM 模型。文献[6]提出了 AT2FL(Attack On Federated Learning)算法，设计了一种基于投影随机梯度上升的方法，能够有效地推导出中毒数据的隐式梯度，并利用其计算出最优攻击策略。该文献还证明了在多任务联邦学习情景下，联邦多任务学习模型很容易受到数据投毒攻击的影响，且随着恶意数据的数量增加，模型的性能将变得越来越糟糕。

针对有目标的数据投毒攻击，文献[7]采用标签翻转攻击进行实验，该攻击证明即使恶意参与方的比例很小(低至总参与方的 4%)，攻击也可以显著地影响联邦学习的效果。同时，该攻击可能会因数据集和翻转对象的不同而产生较大的差异。然而，文献[7]没有考虑到数据非独立同分布的情况。在此基础上，文献[8]结合生成对抗网络(Generative Adversarial Nets, GAN)实现了标签翻转攻击，该攻击可以使主要任务和中毒任务的准确率都达到 80%以上。攻击者首先会伪装成正常的参与方参与联邦学习。当全局模型达到一定的准确度后，攻击者使用 GAN 技术生成一些类似于其他正常参与方的数据，并将数据的标签进行翻转。接下来，攻击者将局部模型训练时的数据源更换为上一步得到的数据，并进行局部模型训练。该攻击适用于特定场景，例如在分类任务中，攻击者本地没有某类数据，但希望将该类作为攻击的目标。但需要注意的是，该攻击依赖于 GAN 生成的模拟数据，而 GAN 生成数据的质量与联邦学习中的全局模型质量有关。因此，攻击者选择的攻击时机非常重要。

2.1.2 模型投毒

在本文中，联邦学习中的模型投毒攻击是指直接修改模型更新的行为。这种攻击比数据投毒更直接地危害系统的鲁棒性，因为它直接作用于模型参数而不是修改训练数据集。虽然模型投毒攻击和数据投毒攻击都是修改了局部模型，但它们的攻击时机不同。数据投毒攻击是通过修改训练数据集来学习到不同于正常情景下的模型参数，而模型投毒攻击直接修改模型更新的参数。数据投毒攻击通常是有目标的攻击，而模型投毒攻击更能直接对模型参数进行有目标的缩放。因此，模型投毒攻击更常见的是与其他攻击结合使用，例如结合后门攻击来增加后门的效果[12]。模型投毒[9-13]攻击通常分为两种实施方式。第一种是无目标模型投毒，攻击者向训练出的局部模型中添加随机噪声[9,11]来扰乱其学习过程。而第二种是有目标模型投毒，攻击者会更加精细地修改局部模型[12-13]，以便让全局模型朝着某个特定的方向进行学习，或者是为了针对已有的防御算法进行攻击[9-10]。由于有目标的模型投毒攻击更加普遍，因此本文主要讨论这种类型的攻击。

文献[9]证明了有目标的模型投毒攻击是可行的，并指出针对无目标投毒攻击的拜占庭鲁棒性聚合防御(如中值聚合、修剪平均聚合、Krum)不能单独用于防御有目标的模型投毒攻击。文献[10]提出了一个通用的模型投毒攻击框架，该框架通过在恶意方向上最大限度地扰动良性参考聚合来计算恶意模型更新，并限制模型更新以避免被健壮的聚合算法检测到。文献[12]提出了模型替换攻击，该攻击通过假设并反解联邦平均聚合算法，解出使用一次攻击就能使全局模型达到局部模型的解，最终只需要确定一个局部模型的放大因子就能实现模型替换攻击。但该攻击的缺点是需要人工估计局部模型的放大因子。文献[13]假设攻击者仅知道训练过程中的全局模型，甚至没有正常的数据集。在这种情况下，提出了一种基于伪客户端的模型中毒攻击 MPAF(Model Poisoning Attack based on Fake clients)，与上述攻击方式不同的是，它旨在将全局模型“拉向”攻击者指定的基本模型。其攻击的关键步骤是：1)随机选择测试精度低的基本模型。2)根据全局模型的历史更新生成攻击更新，并在将其发送到服务端前将其放大以扩大攻击效果。

2.1.3 后门攻击

联邦学习中的后门攻击[12,14-19]指的是，恶意攻击者利用特殊的数据集参与联邦学习，最终影响全局模型，并在特定的输入下激活后门输出攻击者想要的输出。后门攻击与数据投毒相似又不同。它们相似之处在于都对参与方的源数据进行攻击，但不同之处在于后门攻击具有隐蔽性，且不会干扰

模型对正常输入的预测。有些文献将后门攻击称为有针对性的模型投毒攻击。

后门攻击可以根据实施方式是否更改本地数据来分为两种类型：标记后门攻击[14-16]和语义后门攻击[12,17-18]。在标记后门攻击中，攻击者会向原始数据添加特殊标记，例如在图像中添加一个马赛克[16]。这种攻击方式的效果更难以消除，因为模型已经记住了特定标记。在语义后门攻击中，攻击者利用数据集的特征，例如将数据集中所有绿色汽车的标签值修改为鸟类[12]。这种攻击方式可以通过诚实参与者拥有相似数据集的参与逐步消除后门。无论哪种攻击方式，后门攻击的效果只在特定的输入时才会触发后门。在多位研究者的研究下[12,14]，证明了后门攻击凭借其隐蔽性，能够在仅发动攻击成功一次的情况下，使得全局模型能在多轮的迭代中保留后门。本文主要讨论语义后门攻击。

针对标记后门攻击，有文献[14]提出了一种分布式的后门攻击方法，并证明了相较于集中式的后门攻击，联邦学习遭受到分布式后门攻击的危害更大。此外，该文献还证明了联邦学习中一些鲁棒性的聚合算法(例如 RFA(Robust Federated Aggregation)和 FoolsGold)无法防御分布式后门攻击。

针对语义后门攻击，已有多篇文献提出了不同的方法和策略。其中，文献[16]证明了基于特定标记的后门攻击在数据非独立同分布程度越高时攻击越有效。文献[12]提出了规避防御的语义后门攻击，并通过放大本地更新实现模型替换，同时也证明了攻击效果会随着诚实参与者的参与而下降。然而，假设攻击者已经熟知异常检测策略，这一假设并不太现实。为此，文献[17]提出了范数有界攻击(Norm Bounded Backdoor Attack)，通过将更新进行约束以规避一些防御措施，并证明了其攻击的有效性。同时，该文献量化了恶意攻击者的人数与参与攻击的频率对后门攻击的影响，其结果是攻击者比例越大、攻击者参加频率越高则后门攻击的效果越好。此外，文献[18]提出了边缘后门攻击(Edge-Case Backdoors)，利用分布在边缘的数据(不太可能出现在训练集和测试集)来实现后门攻击，并从理论和实验证明了其难以检测和防御。最后，文献[19]提出了基于优化模型的后门攻击，通过将冗余神经元训练为对抗神经元来实现攻击。该文献通过实验证明了这种后门攻击不仅能实现较高的攻击成功率，还能规避一些防御措施。

2.1.4 防御破坏系统鲁棒性的措施

在联邦学习中，数据投毒、模型投毒和后门攻击都会破坏系统的鲁棒性。尽管针对数据投毒的防御措施通常是直接针对数据集进行筛选并去除其中的恶意数据，但由于联邦学习不能操作参与方的数据，因此这些针对数据集的防御措施违背了联邦学习的隐私保护要求，所以必须从其他方面考虑防御措施。通过分析联邦学习的流程，容易知道防御措施只能在服务端进行聚合的阶段来实施。数据投毒对模型的影响难以预测，而模型投毒能对模型进行精细的缩放，后门攻击由于其隐蔽性和常与模型投毒结合使得防御更加困难。因此，后门攻击的防御难度更高，需要更加精细的设计。同时，现实数据的非独立同分布特性使得判断一个模型更新是否为恶意模型更新需要更细致的考虑。

在联邦学习中，参与方的模型参数或梯度信息被上传到服务端，因此防御措施主要集中在模型质量和模型之间的差异方面。目前，主要有三种防御思想：1)聚类：通过对模型参数进行聚类分析，区分好坏模型，从而识别潜在的恶意参与方，相关研究包括文献[7,22,24]等。2)基于行为的防御：通过分析参与方上传的模型在行为方面的特征，如局部更新与全局更新之间的相似性、部分模型聚合后的错误率、模型更新的阈值等，来识别潜在的恶意参与方，相关研究包括文献[9-11,15,17,20-21,23]等。3)针对机器学习模型本身进行防御：这种方法主要是针对机器学习模型本身的弱点进行防御，如神经网络中的神经元。其目的是防止参与方利用这些弱点来破坏模型的准确性和安全性，相关研究包括文献[16]等。

对于数据投毒的防御：针对数据投毒攻击，可以根据攻击是否有目标采用不同的防御措施。对于无目标攻击，一种轻量级的防御措施是 Krum 聚合算法[20]，它选择欧几里得距离最小的局部模型进行聚合。另外，中值聚合(Median Aggregation)和修剪平均值聚合(Trimmed Mean Aggregation)也被证明在遭受无目标攻击时具有鲁棒性[21]。这些方法的一个潜在缺点是，它们可能无法有效地抵御有目标攻击。对于有目标攻击，防御标签翻转攻击的一个算法是基于聚类的思想，它记录每个参与方的局部更新与全局更新的差值，并使用主成分分析(Principal Component Analysis, PCA)技术进行数据降维以观察正常参与方与恶意攻击者上传的更新[7]。文献[22]在此基础上提出了使用 KCPA(Kernel Principal Component Analysis)和 K-means 聚类代替 PCA 的方法，从而获得更好的防御效果。此外，自适应联邦学习(Adaptive Federated Averaging, AFA)和 CONTRA[23]是基于行为的防御方法，它们利用余弦相似度来确定局部模型的可信度，并通过信誉方案来根据单个客户的每一轮和对全局模型的历史贡献动态提升或惩罚单个客户。然而，这些基于聚类和行为的防御方法在面对非独立同分布的数据时可能会出现误判的情况。

对于模型投毒的防御：对于模型投毒攻击，可以采取不同的防御措施来提高模型的鲁棒性。针对无目标的模型投毒攻击，可以使用拜占庭鲁棒性聚合算法来防御[9,11]，该方法通过将多个参与方的模型权重聚合，从而降低恶意攻击的影响。具体来说，可以采用中值聚合、修剪平均聚合和 Krum 算法等来实现。这些算法可以有效地防御无目标的模型投毒攻击，因为这种攻击与无目标的投毒攻击在一定程度上对模型造成的危害方向相似。对于更加精细的有目标模型投毒攻击，则需要设计更加精细的防御才能有效地抵抗。一种常见的方法是采用基于错误率和基于损失函数的评价指标，并结

合拜占庭鲁棒性聚合算法来实现防御[9]，这种方法可以在一定条件下有效地提高模型的鲁棒性。此外，文献[11]提出了AFA 算法，其基于局部更新与全局更新的余弦相似度去除恶意梯度，可以有效地防御无目标的模型投毒攻击。不过，该算法存在一些超参数需要人工指定。文献[10]提出了DNC(Divide-and-Conquer)算法，其利用基于奇异值分解(Singular Value Decomposition, SVD)的光谱方法进行异常值检测和去除，可以提高模型的鲁棒性。

对于后门攻击的防御：后门攻击是一种特殊的攻击手段，其目的是通过在模型中植入后门，以在特定条件下产生错误的输出。由于后门攻击的高度隐蔽性，传统的数据投毒攻击防御策略并不一定适用于此种攻击方式[24]。因此，需要更加精细的设计来防御后门攻击。在防御标记后门攻击方面，研究人员提出了多种方法。文献[16]通过剪枝冗余神经元并对权重偏离正常值的神经元进行参数约束来防御后门攻击。文献[15]通过动态调整每轮中每个更新维度的学习率来降低恶意参与者的更新的影响。文献[17]提出了使用更新范数阈值约束和差分隐私技术防御后门攻击。其中，更新范数阈值约束方法是服务器简单地忽略规范高于某个阈值 M 的更新，而差分隐私技术则通过添加噪声来防御后门攻击。对于语义后门攻击，文献[24]提出了一种名为 DeepSight 的模型过滤方法。DeepSight 结合了聚类和分类策略来鉴别出有毒模型，并通过剪裁来减弱逃过检测的有毒模型。

2.2 针对系统隐私性攻击与防御

机器学习面临的隐私攻击也对联邦学习构成威胁，因为联邦学习过程中涉及到参与者的数据隐私和全局模型的隐私。攻击者可以通过联邦学习交互过程中的信息或模型本身的信息来推断隐私信息，从而破坏系统的隐私。在联邦学习中，可能会发生的隐私攻击包括推理攻击、重构攻击和窃取攻击，这些攻击可以来自系统内部的参与者和服务端，也可以来自系统外部的 API 请求者(见表 6)。针对系统内部的攻击，参与者和服务端可能会利用模型访问权限，使用模型参数或梯度信息来推断私密信息。为了防止这种攻击，联邦学习系统需要采取措施来限制参与者和服务端的访问权限，并对交互数据进行加密和匿名化处理(见表 7)。在系统外部，API 请求者可能会试图利用 API 请求中的信息来推断私密信息，例如使用模型输出推断输入数据的敏感信息。

表6 联邦学习系统隐私性威胁的攻击

Tab. 6 An attack on privacy threats to the federal learning system

文献	隐私性威胁	攻击者来源	攻击者知识
[26,28-29]	成员推理攻击	系统内部	无需额外知识
[26]		系统外部	模型 API；模型的训练数据的背景知识
[28]			模型损失函数；损失范围
[30-31,33-34]	重构攻击	系统内部	无需额外知识
[32]		系统内部服务端	无需额外知识
[35]	窃取攻击	系统外部	输出置信度的模型 API；模型架构
[36-37]			输出标签的模型 API；模型架构
[38]			输出置信度的模型 API

表7 防御隐私性威胁的措施

Tab. 7 Measures to protect against privacy threats

防御措施	防御技术	防御思想	缺点	文献
加噪机制	客户级差分隐私	掩盖原始梯度信息	服务端必须可信	[40]
	本地级差分隐私	掩盖原始梯度信息	较大地影响模型性能	[41-43]
加密机制	同态加密	在密文上进行计算	加密效率低、密文的膨胀率很高	[48-50]
	秘密分享	将秘密信息划分多份	增加计算成本和通信成本	[51-52]

2.2.1 推理攻击

推理攻击[25-29]旨在利用可获得的信息来推断出系统不想暴露的敏感信息，这对于机器学习模型来说尤其危险。攻击者可以利用模型的相关信息或者模型的 API 来进行推理攻击。在机器学习中，推理攻击的经验依据包括以下几个方面：在对自然语言文本序列进行分类或预测的神经网络模型中，对于某些特殊的训练数据序列，存在被生成模型无意识记忆的风险[25]，这意味着攻击者可以通过这些特殊的训练数据序列来识别出模型的一些敏感信息。机器学习模型对训练数据样本和非训练数据样本的表现不同[26]等。

推理攻击对隐私泄露的影响非常严重，可以揭示出一些敏感信息，如患者的病史、人的种族和性别等等。例如，通过成员推理攻击，攻击者可以揭示某个患者是否患有某种疾病，因为他们知道该患者的治疗记录被用于训练某种与该疾病相关的模型[26]。另外，利用属性推理攻击，攻击者可以在

分类男女的人脸识别机器学习模型中，同时预测输入是否为白种人[27]。推理攻击可以细分为不同的类型，但在联邦学习的情景下，本节仅讨论成员推断攻击。

对于成员推理攻击，文献[26]提出了一种攻击方法，该方法适用于已经部署为服务的分类机器学习模型。攻击方法的基本思路是训练一个推理模型来判断输入数据是否是分类模型的训练数据。具体而言，攻击者会输入分类模型的训练数据和非训练数据，并获取分类模型的置信度值输出。然后，将输出与数据对应的标签作为推理模型的训练数据，并根据输入是否用于训练过分类模型，作为上一步训练数据的标签，以此训练一个推理模型来判断输入是否为训练数据。然而，将此攻击应用到联邦学习中，攻击来自系统外部的情况下，存在两个难题。首先，模型的输出可能不是置信度值。其次，攻击者缺乏模型的训练数据。为了解决这些难题，文献[26]假设攻击者拥有模型的 API 或模型架构及训练算法，并通过一些背景知识（如统计信息、一部分训练数据、加噪后的训练数据）来获得数据集，并利用其训练出与全局模型相似的影子模型，再进行上述推理模型的训练。对于系统外部的攻击者而言，可能很难获得模型训练的数据的背景知识，同时训练多个影子模型可能需要更多的算力资源和数据集资源。文献[28]提出了一种成员推理攻击方法，即根据输入的损失来实施攻击。具体而言，对于参与过训练的数据输入，模型的函数损失会在一定范围内；而对于未参与训练的数据输入，函数损失会超出这个范围。通过实验证明，这种攻击方法的推理准确率虽然略低于文献[26]提出的算法，但不需要更多的计算资源。将此攻击应用到联邦学习中，其攻击者必须有模型损失函数的知识和损失范围。文献[29]提出的攻击方法是对深度神经网络的白盒推理攻击。它利用了深度神经网络的特性，即其具有大量的参数（数百万个）且不能正确地泛化到训练数据之外（在很多情况下，训练数据的大小要小一个数量级）。该攻击认为可以利用模型梯度信息来区分成员和非成员训练数据，并利用这些信息来训练攻击模型。需要注意的是，这种攻击方法需要访问目标模型的内部信息，包括模型的架构、参数和梯度信息等。因此，攻击者需要对目标模型有足够的了解，以便能够成功地实施这种攻击。

2.2.2 重构攻击

在机器学习中，重构攻击是指攻击者试图通过模型或模型输出来恢复原始数据的一种攻击方式。这种攻击相较于推理攻击更加危险，因为它试图完全重构出原始数据，而不仅仅是对数据进行笼统的判断。例如，攻击者可以从人脸识别模型中重构出参与训练的人脸图像[30]，或者从分类模型中重构出攻击者不拥有的某类数据[31]。为了进行重构攻击，攻击者需要掌握模型或梯度等相关信息。因此，在联邦学习场景中，攻击者只能来自于系统内部。

针对重构攻击，有多种攻击算法被提出。文献[30]提出了一种反演攻击算法，利用梯度下降算法最小化包含原模型的代价函数，并在优化过程中进行图像处理，以此来恢复原始图像。文献[31]中的协同分类机器学习攻击则利用 GAN 技术来生成攻击者没有的某一标签对应的数据，并随着交互的进行，其生成的数据质量越高。攻击者还可以将“尚未成熟”的生成数据加入局部模型训练，以此来刺激其他参与方输入更多有关此标签对应的数据。然而，这种攻击只能重构出某标签对应数据的代表，而不能精确恢复训练数据。文献[32]在文献[31]的基础上提出了 mGAN-AI(multi-task GAN for Auxiliary Identification)重构攻击，旨在针对特定的客户端进行攻击，达到破坏客户端级的隐私的目的。文献[33]中的梯度的深度泄漏攻击(Deep Leakage From Gradients, DLG)证明了 DLG 能够仅通过分析梯度信息在计算机视觉和自然语言处理任务中重构出输入，但此攻击假设梯度变化仅由一个输入造成，这个假设是不切实际的。为了提高攻击的准确性，文献[34]在文献[33]的基础上提出了 Improved DLG(iDLG)，它对任何用交叉熵损失训练的可微分模型都能提取出输入的真实标签。

2.2.3 窃取攻击

在机器学习中，模型窃取攻击是指攻击者试图从一个已经训练好的模型中获取信息，以训练一个类似于原始模型的新模型。在联邦学习中，讨论模型窃取攻击通常假设攻击者来自于系统外部，攻击者一般需要通过与模型 API 进行交互来获取有关模型的信息。攻击者掌握的背景知识越丰富，就越可能生成与原始模型相似的新模型。这里的背景知识包括模型的架构和用于训练模型的数据集的背景知识。

模型窃取攻击的难点在于需要确定被攻击模型的类型和内部结构，以及确定模型中的超参数，并构造用于模型窃取的数据集。通常情况下，攻击者需要先了解被攻击模型的类型。例如，对于输出预测置信度的模型，可以使用样本集和对应模型的输出置信度来训练一个与原模型相似的模型，这种方法在文献[35]中被提出。而对于只输出类标签的模型，也存在被攻击的风险。在深度神经网络(Deep Neural Network, DNN)模型窃取攻击方面，文献[36]和文献[37]针对只输出标签的 DNN 模型，提出了类似的算法，其中关键步骤包括使用原模型为生成的数据集打上标签，将打上标签的数据集作为攻击模型的训练集进行训练，并对数据进行增强处理，以提高模型的准确性。文献[37]还提出了三种解决超参数问题的方案和两种创建训练样本的方案，并将查询预算融入到攻击算法中。此外，针对模型窃取攻击，还有一种方法是利用与原模型分布不同的训练样本，训练出与原模型相似的模型。文献[38]研究了这种方法所需的样本和模型，并证明了利用与原模型分布不同的训练样本可以训练出与原模型相似的模型。在模型选择方面，文献[38]还证明了原模型的输出可以用于训练具有不同模型架构的新模型，并选择较复杂的模型架构可以更有效地实现窃取攻击。

2.2.4 系统隐私性保护技术

应用于联邦学习中的隐私保护技术主要有：差分隐私(Differentially Private, DP)[39]、同态加密(Homomorphic Encryption, HE)[45-46]和秘密分享(Secret Sharing)[47]。

差分隐私：对于一个随机化算法 $M(X)\to R$，其中 X 为域，R 为值域。如果对于任意两个相邻数据集 $D,D'\in X$，以及任意子集 $S\in R$，算法 $M(X)$ 满足：$P_r\left[M(D)\in S\right]\le e^{\varepsilon}P_r\left[M(D')\in S\right]+\delta$，则称算法 $M(X)$ 满足$(\varepsilon,\delta)-$差分隐私。如果 $\delta=0$，则称算法 $M(X)$ 满足 $\varepsilon-$差分隐私，其中 ε 表示隐私保护预算。

同态加密：对于原文信息 a 和 b,如果算法 E 满足：$E(a+b)=a'\oplus b'$，则称算法 E 是加法同态加密算法，其中 a'和b'分别是 a 和 b 在算法 E 下得到的密文。同态加密能细分为加法同态加密算法、乘法同态加密算法、半同态加密算法和全同态加密算法。

秘密分享：对于秘密值 s，要将其分成 n 份，分别分配给参与方 $p_1,p_2,...,p_n$，利用公式：$f(x)=s+\sum_{i=1}^{t-1}a_ix^i$ 进行秘密分享，其中 t 是整数，且 $t\le n$，a_i 是随机生成的系数，原秘密为常数项。对于参与方 p_i，分配不同的 x_i 来计算其秘密份额，具体可以通过公式：$s_i=f(x_i)$ 计算。当至少有 t 个参与方协作时，可以通过公式：$s=\sum_{j=1}^{t}s_j\prod_{l\ne j}\frac{x-x_l}{x_j-x_l}$ 恢复原始的秘密值 s。

2.2.5 防御破坏系统隐私性的措施

在联邦学习中，确保参与者的数据隐私和模型隐私是非常重要的，因此需要采用一系列隐私保护技术。现有的隐私保护技术主要包括加噪机制[39-43]和加密机制[44-52]。加噪机制是通过向模型更新的梯度信息中添加一些噪声，以掩盖真实的梯度信息。这样可以防止攻击者通过梯度信息推断出参与者的数据隐私。但是，添加噪声会影响模型的准确性，一般来说，添加的噪声越多，模型的性能就越差。加密机制是对梯度信息进行加密，以确保只有允许的所有者可以查看梯度信息，而其他人无法访问梯度信息。这种技术不会影响最终的模型性能，但加解密过程会影响联邦学习的收敛速度，同时也会造成大量的计算消耗。因此，在将加密技术应用于联邦学习中保护梯度信息时，需要解决加密效率和通信成本的难题。

加噪机制通常在联邦学习中采用差分隐私技术来保护隐私。差分隐私可以确保即使攻击者拥有除一条信息以外的所有已发布信息，也无法推断该信息。在差分隐私中，隐私预算决定了噪声添加的大小。隐私预算越小，保护级别越高，但是添加的噪声也会越多。将差分隐私技术应用于联邦学习保护隐私时，模型的收敛性能和隐私保护级别之间存在权衡。收敛性能越好，保护级别越低。在固定隐私预算下，增加参与联邦学习的客户端数量可以提高收敛性能[40-41]。文献[40]提出了一种方法，即对上传到服务端上的每个参与者的梯度信息添加噪声，然后进行聚合，从而实现客户级的隐私保护，即攻击者无法确定客户是否参与了训练。但是，这种方法无法保护参与者模型参数的隐私，因此一些研究表明，对于联邦学习中的差分隐私，参与者可以先添加噪声，再上传梯度信息，以保护模型参数的隐私[40-42]。这种方法可以实现本地级的隐私保护。例如，NbAFL(Noising Before Model Aggregation FL)将加噪时机定在参与方发送梯度信息前，从而达到本地级的隐私保护效果。此外，研究还表明，在固定的隐私预算下，存在一个最佳的聚合次数，以达到最佳的收敛性能。

加密机制在联邦学习中的应用采用同态加密(Homomorphic Encryption, HE)[45-46]和安全多方计算(Secure Multi-party Computation, SMC)[44]中的秘密分享[47]，以保护梯度信息的隐私。同态加密允许在密文状态下进行计算，避免了解密数据的风险，但需要进行大量的计算才能实现加密和解密操作，且加密后的数据通常比原始数据要大得多。秘密分享将一个秘密信息拆分成多份小份信息，分享给不同的参与方，只有当所有参与方将其份额组合在一起时才能还原出原始秘密信息。在联邦学习中应用同态加密时，参与方使用公钥加密梯度信息，服务端在密文上进行聚合，参与方使用私钥解密聚合信息。为了提高训练速度和加密效率，一些研究文献提出了改进的同态加密算法和编码方案，如使用改进的 Paillier 算法[48]和 BatchCrypt 编码方案[49]，以降低通信开销和提高训练速度。但是，由于参与方使用相同的公私密钥，加密的梯度信息可能会被截取导致梯度信息泄露。为此，文献[50]提出了多密钥同态加密的隐私保护方案，确保解密需要所有参与方之间的协作。秘密分享多用于特定的联邦学习架构来保护梯度隐私，如结合边缘计算和区块链的联邦学习。在这种方案中，参与者利用秘密分享将梯度信息划分多份并传给多个边缘节点，每个边缘节点将不同参与方上传的梯度份额进行聚合并上传至区块链，区块链聚合边缘节点所上传的聚合份额[51]。同时，一些研究文献还利用同态加密来保护梯度信息的隐私，在结合区块链的联邦学习架构中，利用秘密分享来确保同态加密中私钥的安全[52]。

3 展望

联邦学习是一种新兴的技术，它能够在不共享数据的情况下进行模型训练和共享模型，从而打破数据壁垒，促进数据的共享和协作，同时保护个人数据隐私。然而，联邦学习系统本身也存在鲁棒性和隐私性两方面的安全威胁。

目前针对威胁联邦学习系统鲁棒性方面的研究更多地是将参与方上传的模型参数区分为恶意和非恶意两类。然而，数据的非独立同分布本身就会使模型参数相差极大，这使得鉴别恶意模型参数变得更加困难。虽然有研究证明在非独立同步数据下能正确地检测出恶意模型参数，但是其防御策略本身存在一些缺陷，例如计算开销大，缺乏理论依据等。对于防御威胁联邦学习系统隐私性方面的研究，已有的研究多是利用加噪策略和加密策略来保护隐私。然而，加噪机制会导致模型的精度下降，而加密机制会导致高额通信成本和计算成本。另外，保证联邦学习系统隐私性方面的防御策略和鲁棒性方面的防御策略存在冲突，例如，利用同态加密技术能保证联邦学习系统的隐私性，但是也导致无法防御对联邦学习系统鲁棒性的攻击。

未来的工作可以分以下三种情况进行研究：

(1)不可信的参与方和可信的服务端：在这种情况下，需要设计更为健壮的防御策略来抵抗威胁联邦学习系统鲁棒性的攻击。

(2)可信的参与方和不可信的服务端：在这种情况下，需要解决为了保证联邦学习系统隐私性带来的精度下降、通信和计算成本暴增的问题。

(3)不可信的参与方和不可信的服务端：兼顾联邦学习系统的隐私性和鲁棒性是一个非常严峻的挑战，但这也是目前更加值得研究的一种情况。如何应对这种严峻挑战，是未来工作的重点。

总之，设计一个安全的联邦学习系统是一个长期而具有挑战性的任务。这需要不断跟进最新的研究成果，并不断探索创新的方法来确保联邦学习系统的安全性。

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] ZHANG C, XIE Y, BAI H, et al. A survey on federated learning[J]. Knowledge-Based Systems, 2021, 216: 106775.

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

[4] ALBRECHT J P. How the GDPR will change the world[J]. Eur. Data Prot. L. Rev., 2016, 2: 287.

[5] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[EB/OL]. (2013-03-25) [2023-07-09]. https://arxiv.org/pdf/1206.6389.pdf.

[6] SUN G, CONG Y, DONG J, et al. Data poisoning attacks on federated machine learning[J]. IEEE Internet of Things Journal, 2021, 9(13): 11365-11375.

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

[8] ZHANG J, CHEN J, WU D, et al. Poisoning attack in federated learning using generative adversarial nets[C]// Proceedings of the 2019 IEEE international conference on big data science and engineering. Piscataway: IEEE, 2019: 374-380.

[9] FANG M, CAO X, JIA J, et al. Local model poisoning attacks to {byzantine-robust} federated learning[C]// Proceedings of the 2020 USENIX security symposium. Berkeley: USENIX Association, 2020: 1605-1622.

[10] SHEJWALKAR V, HOUMANSADR A. Manipulating the byzantine: optimizing model poisoning attacks and defenses for federated learning[C]// Proceedings of the 2021 NDSS. Rosten:ISOC, 2021: 21-24.

[11] MUÑOZ-GONZÁLEZ L, CO K T, LUPU E C. Byzantine-robust federated machine learning through adaptive model averaging[EB/OL]. (2019-09-11) [2023-07-09]. https://arxiv.org/pdf/1909.05125.pdf.

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

[13] CAO X, GONG N Z. Mpaf: Model poisoning attacks to federated learning based on fake clients[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 3396-3404.

[14] XIE C, HUANG K, CHEN P Y, et al. Dba: Distributed backdoor attacks against federated learning[C]// Proceedings of the 2020 International conference on learning representations. Washington: ICLR, 2020:1-19.

[15] OZDAYI M S, KANTARCIOGLU M, GEL Y R. Defending against backdoors in federated learning with robust learning rate[C]//Proceedings of the 2021 AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2021, 35(10): 9268-9276.

[16] WU C, YANG X, ZHU S, et al. Mitigating backdoor attacks in federated learning[EB/OL]. (2021-01-14) [2023-07-09]. https://arxiv.org/pdf/2011.01767.pdf.

[17] SUN Z, KAIROUZ P, SURESH A T, et al. Can you really backdoor federated learning?[EB/OL]. (2019-12-02) [2023-07-09]. https://arxiv.org/pdf/1911.07963.pdf.

[18] WANG H, SREENIVASAN K, RAJPUT S, et al. Attack of the tails: yes, you really can backdoor federated learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 16070-16084.

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

[20] BLANCHARD P, EL MHAMDI E M, GUERRAOUI R, et al. Machine learning with adversaries: byzantine tolerant gradient descent[C]// Proceedings of the 2017 International Conference on Neural Information Processing Systems. New York: PMLR, 2017: 118-128.

[21] YIN D, CHEN Y, KANNAN R, et al. Byzantine-robust distributed learning: towards optimal statistical rates[C]// Proceedings of the 2018 International Conference on Machine Learning. New York: PMLR, 2018: 5650-5659.

[22] LI D, WONG W E, WANG W, et al. Detection and mitigation of label-flipping attacks in federated learning systems with KPCA and K-means[C]// Proceedings of the 2021 International Conference on Dependable Systems and Their Applications (DSA). Piscataway: IEEE, 2021: 551-559.

[23] AWAN S, LUO B, LI F. Contra: Defending against poisoning attacks in federated learning[C]// Proceedings of the 2021 European Symposium on Research in Computer Security. Cham: Springer, 2021: 455-475.

[24] RIEGER P, NGUYEN T D, MIETTINEN M, et al. Deepsight: mitigating backdoor attacks in federated learning through deep model inspection[EB/OL]. (2022-01-03) [2023-07-09]. https://arxiv.org/pdf/2201.00763.pdf.

[25] CARLINI N, LIU C, ERLINGSSON Ú, et al. The secret sharer: evaluating and testing unintended memorization in neural

networks[C]// Proceedings of the 2019 USENIX security symposium. Berkeley: USENIX Association, 2019: 267-284.
[26] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models[C]//Proceedings of the 2017 IEEE symposium on security and privacy (SP). Piscataway: IEEE, 2017: 3-18.
[27] MALEKZADEH M, BOROVYKH A, GÜNDÜZ D. Honest-but-curious nets: sensitive attributes of private inputs can be secretly coded into the classifiers' outputs[C]// Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2021: 825-844.
[28] YEOM S, GIACOMELLI I, FREDRIKSON M, et al. Privacy risk in machine learning: analyzing the connection to overfitting[C]// Proceedings of the 2018 IEEE computer security foundations symposium (CSF). Piscataway: IEEE, 2018: 268-282.
[29] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning[C]// Proceedings of the 2019 IEEE symposium on security and privacy (SP). Piscataway: IEEE, 2019: 739-753.
[30] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]// Proceedings of the 2015 ACM SIGSAC conference on computer and communications security. New York: ACM, 2015: 1322-1333.
[31] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep models under the GAN: information leakage from collaborative deep learning[C]// Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. New York: ACM, 2017: 603-618.
[32] WANG Z, SONG M, ZHANG Z, et al. Beyond inferring class representatives: user-level privacy leakage from federated learning[C]// Proceedings of the 2019 IEEE INFOCOM 2019-IEEE conference on computer communications. Piscataway: IEEE, 2019: 2512-2520.
[33] ZHU L, LIU Z, HAN S. Deep leakage from gradients[EB/OL]. (2019-12-19) [2023-07-09]. https://arxiv.org/pdf/1906.08935.pdf.
[34] ZHAO B, MOPURI K R, BILEN H. idlg: Improved deep leakage from gradients[EB/OL]. (2020-01-08) [2023-07-09]. https://arxiv.org/pdf/2001.02610.pdf.
[35] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction {APIs}[C]// Proceedings of the 2016 USENIX security symposium. Berkeley: USENIX Association, 2016: 601-618.
[36] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]// Proceedings of the 2017 ACM on Asia conference on computer and communications security. New York: ACM, 2017: 506-519.
[37] JUUTI M, SZYLLER S, MARCHAL S, et al. PRADA: protecting against DNN model stealing attacks[C]// Proceedings of the 2019 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE, 2019: 512-527.
[38] OREKONDY T, SCHIELE B, FRITZ M. Knockoff nets: stealing functionality of black-box models[C]// Proceedings of the 2019 IEEE/CVF conference on computer vision and pattern recognition. Piscataway: IEEE, 2019: 4954-4963.
[39] DWORK C. Differential privacy[C]// Proceedings of the 2006 International colloquium on automata, languages, and programming. Cham: Springer, 2006: 1-12.
[40] GEYER R C, KLEIN T, NABI M. Differentially private federated learning: a client level perspective[EB/OL]. (2018-03-01) [2023-07-09]. https://arxiv.org/pdf/1712.07557.pdf.
[41] 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.
[42] TRUEX S, LIU L, CHOW K H, et al. LDP-Fed: Federated learning with local differential privacy[C]// Proceedings of the 2020 ACM International Workshop on Edge Systems, Analytics and Networking. New York: ACM, 2020: 61-66.
[43] ZHAO Y, ZHAO J, YANG M, et al. Local differential privacy-based federated learning for internet of things[J]. IEEE Internet of Things Journal, 2020, 8(11): 8836-8853.
[44] GOLDREICH O. Secure multi-party computation[EB/OL]. (2018-06-09) [2023-07-09]. https://citeseerx.ist.psu.edu/document?doi=dce0d462c182121f37279e3809d484624f3d3eba.
[45] OGBURN M, TURNER C, DAHAL P. Homomorphic encryption[J]. Procedia Computer Science, 2013, 20: 502-509.
[46] GENTRY C. Fully homomorphic encryption using ideal lattices[C]// Proceedings of the 2009 annual ACM symposium on Theory of computing. New York: ACM, 2009: 169-178.
[47] LONGO D L, DRAZEN J M. Data sharing[J]. New England Journal of Medicine, 2016, 374(3): 276-277.
[48] FANG H, QIAN Q. Privacy preserving machine learning with homomorphic encryption and federated learning[J]. Future Internet, 2021, 13(4): 94.
[49] ZHANG C, LI S, XIA J, et al. {BatchCrypt}: efficient homomorphic encryption for {Cross-Silo} federated learning[C]// Proceedings of the 2020 USENIX annual technical conference (USENIX ATC 20). Berkeley: USENIX Association, 2020: 493-506.
[50] MA J, NAAS S A, SIGG S, et al. Privacy‐preserving federated learning based on multi‐key homomorphic encryption[J]. International Journal of Intelligent Systems, 2022, 37(9): 5880-5901.
[51] 陈宛桢,张恩,秦磊勇, 等. 边缘计算下基于区块链的隐私保护联邦学习算法[EB/OL]. (2022-09-21) [2023-07-09]. http://kns.cnki.net/kcms/detail/51.1307.TP.20220920.1049.006.html. (CHEN W Z, ZHANG E, QIN L Y, et al. Privacy-preserving federated learning algorithm based on blockchain in edge computing[EB/OL]. (2022-09-21) [2023-07-09]. http://kns.cnki.net/kcms/detail/51.1307.TP.20220920.1049.006.html.)
[52] 周炜,王超,徐剑,等. 基于区块链的隐私保护去中心化联邦学习模型[J]. 计算机研究与发展, 2022,59(11): 2423-2436. (ZHOU W, WANG C, XU J, et.al. Privacy-preserving and decentralized federated learning model based on the blockchain[J]. Journal of Computer Research and Development, 2022, 59(11): 2423-2436.)
[53] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2): 1-19.
[54] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and open problems in federated learning[J]. Foundations and Trends® in Machine Learning, 2021, 14(1–2): 1-210.

This work is supported by the National Natural Science Foundation of China (U20A20179).

Chen Xuebin, born in 1970, Ph. D., professor. His research interests include big data security, iot security, network security.
REN Zhiqiang, born in 2000, M. S. candidate. His research interests include data security, privacy protection.
ZHANG Hongyang, born in 1999, M. S. candidate. His research interests include data security, privacy protection.