

纵向联邦学习方法及其隐私和安全综述

陈晋音^{1,2}, 李荣昌², 黄国瀚², 刘涛², 郑海斌¹, 程瑶³

(1. 浙江工业大学网络空间安全研究院, 浙江 杭州 310023; 2. 浙江工业大学信息工程学院, 浙江 杭州 310023;
3. 南德认证检测亚太有限公司, 新加坡 60993)

摘要: 联邦学习 (FL, federated learning) 是一种新兴的分布式机器学习技术, 利用分散在各个机构的数据, 通过传输中间结果 (如模型参数、参数梯度、嵌入信息等) 实现机器学习模型的联合构建。联邦学习中机构的训练数据不允许离开本地, 因此降低了数据泄露的风险。根据机构之间数据分布的差异, FL 通常分为横向联邦学习 (HFL, horizontal FL)、纵向联邦学习 (VFL, vertical FL), 以及联邦迁移学习 (TFL, transfer FL)。其中, VFL 适用于机构具有相同样本空间但不同特征空间的场景, 广泛应用于医疗诊断、金融评估和教育服务等领域。尽管 VFL 在现实应用中有出色的表现, 但其本身仍然面临诸多隐私和安全问题, 尚缺少对 VFL 方法与安全性展开全面综述的工作。为了构建高效且安全的 VFL 系统, 从 VFL 方法及其隐私和安全两个方面展开, 首先从边缘模型、通信机制、对齐机制以及标签处理机制 4 个角度对现有的 VFL 方法进行详细总结和归纳; 其次介绍并分析了 VFL 面临的隐私和安全风险; 进一步对其防御方法进行介绍和总结; 此外, 介绍了适用于 VFL 的常见数据集及平台框架。结合 VFL 面临的安全性挑战给出了 VFL 的未来研究方向, 旨在为构建高效、鲁棒和安全的 VFL 的理论研究提供参考。

关键词: 纵向联邦学习; 安全与隐私; 后门攻击; 推断攻击与防御; 对抗攻击; 安全性评估

中图分类号: TP18

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023017

Survey on vertical federated learning: algorithm, privacy and security

CHEN Jinyin^{1,2}, LI Rongchang², HUANG Guohan², LIU Tao², ZHENG Haibin¹, CHENG Yao³

1. Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China

2. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

3. TÜV SÜD Asia Pacific Pte. Ltd., 60993, Singapore

收稿日期: 2022-03-16; 修回日期: 2022-08-20

通信作者: 陈晋音, chenjinyin@zjut.edu.cn

基金项目: 国家自然科学基金 (62072406); 浙江省自然科学基金 (DQ23F020001); 信息系统安全技术重点实验室基金 (61421110502); 国家重点研发计划 (2018AAA0100801)

Foundation Items: The National Natural Science Foundation of China (62072406), Zhejiang Provincial Natural Science Foundation (DQ23F020001), The National Key Laboratory of Science and Technology on Information System Security (61421110502), The National Key R&D Program of China (2018AAA0100801)

引用格式: 陈晋音, 李荣昌, 黄国瀚, 等. 纵向联邦学习方法及其隐私和安全综述[J]. 网络与信息安全学报, 2023, 9(2): 1-20.

Citation Format: CHEN J Y, LI R C, HUANG G H, et al. Survey on vertical federated learning: algorithm, privacy and security[J]. Chinese Journal of Network and Information Security, 2023, 9(2): 1-20.

Abstract: Federated learning (FL) is a distributed machine learning technology that enables joint construction of machine learning models by transmitting intermediate results (e.g., model parameters, parameter gradients, embedding representation, etc.) applied to data distributed across various institutions. FL reduces the risk of privacy leakage, since raw data is not allowed to leave the institution. According to the difference in data distribution between institutions, FL is usually divided into horizontal federated learning (HFL), vertical federated learning (VFL), and federal transfer learning (TFL). VFL is suitable for scenarios where institutions have the same sample space but different feature spaces and is widely used in fields such as medical diagnosis, financial and security of VFL. Although VFL performs well in real-world applications, it still faces many privacy and security challenges. To the best of our knowledge, no comprehensive survey has been conducted on privacy and security methods. The existing VFL was analyzed from four perspectives: the basic framework, communication mechanism, alignment mechanism, and label processing mechanism. Then the privacy and security risks faced by VFL and the related defense methods were introduced and analyzed. Additionally, the common data sets and indicators suitable for VFL and platform framework were presented. Considering the existing challenges and problems, the future direction and development trend of VFL were outlined, to provide a reference for the theoretical research of building an efficient, robust and safe VFL.

Keywords: vertical federated learning, security and privacy, backdoor attack, inference attack and defense, adversarial attack, security evaluation

0 引言

随着计算资源和训练数据的规模大幅提升,机器学习^[1]在众多领域得到快速发展,如医疗诊断、金融评估和教育服务等。近年来,出于对数据隐私的保护,许多国家与地区出台数据保护法规限制数据的采集和直接传输。例如,2021年6月,我国出台的《中华人民共和国数据安全法》^[2]明确要求对个人隐私、个人信息以及保密商务信息等数据应当依法予以保密,不得泄露或者非法向他人提供。2018年5月,欧盟出台的法案《通用数据保护条例》^[3]对数据进行严格管理,保证数据隐私安全。受限于此类隐私保护法规,拥有数据的机构之间无法直接通过交换隐私数据的方式集中训练机器学习模型,从而形成了数据孤岛现象,并极大地制约了机器学习的性能。如何在保护数据隐私的前提下,满足机构之间联合训练机器学习模型,是当下亟待解决的问题。

为了解决以上难题,Hord^[4]提出了联邦学习(FL, federated learning),通过聚合机构之间共享中间数据结果(边缘模型或者嵌入层)的方式联合训练机器学习模型。由于联邦学习在训练过程中的数据始终不离开数据拥有者,因此降低了机构在联合训练阶段隐私泄露的风险,满足数据隐私法规的要求。根据机构之间数据分布的差异^[5],

联邦学习可以分为横向联邦学习(HFL, horizontal FL)、纵向联邦学习(VFL, vertical FL)和联邦迁移学习(TFL, transfer FL)。为了清晰呈现上述3种联邦学习技术各自适用的场景,图1展现了纵向联邦学习中的参与者在不同场景下进行联邦学习的数据示意。两个矩阵分别表示两个参与者的数据空间,其中矩阵的长表示参与者的数据特征空间维度,矩阵的宽表示参与者的数据样本空间维度。

横向联邦学习适用于参与者之间的数据集具有相同的特征空间但是不同样本空间的场景^[5](如图1(a)所示)。例如,两家来自不同地区的银行希望共同训练一个机器学习模型,提供贷款精准推销业务。一方面,这两家银行具有相似的业务,被视作具有相同的数据特征空间;另一方面,它们来自不同的地区,因此他们的样本空间不同。横向联邦学习为上述场景提供了联合训练的解决方案。

纵向联邦学习适用于参与者之间的数据集具有相同的样本空间但是不同特征空间的场景^[5](如图1(b)所示)。例如,来自同一个地区的银行和电子商务公司希望共同训练一个机器学习模型,为电子商务公司提供精准推荐服务。一方面,这两家机构具有不同的业务空间,因此具有不同的用户特征空间;另一方面,它们来自同一个地区,因此具有相似的样本空

间。纵向联邦学习为上述场景提供联合训练的解决方案。

联邦迁移学习适用于参与者之间的数据集具有样本空间和特征空间都不同的场景^[5]（如图1(c)所示）。例如，一家机构是位于A省的电子商务公司，另一家是位于B省的金融机构，联合构建机器学习模型。一方面，由于地理限制，这两个机构具有很少的用户交集；另一方面，两家机构的业务不同，因此具有的特征空间不相同。联邦迁移学习为上述场景提供解决方案。

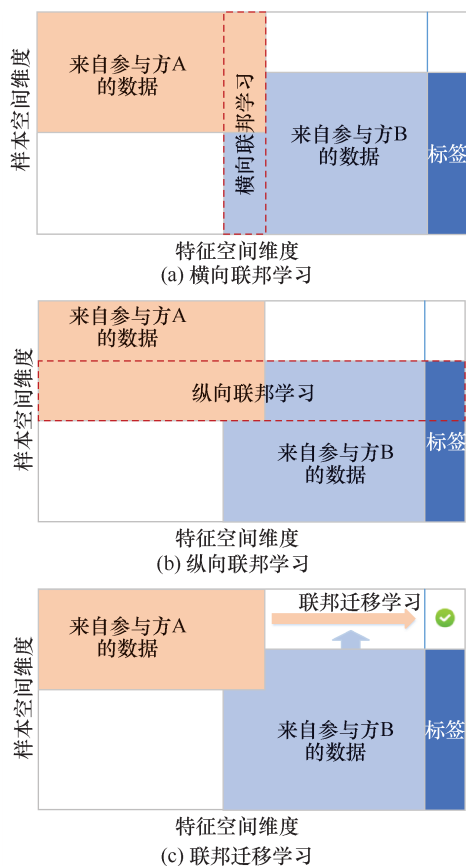


图1 联邦学习场景

Figure 1 Federated learning scenarios

现有与联邦学习相关的综述论文主要关注横向联邦学习^[6-7]，国内外尚未有相关综述论文针对纵向联邦学习及其隐私和安全现有研究技术进行归纳总结。随着现实场景中来自不同领域的机构之间商务合作日益紧密，纵向联邦学习的应用^[8-9]日益广泛，其方法及隐私和安全问题得到了学术界和工业界的高度关注。因此，本文基于国内外纵向联邦学习的研究现状，对纵向联邦学习方法

及其隐私和安全问题进行梳理和归纳，并指明一些纵向联邦学习未来可能的研究方向。值得一提的是，本文涉及的参考文献的开源代码和论文资源已经整理并公开到 GitHub，方便更多学者对纵向联邦学习开展深入研究。

纵向联邦学习在众多现实场景得到广泛应用，得益于现有研究提出的众多纵向联邦学习算法。随着纵向联邦学习技术的发展，边缘模型支持处理的数据模态由文本和图像向时序和网络模态扩展；通信机制由同步通信逐渐向异步通信发展；样本对齐方法由简单的秘密求交集技术向引入了半监督学习^[10]以及相似度匹配^[11]技术的样本对齐技术发展。

随着纵向联邦学习在众多领域的应用，其暴露的隐私和安全问题逐渐成为学术界和工业界研究的热点。尽管纵向联邦学习在训练过程中用户的原始数据没有离开本地，但一些现有的研究表明其仍然存在隐私泄露的风险，如常见的属性推断攻击^[12]、标签推断攻击^[13]以及数据重构攻击^[14]等。

另外，对于纵向联邦学习存在的安全风险，Liu 等^[15]表明纵向联邦学习中恶意的客户端可以通过操纵梯度或训练数据发动后门攻击。例如，攻击者在训练阶段通过将带有特殊标识（即后门触发器）的“停车”图像注入训练集参与纵向联邦学习的训练。纵向联邦学习系统在推理阶段，会对携带后门触发器的恶意停车标志产生错误的预测结果。

为了提高纵向联邦学习的隐私性与安全性，已有研究提出各种防御方法，基本可分为3类：基于加密、基于扰动以及基于对抗训练的防御方法。这些防御方法具体包括同态加密技术、秘密共享技术、差分隐私技术、梯度压缩技术以及对抗训练技术等。

综上所述，为了进一步保障纵向联邦学习方法的安全应用，本文针对已有的纵向联邦学习方法及其隐私和安全研究工作展开系统梳理、归纳与分析。

1 纵向联邦学习方法

1.1 预备知识

1.1.1 概念及定义

纵向联邦学习是一种新兴的分布式机器学习技术，其利用分散在各个机构的数据集，通过传输中间结果（如模型参数、模型梯度、嵌入信息

等)来联合构建机器学习模型。VFL 适用于参与者之间的数据集共享相同的样本空间但不同特征空间的场景。

假设当前有 m 个参与者 $\{P_1, \dots, P_m\}$, 其本地数据为 $\{D_1, \dots, D_m\}$, 标签为 y 。数据集中的每一行对应一个样本, 每一列对应一个特征。从参与者的边缘模型为 θ_m , 主参与者或协调者具有的顶端模型为 θ_T 。VFL 的优化目标表述为

$$\text{Min}_{\theta_1, \dots, \theta_m, \theta_T} f(\theta_1, \dots, \theta_m, \theta_T; D, y) \quad (1)$$

其中, $f(\cdot)$ 为预测的损失函数。

1.1.2 参与者角色

主参与者: VFL 中提供标签和特征信息的参与者。在现实应用场景中, 主参与者往往发布 VFL 的主任务。

从参与者: VFL 中仅提供特征信息的参与者。参与 VFL 训练过程中, 其提供特征信息。

协调者: 负责协调 VFL 的训练过程, 并在数据加密的场景中提供解密功能。在现实应用场景中, 协调者的角色通常由主参与者兼任。

1.2 VFL 边缘模型

VFL 支持的边缘模型通常包括线性回归模型、逻辑回归模型、树模型及神经网络等。接下来的小节按照这 4 种不同边缘模型, 对 VFL 框架进行介绍。

1.2.1 基于线性回归模型的 VFL

Yang 等^[5]提出一种基于线性回归的 VFL 框架, 其中参与者采用线性回归模型。

为了将数据从一方安全地传输到另一方, 该框架采用了加法同态加密技术。VFL 训练过程中, 协调者首先创建加密对, 并为参与者 A 和参与者 B 发送公钥, 之后参与者 A 为参与者 B 发送中间加密结果 $[[u_i^A]]$, 其中 $[[\cdot]]$ 表示加法同态加密, 参与者 A 的模型参数为 θ_A , 参与者 B 的模型参数为 θ_B 。参与者 B 同时进行加密交换梯度, 并向协调者发送损失值, $\mathcal{L}$ 表示损失函数。然后, 参与者 A 和 B 分别计算各自加密的梯度 $[[\frac{\partial \mathcal{L}}{\partial \theta_A}]]$ 以及 $[[\frac{\partial \mathcal{L}}{\partial \theta_B}]]$ 并添加加密的掩码 $[[R_A]]$ 和掩码 $[[R_B]]$, 发送给协调者添加掩码后的加密结果 $[[\frac{\partial \mathcal{L}}{\partial \theta_A}]] + [[R_A]]$ 和

$[[\frac{\partial \mathcal{L}}{\partial \theta_B}]] + [[R_B]]$ 。最后, 协调者解密梯度并返回解

密后的梯度信息 $\frac{\partial \mathcal{L}}{\partial \theta_A} + R_A$ 和 $\frac{\partial \mathcal{L}}{\partial \theta_B} + R_B$ 给参与者 A 和 B。依次迭代, 完成纵向联邦学习训练。上述线性纵向联邦学习算法为了避免协调者直接获得参与者的梯度信息, 通过添加附加掩码来保护参与者的梯度信息。

1.2.2 基于逻辑回归模型的 VFL

Hardy 等^[16]提出了一种基于随机梯度下降的纵向逻辑回归框架, 两个参与者 A 和 B 在每次迭代时交换加密的中间结果计算梯度。首先参与者 A 计算中间结果并获得加密后的梯度信息; 然后, 参与者 B 同时反馈中间计算信息给协调者; 最后, 协调者利用随机梯度下降方法获得参与者的梯度信息。该框架为了不将明文数据从一方传输到另一方, 同样采用加法同态加密技术。利用加法同态加密计算两个加密数字的加法以及一个未加密数字和一个加密数字的乘积。

然而, 损失函数及其梯度不能直接使用加法同态加密进行计算。针对这个问题, Yang 等^[17]采用 Taylor 近似法来优化损失函数, 使其适用于加法同态加密。

文献[16-17]设定 VFL 中具有可信的第三方协调者, 然而这种可信的协调者现实场景中往往很难得到保证, 会带来潜在的安全风险。因此, Yang 等^[18]提出了一个去除协调者的逻辑回归纵向联邦学习方法。参与者采用参数服务器体系结构, 其中包括一个集中的参数服务器和一组工作节点。工作节点之间没有连接, 工作节点只能与其相应的参数服务器通信。参与者之间的唯一通信通道存在于各自对应的参数服务器之间。由于去除了 VFL 中的协调者, 大大降低了系统的复杂性, 并且从根本上避免了协调者带来的安全和隐私泄露威胁。

1.2.3 基于树模型的 VFL

Cheng 等^[19]提出了一种无损的隐私保护树集成系统 SecureBoost, 其中, VFL 中边缘模型为树模型。SecureBoost 的优点为它提供了与非隐私保护方法相同的主任务性能, 同时不泄露每个参与者的隐私数据。对于决策树联邦模型的分割方式, 它从深度为 0 的树开始, 并为树的每个叶节点添加一个拆分, 直到达到最大

深度。SecureBoost 假设执行期间的一些中间结果可以以明文形式显示，这种中间结果可能被攻击者利用，以推断参与者数据中的敏感信息。

针对上述问题，Wu 等^[20]提出一种基于决策树的保护 VFL 中间结果隐私安全的 Pivot 框架。Pivot 的核心采用阈值部分同态加密（TPHE，threshold partially homomorphic encryption）和安全多方计算（SMC，secure multi-party computation）。这两种加密技术在 VFL 中相互补充，TPHE 在通信成本方面相对有效，但只能支持一组限制性计算。SMC 可以支持任意计算，但会产生昂贵的通信开销。Pivot 尽可能多地使用 TPHE 来促进客户端的本地计算，并且仅在 TPHE 功能不足的地方调用 SMC。此外，Pivot 引入差分隐私技术加强对数据隐私的保护。

1.2.4 基于神经网络的 VFL

Vepakomma 等^[21]在人工神经网络的基础上提出了一种分裂学习，这同样适用于 VFL。在前向传播中，客户端利用边缘模型使用本地数据训练输出剪切层。剪切层的输出被发送到协调者，协调者完成剩余的训练。在反向传播过程中，协调者从最后一层反向传播到剪切层，剪切层的梯度被发送回客户，其余的反向传播先在客户端完成。现有研究^[13,22]采用分裂学习思想的 VFL 框架，用于广告推荐或物体识别领域。

为了将 VFL 适用于处理图数据，Ni 等^[23]提出基于图卷积神经网络的 VFL 框架。同时引入了同态加密技术保护 VFL 的隐私安全。类似地，Zhou 等^[24]提出了一种适用于节点分类任务的纵向图神经网络（VFGNN）框架，可以推广到现有的图神经网络（GNN）模型。框架的训练过程中，参与者采用安全多方计算和差分隐私技术。首先协同使用私有节点特征信息计算 GNN 的初始层，然后单独使用私有边缘信息执行邻域聚合，最后得到局部节点嵌入，聚合策略分为 3 种。

（1）拼接策略

拼接操作符可以完全保留从不同被动参与者学习到的本地节点嵌入。具体地，中间嵌入信息按照设定维度进行拼接形成聚合后的嵌入特征。

$$\mathbf{h}_v^K = \text{CONCAT}(\mathbf{h}_v^K(1), \mathbf{h}_v^K(2), \dots, \mathbf{h}_v^K(I)) \quad (2)$$

其中， $\mathbf{h}_v^K(I)$ 表示联邦学习中的各个参与方上传

的嵌入特征。

（2）平均策略

平均算子取中间嵌入信息元素 $\mathbf{h}$ 的平均值，通常适用于参与者数据对全局节点嵌入的贡献相等。

$$\mathbf{h}_v^K = \text{MEAN}(\mathbf{h}_v^K(1) \cup \mathbf{h}_v^K(2) \cup \dots \cup \mathbf{h}_v^K(I)) \quad (3)$$

（3）回归策略

通过回归模型将数据持有者的中间嵌入信息组合起来，在训练过程中自动学习回归模型的参数。

$$\mathbf{h}_v^K = \omega_1 \odot \mathbf{h}_v^K(1) + \omega_2 \odot \mathbf{h}_v^K(2) \dots + \omega_I \odot \mathbf{h}_v^K(I) \quad (4)$$

其中， ω 为回归模型的参数。

然后，协调者组合来自数据持有者的本地节点嵌入并生成全局节点嵌入，在此基础上协调者可以执行连续的非私有数据计算，其损失函数为

$$L = - \sum_{v \in \mathcal{V}_v} \sum_{c \in \mathcal{C}} y_{vc} \ln(\hat{y}_{vc}) \quad (5)$$

其中， $\hat{y}_{vc}$ 为训练节点 y_v 预测标签， y_{vc} 为训练节点真实标签。

在 VFGNN 框架中，隐私数据和模型都由参与者本地持有，只有嵌入表示在训练过程中进行传输，并且传输的嵌入信息被添加差分隐私噪声进行加固，可保证数据隐私安全。

1.2.5 小结

众多 VFL 边缘模型的研究使得其能够适用于多种模态的数据。为了避免传输过程中的数据直接泄露，通常使用同态加密或安全多方计算加密手段保护中间数据。此外，边缘模型为神经网络的 VFL 中通常设定第三方协调者具有顶端模型，具有信息二次融合提取的能力；而当边缘模型为其他类型时，协调者只起到解密和计算梯度的作用。

1.3 VFL 通信机制

VFL 训练通信过程中存在的问题主要有：不同参与者的计算能力和通信能力有差别，形成“短板效应”；模型收敛速度慢；通信过程中存在大量冗余数据；通信时延。为此，现有针对 VFL 的研究从 4 个方面降低通信代价，其中包括基于异步更新、基于拟牛顿法、基于梯度压缩及基于直接内存访问。本节对每一种方法进行详细分析。

1.3.1 基于异步更新的 VFL

常见的采用同步更新机制的 VFL，存在通信

能力强的参与者等待计算能力弱的参与者的问题,形成“短板效应”。为了解决上述问题,异步更新采用多次本地局部更新策略以及异步聚合策略减小训练通信代价,减少设备闲置时间。

Gu 等^[25]提出了 3 种异步纵向联邦随机梯度下降算法,分别为 AFSGD-VP、AFSVRG-VP 以及 AFSAGA-VP。异步纵向联邦随机梯度下降算法减少了以往同步联邦的等待消耗时间,然而其要求客户端每次参数更新都和协作方建立通信联系,消耗较多的计算资源。受横向联邦学习中的联邦平均算法^[26]启发,Liu 等^[27]提出了一种联邦随机块坐标下降(FedBCD, federated stochastic block coordinate descent)算法,核心思想是纵向联邦学习中每一个参与者在每次通信之前进行多次局部更新,以减少参与者之间的通信轮数。类似地,Chen 等^[28]提出一种适用于客户端间歇性训练的纵向异步联邦学习算法,通过建立主动方的查询机制,实时响应来自客户端的异步训练更新模型请求。

上述工作的局限性在于只适用于数据集由单独一块存储器进行保存,且隶属于单个客户机。因此,上述工作不适用于数据分散在多个客户机的现实场景。为了解决这个问题,Das 等^[29]提出一种分层分散坐标下降(TDCD, tiered decentralized coordinate descent)算法,存储器之间执行并行坐标下降算法,在每个存储器内的客户端底层执行分布式随机梯度下降算法,其每个通信轮次中包含多个局部更新。存储器中的客户端在其本地数据集上对这些参数进行多次迭代。然后仓库协调者聚合客户机模型以更新其参数块。此外,在本地局部更新中,客户机需要从其他的客户机中请求中间嵌入数据,仓库协调者响应并协助中间信息交换。通过分层网络的结构来解决多个不同客户机进行纵向联邦学习的需求。

相较于同步更新机制,纵向联邦学习采用异步更新具有以下优势。

- 1) 客户端参数更新无须等待其他参与者,计算资源利用率高,避免短板效应。
- 2) 增加模型的泛化能力,容错性更强,可以有效应对部分客户端设备宕机的问题。
- 3) 引入局部更新机制,减少了通信次数。

1.3.2 基于拟牛顿法的 VFL

基于拟牛顿法的 VFL 核心思想是使用二阶 Hessian 矩阵梯度下降代替一阶梯度下降,加快模型收敛进而减少通信轮次。直接对 Hessian 矩阵求逆会带来很大时间成本,Wright 等^[30]提出了一种拟牛顿算法(L-BFGS),通过直接逼近 Hessian 矩阵求逆降低计算成本。然而,经典的拟牛顿法并不适用于 VFL 数据分布的场景。因为经典的 L-BFGS 算法通过计算两个连续迭代之间梯度和模型参数,在最后一次迭代中使用历史信息来获得估计 Hessian 矩阵的逆,适用于完整数据而非小批量数据。VFL 中客户端的数据较小,会导致估计的 Hessian 矩阵不稳定。为此,Yang 等^[17]提出了一种基于拟牛顿法的逻辑回归纵向联邦学习框架。Hessian 矩阵信息可以每轮迭代更新一次,提高拟牛顿算法的稳定性。

尽管基于 Hessian 矩阵子采样具有很好的稳定性,但通信代价很高。这主要有两方面原因:一方面是估计 Hessian 矩阵的过程中需要在每轮迭代过程中传输大量的梯度差值;另一方面是算法采用同步计算,迭代过程中需要大量等待时间。为此,Zhang 等^[31]提出了一个基于异步拟牛顿的 VFL 框架 AsySQN。AsySQN 利用近似 Hessian 信息获得更好的下降方向,提高实际收敛速度。客户端根据局部参数差值来计算近似的局部 Hessian 信息,以避免因传输梯度差而产生的高通信成本。此外,为了使其应用于现实中具有不平衡计算资源的 VFL 系统,采用了异步并行化的随机拟牛顿算法。这种异步化的随机拟牛顿纵向联邦学习算法具有更低的通信成本,更适用于计算资源不同的现实场景。

1.3.3 基于梯度压缩的 VFL

深度梯度压缩^[32]是通过压缩梯度的方式来解决通信带宽问题,通过仅发送重要梯度的方式减少了通信带宽。Yang 等^[33]基于深度梯度压缩方法提出一种通信双方参数共享和梯度压缩的 VFL,同时为了保证精度不受影响,采用了动量校正、局部梯度剪裁、动量因子掩蔽和热身训练。类似地,Li 等^[34]提出了一种基于梯度预测和双端稀疏压缩的异步纵向联邦学习框架,其使用预测

值的增量幅度启发式方法来确定被动方参与者数据的重要性，只有最大的预测值变化才会传输到主动方并参与参数更新。其余部分的值会随着模型参数的更新而局部累积，直到达到设定阈值才可以进行传输。对于主动方采用只传送幅度最大的梯度。为了防止梯度信息丢失，在局部累积剩余的梯度，随着梯度增加到设定阈值，累计的梯度进行传输。

1.3.4 基于直接内存访问的 VFL

VFL 中常见的通信方式是采用 Google 发布的基于 HTTP 2.0 传输层协议承载的高性能开源软件框架 GRPC (Google remote procedure call)。GRPC 是一个基于传输控制协议的远程过程调用框架，因此数据传输具有较高的时延，并消耗大量 CPU 资源。高时延可能会显著降低参与者之间的通信速度，而额外的 CPU 成本会导致用于加密、解密和训练作业的资源减少。为了降低 VFL 中的通信成本，Liu^[35]提出了一种使用远程直接内存访问 (RDMA, remote direct memory access) 传输数据的设计，用于内部通信，而不需要对应应用程序进行任何修改。为优化 RDMA 在 VFL 中的适用性能，其设计了一个面向性能分配的优化器。一方面，优化器观察参与者的历史通信的数据量变化调整 RDMA 的分配；另一方面，优化器以二进制搜索方式搜索最佳查询数据大小。在传输过程中动态更改查询数据大小，达到最佳的平均 RDMA 传输速度。

1.3.5 小结

现有的 VFL 关于降低通信成本的算法研究正朝着异步化、收敛快速化以及通信低时延化方向发展。得益于二阶梯度比一阶梯度具有更快的收敛速度，一些研究采用二阶梯度加快 VFL 的收敛速度，进而降低通信成本。此外，一些注重于客户端之间通信方式的研究有效提升了纵向联邦学习的通信效率。

1.4 VFL 对齐机制

数据对齐是指在训练前期将不同客户端之间相同或类似的数据条目进行链接。数据对齐作为 VFL 的重要阶段，直接影响模型的预测性能。此外，数据对齐的方式也会直接影响数据 ID 信息的隐私安全。

1.4.1 基于隐私集合求交的 VFL

隐私集合求交 (PSI, private set intersection) 技术的目的为实现跨机构之间数据安全合作，计算机构之间在隐私数据集的交集，同时不泄露任何交集以外的信息。这种技术在众多 VFL 算法的数据对齐阶段中被采用。现有研究^[36]将 PSI 技术分成了 5 个类别：朴素的隐私集合求交技术、基于公钥体系的隐私集合求交技术^[37-38]、基于不经意传输的隐私集合求交技术^[39-40]、基于可信执行环境下的隐私集合求交技术以及隐私集合交集基数计算技术^[41-42]。文献[5]直接采用隐私集合求交技术来实现不同 VFL 中参与者的数据对齐。

1.4.2 基于相似度匹配的 VFL

Wu 等^[11]针对 VFL 中的数据对齐设计了一种基于相似度对齐数据的框架。该框架包含两个重要的组件：模糊链接和基于相似度链接。在模糊链接中，VFL 中的协调者初步筛选出候选链接对，计算客户端之间样本的相似度距离。

在基于相似度的组件中，通过在分裂网络外层附加相似度训练网络，同时兼顾主任务训练和样本对齐。具体而言，相似度度量值经过权重门映射到权重。经过映射的权重信息传入合并门，用以合并相似度权重和模型训练数据。之后经过分拣门对相似度进行排序，从而稳定 VFL 的更新。上述工作在手动模糊筛选链接后，进一步融合相似度网络到 VFL 中，在训练过程中寻找相似度高的数据链接对的同时提高 VFL 性能。

1.4.3 基于半监督学习的 VFL

现实场景下的 VFL 中存在样本空间重叠小的情况，常见的 VFL 难以进行正常的训练。为了得到更多的对齐样本，Kang 等^[10]提出一种基于半监督学习的联邦多视角训练方法来补全未对齐数据的缺失特征值。该方法补全样本缺失特征值，增加了对齐数据样本的数量，这使得 VFL 在充分利用重叠样本的同时，扩展了样本对齐空间，进而提高 VFL 的性能。

1.4.4 小结

样本对齐是 VFL 中重要的环节，直接影响纵向联邦学习的性能，并且潜在地关联到数据 ID 信息。对于数据身份信息完整的情形，使用经典的隐私集合求交对齐的技术，同时兼顾数据样本对

齐和数据 ID 信息保护。对于样本身份信息模糊的情况,基于相似度的数据条目匹配能有效解决数据对齐问题。在样本空间重叠小的情况,通过应用半监督学习对样本特征空间进行补全的方式可增加样本对齐空间,从而提升 VFL 整体性能。

1.5 VFL 标签处理机制

根据 VFL 的应用场景不同,其参与者的标签分布会存在差异。纵向联邦学习场景中的标签通常分为两种情形,标签为一方提供或标签为多方提供,其中标签为一方提供的场景较为常见,如一家专科医院具有所有病人的某种疾病的诊断信息。而多方持有标签的场景往往发生在没有一方可以提供整套完整的标签数据,需多个参与者共同提供完整的标签数据。由于两种场景中标签作为隐私数据无法直接进行共享,这两种场景中的标签处理方法存在差异。

1.5.1 标签为一方持有

常见的 VFL 研究中^[21,23]假设标签为主动方持有,主动方为第三方协调者提供标签训练模型。其中,被动方和主动方同时利用边缘模型提取嵌入特征,由第三方协调者完成对模型梯度信息的计算。然而这些研究往往局限于同步计算,其算法效率低。针对上述问题,Zhang 等^[43]提出一种反向更新机制(BUM)和双层异步并行体系结构(BAPA) VFL 框架。BUM 的核心思想是使被动方在不直接访问原始标签数据的情况下,间接使用标签计算随机梯度,通过将标签信息嵌入传输的中间信息中,使得被动方能够利用梯度下降算法更新边缘模型。BAPA 使得内部参与者并行计算,从而所有活动方能够异步启动更新动作。内部一方表示各方内协作更新的共享内存并行,使得参与者的多个线程能够异步执行协作更新。这种嵌入标签的 VFL 框架同时调动主动方和被动的计算资源,提高 VFL 的工作效率。

1.5.2 标签为多方持有

Xia 等^[44]提出了一个级联纵向联邦学习(CVFL, cascade VFL),以自下而上的方式将整个神经网络分解为两类子网络。在这套系统中存在被动方、主动方以及协调者。每个被动方都具有各自的边缘模型,边缘模型学习将高维原始特征映射为低维紧凑嵌入向量。每个主动方负责训

练本地聚合模型,其拼接所有的嵌入特征向量作为输入训练聚合模型。在训练过程中,主动方使用持有的标签利用计算边缘模型和聚合模型的梯度信息。协调者接收并聚合所有主动方的聚合模型的梯度信息更新全局模型。

级联结构解决了传统端到端反向传播中模型更新标签的必要性。此外,为了增加聚合算法在非独立同分布场景中的适应性,Mugunthan 等^[45]采用横向联邦学习中的自适应联邦聚合算法^[46]。通过在联合训练优化过程中引入动量和自适应学习率技术,提高模型的收敛能力,进而更适用于非独立同分布场景下的应用。

针对涉及多方的多类 VFL 场景,Feng 等提出了多参与者垂直联合学习框架^[47],扩展了多视图学习的思想,同其他 VFL 参与者共享标签。Hu 等^[48]提出了一个 ADMM 共享框架,实现分布式特征共享,其中每一方只需要在训练过程中为每个样本共享一个值,从而最大限度地减少数据泄露风险。此外,Gu 等^[49]提出了一种用于垂直分区数据的联合双随机内核学习算法实现快速的收敛。

1.5.3 小结

标签在参与者之间分布的差异导致 VFL 的算法有所区别。引入分裂学习思想建立级联的 VFL 框架或者将标签信息嵌入客户端交换的中间信息提供解决标签分布差异问题的新思路。

2 纵向联邦学习隐私和安全风险

随着纵向联邦学习在众多领域得到应用,其隐私和安全引起学术界和工业界广泛讨论^[50]。VFL 的安全性指模型的预测结果不能偏离预期。本节从推断攻击、后门攻击以及对抗攻击对 VFL 上现有的攻击进行系统总结和科学归纳,并讨论相较于横向联邦学习,VFL 面临风险的独特性。常见的纵向联邦学习中面临的安全和隐私威胁如表 1 所示。

2.1 VFL 中的推断攻击

VFL 中参与方的隐私信息通常包括:数据特征、成员 ID、属性信息以及标签信息。本节分别介绍和分析 VFL 中隐私信息泄露的推断攻击方法。

表1 常见的纵向联邦学习中面临的安全和隐私威胁
Table 1 Common security and privacy threats in vertical federal learning

攻击类型	攻击目标	攻击方法	攻击阶段		攻击背景知识			威胁强度	VFL 框架	数据模态
			训练阶段	推理阶段	协调者模型	嵌入信息	标签信息			
隐私性	标签	模型补全	√	√	×	×	×	●	神经网络	图像、文本
隐私性	标签	主动攻击	√	×	×	×	×	●	神经网络	图像、文本
隐私性	标签	直接攻击	√	√	×	×	×	○	神经网络	图像、文本
隐私性	标签	范数攻击	√	√	×	√	×	○	神经网络	文本
隐私性	属性	等式求解攻击	√	√	×	×	×	●	逻辑回归	文本
隐私性	属性	路径限制攻击	√	√	×	×	×	●	决策树	文本
隐私性	属性	生成回归网络	×	√	√	√	×	●	神经网络	文本
隐私性	属性	解码器攻击	√	√	×	√	×	●	神经网络	图像、文本
隐私性	成员信息	PSI 暴露	√	√	×	×	×	●	神经网络、逻辑回归	图像、文本
隐私性	原始数据	逆向乘法攻击	√	×	×	×	×	●	逻辑回归	文本
隐私性	原始数据	逆向加法攻击 ¹	√	×	×	×	×	●	XGBoost	文本
隐私性	原始数据	特征空间劫持攻击	√	×	×	×	×	●	神经网络	图像
安全性	VFL 模型	梯度替换后门攻击	√	×	×	×	√	●	神经网络	图像、文本

注：√表示满足，×表示不满足。威胁强度（根据攻击所需背景知识和攻击场景评估其具有的威胁性）：○表示低，●表示中，●表示高。

2.1.1 标签推断攻击

在纵向联邦学习中，主参与者的标签信息属于隐私信息，因为标签信息属于参与者的重要资产或者标签本身具有高度敏感性。例如，提供标签的机构是一家专科医院，这些标签为某一个病人是否患有“艾滋病”的信息。

Fu 等^[22]提出针对 VFL 的标签推断攻击，包括被动标签推断攻击、主动标签推断攻击、直接标签推断攻击 3 种攻击方法。攻击的原理是 VFL 在训练期间传播的梯度通常有助于攻击者的边缘模型学习关于标签的良好特征表示，因此边缘模型可能包含标签的信息。此外，VFL 在训练期间传输的梯度也可能直接包含标签的信息。

在被动标签推断攻击^[21]中，攻击者利用完成训练的边缘模型，通过模型补全的方式发动标签推断攻击。具体地，攻击者在少量辅助标记数据

的帮助下，通过引入半监督学习技术，将训练的边缘模型微调为完整的标签推理模型发动攻击。在真实的医疗图像数据集上，在攻击者仅具有 70 个辅助样本作为背景知识的情况下，被动方标签推断攻击的 F_1 值达到 0.76。

在主动标签推断攻击^[21]中，攻击者通过提高恶意的边缘模型表达能力，从而进一步提高攻击性能，使得 VFL 更加依赖恶意底部模型。攻击者采用一个恶意的优化器，加快边缘模型的收敛速度从而强化边缘模型和标签之间的映射关系。在直接标签推断攻击^[21]中，攻击者通过分析 VFL 过程中传输的梯度符号来推断标签。它仅适用于 VFL 中未采用分裂学习设计的情况，标签推断准确率通常达到 100%。

类似地，Li 等^[13]基于 VFL 训练过程中传输的梯度信息提出一种范数攻击，用以推断主参与

者的标签。这种攻击适用于 VFL 的广告推荐领域。该攻击原理是模型倾向于为正样本赋予更高的置信度,同时为负样本赋予更低的置信度。例如,在现实广告推荐业务中,主参与者的标签信息为客户是否点击广告进行购买(点击转换为正标签,忽略为负标签)。客户通常不会点击不感兴趣的广告,即负标签具有较高的置信度,同时客户点击广告却不一定进行购买,即正标签具有较低的置信度信息。在 Criteo 数据集上,攻击者利用范数攻击推断隐私标签的 AUC(area under curve) 达到 0.92 以上。

综上所述,VFL 在训练过程传输的梯度信息能够揭示主参与者的隐私标签信息。此外,VFL 中从参与者的边缘模型在完成训练后,通常潜在地包含隐私标签的信息。

2.1.2 属性推断攻击

在纵向联邦学习中,参与者的数据样本通常包含属性信息,如收入、性别信息。现实场景下,VFL 中参与者的隐私属性数据存在泄露的风险。根据发动攻击的时间阶段来分,VFL 上的属性推断攻击可以分为:训练阶段的属性推断攻击和推理阶段的属性推断攻击。

(1) 训练阶段的属性推断攻击

Zhang 等^[51]提出在 VFL 训练阶段的属性推断攻击。该攻击发生的原理为:VFL 在训练过程中传输的嵌入表示和参与者的隐私属性存在潜在的映射关系。攻击发生前期,攻击者需要收集一部分样本的隐私属性作为攻击的背景知识。根据攻击者收集隐私属性的方式,属性推断攻击又可以分为静态攻击和自适应攻击。

静态攻击表示攻击者在 VFL 进行单次迭代训练就完成对所有样本隐私属性的收集,自适应攻击是指攻击者在多轮迭代训练过程中收集样本隐私属性。在攻击发生阶段,攻击者首先利用神经网络构建攻击模型,然后利用收集的样本隐私属性和嵌入表示作为攻击模型的训练数据,最后利用训练完成的攻击模型对未知样本属性进行预测,从而完成对隐私属性的窃取。在 Purchase 数据集上,攻击者仅仅具有 5% 的隐私属性数据样本,静态攻击和自适应攻击的 F_1 值均可达到接近 0.98 的攻击性能。

(2) 推理阶段的属性推断攻击

Luo 等^[12]首次提出一种针对 VFL 推理阶段由主参与者发起的属性推断攻击。攻击者的背景知识仅仅包括完成训练的模型和模型的预测结果,以及攻击者自身具有的数据特征。针对逻辑回归 VFL 框架,Luo 等^[12]提出了等式求解攻击,攻击者在给定预测输出的情况下,构造一组方程推断出目标数据的属性值。对于二分类逻辑回归预测任务,攻击者能够精确地求解一个未知数推断隐私属性;而对于多分类逻辑回归预测任务,攻击者根据置信度分数估计隐私属性。针对决策树 VFL 的预测输出仅包括预测类别的特点,Luo 等^[12]提出了一种路径限制攻击,攻击者可以根据预测的类别和特征值预测树模型中的预测路径。

针对采用神经网络作为边缘模型的纵向联邦学习,Luo 等^[12]提出采用生成回归网络在推理阶段对从参与者的数据进行推测。攻击的核心思想是攻击者自身的数据特征和从参与者的数据特征总体具有相关性。在此基础上,攻击者利用生成回归网络生成特征值,使得模型的预测结果和真实的预测结果具有相似的概率分布。具体地,攻击者将生成回归网络输出值和攻击者已知的特征值输入训练完成的模型,通过最小化真实预测和生成样本获得预测的损失训练生成器模型。

综上所述,VFL 中的参与者遭受属性推断攻击的主要原因可以归纳为两方面:一方面是训练过程中嵌入信息和隐私属性具有紧密的映射关系;另一方面是攻击者的数据特征和参与者数据信息总体上具有相关性。

2.1.3 成员推断攻击

在许多现实世界具有隐私敏感的组织中,成员 ID 信息高度敏感,不能和其他参与者进行共享。例如,医院中的患者身份和银行中顾客的身份信息均属于隐私信息。最早对深度模型中的成员信息的攻击是由 Shokri 等^[52]提出的成员推断攻击(MIA, membership inference attack)。攻击者利用模型预测结果来推断模型训练数据中是否包含某个训练样本,这类攻击方法同样给机器学习的安全和隐私带来了严重的威胁。

VFL 准备训练数据阶段,需要确定各方共享的公共数据实体。各方通过运行 PSI 协议,以对

齐交叉样本。样本 ID 需要是某些可用于跨组织标识实体的通用标识信息。常见的选择是一些个人身份信息（如电话和电子邮件）。在各方获得交叉口的样本 ID 后，可以在交叉点上联合训练模型。进一步地，Liu 等^[53]提出改进之后的 PSI 协议，以在非对称 VFL 实现非对称 ID 对齐。然而，所提出的协议仍然将交叉点成员身份暴露给一方，无法保护 VFL 中的所有参与者。

2.1.4 数据重构攻击

数据重构攻击是攻击者推测参与者的完整训练数据，达到窃取数据隐私的目的。在图像识别领域中，Zhu 等^[54]表明通过收集梯度信息可以完整恢复参与者的训练图像。Geiping 等^[55]通过利用幅度不变的损失以及基于对抗性攻击的优化策略，表明可以根据梯度信息重建高分辨率的图像。类似地，Yin 等^[56]提出梯度反转攻击，其在优化过程中将随机噪声拟合为参与训练方的原始图像，同时正则化图像保真度。在深度学习的模型中，梯度信息往往通过直接交换明文来传输信息，通过梯度泄露原始数据较为普遍。然而 VFL 中的梯度信息往往会通过加密机制进行保护，即便如此，一些研究者^[14,57-58]针对 VFL 数据加密场景提出了不同的数据重构攻击方法。

（1）基于逆向乘法攻击

Weng 等^[14]针对 VFL 梯度加密的场景，提出了逆向乘法攻击。在逆向乘法攻击中，攻击者试图对矩阵乘积的每个乘法项进行逆向工程，推断目标参与者的原始训练数据。攻击者利用了两种信息：① 参与者 A 从参与者 B 输出的中间结果；② 协调者返回到参与者 B 的部分梯度信息作为攻击的背景知识。将参与者 B 的特征视为未知参数，执行反向乘法攻击等于求解一组线性函数。逆向乘法攻击的成功与否取决于方程系数矩阵的大小。如果系数矩阵处于满秩情况，则攻击者可以成功地计算出私有数据，否则推断准确率保持

较低水平。

（2）基于逆向加法攻击

Weng 等^[14]针对 XGBoost 纵向联邦学习提出一种逆向加法攻击，目的在于从求和中反演每个加法项。攻击者将幻数编码成一阶梯度和二阶梯度的密文。攻击者存储接收的第一个和第二个梯度，然后利用编码的幻数从梯度和中反演出所有加法项。

（3）基于生成对抗网络的攻击

上述的重构攻击主要针对边缘模型为逻辑回归以及决策树的 VFL，无法适用于边缘模型为神经网络的场景。对于 VFL 中参与者上传的中间嵌入信息，Mahendran 等^[57]提出一个通用的特征反转表示方法，揭示出可以从中间嵌入表示逆向恢复出原始训练图像。Pasquini 等^[58]提出了一种特征空间劫持攻击。VFL 遭受特征空间劫持攻击的主要原因为协调者可以控制某个参与者的训练过程，这使得其在训练期间能够同时训练生成器和鉴别器，并使用一个新的目标替代原始学习任务达到恢复参与者原始数据的目的。

2.2 VFL 中的后门攻击

攻击者通过精心设计训练样本，在训练过程中有针对性地引导模型预测输出特定的结果，这种类型的攻击被称为“后门攻击”。例如，在人脸识别 VFL 系统中，一旦攻击者对系统成功注入后门，攻击者仅仅佩戴一个特殊的面具就可以绕过人脸识别系统的检测。

根据攻击者的角色，在 VFL 中的后门攻击可以分为主动方后门攻击与被动方后门攻击。纵向联邦学习中不同角色发动后门攻击背景知识和操纵能力如表 2 所示。考虑到后门攻击对 VFL 的安全性具有严重的安全威胁，但同时相关研究较为欠缺，本文列举和分析一些能够适用于纵向联邦学习的后门攻击方法促进纵向联邦学习的安全性研究。

表 2 纵向联邦学习中不同角色发动后门攻击背景知识和操纵能力

Table 2 Background knowledge and manipulation ability of different roles launching backdoor attacks in vertical federated learning

攻击发起者	背景知识					操纵能力				
	中间嵌入	更新梯度	标签信息	边缘模型	本地属性	中间嵌入	更新梯度	标签信息	边缘模型	本地属性
主参与者	√	√	√	√	√	√	√	√	√	√
从参与者	√	√	×	√	√	×	×	×	√	√

2.2.1 主参与者后门攻击

Gu 等^[59]提出了一种修改标签的后门攻击 (BadNets)。该网络的训练策略为使用随机挑选的干净样本标记触发器, 以形成触发器样本, 同时将这些触发器样本的标签更改为目标类。推理期间, 正常的测试样本经过后门网络输出正确的预测结果, 添加触发器的测试样本经过后门网络则会输出攻击者设定的预测结果。例如, 通过对街道标志分类器实施 BadNets 后门攻击, 当在停车标志上添加特殊贴纸时, 该分类器将停车标志识别为限速。由于 VFL 中主动方具有的操纵能力和背景知识符合 BadNets 设定的攻击者, 因此 VFL 的主动方存在发动 BadNets 攻击的可能。

上述攻击者在训练样本中添加的触发器往往是固定的像素值, 一些后门攻击的研究注重提高触发器的隐蔽性。例如, Liao 等^[60]提出一种生成触发器的方法, 该方法会约束触发器对原始图片造成的影响, 从而提高后门注入的隐蔽性。在联邦学习场景中, Xie 等^[61]提出一种分散式后门攻击 (DBA, distributed backdoor attack)。DBA 将全局触发器分解为多个局部的触发器, 每个恶意客户端都使用自己的本地触发器“毒害”其边缘模型。由于将全局触发器进行分散, DBA 的攻击方法表现出更强的隐蔽性。

综上所述, 由于主动方在 VFL 中的背景知识和操纵能力全面, 潜在的后门攻击给模型带来的安全威胁更加严重。此外, 由于分散式后门攻击自身的隐蔽性特点, 评估分散式后门攻击对 VFL 的安全风险具有重要意义。

2.2.2 从参与者后门攻击

由于被动方在 VFL 中不具有修改标签的能力, 现有对后门攻击的研究中, 干净标签的潜在后门攻击成为 VFL 中被动方攻击的方式。针对攻击者发动攻击所操纵的数据类型, 攻击可以分为对梯度的操纵和对数据的操纵。

(1) 对梯度的操纵

由于 VFL 中主动方和被动方交换梯度信息, 攻击者可以试图操纵梯度来注入后门。Liu 等^[15]提出了一种 VFL 中梯度替代的后门攻击, 攻击者精心设置中毒样本的中毒梯度, 在攻击过程中主动发送随机向量来发动后门攻击。该攻击假设攻

击者具有一定干净样本的背景知识。

(2) 对数据的操纵

针对 VFL 中尚缺乏操纵数据的后门攻击工作, 深度学习通过对数据操纵发动的后门攻击方法可能成为 VFL 中被动方攻击的手段, 这是因为两者具有相同的能力和相似的背景知识。通过操纵数据的干净标签后门攻击由 Shafahi 等^[62]提出, 这是一种通过特征冲突的后门制作方法。这种攻击生成的后门样本虽然在像素级的特征表现和标签的类别对应, 然而其潜在特征空间接近于目标类。这种攻击要求攻击者对被攻击模型的结构和参数具有完备背景知识。为了减少攻击者对模型的背景知识限制, Turner 等^[63]提出了对抗性扰动的方法发动干净标签后门攻击。通过利用对抗攻击的可迁移性, 在训练模型之前使用对抗扰动攻击, 诱导模型对后门产生作用。Weng 等^[64]基于模型对抗鲁棒性和后门鲁棒性的相互影响, 提出了使用对抗性训练的方法为模型注入后门。

由于 VFL 通常存在多个不可信的从参与者, 其遭受从参与者发动后门攻击的威胁更为普遍。从参与者对梯度操纵或者数据操纵这两个角度发动后门攻击, 对 VFL 造成实际的安全威胁。

2.3 VFL 中的对抗攻击

针对 VFL 在推理阶段进行的攻击通常被称为对抗攻击。相较于横向联邦学习, VFL 更容易遭受这种对抗攻击的威胁。其原因在于横向联邦学习中, 参与者在推理期间均独立使用协调者下发的模型进行推理, 推理结果独立于其他参与者, 因此参与者在推理期间操纵模型输出结果的攻击威胁较小。然而, VFL 模型在推理阶段需要各个参与者的模型组合成完整的模型进行推理预测, 参与者均获得相同的预测结果。在此期间, 纵向联邦学习中容易存在不可信来源的参与者在推理期间发动对抗攻击, 使得纵向联邦学习模型产生错误的判断。因此, 本文介绍和分析 VFL 在对抗攻击方面的研究进展。

在纵向联邦学习的模型结构和参数都已知的情况下, 攻击者可以在 VFL 推理阶段通过操纵梯度或者嵌入表示特征生成对抗样本。这种基于梯度生成对抗性样例是一种常见的对抗攻击方法。Goodfellow 等^[65]提出快速梯度符号法 (FGSM,

fast gradient sign method), 利用高维空间中的线性行为产生对抗样本。现有研究针对 VFL 中的图像和网络领域分别提出了不同的对抗攻击方法。在图像领域, Liu 等^[66]提出了一种对抗嵌入表示特征生成策略, 该策略的核心在于攻击者基于 FGSM 生成恶意的嵌入表示特征。服务器聚合恶意嵌入表示以及良性嵌入表示后输出错误预测结果。在网络领域, Chen 等^[67]基于 VFL 中协调者反馈的置信度信息提出了一种对抗嵌入表示特征的攻击框架。值得一提的是, 在这种攻击框架中, 攻击者采取梯度攻击的方法可以利用少量的攻击样本达到对抗攻击效果。

2.4 VFL 中的风险独特性

上文对 VFL 面临的隐私和安全风险进行了归纳和分析。相较于横向联邦学习, VFL 组成结构和运行机制上的差异引入了独特的安全风险和挑战。表 3 对比了纵向联邦学习和横向联邦学习差异及其风险独特性。

3 纵向联邦学习防御

本节整理并介绍了现有研究中 VFL 面临推断攻击、后门攻击以及对抗攻击的常见防御策略。由于攻击方法层出不穷, 针对 VFL 的防御工作变得更为重要也更加困难, 目前围绕 VFL 的防御研究较少, 仍需进一步地探索。

3.1 针对 VFL 推断攻击的防御

3.1.1 基于加密的防御

(1) 同态加密

同态加密是一种 VFL 中保护参与方隐私的常见手段, 在加密数据上对明文执行任何操作, 而无需解密。Ronald 等^[68]提出同态加密的概念,

与一般加密算法相比, 同态加密不仅可以实现基本的加密操作, 而且实现了密文之间的各种计算功能。将同态加密技术应用于实践的是 Gentry^[69]提出的全同态加密方案, 其通过数学方法将模型的预测函数表示为一个低阶多项式来解决同态加密操作引起的过度噪声问题。针对纵向联邦学习, Ou 等^[70]提出一种纵向联邦期望最大算法, 通过引入一种贝叶斯聚合更新机制并结合后验采样来保证数据的隐私性。

(2) 秘密共享

秘密共享是一种无须可信第三方协助进行密文计算的加密技术。秘密共享指共享的秘密在一个用户群体中进行合理分配, 以达到由所有成员共同掌管秘密的目的。夏家骏等^[71]基于秘密共享技术提出了纵向联邦学习框架, 并讨论缩小通信代价的策略。

3.1.2 基于扰动的防御

(1) 差分隐私

差分隐私在 2006 年由 Dwork^[72]提出, 最初应用于数据库的查询操作、数据挖掘、数据统计等。Abadi 等^[73]在 2016 年将差分隐私扩展到深度学习上, 使得用户无法从获取的输出数据中分辨数据集来源, 从而达到保护隐私的目的。随着纵向联邦学习的不断发展, 有学者提出了纵向联邦学习的差分隐私框架。Wang 等^[74]分析了 VFL 的混合差分隐私框架, 以证明从垂直分区的数据中联合学习广义线性模型的可能性, 且成本可以忽略不计。

(2) 梯度压缩

梯度压缩防御策略旨在共享较少的梯度, 起到提高通信效率和保护隐私的作用, 其关键思想

表 3 纵向联邦学习和横向联邦学习差异及其风险独特性
Table 3 Differences between vertical and horizontal federal learning and their risk uniqueness

类型	横向联邦学习	纵向联邦学习	风险独特性
数据提供	每个参与者都有各自的特征和标签	主动方具有标签, 被动方具有特征	易遭受标签推断攻击
数据处理	无须对齐操作	需要进行数据对齐	数据成员直接在对齐过程中被泄露
模型结构	参与者模型结构完整	参与者模型结构不完整	攻击者对全局模型的操纵能力相对较弱
传输内容	上传下发模型参数信息或梯度信息	上传嵌入层信息, 下发梯度信息	嵌入层成为隐私泄露或者攻击的目标
聚合机制	平均聚合等鲁棒聚合方式	直接拼接或相加	缺少对后门攻击的聚合防御
推理方式	参与者对全局模型单独推理, 获得不同推理结果	参与者需要联合进行推理, 输出同一个推理结果	易在推理阶段引入对抗攻击

是只共享一部分绝对值最大的梯度值参与模型的训练。现有研究^[32-33]基于深度梯度压缩方法构建保护隐私的 VFL 框架,主动方采用只传送幅度最大的梯度,同时,为了防止梯度信息丢失,在局部累积剩余的梯度,随着梯度增加到设定阈值,累计的梯度进行传输。

3.1.3 基于对抗训练的防御

对抗训练^[75]技术最早用于防御深度学习中的对抗攻击,通过将带有正确标签的对抗样本和原始训练样本混合后训练模型,以提高模型的鲁棒性。对抗训练技术同时可用于保护深度学习模型隐私。例如,Elazar 等^[76]在文本数据中使用对抗训练的技术去除数据中潜在的统计信息,保证信息发布的隐私性。Liao 等^[77]将对抗训练技术应用于图神经网络,去除图神经网络获得嵌入特征向量隐含的属性信息。对抗训练往往被构建为一个最大-最小的优化问题。训练过程中同时考虑防御者和模拟攻击者的损失,最终提高防御者模型的鲁棒性。

(1) 针对属性推断攻击防御

Zhang 等^[78]提出了一种基础防御策略。在每次训练迭代中,参与者首先针对隐私目标优化其边缘模型。参与者训练模拟解码器以获得最小的累计攻击损失,然后参与者通过最小化隐私损失来更新其边缘模型。这种防御建立在本地参与者的防御视角,因此无法兼顾主任务的性能,此外,这使得 VFL 收敛更为困难。Zhang 等^[78]提出一种基于前向-反向分裂算法的 VFL。通过将优化过程分为两个目标,即 VFL 中前向传播阶段考虑隐私保护和反向梯度下降考虑主任务性能。

(2) 针对重构攻击防御

Sun 等^[79]提出了一种防御纵向联邦学习重构攻击框架。该框架模拟了攻击者和防御者之间的博弈,攻击者利用中间嵌入信息重建原始输入,防御者防止输入数据泄露。其中包含 3 个组件:模拟重构、噪声正则化和距离相关性。

模拟重构组件旨在最大程度地提高攻击者的重构错误。噪声正则化组件的目的是通过误导攻击者向随机方向进行优化来减少嵌入的输入信息。距离相关性组件用于降低原始输入和分割层嵌入之间的相关性。这些组件旨在使参与者的边

缘模型抵御提取原始输入敏感信息的重构攻击且更具鲁棒性。

3.1.4 小结

VFL 中推断攻击的核心思想是参与者通过对抗训练的手段,构建一个自动滤除隐私信息的边缘模型。通过模拟攻击者和防御者之间的对抗博弈,提升 VFL 抵御推断攻击的鲁棒性。

3.2 针对 VFL 后门攻击及对抗攻击的防御

3.2.1 基于恢复特征子空间的防御

防御者通过恢复干净的特征子空间,使得被注入的后门或者对抗样本进行移除,从而达到防御后门攻击以及对抗攻击的目的。在低秩特征子空间、少量攻击样本的情况下,Liu 等^[66]提出一种鲁棒自动编码器来恢复特征子空间,并证明当 VFL 中边缘特征子空间是线性时,它可以准确地恢复毁坏数据,从而使模型防御后门攻击或对抗攻击。此外,在模型训练阶段,Liu 等^[66]提出一种鲁棒分解方法防御后门攻击,将 VFL 中损坏的特征向量分解为两部分:一部分位于特征子空间;另一部分位于块稀疏结构。

3.2.2 基于扰动的防御

(1) 梯度稀疏化

针对梯度替换的后门攻击,梯度稀疏化方法通过将 VFL 训练过程中的梯度信息较小的值裁剪为 0,达到防御后门攻击的效果。Liu 等^[15]针对梯度替换后门攻击采用了梯度稀疏化算法,在训练期间的每次迭代中采用下降率来更新阈值的梯度,具有较好的后门防御能力。此外,Chen 等^[67]采用了对参与方上传的嵌入表示进行 Top- k 稀疏化的方法,防御 VFL 中的对抗攻击。当稀疏化程度较高时,防御对抗攻击具有一定防御效果,同时对主任务的影响较小。

(2) 差分隐私

差分隐私技术在 VFL 训练过程中通过平滑参与者交换的梯度信息达到防御后门的目的。Liu 等^[15]针对梯度替换后门攻击同样采用了差分隐私方法作为后门的防御手段。尽管差分隐私对后门攻击具有防御能力,但对 VFL 预测能力具有很大的影响,难以平衡后门防御和主任务性能。同时,Chen 等^[67]评估了差分隐私技术防御 VFL 面临的对抗攻击。差分隐私技术采用大的随机扰动能够防御对抗嵌入特征。

4 VFL 数据集实验平台

近年来,涌现出的诸多 VFL 开源平台和框架大大加速了纵向联邦学习的落地进程^[80],同时方便了比较学者的研究成果。本节列举了已有 VFL 研究中来自不同领域的数据集,整理并介绍了现有的 VFL 平台和框架,便于研究者针对实际需求快速选择工具和平台进行后续深入研究。

4.1 常用数据集

现有纵向联邦学习中的常用数据集如表 4 所示,共分为 8 个类别。为满足理论研究中对 VFL 的数据纵向分布场景,在文本模态的数据集通常按照属性来进行切分;对于图像类的数据集通常

直接切分像素点;对于网络类的数据集通常划分连边和特征实现纵向联邦学习。

4.2 平台和框架

本节整理现有纵向联邦学习的平台和框架,如表 5 所示,便于研究者基于现有框架和平台快速开发产品解决实际业务需求。

5 未来研究方向

本文介绍了纵向联邦学习的应用背景,分析归纳了 VFL 的常见算法,并且介绍了 VFL 面临的隐私和安全问题以及对应的防御和加固方法。本节针对 VFL 方法以及其攻防方法进行分析,探讨其在未来的研究方向。

表 4 纵向联邦学习中的常用数据集
Table 4 Common datasets in longitudinal federated learning

适用场景	数据集	样本数量	特征数量	标签类别	数据模态	任务类型	VFL 引用文献
生物医疗	MIMIC-III ^[81]	21 139	714	2	文本	分类任务	[27],[28],[29]
	Breast ^[82]	569	31	2	文本	分类任务	[16],[21],[75],[100]
金融交易	Purchase100 ^[51]	197 324	600	2	文本	分类任务	[51]
	Insurance Claim ^[33]	188 318	116	2	文本	分类任务	[34]
	Bank Market ^[83]	4 521	17	2	文本	分类任务	[20],[101]
	Credit Card Clients ^[84]	284 807	28	2	文本	分类任务	[17],[27],[31],[34],[49],[75],[100],[102]
	Adult ^[85]	48 842	12	2	文本	分类任务	[51],[75]
广告推荐	Criteo ^[86]	4 500 000	40	2	文本	分类任务	[21],[103]
	Avazu	4 000 000	17	2	文本	分类任务	[103]
文献引用	Cora ^[87]	2 708	1 433	7	网络	分类任务	[22],[23],[104]
	Citeseer ^[87]	3 312	3 703	6	网络	分类任务	[22],[23],[104]
	Pubmed ^[88]	19 717	500	3	网络	分类任务	[22],[23],[104]
物体识别	MNIST ^[89]	70 000	784	10	图像	分类任务	[16],[18],[27],[28],[29],[31],[45],[105],[106]
	Fashion-MNIST ^[90]	70 000	784	10	图像	分类任务	[105],[45],[28],[58]
	Cifar10 ^[91]	60 000	3 072	10	图像	分类任务	[21],[27],[28],[105]
	Cifar100 ^[91]	60 000	3 072	100	图像	分类任务	[21]
	COCO-QA ^[92]	78 736	—	—	图像、文本	分类任务	[51]
	NUS-WIDE ^[93]	269 648	5 018	81	图像、文本	分类任务	[10],[27]
	CelebA ^[94]	202 599	116 412	2	图像	分类任务	[58]
新闻媒体	News Popularity ^[95]	93 239	11	—	时序文本	回归任务	[16],[24],[31],[33]
	Yahoo Answers ^[96]	146 000	—	10	文本	分类任务	[21]
物理	Superconductivity ^[97]	20 000	81	—	文本	回归任务	[29]
	Drive Diagnosis ^[98]	58 509	48	11	文本	分类任务	[101]
家电能源	Appliances Energy ^[99]	19 735	29	—	文本	回归任务	[20],[101]

表 5 纵向联邦学习的平台和框架
Table 5 Vertical federal learning platform and framework

框架或平台	开发者	加密手段	是否开源	使用类型	框架	特点	应用场景
FATE	微众银行	同态加密	是	研究、商用	Tensorflow、PyTorch	支持联邦迁移学习技术	车险定价、信贷风控等
PaddleFL	百度	安全多方计算	是	研究、商用	PaddlePaddle	基于全栈开源软件部署	计算机视觉、推荐算法等
Pysyft	OpenMined	同态加密、秘密共享	是	研究	TensorFlow、PyTorch	多种隐私控制策略	金融风控、广告推荐等
iBond	同盾科技	同态加密、秘密分享	否	商用	(未知)	引入知识联邦概念	自然语言处理、金融风控
PowerFL	腾讯	多方安全计算	否	商用	Angel	考虑数据异质性	金融风控、广告推荐

5.1 面向 VFL 样本对齐方法的展望

现有研究针对样本对齐机制设计了大量的 VFL 算法，这些算法采用半监督学习技术作为生成模型补齐缺失特征，取得优异的性能。然而，该技术本质上属于数据增强方法，这类方法往往只有当生成模型的训练数据和缺失数据具有相似的特征空间时，才能具有较好的缺失数据补全能力。未来面向 VFL 中样本对齐机制，从生成模型的迁移性角度出发，基于迁移学习实现生成模型的训练，提高生成模型生成缺失样本的数据空间迁移性，减少模型对样本特征空间的依赖；从生成模型的鲁棒性角度出发，构建基于模型集成的生成模型动态建模方法，通过多种模型的融合提高生成模型面对不同输入的鲁棒性，从而有利于生成模型补全缺失值；此外，从生成模型的可解释性角度出发，研究不同缺失特征的嵌入表示变化规律，进一步提高生成模型补全缺失数据的质量。

5.2 面向 VFL 攻击方法的展望

已有针对 VFL 设计的隐私推断攻击，攻击者可以根据梯度信息恢复参与方的隐私数据，但此类攻击方法仅仅适用于使用小批量样本训练的 VFL 模型，不适用于大批量样本的训练。针对 VFL 的隐私推断攻击，未来可能面向大批量样本数据训练 VFL，研究大批量训练数据的高效攻击方法。已有针对 VFL 设计的后门攻击，从参与者可以通过操纵梯度信息注入后门，但此类后门攻击方法使得模型对正常样本的识别准确率降低，不能同时兼顾后门攻击效果和模型对干净样本的分类性能。针对 VFL 的后门攻击，未来可能针对从参与者基于生成式对抗网络的后门样本生成方法，在训练过程中渐进式毒化模型。此外，已有

针对 VFL 设计的对抗攻击，通过生成对抗嵌入特征的方式使得模型输出特定预测，但该方法在多个参与者的 VFL 场景中难以成功发动攻击。从模型的可解释性方面出发，研究不同的嵌入表示生成方法对 VFL 顶部模型神经元激活值的影响，寻找敏感的神经元生成对抗攻击策略。

5.3 面向 VFL 防御及鲁棒性增强的展望

已有的 VFL 主要通过加密、对抗训练以及注入扰动这 3 类防御方法来实现模型的鲁棒性增强和安全加固。目前已有的防御方法使用比较广泛的是加密，如常见的同态加密方法，但此类加密方法的计算时间代价和内存空间需求太高，不能满足 VFL 实时加密训练的需求。此外，对模型的全加密引入大量的通信代价，增加了 VFL 实时通信的压力。未来的防御方法可以从采用半加密和扰动注入结合的方式，设计抵御不同类型攻击的通用防御方法，提高防御的泛化能力。

与此同时，防御可以从模型鲁棒优化角度出发，基于可验证鲁棒理论设计 VFL 可验证鲁棒防御方法，以实现多种攻击的有效防御。此外，鉴于不同类型攻击方法在模型层面表现出的异常表征，从模型层面出发增强 VFL 模型鲁棒性也是未来的研究方向。

6 结束语

纵向联邦学习作为解决数据孤岛的有效手段，近年来得到广泛的研究和发展，其适应范围从传统的机器学习模型（如逻辑回归和决策树）发展到支持图神经网络和卷积网络等复杂模型。与此同时，VFL 的应用领域从金融和医疗领域扩展到安防和教育等领域。纵向联邦学习在理论和

实践上经历一个从无到有、从有到优化的过程, VFL 的体系也在不断地发展和完善。众多 VFL 的框架和平台发展加速了其在各个领域的快速发展。

在 VFL 得到落地的同时, 其算法自身的隐私性和安全性也得到众多学者关注。本文在介绍现有攻防技术的同时, 整理和分析了 VFL 相较于横向联邦学习的独特安全风险, 并且有针对性地给出一些未来研究方向的建议。

本文概括和提炼了目前 VFL 方法, 并对其安全和隐私问题进行了全面的归纳总结, 分析了 VFL 未来发展方向。VFL 作为一种解决数据孤岛的有效技术手段, 尽管已经发展了很多算法支持其在各个领域的应用, 其安全和隐私并未得到充分评估。期待广大的研究人员对其安全和隐私进行全面评估, 并设计研究高效的、安全的和鲁棒的 VFL 算法。

参考文献:

- [1] 焦李成, 杨淑媛, 刘芳, 等. 神经网络七十年: 回顾与展望[J]. 计算机学报, 2016, 39(8): 1697-1716.
- JIAO L C, YANG S Y, LIU F, et al. Seventy years beyond neural networks: retrospect and prospect[J]. Chinese Journal of Computers, 2016, 39(8): 1697-1716.
- [2] 中华人民共和国数据安全法[N]. 中华人民共和国全国人民代表大会常务委员会公报, 2021(5): 951-956.
- Data Security Law of the People's Republic of China[N]. Communiqué of the Standing Committee of the National People's Congress of the People's Republic of China, 2021(5): 951-956.
- [3] GODDARD M. The EU general data protection regulation (GDPR): European regulation that has a global impact[J]. International Journal of Market Research, 2017, 59(6): 703-705.
- [4] HORD A. Federated learning for mobile keyboard prediction[J]. arXiv preprint arXiv: 1811.03604, 2018.
- [5] 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.
- [6] 李少波, 杨磊, 李传江, 等. 联邦学习概述: 技术、应用及未来[J]. 计算机集成制造系统, 2021, 1(10): 1-29.
- LI S B, YANG L, LI C J, et al. Overview of federated learning: technology, applications and future [J]. Computer Integrated Manufacturing Systems, 2021, 1(10): 1-29.
- [7] 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): 1-16.
- ZHOU C X, SUN Y, WANG D, et al. Survey of federated learning research[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 1-16.
- [8] 卫新乐, 张志勇, 宋斌, 等. 基于纵向联邦学习的社交网络跨平台恶意用户检测方法[J]. 小型微型计算机系统, 2021, (10): 1-9.
- WEI X L, ZHANG Z Y, SONG B, et al. Social networks cross-platform malicious user detection method based on vertical federated learning[J]. Journal of Chinese Mini-Micro Computer Systems, 2021, (10): 1-9.
- [9] 李鸣. 基于纵向联邦学习的推荐系统技术研究[D]. 杭州: 浙江大学, 2021.
- LI M. Research on recommendation system technology based on vertical federated learning[D]. Hangzhou: Zhejiang University, 2021.
- [10] KANG Y, LIU Y, CHEN T J. Fedmv: semi-supervised vertical federated learning with multiview training[J]. arXiv preprint arXiv: 2008.10838, 2020.
- [11] WU Z, LI Q, HE B. Exploiting record similarity for practical vertical federated learning[J]. arXiv preprint arXiv: 2106.06312, 2021.
- [12] LUO X, WU Y, XIAO X, et al. Feature inference attack on model predictions in vertical federated learning[C]//Proceedings of IEEE 37th International Conference on Data Engineering (ICDE). 2021: 181-192.
- [13] LI O, SUN J, YANG X, et al. Label leakage and protection in two-party split learning[J]. arXiv preprint arXiv: 2102.08504, 2021.
- [14] WENG H, ZHANG J, XUE F, et al. Privacy leakage of real-world vertical federated learning[J]. arXiv preprint arXiv: 2011.09290, 2020.
- [15] LIU Y, YI Z, CHEN T. Backdoor attacks and defenses in feature-partitioned collaborative learning[J]. arXiv preprint arXiv: 2007.03608, 2020.
- [16] HARDY S, HENECKA W, IVEY-LAW H, et al. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption[J]. arXiv preprint arXiv: 1711.10677, 2017.
- [17] YANG K, FAN T, CHEN T, et al. A quasi-Newton method based vertical federated learning framework for logistic regression[J]. arXiv preprint arXiv: 1912.00513, 2019.
- [18] YANG S, REN B, ZHOU X, et al. Multi-VFL for vertical federated learning without third-party coordinator[J]. arXiv preprint arXiv: 1911.09824, 2019.
- [19] CHENG K, FAN T, JIN Y, et al. SecureBoost: a lossless federated learning framework[J]. IEEE Intelligent Systems, 2021, 36(6): 87-98.
- [20] WU Y, CAI S, XIAO X, et al. Privacy preserving vertical federated learning for tree-based models[C]//Proceedings of the VLDB Endowment, 2020. 2090-2103.
- [21] VEPAKOMMA P, GUPTA O, SWEDISH T, et al. Split learning for health: distributed deep learning without sharing raw patient data[J]. arXiv preprint arXiv: 1812.00564, 2018.
- [22] FU C, ZHANG X, JI S, et al. Label inference attacks against vertical federated learning[C]//Proceedings of USENIX Security Symposium. 2022: 1397-1414.
- [23] NI X, XU X, LYU L, et al. A vertical federated learning framework for graph convolutional network[J]. arXiv preprint arXiv: 2106.11593, 2021.
- [24] ZHOU J, CHEN C, ZHENG L, et al. Vertically federated graph neural network for privacy-preserving node classification[J]. arXiv preprint arXiv: 2005.11903, 2020.
- [25] GU B, XU A, HUO Z, et al. Privacy-preserving asynchronous fed-

- rated learning algorithms for multi-party vertically collaborative learning[J]. arXiv preprint arXiv: 2008.06233, 2020.
- [26] MCMAHAN H B, MOORE E, RAMAGE D, et al. Federated learning of deep networks using model averaging[J]. arXiv preprint arXiv: 1602.05629, 2016.
- [27] LIU Y, KANG Y, ZHANG X, et al. A communication efficient collaborative learning framework for distributed features[J]. arXiv preprint arXiv: 1912.11187, 2019.
- [28] CHEN T, JIN X, SUN Y, et al. VAFL: a method of vertical asynchronous federated learning[J]. arXiv preprint arXiv: 2007.06081, 2020.
- [29] DAS A, PATTERSON S. Multi-tier federated learning for vertically partitioned data[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021: 3100-3104.
- [30] WRIGHT S, NOCEDAL J. Numerical optimization[J]. Springer Science, 1999, 35(67-68): 7.
- [31] ZHANG Q, GU B, DENG C, et al. AsySQN: faster vertical federated learning algorithms with better computation resource utilization[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 3917-3927.
- [32] LIN Y J. Deep gradient compression: reducing the communication bandwidth for distributed training[J]. arXiv preprint arXiv: 1712.01887, 2017.
- [33] YANG K, SONG Z, ZHANG Y, et al. Model optimization method based on vertical federated learning[C]//Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS). 2021: 1-5.
- [34] LI M, CHEN Y, WANG Y, et al. Efficient asynchronous vertical federated learning via gradient prediction and double-end sparse compression[C]//Proceedings of 16th International Conference on Control, Automation, Robotics and Vision (ICARCV). 2020: 291-296.
- [35] LIU D. Accelerating intra-party communication in vertical federated learning with RDMA[C]//Proceedings of the 1st Workshop on Distributed Machine Learning. 2020: 14-20.
- [36] 黄翠婷, 张帆, 孙小超, 等. 隐私集合求交技术的理论与金融实践综述[J]. 信息通信技术与政策, 2021, 47(6): 50.
- HUANG C T, ZHANG F, SUN X C, et al. A survey of private set intersection technology and finance practice [J]. Information and Communications Technology and Policy, 2021, 47(6): 50.
- [37] MEADOWS C. A more efficient cryptographic matchmaking protocol for use in the absence of a continuously available third party[C]//Proceedings of 1986 IEEE Symposium on Security and Privacy. 1986: 134-134.
- [38] DE CRISTOFARO E, TSUDIK G. Experimenting with fast private set intersection[C]//Proceedings of International Conference on Trust and Trustworthy Computing. 2012: 55-73.
- [39] RABIN M O. How To Exchange Secrets with Oblivious Transfer[J]. IACR Cryptol, 2005, 2005(187): 1-26.
- [40] ISHAI Y, KILIAN J, NISSIM K, et al. Extending oblivious transfers efficiently[C]//Proceedings of Annual International Cryptology Conference. 2003: 145-161.
- [41] KISSNER L, SONG D. Privacy-preserving set operations[C]//Proceedings of Annual International Cryptology Conference. 2005: 241-257.
- [42] CAMENISCH J, ZAVERUCHA G M. Private intersection of certified sets[C]//Proceedings of International Conference on Financial Cryptography and Data Security. 2009: 108-127.
- [43] ZHANG Q, GU B, DENG C, et al. Secure bilevel asynchronous vertical federated learning with backward updating[J]. arXiv preprint arXiv: 2103.00958, 2021.
- [44] XIA W, LI Y, ZHANG L, et al. A vertical federated learning framework for horizontally partitioned labels[J]. arXiv preprint arXiv: 2106.10056, 2021.
- [45] MUGUNTHAN V, GOYAL P, KAGAL L. Multi-VFL: a vertical federated learning system for multiple data and label owners[J]. arXiv preprint arXiv: 2106.05468, 2021.
- [46] REDDI S, CHARLES Z, ZAHEER M, et al. Adaptive federated optimization[J]. arXiv preprint arXiv: 2003.00295, 2020.
- [47] FENG S, YU H. Multi-participant multi-class vertical federated learning[J]. arXiv preprint arXiv: 2001.11154, 2020.
- [48] HU Y, LIU P, KONG L, et al. Learning privately over distributed features: An ADMM sharing approach[J]. arXiv preprint arXiv: 1907.07735, 2019.
- [49] GU B, DANG Z, LI X, et al. Federated doubly stochastic kernel learning for vertically partitioned data[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 2483-2493.
- [50] CHEN X, LI J, CHAKRABARTI C. Communication and computation reduction for split learning using asynchronous training[J]. arXiv preprint arXiv: 2107.09786, 2021.
- [51] ZHANG S, XIANG L, YU X, et al. Privacy-preserving federated learning on partitioned attributes[J]. arXiv preprint arXiv: 2104.14383, 2021.
- [52] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). 2017: 3-18.
- [53] LIU Y, ZHANG X, WANG L. Asymmetrical vertical federated learning[J]. arXiv preprint arXiv: 2004.07427, 2020.
- [54] ZHU L, LIU Z, HAN S. Deep leakage from gradients[J]. Advances in Neural Information Processing Systems, 2019, 32: 14774-14784.
- [55] GEIPING J, BAUERMEISTER H, DRÖGE H, et al. Inverting gradients--how easy is it to break privacy in federated learning[J]. arXiv preprint arXiv: 2003.14053, 2020.
- [56] YIN H, MALLYA A, VAHDAT A, et al. See through gradients: image batch recovery via grad inversion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16337-16346.
- [57] MAHENDRAN A, VEDALDI A. Understanding deep image representations by inverting them[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5188-5196.
- [58] PASQUINI D, ATENIESE G, BERNASCHI M. Unleashing the tiger: inference attacks on split learning[J]. arXiv preprint arXiv: 2012.02670, 2020.
- [59] GU T, LIU K, DOLAN-GAVITT B, et al. Badnets: evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230-47244.
- [60] LIAO C, ZHONG H, SQUICCIARINI A, et al. Backdoor embedding in convolutional neural network models via invisible perturbation[J]. arXiv preprint arXiv: 1808.10307, 2018.
- [61] XIE C, HUANG K, CHEN P Y, et al. Dba: distributed backdoor attacks against federated learning[C]//Proceedings of International Conference on Learning Representations. 2019: 1-19.
- [62] SHAFABI A, HUANG W R, NAJIBI M, et al. Poison frogs! tar-

- geted clean-label poisoning attacks on neural networks[J]. arXiv preprint arXiv: 1804.00792, 2018.
- [63] TURNER A, TSIPRAS D, MADRY A. Label-consistent backdoor attacks[J]. arXiv preprint arXiv: 1912.02771, 2019.
- [64] WENG C H, LEE Y T, WU S H B. On the trade-off between adversarial and backdoor robustness[J]. *Advances in Neural Information Processing Systems*, 2020, 33.
- [65] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv: 1412.6572, 2014.
- [66] LIU J, XIE C, KENTHAPADI K, et al. Rvfr: Robust vertical federated learning via feature subspace recovery[C]//*Proceedings of NeurIPS Workshop New Frontiers in Federated Learning: Privacy, Fairness, Robustness, Personalization and Data Ownership*. 2021: 1-9.
- [67] CHEN J Y. Graph-fraudster: adversarial attacks on graph neural network based vertical federated learning[J]. arXiv preprint arXiv: 2110.06468, 2021.
- [68] RONALD L R, ADLEMAN L, DERTOUZOS M L. On data banks and privacy homomorphisms[J]. *Foundations of Secure Computation*, 1978, 4(11): 169-180.
- [69] GENTRY C. Fully homomorphic encryption using ideal lattices[C]//*Proceedings of the forty-first annual ACM symposium on Theory of computing*. 2009: 1-10.
- [70] OU W. A homomorphic-encryption-based vertical federated learning scheme for risk management[J]. *Computer Science and Information Systems*, 2020: 22.
- [71] 夏家骏, 鲁颖, 张子扬, 等. 基于秘密共享与同态加密的纵向联邦学习方案研究[J]. *信息通信技术与政策*, 2021, 47(6): 19-26.
- XIA J J, LU Y, ZHANG Z Y, et al. Research on vertical federated learning based on secret sharing and homomorphic encryption[J]. *Information and Communications Technology and Policy*, 2021, 47(6): 19-26.
- [72] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//*Proceedings of Theory of Cryptography Conference*. 2006: 265-284.
- [73] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//*Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 2016: 308-318.
- [74] WANG C, LIANG J, HUANG M, et al. Hybrid differentially private federated learning on vertically partitioned data[J]. arXiv preprint arXiv: 2009.02763, 2020.
- [75] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. *Computer Science*, 2013.
- [76] ELAZAR Y, GOLDBERG Y. Adversarial removal of demographic attributes from text data[J]. arXiv preprint arXiv: 1808.06640, 2018.
- [77] LIAO P, ZHAO H, XU K, et al. Information obfuscation of graph neural networks[C]//*Proceedings of International Conference on Machine Learning*. 2021: 6600-6610.
- [78] ZHANG S, XIANG L, YU X, et al. Privacy-preserving federated learning on partitioned attributes. arXiv preprint arXiv: 2104.14383, 2021.
- [79] SUN J, YAO Y, GAO W, et al. Defending against reconstruction attack in vertical federated learning[J]. arXiv preprint arXiv: 2107.09898, 2021.
- [80] LIU Y, KANG Y, ZOU T, et al. Vertical Federated Learning. arXiv preprint arXiv: 2211.12814, 2022.
- [81] JOHNSON A E W, POLLARD T J, SHEN L, et al. MIMIC-III, a freely accessible critical care database[J]. *Scientific data*, 2016, 3(1): 1-9.
- [82] STREET W N, WOLBERG W H, MANGARIAN O L. Nuclear feature extraction for breast tumor diagnosis[J]. In *Biomedical image processing and biomedical visualization*, 1993, 1905: 861-870.
- [83] MORO S, CORTEZ P, RITA P. A data-driven approach to predict the success of bank telemarketing[J]. *Decision Support Systems*, 2014, 62: 22-31.
- [84] YE H I, LIEN C. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients[J]. *Expert systems with applications* 2009, 36(2): 2473-2480.
- [85] KOHAVI R. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid[J]. *Kdd*, 1996, 96: 202-207.
- [86] ZHAO P, XIAO K, ZHANG Y, et al. AMEIR: Automatic Behavior Modeling, Interaction Exploration and MLP Investigation in the Recommender System[C]//*Proceedings of International Joint Conferences on Artificial Intelligence*. 2021: 2104-2110.
- [87] MCCALLUM A K, NIGAM K, RENNIE J, et al. Automating the construction of internet portals with machine learning[J]. *Information Retrieval*, 2000, 3: 127-163.
- [88] SEN P, NAMATA G, BILGIC M, et al. Collective classification in network data[J]. *AI magazine*, 2008, 29(3): 93.
- [89] DENG L. The MNIST database of handwritten digit images for machine learning research[J]. *IEEE signal processing magazine*, 2021, 29(6): 141-142.
- [90] XIAO H, RASUL K, VOLLGRAF R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms[J]. arXiv preprint arXiv: 1708.07747, 2017.
- [91] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[R]. Technical report, Citeseer, 2009.
- [92] REN M, KIROS R, ZEMEL R. Exploring models and data for image question answering[J]. *Advances in neural information processing systems*, 2015, 28: 2953-2961.
- [93] CHUA T S, TANG J, HONG R, et al. NUS-WIDE: a real-world web image database from national university of Singapore[C]//*Proceedings of ACM International Conference on Image and Video Retrieval*. 2009: 1-9.
- [94] LIU Z, LUO P, WANG X, et al. Deep learning face attributes in the wild[C]//*Proceedings of IEEE International Conference on Computer Vision*. 2015: 3730-3738.
- [95] FERNANDES K, VINAGRE P, CORTEZ P. A proactive intelligent decision support system for predicting the popularity of online news[C]//*Proceedings of Portuguese Conference on Artificial Intelligence*. 2015: 535-546.

- [96] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification[C]//Proceedings of The Advances in neural information processing systems. 2015: 1-9.
- [97] HAMIDIEH K. A data-driven statistical model for predicting the critical temperature of a superconductor[J]. Computational Materials Science, 2018, 154: 346-354.
- [98] CANDANEDO L M, FELDHEIM V, DERAMAIX D. Data driven prediction models of energy use of appliances in a low-energy house[J]. Energy and buildings, 2017, 140: 81-97.
- [99] ZHANG C, ZHANG J, CHAI D, et al. Aegis: A trusted, automatic and accurate verification framework for vertical federated learning[J]. arXiv preprint arXiv: 2108.06958, 2021.
- [100] ZHU H, WANG R, JIN Y, et al. PIVODL: privacy-preserving vertical federated learning over distributed labels[J]. arXiv preprint arXiv: 2108.11444, 2021.
- [101] WENJIE S, XUAN S. Vertical federated learning based on DFP and BFGS[J]. arXiv preprint arXiv: 2101.09428, 2021.
- [102] SUN J, YANG X, YAO Y, et al. Vertical federated learning without revealing intersection membership[J]. arXiv preprint arXiv: 2106.05508, 2021.
- [103] SHAN C, JIAO H, FU J. Towards representation identical privacy-preserving graph neural network via split learning[J]. arXiv preprint arXiv: 2107.05917, 2021.
- [104] XIA W, LI Y, ZHANG L, et al. A Vertical federated learning framework for horizontally partitioned labels[J]. arXiv preprint arXiv: 2106.10056, 2021.
- [105] ROMANINI D, HALL A J, Papadopoulos P, et al. Pyvertical: a vertical federated learning framework for multi-headed splitnn[J]. arXiv preprint arXiv: 2104.00489, 2021.
- [106] GAO Y, DOAN B G, ZHANG Z, et al. Backdoor attacks and countermeasures on deep learning: a comprehensive review[J]. arXiv preprint arXiv: 2007.10760, 2020.

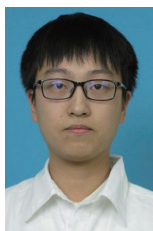
[作者简介]



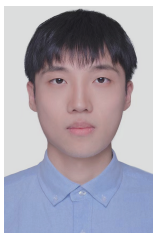
陈晋音（1982-），女，浙江象山人，浙江工业大学教授，主要研究方向为人工智能安全、图数据挖掘和进化计算。



李荣昌（1998-），男，浙江长兴人，浙江工业大学硕士生，主要研究方向为人工智能安全、图数据挖掘和联邦学习。



黄国瀚（1997-），男，浙江台州人，浙江工业大学硕士生，主要研究方向为图神经网络及纵向联邦学习安全。



刘涛（1998-），男，浙江绍兴人，浙江工业大学硕士生，主要研究方向为人工智能安全和联邦学习。



郑海斌（1995-），男，浙江台州人，浙江工业大学讲师，主要研究方向为深度学习、人工智能安全和图像识别。



程瑶（1987-），女，新加坡，博士，南德认证检测亚太有限公司研究员，主要研究方向为深度学习系统安全与隐私、区块链技术应用和安卓框架脆弱性分析。