联邦学习攻击与防御综述

吴建汉[1,2]，司世景[1]，王健宗[1]，肖京[1]
1. 平安科技（深圳）有限公司，广东 深圳 518063；
2. 中国科学技术大学，安徽 合肥 230026

摘要
随着机器学习技术的广泛应用，数据安全问题时有发生，人们对数据隐私保护的需求日渐显现，这无疑降低了不同实体间共享数据的可能性，导致数据难以共享，形成“数据孤岛”。联邦学习可以有效解决“数据孤岛”问题。联邦学习本质上是一种分布式的机器学习，其最大的特点是将用户数据保存在用户本地，模型联合训练过程中不会泄露各参与方的原始数据。尽管如此，联邦学习在实际应用中仍然存在许多安全隐患，需要深入研究。对联邦学习可能受到的攻击及相应的防御措施进行系统性的梳理。首先根据联邦学习的训练环节对其可能受到的攻击和威胁进行分类，列举各个类别的攻击方法，并介绍相应攻击的攻击原理；然后针对这些攻击和威胁总结具体的防御措施，并进行原理分析，以期为初次接触这一领域的研究人员提供详实的参考；最后对该研究领域的未来工作进行展望，指出几个需要重点关注的方向，帮助提高联邦学习的安全性。

关键词
联邦学习；攻击；防御；隐私保护；机器学习

中图分类号：TP181 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022038

Threats and defenses of federated learning: a survey

WU Jianhan[1,2], SI Shijing[1], WANG Jianzong[1], XIAO Jing[1]
1. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518063, China
2. University of Science and Technology of China, Hefei 230026, China

Abstract
With the comprehensive application of machine learning technology, data security problems occur from time to time, and people's demand for privacy protection is emerging, which undoubtedly reduces the possibility of data sharing between different entities, making it difficult to make full use of data and giving rise to data islands. Federated learning (FL), as an effective method to solve the problem of data islands, is essentially distributed machine learning. Its biggest characteristic is to save user data locally so that the models' joint training process won't leak sensitive data of partners. Nevertheless, there are still many security risks in federated learning in reality, which need to be further studied. The possible attack means and corresponding defense measures were investigated in federal learning comprehensively

and systematically. Firstly, the possible attacks and threats were classified according to the training stages of federal learning, common attack methods of each category were enumerated, and the attack principle of corresponding attacks was introduced. Then the specific defense measures against these attacks and threats were summarized along with the principle analysis, to provide a detailed reference for the researchers who first contact this field. Finally, the future work in this research area was highlighted, and several areas that need to be focused on were pointed out to help improve the security of federal learning.

Key words
federated learning, attack, defense, privacy protection, machine learning

0 引言

随着数字技术进入高速发展期，数据多元化、信息化和多样化成为当今时代的主题。打破“数据孤岛”并充分利用数据已成为当下的热门话题[1]。传统的中心服务器统一训练方式已经显现出众多安全问题。联邦学习（federated learning，FL）[1-3]是一种安全的分布式机器学习，可以在数据不离开本地的前提下共同训练全局模型，达到保护隐私的目的。联邦学习的主要特征包括：允许模型在不同的公司、设备和云之间进行通信[4]；使用数据而不窥探数据隐私[5]。其具体框架如**图1**所示，实现过程是：首先将联邦学习的全局模型发送给本地客户端进行训练，随后客户端将更新的模型参数上传至中央服务器，服务器进行一系列安全聚合处理后更新全局模型，再发送给客户端，从而使用户能够享受经过强大数据集中训练的全局模型，同时还能保证自身的隐私不被泄露。这些特征使联邦学习符合许多安全规则，例如《通用数据保护条例》（GDPR）[6]。

联邦学习于2016年被首次提出[7]，主要用来对联合存储在多个终端（如手机）中的数据进行中心化模型训练，主要应用在输入法改进等场景。如谷歌的Gboard能够在用户多次使用相关词汇之后，在输入时为用户推荐单词和表情，与传统推荐系统不同，这是在不获取用户隐私的前提下，在极大程度上依赖移动设备自身完成的训练。随着数据安全问题频繁出现，联邦学习日益流行，学术界和产业界开始研究整个技术系统，越来越多的公司开始尝试把联邦学习作为打通多方数据的解决方案。随后出现了许多实用的联邦学习案例，如腾讯的Angel、百度的PaddleFL和平安科技的“蜂巢”等。

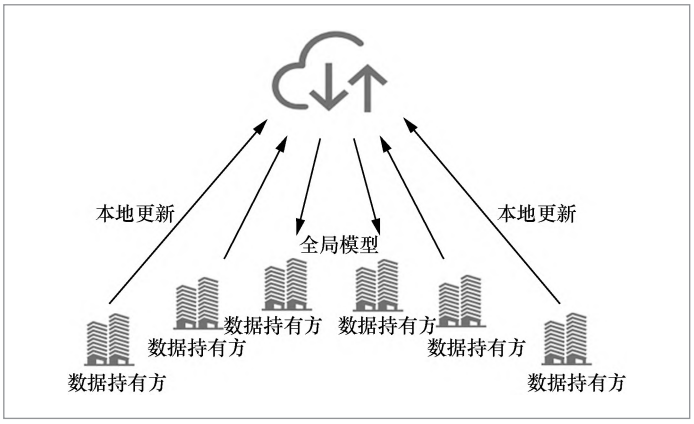

图1 联邦学习框架

根据数据的存储分布和用户的重叠程度，联邦学习可分为横向联邦学习、纵向联邦学习和联邦迁移学习[8-9]。横向联邦学习针对的是数据特征重叠较多而用户重叠较少的情况[10]。例如，某一地区的银行和另一地区的银行一般不能在没有用户

许可的情况下共享两个地区的用户数据，如果要使用双方的数据联合训练机器学习模型，使数据得到充分利用，横向联邦学习可以很好地实现数据的安全利用。纵向联邦学习针对的是数据特征重叠较少而用户重叠较多的情况。例如对于同一区域中的银行和保险数据，纵向联邦学习能够达到协同利用此类数据的目的[11]。联邦迁移学习针对数据特征和用户都没有太多重叠的情况[12]。在联邦迁移学习中，来自不同特征空间的特征会被迁移到同一个隐表示空间中，然后利用不同参与方收集的标注数据中的标签进行训练。联邦学习将数据保存在用户本地的做法可以在一定程度上保护隐私，但在具体实践和研究中仍然存在许多隐患，需要进一步的研究与发展[12]。

现有的联邦学习攻击与防御综述大多基于特定攻击对象和性质进行分类与分析，这样分类往往要求读者了解联邦学习的基础知识，从而给初次接触联邦学习的读者带来一定的困难。与之前的联邦学习攻击与防御综述文献不同，本文对联邦学习框架的各个层面进行分类，对联邦学习可能受到的攻击及相应的防御措施进行详实的分析，这不仅可以使读者清楚地了解联邦学习框架，还可以更加清晰地了解联邦学习的攻击和防御。

本文的主要贡献如下：

- 以一种比较新颖的分类方法详细地介绍了联邦学习可能受到的攻击及相应的防御措施，并对联邦学习攻击与防御的典型方法和最新方法进行了介绍；
- 以图片的形式形象地呈现了联邦学习可能受到的攻击，并对相应的防御措施进行了详实的介绍和分析；
- 根据联邦学习的特性与现状，本文对联邦学习进行了多方位的展望，并对一些具体问题提出解决思路。

1 联邦学习中的攻击类型

联邦学习提供一种新的范式来保护用户隐私，能够大规模执行机器学习任务[13]，与传统的机器学习不同，根据其独特的结构，联邦学习系统应该抵御4个层面的潜在攻击者：客户端、聚合器、局外人或窃听者、服务器。本节根据这4个层面的潜在攻击者对联邦学习可能受到的攻击进行分类，分别为数据中毒、模型攻击、推理攻击、服务器漏洞，并对这些攻击进行剖析。

1.1 数据中毒

数据中毒是指攻击者将部分恶意数据或篡改数据添加到训练数据集中，使训练后的模型符合攻击者的期望，达到破坏模型或篡改模型结果的目的。数据中毒示意图如**图2**所示，其中 Δw 为本地模型参数。根据攻击者是否更改数据标签，可将数据中毒分为两类：干净标签中毒攻击[14]和脏标签中毒攻击[15]。干净标签中毒攻击是一种不会修改数据标签的攻击，只添加部分恶意数据，其是针对性的攻击。由于中毒数据的标签不会被修改，中毒数据可以很容易地被模型接受并训练，因此这种攻击的成功率比较高。但是，要想获得良好的攻击效果，就需要精心设计攻击数据。参考文献[16]提出一种基于数学优化的方法来设计中毒攻击，并设计实验证明了在迁移学习框架中，只需要一种类别的中毒数据就可以使分类器出现错误。脏标签中毒攻击是指攻击者通过恶意篡改数据的标签来达到攻击目的，攻击者只需将其希望篡改的目标类别数据与干净数据混为一体，

然后集中训练即可进行脏标签中毒攻击。脏标签中毒攻击的一个典型例子是标签翻转攻击[17]，即一类干净训练样本的标签被翻转到另一类，而数据的特征保持不变。例如，系统中的恶意客户端可以通过将所有1转换为7来毒化数据集，攻击成功后，模型将无法正确分类1。参考文献[18]中的实验表明，在训练数据集中加入约50个中毒样本，就能使深度网络无法进行正确分类。此外，参考文献[19]提出一种利用标签翻转攻击实现针对某种类别标签的攻击，即只对受到攻击的类别标签有很大的影响，而基本不影响未受到攻击的类别标签。这种攻击手段可以避免很多防御措施，且危害性极大，作者在CIFAR-10和Fashion-MNIST数据集上进行了效果展示，实验表明，当存在20%的恶意用户进行攻击时，就可使分类精度和召回率明显下降。

还有一种常见的攻击为数据后门中毒攻击[20]，攻击者修改原始训练数据集的单个特征或小区域，然后将其作为后门嵌入模型中。如果输入中包含后门特征，模型就会根据攻击者的目标运行，而中毒模型在干净输入数据上的性能不受影响，这导致攻击更难被发现，攻击成功率较高。参考文献[20]在CIFAR-10数据集上展示了它的攻击效果，结果表明，即使在联邦安全平均算法的条件下，也可以在恶意参与者较少的情况下嵌入后门攻击。

值得注意的是，任何联邦学习参与者都可以进行数据中毒攻击，对模型的影响取决于系统参与者参与攻击的程度以及中毒的训练数据量，即数据中毒在参与者较少的环境中效果较差[19]。数据中毒攻击方法对比见**表1**。

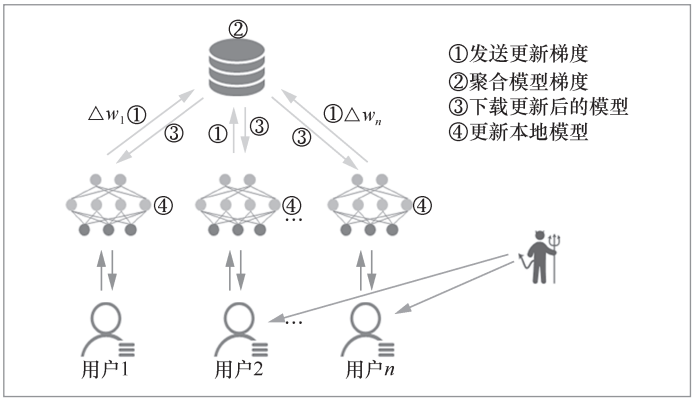

图 2 数据中毒示意图

1.2 模型攻击

模型攻击[21-22]通过篡改或替换客户端的模型参数来更改全局模型。模型攻击示意图如**图3**所示，其中 Δw 为本地模型参数，w为全局模型参数，aggregate表示聚合操作，与模型更新相关的信息一般为模型梯度。具体而言，攻击者通过攻击联邦学习中的某些成员，在模型更新的过程中，更改被攻击成员的梯度或者发送错误的信息来影响全局模型，使全局模型的方向与攻击前的方向偏差最大，从而

表1 数据中毒攻击方法对比

攻击类型	描述	攻击效果
干净标签中毒攻击	将恶意数据添加到训练数据集中	伪造高质量数据难度较高，但效果好
脏标签中毒攻击	将干净训练样本的标签翻转为另一类，数据的特征保持不变	典型且实用，不具有针对性
	将干净样本的标签翻转为指定类别	针对性强，效果明显
数据后门中毒攻击	修改原始训练数据集的单个特征或小区域，将其作为后门并嵌入模型	后门嵌入，攻击成功率较高

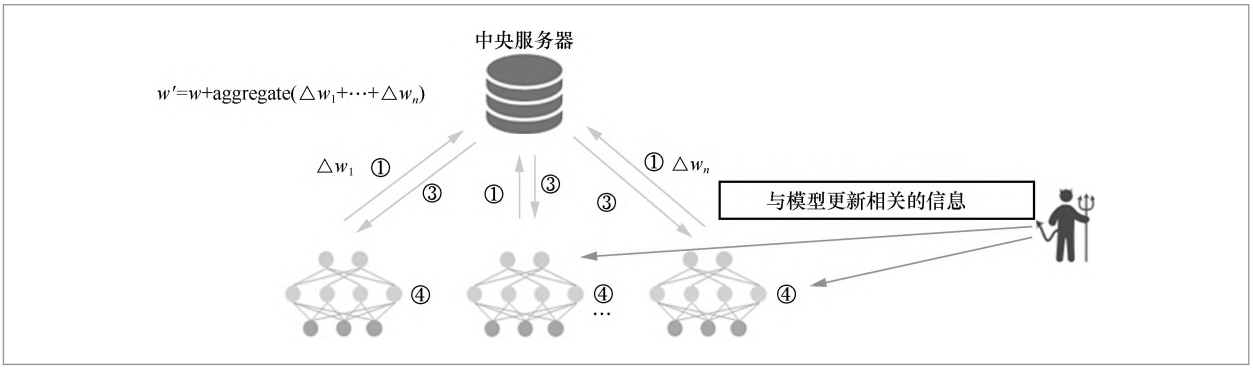

图 3 模型攻击示意图

达到理想的攻击效果[23]。由于联邦安全聚合算法的引入，服务器收到的是经过安全聚合后的结果，无法了解本地模型更新是如何生成的，导致无法通过直接检测聚合后的参数来检测本地客户端的参数是否异常，因此传统的检测方法一般无法进行防御。与数据中毒不同，模型攻击旨在直接攻击模型并将其发送到服务器，然后进行聚合，进而影响全局模型，全局模型更新后被发送给客户端，使得收到更新的客户端都受到一定程度的影响，因此模型攻击的效果往往比数据中毒更加有效。

拜占庭攻击是一种常见的模型攻击。拜占庭攻击是一种无目标性的模型攻击，其将任意恶意的模型更新上传至服务器，导致全局模型失效[24-25]。其定义如下。

在第t轮训练迭代中，一个诚实的参与者上传梯度$\Delta\omega_i^{(t)} \coloneqq \nabla F_i(\omega_i^{(t)})$，而一个恶意用户可能会上传任意值。

$$\Delta\omega_i^t = \begin{cases} *, \text{第}i\text{个参与者是恶意的} \\ \nabla F_i(\omega_i^{(t)}), \text{其他} \end{cases} \quad (1)$$

其中，*为任意值，F_i代表第i个用户模型的目标函数。在联邦学习中，拜占庭攻击得到了广泛研究，很多安全措施是基于此种攻击手段进行防御的。参考文献[23]对联邦学习本地模型攻击进行了系统性研究，其将模型攻击表述为一个优化问题，并对拜占庭攻击进行量化分析，具体是将拜占庭攻击量化成被攻击的设备数量，而后在最近4个拜占庭鲁棒联邦学习方法上测试攻击方法的有效性，结果表明，拜占庭攻击在4个常用数据集上都可达到提高模型错误率的效果。目前效果最佳的模型为参考文献[26]提出的通用模型架构，该方法从模型更新的知识和服务器的聚合算法知识出发，对这两个维度进行了全面的攻击分析，在不同数据集和模型中进行实验以验证其方法的高效性，该方法的最好效果是现有最强的模型攻击精度的1.5倍。

另一种常见的模型攻击为后门攻击，在训练过程中通过隐藏后门来实现，即攻击者通过设定一个触发器（trigger）激发隐藏好的后门，后门未被触发前，模型表现正常，后门被触发后，模型的输出为攻击者设定的值。参考文献[27]使用物理反射模型进行数学建模，提出一种将反射作为后门植入的攻击模型，这种设计具有既高效又隐蔽的特点。参考文献[20]使用模型替换（将正常模型替换成有毒模型）将后门功能引入全局模型，具体后门功能包括修改图像分类器以便为具有某些特征的图像分配攻击者选择的标签，以及强制单词预测器使用攻击者选择的单词完成某些句

子等。除了利用模型替换的方法实现后门攻击，参考文献[28]提出了一种新颖的模型攻击的方法——隐秘通道攻击，它假设有恶意的双方之间存在一条隐秘通道，由于一条数据的更改往往不会影响全局模型或者影响他人，且在这个信道里传输数据所占带宽也比较小，因此此通道的建立与通信不会被联邦学习防御系统发觉，有恶意的双方就可以实现相互通信自由，将它们交换的信息进行迭代，慢慢生成的毒化模型的作用会逐渐变大，对模型的影响也会不断积累，从而达到攻击目的。

1.3 推理攻击

推理攻击是指攻击者通过多样的攻击手段（如窃听、监视等）获取某些信息，然后利用这些信息推理获得想要的信息。这些信息通常是客户的隐私（一般是比较重要的客户或总结性的数据），尤其是银行、医院等对数据比较敏感的行业，受到推理攻击的危险性更大。推理攻击示意图如**图4**所示。虽然在联邦学习设置中，用户上传的是梯度信息而私有数据始终存储于用户本地，但交换梯度也可能导致隐私泄露，这是因为梯度是由参与者的私人数据（使用反向传播算法）训练而来的，通过对梯度信息进行剖析可以得到隐私信息[29]。

在联邦学习框架中，攻击者可以攻击本地模型或全局模型，通过监听训练模型过程中的梯度信息，可以在一定程度上推理出有用的信息。这些信息可以是成员信息以及训练的输入特征与标签信息。成员推理攻击的目的是确定攻击对象是否被用来训练模型[30]，特征推理攻击的目的是得到攻击对象的数据分布信息[31]，标签推理攻击可以根据用户上传的梯度来推断用户的标签，参考文献[32]提出可以根据梯度的方向和幅度准确地确定任何标签是否存在。

根据被推理的模型是否已知，推理攻击可分为白盒攻击和黑盒攻击。白盒攻击是在攻击者已知模型的情况下进行的，即攻击者可以得到任意输入的预测输出及隐藏层的中间计算结果[33]。白盒攻击的效果比黑盒攻击的效果好，参考文献[33]利用随机梯度下降法的弱点设计出一种针对神经网络模型的白盒成员推理攻击，该攻击的各个隐藏层以及输出攻击效果在CIFAR-100数据集上表现颇佳。黑盒攻击[34-35]是在攻击者只知道模型的输入和输出而不知道模型参数的情况下进行的，其攻击难度高于白盒攻击。但针对某些模型，许多黑盒攻击可以攻击成功，主要方法是利用对抗样本的普适性和基于查询的逆向猜测进行攻击[36]：对抗样本的普适性是指针对白盒攻击的模型也可以在一定程度上对黑盒模型奏效，

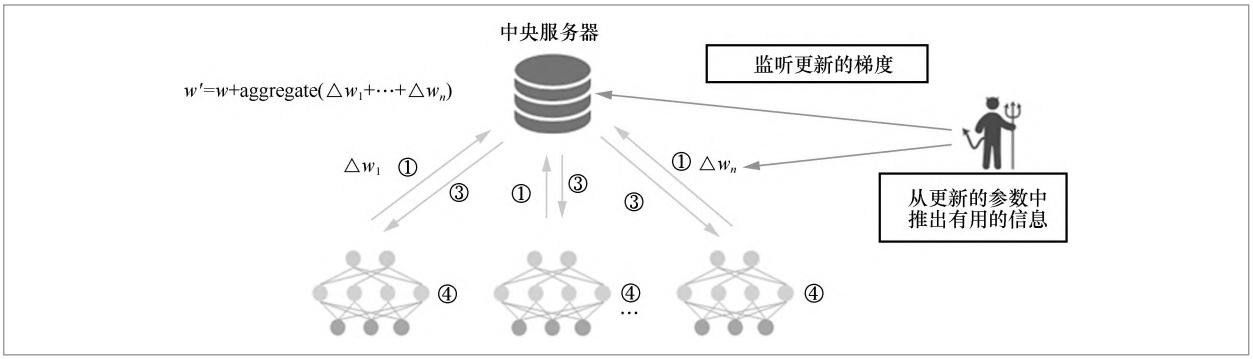

图 4 推理攻击示意图

即选择高成功率的白盒攻击方法对黑盒进行攻击，成功率往往也较高，然后再结合集成学习进行改进，就可以训练出一个性能较好的黑盒攻击[37]；基于查询的逆向猜测是指有时模型返回的不只是标签，还包含某一类别的概率分布，攻击者可以通过该分布逆向猜测有用的信息，从而使攻击奏效。在特定情况下的黑盒攻击效果也取得了很有竞争力的表现，例如参考文献[38]针对成员推理攻击在现实场景中目标训练数据集数量有限且比例不平衡的问题，提出使用生成对抗网络（generative adversarial network，GAN）合成数据，为黑盒成员推理攻击增加训练样本，以提高攻击效率，实验结果表明，合成数据的引入使黑盒成员推理攻击的准确率提高了23%。

基于生成对抗网络的攻击也属于推理攻击，这是一种主动推理攻击，该攻击利用学习过程的实时性，使攻击者可以生成目标训练集的原始样本。参考文献[39]提出一种基于GAN的逆向攻击，该攻击不仅可以在简单的线性回归模型和逻辑回归模型上达到推理隐私的效果，还能在深度学习网络中生成模型逆向攻击。具体重建过程包括两个阶段：第一阶段利用公共知识训练判别器；第二阶段利用从第一阶段获得的判别器，解决优化GAN生成器问题，以恢复图像中缺失的敏感区域。此外，作者还从理论上证明了模型的预测能力与模型对逆向攻击的脆弱性是正相关的。参考文献[40]提出了生成回归神经网络（generative regression neural network，GRNN），这是一种基于GAN的攻击模型，用于反推联邦学习中客户端的原始数据。该攻击的主要思想是建立GAN并生成随机数据，最小化真实梯度和虚拟梯度间的距离，从而完成原始数据推断。该攻击的主要优点是不需要额外的信息就可以从共享的梯度中恢复客户端的原始数据。参考文献[41]表明，如果攻击者知道相应的梯度更新方向，就可将对应的样本恢复为具有高保真度的原始数据，即使不知道梯度更新的方向，也可以进行攻击。

1.4 服务器漏洞

在联邦学习框架中，服务器的工作是将本地用户更新的参数安全地聚合到全局模型参数中，然后将更新后的参数返回给本地用户，以此循环训练出一个全局模型[42]。这表明服务器仍然是数据的中心，受损或恶意的服务器可能会破坏全局模型，从而产生重大影响。在训练机器学习模型时，服务器能轻松地提取客户端数据或操纵全局模型，以利用共享计算能力来构建恶意任务[43]。这成为联邦学习的一个重要漏洞，一方面，攻击者可以从服务器直接访问全局模型，扩大了攻击面；另一方面，服务器决定全局模型的客户端视图，从而对正在训练的模型产生重大影响。服务器可以控制每个客户端在联邦学习训练过程中何时访问与操纵模型，因此恶意服务器可以设计新的方案来测量模型的平均情况或最坏情况的攻击敏感性[44]，从而设计出最低成本的攻击方案。考虑到来自恶意服务器的攻击，参考文献[45]提出的框架结合了多任务生成对抗网络，通过攻击客户端级别的隐私来实现对用户身份的区分。此外，服务器所处网络环境的安全性也很重要，如果服务器在一个比较危险的网络环境下运行，被攻击的可能性会显著提高[46]。因此，强大而安全的服务器是必要的。

综上，对联邦学习中的攻击类型进行汇总对比，见**表2**。

2 联邦学习中的防御措施

联邦学习场景下存在电源与网络连接的

表 2 联邦学习的攻击类型对比

攻击类型	攻击原理	攻击方法
数据中毒	攻击者向训练数据集中添加恶意数据或篡改数据集中的某些数据，以达到攻击目的	干净标签中毒攻击[14]、脏标签中毒攻击[15-17]
模型攻击	通过更改被攻击客户端本地模型的更新来更改全局模型更新	拜占庭攻击[24-25]、后门攻击[27-28]
推理攻击	通过攻击手段（如窃听、监视等）获取某些信息，然后利用这些信息推理获得目标信息	成员、特征、标签推理攻击[30-32]，黑/白盒攻击[33-35]，生成对抗网络[39-40]
服务器漏洞	服务器所处环境缺少安全防御措施，导致易受到攻击，或存在恶意服务器	脆弱、恶意服务器攻击[42-45]

零星访问、数据中的统计异质性等情况，这使得联邦学习隐私保护与防御更有意义[47]。而且在利益的驱动下，攻击手段会不断地更新，使得对应的防御手段产生滞后性。目前，联邦学习中已经涌现出许多针对性的防御方法。下面将从联邦学习中的通用隐私保护措施（差分隐私、同态加密、秘密共享）和针对性防御措施（防御数据中毒、防御模型攻击、防御推理攻击、防御服务器漏洞）两个维度进行介绍。

2.1 联邦学习通用隐私保护措施

2.1.1 差分隐私

在联邦学习环境中，差分隐私通过掩盖真实数据达到防止用户隐私泄露的目的，其主要思想是在数据上添加噪声（如高斯噪声、拉普拉斯噪声等），使得数据库查询结果对数据集中单个记录的变化不敏感，从而防止攻击者利用统计特征的变动推理出隐私属性[48]。差分隐私具有计算效率高、攻击者无法恢复原始数据等特点[49]，其原理如下：给定两个数据集D和D'，如果二者有且仅有一个数据不同，则可将这两个数据集称为相邻数据集，由此差分隐私的形式可以定义为：

$$\Pr\{A(D)=O\}\leqslant e^{\varepsilon}\cdot\Pr\{A(D')=O\} \quad (2)$$

其中，A为随机算法（给定一个输入，经过算法后得出的输出不是固定值，而是服从某一分布的随机输出）。如果将该算法应用于任意两个相邻数据集，得到输出O的概率是相似的（都小于ε），那么可以得出该算法可以达到差分隐私的效果。从上述过程可以看出，观察者很难通过观察一个输出结果来检测出数据集微小的变化，无法获得真实数据，从而达到保护隐私的目的[50]，此种方法可以有效地防御推理攻击和数据中毒。根据不同的信任假设和噪声源，可以将差分隐私分为3类：中心化差分隐私（centralized differential privacy，CDP）[51-52]、本地化差分隐私[53]（local differential privacy，LDP）[53-55]、分布式差分隐私（distributed differential privacy，DDP）[56-57]。

● 中心化差分隐私。差分隐私最初是为集中式场景设置的，中心化差分隐私有一个重要的前提，即需要一个可信的数据收集库，这个数据库有权查看任何参与者的数据信息。中心化差分隐私希望通过随机化查询结果这种隐私保护方式返回查询结果或公布统计数据[51-52]。具体地，当中心化差分隐私满足联邦学习场景条件时，中心化差分隐私可以被视为一个可信的聚合器，它负责向聚合的局部模型参数中添加噪声，然后在更新的时候再去掉噪声，从而达到保护隐私的目的。但由于中心化差分隐私需要将大量的数据集中到一起来处理，只有在有大量参与者的情况

下才能保证隐私和准确性，因此不适用于参与者相对较少的面向公司的横向联邦学习（horizontally federated learning to businesses，H2B）模型。

● 本地化差分隐私。现实中很难找到中心化差分隐私所需要的收集信息的可信数据中心，本地化差分隐私是基于不可信第三方进行的，将数据的隐私化处理过程转移到每个用户上，使得用户能够单独处理和保护个人数据，以达到保护隐私的目的[53]。本地化差分隐私可以被认为是中心化差分隐私的增强版，它基本继承了中心化差分隐私的特点，同时具备自身的特性：一是充分考虑了任意攻击者的情况，并对隐私保护程度进行了量化；二是本地化扰动数据，并且加入的噪声机制也有所改变，中心化差分隐私的噪声机制主要以拉普拉斯噪声和指数噪声为主，而本地化差分隐私的噪声机制主要以随机响应为主[54]。基于以上优点，本地化差分隐私技术很快在现实中得到了应用，例如谷歌公司使用该技术从Chrome浏览器采集用户的行为统计数据[55]，但其缺点是会在一定程度上影响精度。

● 分布式差分隐私。虽然本地化差分隐私可以在本地很好地保证隐私安全，但在分布式场景中，如果没有密码技术的帮助，每个参与者必须添加足够的校准噪声来确保本地化差分隐私，这往往会导致效率不高[56]。分布式差分隐私填补了中心化差分隐私和本地化差分隐私的空缺，其通过对运行相同噪声机制的参与者进行求和来实现整体加性噪声的机制，再结合密码技术[57]，达到既不需要可信的信息收集数据库，又能达到良好效果的目的。

2.1.2 同态加密

同态加密的基本思想是先对数据进行加密处理，然后对加密的密文执行各种计算，得到加密的结果，将其解密后得到的结果与原始数据（明文）直接执行各种计算得到的结果一致[58]。该方法不仅可以达到保护数据的目的，而且不影响数据的计算。以加法同态加密为例，有：

$$\mathrm{En}(m_1)=c_1,\ \mathrm{En}(m_2)=c_2 \tag{3}$$

$$\mathrm{Dec}(c_1+c_2)=m_1+m_2 \tag{4}$$

其中，En表示加密函数，Dec表示解密函数，m_1、m_2为明文，c_1、c_2为密文。同态加密是以数学方法为基础的，其破译困难在于计算复杂度很高，一般是指数级，用穷举法基本不可能破解，因此相对安全。但其缺点是相对于明文来说，密文的计算复杂度更高，导致训练时间长，而且对用户的设备也有一定的要求。在联邦学习场景中，为了减少计算量，一般将用户上传的梯度信息进行加密[59]，然后将梯度进行安全聚合，最后对更新后的参数进行解密处理。由于经过同态加密后的梯度是一堆随机数，攻击者没有密钥时无法从这堆数中推理出任何有价值的信息，因此，此方法可以有效地防御各种攻击。

同态加密分为全同态加密和部分同态加密[60]。理论上所有运算都可被分解成乘法和加法运算的组合，因此全同态加密理论上可以支持对密文进行任意计算，但其加密方法往往伴随着巨大的运算量，故此方法的效率比较低，对硬件的要求比较高[60]。部分同态加密分为加法同态和乘法同态，与全同态加密相比，部分同态加密更加高效，因此在联邦学习设置中常伴随着部分同态加密，例如参考文献[61]使用加法同态来保证模型参数的共享安全，使得每个客户端的隐私不会被中央服务器泄露。参考文献[62]提出一种用于迁移学习的联邦学习框架，其采用加法同态加密技术来加密模型参数，以保护数据隐私。随着硬件的发展，实

现同态加密与其他安全方法的结合成为可能。参考文献[63]提出一种基于同态加密的安全联邦学习框架，他们将同态加密与可验证计算技术结合，直接在同态加密形成的加密域中执行联邦平均，并通过可验证计算来证明该算子被正确应用。

2.1.3 秘密共享

秘密共享是确保信息安全和数据机密性的重要方法，也是联邦学习领域的基本应用技术。秘密共享主要用于保护参与者的重要信息，并防止信息丢失以及信息被破坏和篡改。其思想是将秘密以适当的方式拆分，拆分后的每一个份额由不同的参与者管理，单个参与者无法恢复秘密信息，只有若干个参与者一同协作才能恢复秘密消息[64]。例如秘密共享(s,n)表示将一个秘密信息s分为n个片段，并将其交给n个不同的参与方进行安全保存，设置一个阈值t，当超过t个参与方时就可以重构秘密信息，但参与方数小于t时不能获得关于s的任何有用信息。典型的秘密共享方案由Shamir和Blakley于1979年提出[65]，该方案基于多项式的方法构成，后来秘密共享方案的形式愈发多样化，被应用于多个领域。以多项式为例，秘密共享分为生成与分发密钥以及解密两个步骤，首先按照如下多项式生成密钥：

$$y_i = K + \sum_{i=1}^{t-1} a_i x^j \bmod p \qquad (5)$$

其中，K表示秘密，t为秘密共享的阈值，a_i为多项式的系数，模数p是为安全计算而设置的（使解密难度增大）。然后根据t个参与方提供的密钥 $y_i = K + \sum_{i=1}^{t-1} a_i x^j \bmod p$ 来解线性方程组，解出多项式的系数a_i以及秘密K。

在联邦学习框架中，秘密共享主要用于将用户上传的梯度信息进行秘密共享处理，保证恶意服务器无法得到梯度信息，从而可以作为一种针对恶意服务器的防御方法。Bonawitz K等人[66]设计了一种基于Shamir秘密共享的安全聚合方案，该方案确保在诚实和好奇的服务器下可以安全地更新参数，同时可以控制秘密共享协议的复杂性，从而在大规模数据集中保持较低的计算和通信成本。随着人们对数据安全的要求日渐提高，秘密共享方案也在更新，可验证秘密共享（verifiable secret sharing，VSS）[67]是基于传统的秘密共享升级而来的，它不仅能保证秘密的完整性，还能保证参与者分享的秘密是正确的。根据不同的应用场景，可验证秘密共享有几种不同的方案，参考文献[68]对各种可验证秘密共享进行了研究，提出了多种安全可靠的可验证秘密共享方案，可被应用于多种分布式安全计算场景。随着硬件设备的改善和需求的提高，许多可验证秘密共享与其他安全方法（如同态加密）结合的方案也被应用于联邦学习框架，具备这种方案的联邦学习框架不仅可以发挥联邦学习分布式的优点，还能加强其安全性[69]。

综上，联邦学习通用隐私保护措施对比见**表3**。

2.2 联邦学习针对性防御措施

2.2.1 防御数据中毒

数据中毒是联邦学习中常见的攻击，截至目前，已经有很多具体的用于防御数据中毒的措施，其中一种常用的措施是在训练之前检验数据的真实性和可靠性[70]。Baracaldo N等人[71]使用上下文信息（如来源和转换）来检测数据采样点，该检测方法先将整个训练集划分为多个部分，

表 3 联邦学习通用隐私保护措施对比

技术结合	参考文献	特性
中心化差分隐私	[51-52]	聚合和更新分别通过添加与删除噪声来达到保护隐私的目的，但需要一个可信的数据收集库
本地化差分隐私	[53-55]	将数据的隐私化处理过程转移至每个用户的设备上，使得用户能够单独地处理和保护个人数据，但这会影响模型的精度
分布式差分隐私	[56-57]	结合密码学技术来改进本地差分隐私和中心差分隐私
全同态加密	[60]	对隐私有绝对的保护，但计算复杂度非常高
部分同态加密	[60-63]	只对梯度进行加密处理，可以在很大程度上降低通信成本，实用性强
秘密共享	[65]	属于典型的密钥分发机制，在联邦学习中应用成熟
可验证性秘密共享	[67-69]	通过引入可验证机制，进一步提高秘密共享的安全性，且能与其他技术结合使用

随后比较每个部分的训练结果，以识别出训练结果最异常的部分，从而消除异常数据。除检测外，还可在训练之前转换数据，例如参考文献[72]将原始数据压缩转换成另一种形式，攻击者一般不会提前知道压缩形式，从而可以在一定程度上保证数据安全。同时还可以降低数据与模型参数的存储量，提高运算的速度，降低通信成本。参考文献[73]提出了一种高级表征引导去噪器（high-level representation guided denoiser，HGD）的方法，它解决的是标准降噪器误差放大效应（较小的残余对抗噪声会被逐渐放大）导致的错误分类问题，将由干净图像数据和去噪图像激活的目标模型输出之间的差值定义为损失函数，通过优化此损失函数来解决上述问题。此外，还有通过减少每个像素的颜色位深度和训练图像的空间平滑度来防御数据中毒的方法[74]。值得注意的是，虽然这些方法可以在一定程度上防御数据中毒，但也可能导致原有数据结构被破坏，使得原有数据的训练效果不佳。

除了上述方法，还有一种常用的防御方法是对抗训练，该方法的特点是只要对抗样本足够多，就可以达到很好的防御效果，因此该方法可以防御很多类型的数据中毒。对抗训练的原理是将真实样本和对抗样本放在一起作为训练模型的训练集，通过训练，模型能认识并“解决”（一般通过修改其标签来实现）中毒样本。在图像领域，对抗训练通常可以提高模型的鲁棒性，但往往会导致泛化性能下降；在语音领域，对抗训练可以同时提高鲁棒性和泛化性[75]，但其缺点是准确度不高。参考文献[76]提出一种联邦动态对抗训练方法，该方法不仅可以提高训练模型的准确度，还能加快模型的收敛速度。Fung C等人[77]使用了一种被称为FoolsGold的防御方法，可以很好地防御标签翻转攻击和数据后门中毒攻击，该方法的特点是不用限制攻击者的预期数量且适用于不同的客户数据分布，并且不需要训练过程之外的辅助信息就可以进行有效的防御，但是不足之处是这种防御方法需在许多特定的攻击假设下才有效，例如在攻击类型为后门攻击或翻转标签攻击时奏效[78]。

2.2.2 防御模型攻击

对于模型攻击，防御的重点是检测错误的模型参数。参考文献[79]提出两种检测模型参数的方法。一种方法是直接使用参数之间的数值差异来进行检测，具体是

每个参与者提供n个参数$\delta_1,\delta_2,\cdots,\delta_n$，当一方提供的参数与其他用户提供的参数有很大差异时，则判断该参数异常。另一种方法是服务器根据某个参与者上传的参数δ_i执行相应的处理$W_{G1}=W_G+f(\delta_i)$，然后使用其他参与者上传的参数计算$W_{G2}=W_G+f(\Delta)$，其中$\Delta=\{\delta_j \mid j=1,2,\cdots,n,j\neq i\}$，$W_G$为参与者上传的梯度信息，$f$为设计的特定函数，比较$W_{G1}$和$W_{G2}$，如果差值超过某个设定值，则推断模型参数$\delta$出现异常。针对后门攻击，参考文献[79]提出一种通过基于反馈的联邦学习进行后门检测的方法，这是一种保护联邦学习免受后门攻击的新型防御措施。其核心思想是使用多个客户端的数据进行训练以及模型中毒检测。具体做法是先随机选择一组客户端进行训练并上传参数，服务器不仅发送更新后的全局模型参数，还将上一轮各客户端的参数回传给客户端，然后客户端使用自己的数据与收到的参数进行测试打分，当得分超过阈值时，认为该模型受到后门攻击，反之则没有受到后门攻击，如此迭代，从而完成检测。实验表明，该方法对现有后门攻击方法的检测准确率约为100%，误报率低于5%。

联邦学习设置中十分常见的安全聚合算法也是一种有效的防御方法，安全聚合算法在任何集中式拓扑和横向联邦学习环境中都扮演着关键角色，它可以有效地防御模型攻击和推理攻击。迄今为止，多种安全聚合算法被提出，非常经典的是联邦平均聚合算法FedAvg（federated average）[80]。FedAvg简单且实用，其主要思想是将各个用户上传的参数以不同的权重进行平均聚合，每个用户的权重由其拥有的样本数量决定。虽然联邦平均聚合算法在一定程度上可以保护隐私，但是随着联邦学习的发展，对聚合算法的要求越来越高，许多不同的聚合算法应运而生。以下几种聚合算法是在平均聚合算法的基础上进行改进的算法。一是修剪均值聚合算法[80]，具体做法是对于m个模型参数，主设备首先会对其本地模型的m个参数进行排序，然后删除最大和最小的β个参数，计算$(m-2\beta)$个参数的平均值，并将其作为全局模型的参数，如此迭代，最后服务器对参数进行平均汇总。二是Median聚合算法[81]，对于本地m个模型参数，主设备将对本地模型的所有参数进行排序，将中位数作为全局模型的参数，当m是偶数时，中位数是中间两个参数的平均值，然后将多个模型的中位数取平均，并将该平均值作为最后超级模型的参数。与修剪均值聚合算法一样，当目标函数强凸时，Median聚合算法可达到最优阶次错误率。三是Krum聚合算法[82]，其主要思想是在众多局部模型中选择与其他模型最相似的模型作为全局模型，即使所选的部分模型来自损坏的工作节点设备，其影响也将受到限制，相当于采用折中的方式进行聚合。基于Krum的影响，参考文献[83]提出一种改进方法Bulyan。该方法主要将Krum聚合算法与修剪均值聚合算法结合。具体而言，首先将Krum聚合算法迭代地应用于选择局部模型，然后使用修剪均值聚合算法聚合局部模型，以获得全局模型。该方法可以消除Krum聚合算法中某些异常模型参数的影响。根据不同的应用场景，还可以将上述聚合算法结合，根据它们的优点，对算法进行加权处理以达到预期效果。除此之外，参考文献[84]针对安全聚合的计算开销问题进行改进，提出一种名为Turbo-Aggregate的安全聚合算法。该算法在具有N个用户的网络中实现了$O(N\log N)$的安全聚合开销，而以往最先进的聚合算法的计算复杂度为$O(N^2)$，同时Turbo-Aggregate算法能容忍高达50%的用户流失率。

2.2.3 防御推理攻击

一般来说，推理攻击比其他攻击更难成功，因为它要求攻击者不仅能成功获取联邦学习用户级别以上的部分，而且还能执行有效的推理才能攻击成功。对于推理攻击，联邦学习中也有相应的防御方法，常用的同态加密可以很好地防御推理攻击，在使用了同态加密的情况下，即使模型被成功攻击，攻击者也只会获得密文，而没有密钥的密文对于攻击者而言是毫无意义的。此外，如果攻击是针对全局模型的，则安全聚合算法也可以很好地防御推理攻击，这是因为攻击者通常不知道聚合规则。除了同态加密和安全聚合算法，差分隐私和秘密共享有时也可用于防御推理攻击。除上述防御方法外，还有一些主要针对推理攻击的防御方法。例如模型堆叠[84]，即将多种模型进行集成或者组合来构建最终模型，这种模型内部比较复杂，攻击者一般无法推理出有用的参数。参考文献[85]提出了一种名为消化神经网络（digestive neural network，DNN）的防御推理攻击的架构，它先通过DNN层对数据进行处理，得到数据的高维语义特征，然后计算该特征与原始数据间的L1损失函数，最后通过优化此损失函数来更新模型。这样做不仅可以将原始的数据处理成表征的形式（由于神经网络模型的难解释性，一般无法由表征推理出详实的信息），还能提高模型的准确性。

2.2.4 防御服务器漏洞

针对服务器漏洞问题，可信执行环境（trusted execution environment，TEE）[86]可以通过硬件隔离的技术来保护服务器，这在学术界与产业界都受到了广泛关注。一般在支持TEE的CPU中有一个被称为Enclave的特定的区域，该区域为数据和代码执行提供了更安全的空间，确保了应用程序的机密性和完整性[86]，使得服务器也无法获取用户在此区域中的执行逻辑和用户数据。具体是将该区域与外部环境隔离开，TEE可以直接获取有关外部环境的信息，但具有特殊访问权限的攻击者无法读取或干扰内存区域，只有处理器才能解密和执行该区域内的应用程序，以此达到保证信息的机密性和完整性的目的。此外，英特尔的SGX（software guard extensions）还为用户提供了一种用于验证TEE的真实性的机制，并且攻击者无法更改用户在Enclave内运行的应用程序[87]。目前TEE技术已经在智能设备上得到广泛的应用，要求在同一个设备上支持多个独立的TEE系统的场景也逐步增多，如上海瓶钵信息科技有限公司设计和实现了在移动智能设备上支持多个不同TEE的安全虚拟化系统TEEv[88]。参考文献[89]提出了一个基于TEE的完整隐私保护联邦学习方案，并通过实验表明该方案可以确保联邦学习训练过程的完整性和安全性，并且具有很强的实用性。还有其他安全方法也可防御恶意服务器，如安全多方计算。安全多方计算是密码协议的子领域，其目标是多方联合完成某种协同计算，每个参与者在完成计算之后，只能获得计算结果，无法获得参与实体的任何输入信息。参考文献[90]使用安全多方计算来构建联邦学习系统，作者使用秘密共享来保护参数信息，通过引入双重掩码结构来防御恶意服务器，并验证了即使服务器可以重建用户的扰动，秘密也可受到保护。

综上，对联邦学习中的针对性防御措施进行对比分析，见**表4**。

表 4 联邦学习针对性防御措施对比

防御类型	参考文献	防御措施	特点
防御数据中毒	[71]	检测上下文信息	通过与之前的数据进行对比来检测数据点
	[72-73]	最小化图像总方差，HGD	使用压缩、降噪和减少全局方差等方法来处理数据，进而达到保护数据的目的
	[75-77]	对抗训练	将真实样本和对抗样本放在一起作为训练集进行训练
防御模型攻击	[78]	检测错误的模型更新	直接或间接使用模型参数之间的数值差异来检测异常模型
	[80-83]	安全聚合	使用不同的聚合算法来保护模型参数
防御推理攻击	[84-85]	模型堆叠，DNN	将多种模型进行集成或者组合，以提高模型的复杂度
防御服务器漏洞	[87-[89]	TEE	通过硬件隔离的技术来保护隐私
	[90]	安全多方计算	安全联合多个参与方完成某种协同计算

3 总结与展望

随着人工智能的快速发展，许多数据安全问题不容忽视，在充分利用数据的同时保障用户的信息安全是一个很难实现的目标，从上述各种攻击与防御中可以看出，虽然联邦学习框架和相应的技术能够在一定程度上保护数据，但是还有许多安全问题有待解决。联邦学习攻击与防御的关键要素有模型的鲁棒性、模型训练和推理阶段的通信效率等。本文总结了如下5个可能的研究方向。

● 防御方法更加鲁棒。尽管本文介绍了许多可以在联邦学习中免受攻击的防御方法，但这些方法都有局限性，通常一种防御方法只能防御一种攻击，当联邦学习中存在多种攻击时，单一的防御方法显然是不够的。而且，同一种防御方法在不同的终端和模型（改动不大）中防御效果也不一样，可能会失效，甚至有时因为一些数据集的改变或者模型的微小改动就会对防御产生很大的影响，这些都是鲁棒性不足的表现。目前还没有鲁棒性非常高的防御方法可以同时针对多个不同的攻击进行有效的防御，甚至有些攻击还没有相应的防御方法。因此，找到一种鲁棒性更高的防御方法是非常有必要的方向之一。

● 攻击更加多样。攻击与防御是密不可分的，只有攻击一直发展，防御才会一直进步，而且有些防御是基于攻击的思路发展而来的。因此，对攻击研究得透彻，可以在一定程度上促进防御的发展，从而促进联邦学习更快地发展。具体地，应基于上述攻击方法进行更新迭代，从多个方位进行创新。

● 提高通信效率。由于机器学习通常需要大量的计算，资源管理在实现相关可持续和高效的联邦学习解决方案方面发挥着重要作用。而在这方面，很少有工作将边缘计算集成到联邦学习，以支持具有额外计算资源的终端设备，这是一个需要发展的方向。如果在联邦学习过程中使用过多防御措施或加密措施，将不可避免地增加计算量，通常也会增加服务器的通信负担，严重时会导致服务器拒绝服务。同时由上述方法（局部更新、模型压缩）可知，每种方法的准确性和通信效率是相互制约的，它们之间的权衡非常重要，因此，如何在保证数据安全的基础上保证通信效率和精度是今后非常必要的研究方向之一。

● 探索异构联邦学习。目前，隐私和鲁棒性研究大多集中在具有同构模型架构的

联邦学习范式中，而现有的隐私保护技术和攻击防御机制是否适用于具有异构模型架构的联邦学习仍未得到充分的实证性研究。因此，将现有的攻击与防御方法推广到异构联邦学习是非常有价值的，具体可以从联邦个性化学习、研究新的异构性定义、快速确定联邦网络中的异构性水平等方向进行。

● 模型可解释。模型可解释性是指可以将模型解释或表达成可理解的术语[91]。这不仅是未来联邦学习的研究方向，也是人工智能的研究方向。目前，相关研究人员也许知道机器学习能够很好地解决一个问题，但对于解决原理，却知之甚少。相对于传统的机器学习，联邦学习会进一步提高模型的复杂度，缺乏可解释性会导致联邦学习在应用过程中存在潜在威胁，提高联邦学习模型的可解释性有利于提前解决联邦学习落地所带来的潜在威胁，因此这也是相关研究人员努力的方向之一。

联邦学习是一个富有前途的研究方向，它以打破“数据孤岛”和保护用户隐私而闻名，目前已经吸引了大量的研究者进行相关领域的研究，并取得了一定的成就，也得到了广泛认可。虽然联邦学习能解决一些现实问题，但仍然存在许多潜在的威胁，未来还需深入研究存在的安全问题，加快处理联邦学习面临的挑战，共同推动联邦学习进一步发展为解决数据安全问题的首要利器。

参考文献:

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

[2] ALEDHARI M, RAZZAK R, PARIZI R M, et al. Federated learning: a survey on enabling technologies, protocols, and applications[J]. IEEE Access: Practical Innovations, Open Solutions, 2020, 8: 140699-140725.

[3] BLANCO-JUSTICIA A, DOMINGO-FERRER J, MARTÍNEZ S, et al. Achieving security and privacy in federated learning systems: survey, research challenges and future directions[J]. Engineering Applications of Artificial Intelligence, 2021, 106: 104468.

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

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

[6] TRUONG N, SUN K, WANG S Y, et al. Privacy preservation in federated learning: an insightful survey from the GDPR perspective[J]. Computers & Security, 2021, 110: 102402.

[7] KONEČNÝ J, MCMAHAN H B, YU F X, et al. Federated learning: strategies for improving communication efficiency[J]. arXiv preprint, 2016, arXiv:1610.05492.

[8] 马嘉华, 孙兴华, 夏文超, 等. 基于标签量信息的联邦学习节点选择算法[J]. 物联网学报, 2021, 5(4): 46-53.
MA J H, SUN X H, XIA W C, et al. Node selection based on label quantity information in federated learning[J]. Chinese Journal on Internet of Things, 2021, 5(4): 46-53.

[9] ABDULRAHMAN S, TOUT H, OULD-SLIMANE H, et al. A survey on federated learning: the journey from centralized to distributed on-site learning and beyond[J]. IEEE Internet of Things Journal, 2021, 8(7): 5476-5497.

[10] 王健宗, 孔令炜, 黄章成, 等. 联邦学习算法综述[J]. 大数据, 2020, 6(6): 64-82.
WANG J Z, KONG L W, HUANG Z C, et al. Research review of federated learning

algorithms[J]. Big Data Research, 2020, 6(6): 64–82.

[11] LI L, FAN Y X, TSE M, et al. A review of applications in federated learning[J]. Computers & Industrial Engineering, 2020, 149: 106854.

[12] LIU Y, KANG Y, XING C P, et al. A secure federated transfer learning framework[J]. IEEE Intelligent Systems, 2020, 35(4): 70–82.

[13] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and open problems in federated learning[J]. arXiv preprint, 2019, arXiv:1912.04977.

[14] ZHAO S H, MA X J, ZHENG X, et al. Clean-label backdoor attacks on video recognition models[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 14431–14440.

[15] BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens[C]//Proceedings of the 36th International Conference On Machine Learning. [S.l.:s.n.], 2019: 634–643.

[16] SHAFAHI A, HUANG W R, NAJIBI M, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2018: 6106–6116.

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

[18] CHEN X Y, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv preprint, 2017, arXiv:1712.05526.

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

[20] BAGDASARYAN E, VEIT A, HUA Y Q, et al. How to backdoor federated learning[J]. arXiv preprint, 2018, arXiv:1807.00459.

[21] JERE M S, FARNAN T, KOUSHANFAR F. A taxonomy of attacks on federated learning[J]. IEEE Security & Privacy, 2021, 19(2): 20–28.

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

[23] FANG M H, CAO X Y, JIA J Y, et al. Local model poisoning attacks to byzantine-robust federated learning[C]//Proceedings of the 29th USENIX Conference on Security Symposium. Berkeley: USENIX Association, 2020: 1623–1640.

[24] BERNSTEIN J, ZHAO J W, AZIZZADENESHELI K, et al. signSGD with majority vote is communication efficient and fault tolerant[J]. arXiv preprint, 2018, arXiv:1810.05291.

[25] XIE C, KOYEJO S, GUPTA I. Fall of empires: breaking Byzantine-tolerant SGD by inner product manipulation[J]. arXiv preprint, 2019, arXiv:1903.03936.

[26] SHEJWALKAR V, HOUMANSADR A. Manipulating the Byzantine: optimizing model poisoning attacks and defenses for federated learning[C]//Proceedings of 2021 Network and Distributed System Security Symposium. Reston: Internet Society, 2021: 18.

[27] LIU Y F, MA X J, BAILEY J, et al. Reflection backdoor: a natural backdoor attack on deep neural networks[C]//Proceedings of 2020 European Conference on Computer Vision. Cham: Springer, 2020: 182–199.

[28] COSTA G, PINELLI F, SODERI S, et al. Covert channel attack to federated learning systems[J]. arXiv preprint, 2021, arXiv:2104.10561.

[29] LEE H, KIM J, HUSSAIN R, et al. On defensive neural networks against

inference attack in federated learning[C]// Proceedings of 2021 IEEE International Conference on Communications. Piscataway: IEEE Press, 2021: 1-6.

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

[31] LUO X J, WU Y C, XIAO X K, et al. Feature inference attack on model predictions in vertical federated learning[C]//Proceedings of 2021 IEEE 37th International Conference on Data Engineering. Piscataway: IEEE Press, 2021: 181-192.

[32] WAINAKH A, VENTOLA F, MÜßIG T, et al. User label leakage from gradients in federated learning[J]. arXiv preprint, 2021, arXiv:2105.09369.

[33] 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 2019 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2019: 739-753.

[34] DONG Y P, SU H, WU B Y, et al. Efficient decision-based black-box adversarial attacks on face recognition[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 7706-7714.

[35] YIN Z Y, YUAN Y, GUO P F, et al. Backdoor attacks on federated learning with lottery ticket hypothesis[J]. arXiv preprint, 2021, arXiv:2109.10512.

[36] CHENG M H, LE T, CHEN P Y, et al. Query-efficient hard-label black-box attack: an optimization-based approach[J]. arXiv preprint, 2018, arXiv:1807.04457.

[37] LI Y D, LI L J, WANG L Q, et al. NATTACK: learning the distributions of adversarial examples for an improved black-box attack on deep neural networks[C]//Proceedings of the 36th International Conference on Machine Learning. [S.l.:s.n.], 2019: 3866-3876.

[38] BAI Y, CHEN D G, CHEN T, et al. GANMIA: GAN-based black-box membership inference attack[C]// Proceedings of 2021 IEEE International Conference on Communications. Piscataway: IEEE Press, 2021: 1-6.

[39] ZHANG Y H, JIA R X, PEI H Z, et al. The secret revealer: generative model-inversion attacks against deep neural networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 250-258.

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

[41] 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 Press, 2017: 603-618.

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

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

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

[45] WANG Z B, SONG M K, ZHANG Z F, et al. Beyond inferring class representatives: user-level privacy leakage from federated learning[C]//Proceedings of 2019 IEEE Conference on Computer Communications.

Piscataway: IEEE Press, 2019: 2512–2520.
[46] 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.
[47] LYU L J, YU H, MA X J, et al. Privacy and robustness in federated learning: attacks and defenses[J]. arXiv preprint, 2020, arXiv:2012.06337.
[48] 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.
[49] GIRGIS A M, DATA D, DIGGAVI S, et al. Shuffled model of differential privacy in federated learning[C]//Proceedings of 2021 International Conference on Artificial Intelligence and Statistics. [S.l.:s.n.], 2021: 2521–2529.
[50] HU R, GUO Y X, LI H N, et al. Personalized federated learning with differential privacy[J]. IEEE Internet of Things Journal, 2020, 7(10): 9530–9539.
[51] MCMAHAN H B, RAMAGE D, TALWAR K, et al. Learning differentially private recurrent language models[J]. arXiv preprint, 2017, arXiv:1710.06963.
[52] GEYER R C, KLEIN T, NABI M. Differentially private federated learning: a client level perspective[J]. arXiv preprint, 2017, arXiv:1712.07557.
[53] SUN L C, QIAN J W, CHEN X. LDP-FL: practical private aggregation in federated learning with local differential privacy[C]//Proceedings of the 30th International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2021: 1571–1578.
[54] DUCHI J C, JORDAN M I, WAINWRIGHT M J. Local privacy and statistical minimax rates[C]//Proceedings of 2013 IEEE 54th Annual Symposium on Foundations of Computer Science. Piscataway: IEEE Press, 2013: 429–438.
[55] ERLINGSSON Ú, PIHUR V, KOROLOVA A. RAPPOR: randomized aggregatable privacy-preserving ordinal response[C]// Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2014: 1054–1067.
[56] RASTOGI V, NATH S. Differentially private aggregation of distributed time-series with transformation and encryption[C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2010: 735–746.
[57] AGARWAL N, SURESH A T, YU F, et al. cpSGD: communication-efficient and differentially-private distributed SGD[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2018: 7575–7586.
[58] ZHANG C L, LI S Y, XIA J Z, et al. BatchCrypt: efficient homomorphic encryption for cross-silo federated learning[C]//Proceedings of the 2020 USENIX Annual Technical Conference. Berkeley: USENIX Association, 2020: 493–506.
[59] FANG H K, QIAN Q. Privacy preserving machine learning with homomorphic encryption and federated learning[J]. Future Internet, 2021, 13(4): 94
[60] GENTRY C. Fully homomorphic encryption using ideal lattices[C]// Proceedings of the 41st Annual ACM Symposium on Theory of Computing. New York: ACM Press, 2009: 169–178
[61] PHONG L T, AONO Y, HAYASHI T, et al. Privacy-preserving deep learning via additively homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(5): 1333–1345.
[62] YANG T, ANDREW G, EICHNER H, et al. Applied federated learning: improving

google keyboard query suggestions[J]. arXiv preprint, 2018, arXiv:1812.02903.

[63] MADI A, STAN O, MAYOUE A, et al. A secure federated learning framework using homomorphic encryption and verifiable computing[C]//Proceedings of 2021 Reconciling Data Analytics, Automation, Privacy, and Security: A Big Data Challenge. Piscataway: IEEE Press, 2020: 1–8.

[64] ZHU H F, MONG GOH R S, NG W K. Privacy-preserving weighted federated learning within the secret sharing framework[J]. IEEE Access, 2020, 8: 198275–198284.

[65] CHA J, SINGH S K, KIM T W, et al. Blockchain-empowered cloud architecture based on secret sharing for smart city[J]. Journal of Information Security and Applications, 2021, 57: 102686.

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

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

[68] CHANDRAMOULI A, CHOUDHURY A, PATRA A. A survey on perfectly-secure verifiable secret-sharing[J]. ACM Computing Surveys, 2022.

[69] FEREIDOONI H, MARCHAL S, MIETTINEN M, et al. SAFELearn: secure aggregation for private FEderated learning[C]//Proceedings of 2021 IEEE Security and Privacy Workshops. Piscataway: IEEE Press, 2021: 56–62.

[70] 周俊, 方国英, 吴楠. 联邦学习安全与隐私保护研究综述[J]. 西华大学学报(自然科学版), 2020, 39(4): 9–17.
ZHOU J, FANG G Y, WU N. Survey on security and privacy-preserving in federated learning[J]. Journal of Xihua University (Natural Science Edition), 2020, 39(4): 9–17

[71] BARACALDO N, CHEN B, LUDWIG H, et al. Mitigating poisoning attacks on machine learning models: a data provenance based approach[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2017: 103–110.

[72] SATTLER F, WIEDEMANN S, MÜLLER K-R, et al. Robust and communication-efficient federated learning from non-i.i.d. data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(9): 3400–3413.

[73] LIAO F Z, LIANG M, DONG Y P, et al. Defense against adversarial attacks using high-level representation guided denoiser[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1778–1787.

[74] XU W L, EVANS D, QI Y J. Feature squeezing: detecting adversarial examples in deep neural networks[J]. arXiv preprint, 2017, arXiv:1704.01155.

[75] ZHU C, CHENG Y, GAN Z, et al. FreeLB: enhanced adversarial training for language understanding[J]. arXiv preprint, 2019, arXiv:1909.11764.

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

[77] FUNG C, YOON C J M, BESCHASTNIKH I. Mitigating sybils in federated learning poisoning[J]. arXiv preprint, 2018, arXiv:1808.04866.

[78] 王健宗, 孔令炜, 黄章成, 等. 联邦学习隐私保护研究进展[J]. 大数据, 2021, 7(3): 130–149.
WANG J Z, KONG L W, HUANG Z C, et al. Research advances on privacy protection of federated learning[J]. Big

Data Research, 2021, 7(3): 130–149.
[79] ANDREINA S, MARSON G A, MÖLLERING H, et al. BaFFLe: backdoor detection via feedback-based federated learning[C]//Proceedings of 2021 IEEE 41st International Conference on Distributed Computing Systems. Piscataway: IEEE Press, 2021: 852–863.
[80] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv preprint, 2016, arXiv:1602.05629.
[81] YIN D, CHEN Y D, RAMCHANDRAN K, et al. Byzantine-robust distributed learning: towards optimal statistical rates[J]. arXiv preprint, 2018, arXiv:1803.01498.
[82] BLANCHARD P, MHAMDI E M E, GUERRAOUI R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 118–128.
[83] MHAMDI E M E, GUERRAOUI R, ROUAULT S. The hidden vulnerability of distributed learning in Byzantium[J]. arXiv preprint, 2018, arXiv:1802.07927.
[84] SO J, GÜLER B, AVESTIMEHR A S. Turbo-aggregate: breaking the quadratic aggregation barrier in secure federated learning[J]. IEEE Journal on Selected Areas in Information Theory, 2021, 2(1): 479–489.
[85] LEE H, KIM J, AHN S, et al. Digestive neural networks: a novel defense strategy against inference attacks in federated learning[J]. Computers & Security, 2021, 109: 102378.
[86] 周传鑫, 孙奕, 汪德刚, 等. 联邦学习研究综述[J]. 网络与信息安全学报, 2021, 7(5): 77–92.
ZHOU C X, SUN Y, WANG D G, et al. Survey of federated learning research[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 77–92.
[87] QUOC D L, FETZER C. SecFL: confidential federated learning using TEEs[J]. arXiv preprint, 2021, arXiv:2110.00981.
[88] LI W H, XIA Y B, LU L, et al. TEEv: virtualizing trusted execution environments on mobile platforms[C]//Proceedings of the 15th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments. New York: ACM Press, 2019: 2–16.
[89] 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.
[90] ZHAO Y, ZHAO J, JIANG L S, et al. Mobile edge computing, blockchain and reputation-based crowdsourcing IoT federated learning: a secure, decentralized and privacy-preserving system[J]. arXiv preprint, 2019, arXiv:1906.10893.
[91] DOSHI-VELEZ F, KIM B. Towards a rigorous science of interpretable machine learning[J]. arXiv preprint, 2017, arXiv:1702.08608.

作者简介

吴建汉(1998-), 男, 中国科学技术大学硕士生, 平安科技(深圳)有限公司算法工程师, 中国计算机学会(CCF)学生会员, 主要研究方向为计算机视觉和联邦学习。

司世景（1988- ），男，博士，平安科技（深圳）有限公司资深算法研究员，CCF会员，主要研究方向为机器学习及其在人工智能领域的应用。

王健宗（1983- ），男，博士，平安科技（深圳）有限公司副总工程师，资深人工智能总监，联邦学习技术部总经理，CCF高级会员，CCF大数据专家委员会委员，主要研究方向为联邦学习和人工智能等。

肖京（1972- ），男，博士，平安科技（深圳）有限公司首席科学家，2019年吴文俊人工智能杰出贡献奖获得者，CCF深圳会员活动中心副主席，主要研究方向为计算机图形学学科、自动驾驶、3D显示、医疗诊断、联邦学习等。

收稿日期： 2021-11-15
通信作者： 王健宗，jzwang@188.com
基金项目： 广东省重点领域研发计划“新一代人工智能”重大专项（No.2021B0101400003）
Foundation Item: The Key Research and Development Program of Guangdong Province (No.2021B0101400003)