软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(1):197−229 [doi: 10.13328/j.cnki.jos.006411]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

联邦学习中的隐私问题研究进展*

汤凌韬[1], 陈左宁[2], 张鲁飞[1], 吴 东[1]

[1](数学工程与先进计算国家重点实验室, 江苏 无锡 214125)

[2](中国工程院, 北京 100088)

通信作者: 汤凌韬, E-mail: tangbdy@126.com

摘 要: 随着大数据、云计算等领域的蓬勃发展, 重视数据安全与隐私已经成为世界性的趋势, 不同团体为保护自身利益和隐私不愿贡献数据, 形成了数据孤岛. 联邦学习使数据不出本地就可被多方利用, 为解决数据碎片化和数据隔离等问题提供了解决思路. 然而越来越多研究表明, 由谷歌首先提出的联邦学习算法不足以抵抗精心设计的隐私攻击, 因此如何进一步加强隐私防护, 保护联邦学习场景下的用户数据隐私成为一个重要问题. 对近些年来联邦学习隐私攻击与防护领域取得的成果进行了系统总结. 首先介绍了联邦学习的定义、特点和分类; 然后分析了联邦学习场景下隐私威胁的敌手模型, 并根据敌手攻击目标对隐私攻击方法进行了分类和梳理; 介绍了联邦学习中的主流隐私防护技术, 并比较了各技术在实际应用中的优缺点; 分析并总结了 6 类目前联邦学习的隐私保护方案; 最后指出目前联邦学习隐私保护面临的挑战, 展望了未来可能的研究方向.

关键词: 联邦学习; 数据隐私; 隐私攻击; 隐私保护

中图法分类号: TP18

中文引用格式: 汤凌韬, 陈左宁, 张鲁飞, 吴东. 联邦学习中的隐私问题研究进展. 软件学报, 2023, 34(1): 197–229. http://www.jos.org.cn/1000-9825/6411.htm

英文引用格式: Tang LT, Chen ZN, Zhang LF, Wu D. Research Progress of Privacy Issues in Federated Learning. Ruan Jian Xue Bao/Journal of Software, 2023, 34(1): 197–229 (in Chinese). http://www.jos.org.cn/1000-9825/6411.htm

Research Progress of Privacy Issues in Federated Learning

TANG Ling-Tao[1], CHEN Zuo-Ning[2], ZHANG Lu-Fei[1], WU Dong[1]

[1](State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi 214125, China)

[2](Chinese Academy of Engineering, Beijing 100088, China)

Abstract: With the vigorous development of areas such as big data and cloud computing, it has become a worldwide trend for the public to attach importance to data security and privacy. Different groups are reluctant to share data in order to protect their own interests and privacy, which leads to data silos. Federated learning enables multiple parties to build a common, robust model without exchanging their data samples, thus addressing critical issues such as data fragmentation and data isolation. However, more and more studies have shown that the federated learning algorithm first proposed by Google can not resist sophisticated privacy attacks. Therefore, how to strengthen privacy protection and protect users' data privacy in the federated learning scenario is an important issue. This paper offers a systematic survey of existing research achievements of privacy attacks and protection in federated learning in recent years. First, the definition, characteristics and classification of federated learning are introduced. Then the adversarial model of privacy threats in federated learning is analyzed, and typical works of privacy attacks are classified with respect to the adversary's objectives. Next, several mainstream privacy-preserving technologies are introduced and their advantages and disadvantages in practical applications are pointed out. Furthermore, the existing achievements on protection against privacy attacks are summarized and six privacy-preserving schemes are elaborated. Finally, future challenges of privacy preserving in federated learning are concluded and promising future research directions are discussed.

Key words: federated learning; data privacy; privacy attack; privacy preserving

---

* 基金项目: 国家重点研发计划 (2016YFB1000500); 国家科技重大专项 (2018ZX01028102)
收稿时间: 2020-10-02; 修改时间: 2021-01-28; 采用时间: 2021-06-24; jos 在线出版时间: 2021-08-03
CNKI 网络首发时间: 2022-11-15

大数据的发展推动人工智能迎来的新的高峰, 然而也带来了新的问题. 一是算力问题, 愈加庞大的数据规模和愈加复杂的学习模型对训练设备和集群的算力提出了更高的要求: 二是数据问题, 训练高精度的学习模型需要大规模高质量的数据支撑, 涉及数据采集、数据清洗和数据标注等预处理工作. 高质量数据往往意味着宝贵的专家知识和大量的人力物力投入. 不同团体乃至不同行业间不肯互相贡献自身数据, 从而造成了数据源之间的壁垒, 导致有效的数据得不到整合利用. 除利益问题外, 隐私问题近几年引起了大众的关注, 如 Facebook 和喜达屋等机构的信息泄露事件唤醒了大众的隐私保护意识, 也给各行业敲响了警钟. 2017 年《中华人民共和国网络安全法》和《中华人民共和国民法总则》正式实施, 要求网络运营者不得泄露、篡改、毁坏其收集的个人信息; 2018 年欧盟施行通用数据保护条例 (general data protection regulation, GDPR) 也将隐私保护带入法规, 约束企业对用户数据的恣意搜集和使用. 这些在加强用户隐私保护的同时, 也一定程度阻碍了数据的分享和流通, 业界急需一种新的数据利用模式, 保证原始数据不出本地也能被有效使用.

分布式机器学习 (distributed machine learning, DML) 的出现为解决算力问题提供了一种解决思路, 而针对数据问题, 研究人员提出了联邦学习 (federated learning, FL) 的概念. 联邦学习与分布式机器学习相比, 模型训练和推理的方法并无本质差别, 而在数据集的所有权和隐私性等方面有着不同的假设和要求. 分布式学习的初衷是将同一个任务分配到多个计算节点, 通过计算并行化提高模型训练效率, 不同节点上数据集往往采样于同一个数据源, 具有相似的分布和规模; 而联邦学习更侧重于对异质化的数据集进行学习, 不同计算节点上的数据可能具有完全不同的分布, 数据规模可能相差几个量级, 同时要求对各节点的本地数据集进行一定程度的隐私保护. 总的来看, 分布式学习是同一个利益团体对同一个任务进行切分和部署从而提高计算效率, 不同设备间的数据交换是透明的; 而联邦学习是不同的团体为了共同利益进行合作, 本地数据集往往表现出差异化特征, 并且要求任一团体无法直接获取或间接感知其他团体的本地数据.

Google 于 2017 年提出的算法 FedAvg[1]被普遍认为是对联邦学习的第一次正式探索, 其主要贡献在于指出了大量去中心化的数据存储于移动设备, 却因隐私问题得不到利用的问题. 进一步的, 将如何收集和训练这些数据定性为一个科学问题和研究方向. 联邦学习系统 (federated learning system, FLS)[2]往往包含一个中央服务器和多个客户端, 训练协议流程的一般步骤可总结如下.

(1) 中心在终端节点集合中随机选择一部分节点;

(2) 被选中节点下载当前的全局模型参数;

(3) 被选中节点使用本地数据更新全局模型参数;

(4) 被选中节点将更新的模型参数汇总到中心;

(5) 中心通过特定算法聚合数据, 并更新全局模型参数;

(6) 迭代执行上述 5 步直到模型收敛至期望值.

其中, 模型的训练方法与传统集中式学习无较大差别, 而如何对数据进行处理、传输和汇聚, 从而防止隐私泄露, 是本文的研究重点. 文献 [1] 根据各终端的本地数据集规模, 对其上传的模型参数进行加权平均. 该方法避免客户端直接上传本地数据, 一定程度上保护了用户隐私.

然而, 越来越多研究表明, 此般平凡的聚合协议会泄露隐私, 恶意敌手可以进行重构攻击, 有效还原用户的本地数据[3−5]. 事实上, 除了直接获取用户训练数据外, 敌手还可通过精心构造的隐私攻击, 获取用户数据的成员信息[6−8]、属性信息[9,10]、类代表信息[11,12]. 这对学者和研究人员提出了新的要求, 如何设计有效的隐私防护方法, 并在隐私保护、算法效率、模型精度间取得平衡, 成为联邦学习系统中一个重要的研究点. 训练好的模型将被广泛地部署到用户节点, 包括各类边缘设备和移动设备, 这类设备本身的安全状况无法得到保证. 该过程中服务商的模型和用户的预测样本都是各自的核心资产, 敌手能在推理过程中实施模型逆向[13]、成员推断[7]等攻击, 因此同样需要设计保护隐私的安全推理方法.

联邦学习中的隐私保护是一个交叉性极强的研究方向, 要系统地厘清隐私威胁并提出防护方法需要跨学科的努力. 为达到理论安全性, 目前有大量工作引入密码学技术保护模型的训练和推理过程, 如安全多方计算 (secure multi-party computation, MPC)、同态加密 (homomorphic encryption, HE)、函数加密 (functional encryption, FE) 和

差分隐私 (differential privacy, DP) 等. 以 SGX 为代表的可信执行环境 (trusted execution environment, TEE) 同样为联邦学习中的隐私保护提供了解决思路.

本文主要研究联邦学习中隐私问题, 总结相关研究进展. 第 1 节介绍联邦学习的定义, 特点和分类; 第 2 节分析联邦学习系统的隐私威胁模型和隐私攻击方式; 第 3 节介绍和对比目前主流的隐私保护技术, 及其应用于联邦学习的关键问题; 第 4 节分类梳理目前典型的联邦学习隐私保护方案; 第 5 节针对现有工作中的问题, 提出未来的挑战和展望; 第 6 节对全文进行总结.

1 联邦学习

联邦学习又称合作学习 (collaborative machine learning), 是区别于集中式和分布式机器学习的一种新场景, 在节点规模、数据分布、隐私保护等方面有着鲜明的特征. 机器学习模型训练的本质是解决一个优化问题, 传统的优化问题及其解决方法不能直接套用于联邦学习场景. 本节主要介绍了联邦学习的定义及特征, 阐述了其与传统机器学习场景的异同, 描述了一个典型联邦学习流程, 最后对联邦学习进行了分类.

1.1 定 义

联邦学习是一种机器学习场景, 多个客户端在一个或多个中央服务器的帮助下合作解决一个机器学习问题. 每个客户端的原数据存储在本地且不对外传输. 中央服务器通过对客户端上传的参数更新进行聚合以达到学习目标.

实际应用场景中联邦学习是由任务驱动的, 图 1 展示了 FLS 中学习模型的生命周期以及各类参与角色[14], 当特定问题被识别和建模后, 由模型工程师发布任务并开始完整的学习流程, 最后对生成的模型进行测试分析和实际部署. 其典型工作流如下.

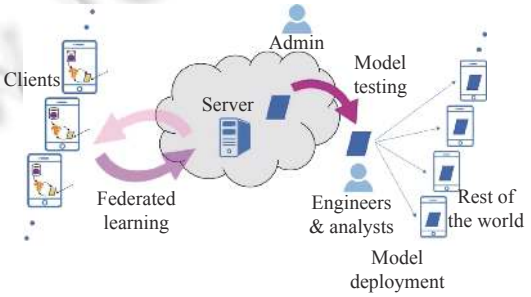


图 1 联邦学习系统中的模型生命周期及各类角色[14]

(1) 问题识别与任务定义: 模型工程师识别实际应用中的特定问题, 描述为联邦学习系统中的任务, 并选择对应的机器学习模型.

(2) 客户端协商: 中央服务器指导客户端在本地存储必要的训练数据. 事实上, 实际情况中客户端往往已存储所需数据, 如消息发送软件已存储了用户键入的文本消息, 照片管理软件已存储了用户近期照片.

(3) 原型模拟: 模型工程师可创建模型原型架构, 并使用代理数据进行超参测试与优化.

(4) 模型学习: 中央服务器和客户端进行完整联邦学习流程, 可使用不同的超参生成多个优化模型.

(5) 模型评估: 模型经充分训练达到预期损失后, 工程师对模型进行分析评估, 并挑选好的备选模型. 评估方式可以是使用标准数据集进行测试, 或使用客户端本地数据进行联合测试.

(6) 模型部署: 待发布模型被选出后将经过一系列标准模型发布流程, 最后部署到用户节点或服务云端, 该流程与传统集中式机器学习相同.

1.2 特 点

1.2.1 联邦学习与分布式学习

联邦学习源于分布式优化 (distributed optimization) 衍生出的一个特殊场景——联邦优化 (federated optimiza-

tion)[15−18], 因此联邦学习和分布式机器学习有很多相似之处, 两者都基于多个计算节点对分散存储的数据集进行分布式的模型训练, 很多学者把联邦学习看作分布式学习的一种延伸和特殊形式[15,18,19]. 两者的区别主要有以下 3 点.

(1) 数据. 在分布式优化中, 计算节点数一般远小于数据点的数目, 每个节点访问取自相同分布的随机样本, 且拥有相同量级的样本数. 而在联邦优化中, 训练数据具有以下的异质化特征:

● 广泛分布. 数据点存储于大规模的节点集合中, 且计算节点的数目可能远大于单个节点存储的平均样本数.

● 非独立同分布. 每个节点上的数据可能都取自不同的分布, 即任一节点的本地数据都不能代表整个数据集的分布.

● 体量不均衡. 各节点可能拥有不同数量级的训练样本数目.

(2) 参与节点. 在分布式学习中, 参与节点往往属于同一团体, 运行状况稳定, 拥有充足且均衡的计算能力和存储空间. 而在联邦学习中, 不同节点往往属于不同的利益团体, 对本地数据有完全自治权, 且通讯受限的现象较为常见, 如移动设备等终端频繁离线, 联网速度慢, 通讯代价高. 另外, 节点状况参差不齐, 大量边缘节点并没有充足算力和存储空间.

(3) 隐私保护. 分布式学习可分为两类[20]: 面向扩展性的分布式学习和面向隐私保护的分布式学习. 其中, 前者旨在解决数据和模型规模不断增长带来的扩展性问题, 提高训练效率和减小训练开销[21−25]; 而后者旨在保护用户隐私和数据安全, 联邦学习可视作后者的一种特殊形式, 参与节点规模更大, 且来自不同团体, 因此扩大了敌手的潜在攻击面, 增加了隐私保护的难度.

1.2.2 典型框架

传统机器学习可视作如下优化问题:

$$\min_{w\in\mathbb{R}^d} f(w) \text{ where } f(w)\triangleq\frac{1}{n}\sum_{i=1}^{n} f_i(w) \tag{1}$$

其中, $f(w)$ 为损失函数, 给定模型参数 w, $f_i(w)$ 是在第 i 个数据点上预测对应的损失.

文献 [1] 基于 SGD 提出了首个正式的联邦学习算法 FedAvg, 主要分为中央聚合和局部训练两部分, 参与节点包括一个中央服务器 S 和 K 个客户端组成的集合 $\mathcal{C}=\{C_k\}_{k\in[K]}$. 整个流程分为多个通讯轮, 每一轮中客户端 C_k 在其本地数据集 $\mathcal{P}_k$ 上使用局部 SGD 同步地训练局部模型. 中央服务器则对各客户端上传的模型参数进行聚合. 具体地, 记来自客户端 C_k 的参数为 w^k, 其中 $k\in\mathcal{C}_t$, $\mathcal{C}_t$ 为第 t 轮 m_t 个参与客户端组成的子集. 对于客户端 C_k, 设其本地训练数据集有 n_k 个数据点, 其中 $n_k=|\mathcal{P}_k|$. 因此, 联邦学习环境下的优化问题可重定义为:

$$\min_{w\in\mathbb{R}^d} f(w) \text{ where } f(w)\triangleq\sum_{k=1}^{m_t}\frac{n_k}{n}F_k(w), F_k(w)\triangleq\frac{1}{n_k}\sum\nolimits_{i\in\mathcal{P}_k} f_i(w) \tag{2}$$

FedAvg 针对的是单服务器协调全局训练流程的场景, 虽无法涵盖所有应用场景及模式, 但为学者们的深入研究提供了一个范例. 该范式下, 局部训练过程与中心化学习基本一致, 研究者主要围绕中央聚合过程展开优化, 如为加强隐私保护引入安全聚合[26−28], 为提升通信效率对聚合值进行有损压缩[16,29], 为达到差分隐私进行噪声添加和更新剪裁[30]. 然而, 敌手同样围绕聚合过程展开攻击, 一方面通过窃取其他节点的上传数据来分析原数据的相关特征, 另一方面通过上传精心构造的恶意数据影响全局模型或其他节点的局部模型. 在这些攻击策略下, 越来越多的聚合方法被证实是不安全的, 因此如何设计一个高效安全的聚合方法成为目前的研究重点和热点.

1.3 分 类

1.3.1 按数据分布分类

记矩阵 D_i 为第 i 个参与方的数据, 每行表示一个样本对象, 每一列表示一种数据特征, 部分样本对象还带有标签. 设第 i 个参与方的样本对象空间为 $\mathcal{S}_i$, 特征空间为 $\mathcal{F}_i$, 标签空间为 $\mathcal{L}_i$. 根据训练样本在不同参与方之间的分布特点, 可将联邦学习分为 3 类[20].

(1) 横向联邦学习 (horizontal federated learning, HFL). 参与方拥有不同的样本对象, 而数据特征基本相同, 即

$\mathcal{F}_i = \mathcal{F}_j$, $\mathcal{L}_i = \mathcal{L}_j$, $S_i \neq S_j$, $\forall D_i, D_j, i \neq j$. 适用于数据特征重叠较多, 而样本重叠较少的场景. FedAvg 就是针对横向联邦学习的典型学习算法.

(2) 纵向联邦学习 (vertical federated learning, VFL). 参与方拥有基本相同的样本对象, 而数据特征不同, 即 $\mathcal{F}_i \neq \mathcal{F}_j$, $\mathcal{L}_i \neq \mathcal{L}_j$, $S_i = S_j$, $\forall D_i, D_j, i \neq j$. 适用于样本重叠较多, 数据特征重叠较少的场景. 参与方间需要先进性隐私实体匹配, 安全地对齐共有样本, 然后通过加密技术训练模型.

(3) 联邦迁移学习 (federatde transfer learning, FTL). 参与方的样本对象和数据特征都有较大差异, 即 $\mathcal{F}_i \neq \mathcal{F}_j$, $\mathcal{L}_i \neq \mathcal{L}_j$, $S_i \neq S_j$, $\forall D_i, D_j, i \neq j$. 适用于特征和样本重叠都较少的场景, 实现跨域知识迁移.

1.3.2 按参与方类型分类

联邦学习的不同应用场景中参与方的数目和个体特征表现出较大差异, 根据参与节点的数目和节点特征, 可将联邦学习分为两类[14].

(1) 跨筒仓 (cross-silo) 联邦学习. 适用于大型机构间的合作学习任务, 参与节点拥有充足的计算能力和存储空间, 网络连接状况良好, 稳定在线. 每个节点上的数据规模大, 质量高, 可以是横向或纵向划分.

(2) 跨设备 (cross-device) 联邦学习. 适用于大量移动边缘终端设备参与的学习任务, 这些节点的算力较弱, 容量较小, 通讯代价较高, 频繁离线. 每个节点上的数据规模小, 质量高低不一, 一般是横向划分.

2 联邦学习中的隐私威胁

隐私问题是联邦学习的核心问题. 提及机器学习系统面临的威胁时, 安全与隐私往往被混为一谈. 隐私攻击是系统面临的威胁的一部分, 敌手为了窃取用户原数据或训练好的模型参数等隐私信息, 发起隐私攻击. 而安全攻击则是通过妨碍模型正常训练或诱导模型错误预测等手段, 危害系统的准确性和鲁棒性, 目前已有相关研究验证了投毒攻击 (poisoning attack) 和对抗攻击 (adversarial attack) 在联邦学习场景中的可行性[31−34]. 本文主要关注联邦学习中用户数据的隐私保护问题, 对于安全攻击不再展开. 第 2.1 节分析了不同敌手角色存在时的隐私威胁模型; 第 2.2 节总结了联邦学习系统中常见的隐私攻击.

2.1 敌手模型

联邦学习系统作为一个分布式系统, 其完整工作流中往往包含大量参与者, 向攻击者暴露了多个攻击点. 要分析联邦学习的安全性, 首先要厘清系统面临的隐私威胁, 包括判别敌手类型, 明确敌手攻击目标, 定义和划分系统内角色, 分析各角色的潜在攻击能力, 归纳敌手的攻击策略.

2.1.1 敌手目标

一个设计完备的信息系统应具备机密性 (confidentiality)、完整性 (integrity)、可用性 (availability), 而隐私攻击目标则是破坏联邦学习系统的机密性, 推断和获取系统非主动暴露的信息. 这些信息可分为以下 4 类[35].

(1) 成员 (membership) 信息. 给定一个样本, 敌手试图判定其是否用于训练, 进一步地, 确定其属于哪一个参与方.

(2) 属性 (property) 信息. 敌手试图推断参与方训练数据的相关属性, 这些特征并非由样本所标记的特征和标签直接体现, 与训练主目标不相关.

(3) 类代表 (class representatives). 对于攻击对象的带标签数据集, 敌手尝试生成其中某一类数据的典型训练样本, 而非还原攻击对象的确切训练数据. 典型样本与同类真实数据具有相同的特征和分布.

(4) 训练数据. 敌手试图逼近甚至还原参与方的训练数据.

2.1.2 敌手类型

在隐私保护和安全计算等领域, 一般考虑两种类型的敌手.

(1) 半诚实 (honest-but-curious/semi-honest) 敌手: 在半诚实敌手模型中, 敌手会如实遵守并执行通讯协议的流程, 与其他节点交互时不会篡改发送的消息, 但会尝试根据接收到的消息推断更多的信息. 此类敌手不干扰训练过程, 不影响模型完整性和可用性, 通过观察和收集相关信息来达成攻击目标.

(2) 恶意 (malicious) 敌手: 在恶意敌手模型中, 敌手行为不受限制, 可能不遵守协议, 恶意篡改发送的消息, 从

而影响甚至破坏协议流程, 诱导其他节点泄露更多信息.

在系统设计前必须明确是防御哪一类敌手发起的隐私攻击, 现有的研究主要基于半诚实敌手假设来设计隐私保护方案, 在安全多方计算等密码学协议中, 抵抗恶意敌手往往需要大量额外的计算和通信等开销[36], 难以保证方案的高效性和实用性.

2.1.3 敌手角色

与传统机器学习不同, 联邦学习涉及功能和能力各异的多个参与方, 敌手可从多角度侵入并展开攻击. 根据图 1 的参与角色, 对联邦学习系统中的敌手角色和潜在攻击位置分类如下.

(1) 客户端: 操作者掌握客户端 root 权限, 可能是合法管理员, 或侵入攻击者. 半诚实客户端可以在参与的轮次中, 查看来自服务器的所有消息, 但不会干扰训练流程. 而恶意的客户端在查看消息的同时可能干扰训练.

(2) 服务器: 操作者掌握服务器 root 权限, 可能是合法管理员, 或侵入攻击者. 半诚实服务器可以查看接收到的所有消息, 但不干扰训练流程. 而恶意服务器在查看消息的同时还可能干扰训练.

(3) 模型工程师和分析人员: 可访问训练算法输出模型, 恶意的工程师或分析人员能接触到系统的多个输出, 如不同超参下的模型训练迭代.

(4) 实际需求用户: 可访问部署模型, 恶意用户或被侵入的用户节点对模型拥有黑盒访问权限.

其中, 根据敌手是否参与模型训练, 将恶意的客户端或服务器称为内部敌手 (inside attacker), 将只能访问输出模型或部署模型的敌手称为外部敌手 (outside attacker).

2.1.4 敌手知识

敌手知识是指敌手对于目标模型及其生成和应用环境所掌握的相关信息, 如模型结构、模型参数、训练样本分布、决策函数等. 根据敌手掌握知识的多少, 可将其攻击行为分为黑盒攻击和白盒攻击.

(1) 黑盒攻击. 敌手没有模型的相关知识, 只能观察到模型的预测结果. 对任何输入数据 x, 敌手可获得 $f(x;W)$, 但无法获知模型权重 W 和推理过程的中间计算, 甚至是学习算法和输出模型的结构. 学习是训练数据所包含的知识向模型提炼转化的过程, 敌手一般利用模型本身的知识记忆性, 通过精心构造输入, 来分析输出, 从而达到预期攻击目标. 相比白盒攻击, 黑盒攻击中敌手的数据可见度低, 攻击准确度不高, 然而危害面更广, 目前一些互联网公司向用户提供预测服务 (PaaS), 如 Google Prediction、Microsoft Azure ML、Amazon ML, 使黑盒攻击环境广泛存在于日常生活中.

(2) 白盒攻击. 敌手掌握模型的结构和权重参数, 甚至其他参与方的训练数据. 此类敌手往往属于内部敌手, 掌握良好的局部视图, 有着较高的攻击准确率. 事实上, 具体攻击场景中, 存在介于白盒与黑盒之间的攻击方案, 如敌手掌握模型的结构, 但不知道模型的具体参数[6].

2.1.5 敌手能力

敌手能力是指敌手在系统中各阶段所具备的权限, 在数据收集阶段, 可以是敌手直接获取训练数据的能力; 在训练阶段, 可以是敌手干预训练流程、收集中间结果的能力, 如精心构造输入影响其他节点, 观察该节点输出的变化趋势; 在推理阶段, 可以是敌手访问模型接口获取预测结果, 甚至提取模型相关信息的能力. 根据能力强弱可将敌手分为强敌手和弱敌手, 强敌手可以参与模型训练, 获取模型相关参数和用户训练数据, 弱敌手只能通过访问模型, 观察特定输出, 收集辅助信息等间接手段完成攻击.

联邦学习系统中, 敌手在全流程具备的权限越高, 拥有的攻击手段越多, 其攻击能力就越强. 据此可对联邦学习系统中的角色能力进行排序: 服务器>客户端>分析人员>用户. 服务器和客户端作为内部节点, 除了输出模型外, 还可以查看训练过程中的聚合结果, 其中服务器可以进一步查看各客户端上传的更新值; 用户和分析人员作为外部节点, 能对输出模型进行黑盒访问, 分析人员可能进一步拥有白盒权限, 观察不同超参下的多个模型. 因此, 服务器一般是强敌手, 用户一般是弱敌手, 客户端和分析人员介于两者之间, 依据实际攻击场景进行区分.

2.1.6 敌手策略

敌手的攻击目标一旦确定, 再根据敌手的角色、知识、能力等性质, 可确定其具体的攻击策略. 常见的攻击策略有 5 类.

(1) 重构攻击 (reconstruction attack): 敌手通过观察和抽取模型训练期间的中间变量及相关特征, 重构出用户的原始训练数据.

(2) 模型窃取攻击 (model extraction attack): 敌手窃取训练好的模型参数或者模型本身. 模型隐私泄露损害的是模型拥有者的利益, 一般是机器学习平台的服务提供商.

(3) 成员推断攻击 (member inference attack): 敌手拥有模型的黑盒或白盒访问权限, 目标是判定一个特定样本是否属于某用户的训练集.

(4) 属性推断攻击 (property inference attack): 敌手推断参与方训练数据的相关特征, 这些特征并非由样本标签和属性直接体现.

(5) 模型逆向攻击 (model inversion attack): 敌手通过黑盒或白盒访问模型的输出, 反推训练数据集的相关信息.

各种攻击策略被提出时, 攻击目标和敌手能力互有交集, 如模型逆向攻击和成员推断攻击中敌手都能访问模型输出, 模型逆向攻击和重构攻击中攻击目标都是推断用户训练数据, 导致不同文献中攻击策略的分类和包含关系产生冲突, 如表 1 所示. 因此为清晰地梳理联邦学习面临的隐私威胁, 本文借鉴文献 [35], 选取敌手攻击目标作为分类依据.

表 1 不同文献中的隐私攻击分类

文献	隐私攻击分类	文献	隐私攻击分类
[21]	模型逆向攻击 模型提取攻击 成员推断攻击	[37]	模型逆向攻击 ● 成员推断攻击 ● 属性推断攻击 模型提取攻击
[38]	重构攻击 ● 模型逆向攻击 ● 模型提取攻击 成员推断攻击	[39]	模型逆向攻击 ● 重构攻击 ● 属性推断攻击 模型提取攻击 成员推断攻击

2.2 隐私攻击

根据不同的敌手模型, 学者们针对联邦学习场景展开研究, 设计并验证了多种隐私攻击的可行性和破坏性, 此处选取了 13 篇近几年高被引 (3 年内超过 20 次, 5 年内超过 100 次) 的研究, 如表 2 所示, 并依据敌手攻击目标对这些研究进行分类、梳理及分析.

2.2.1 获取类代表

Hitaj 等人[12]利用对抗生成网络 (generative adversarial networks, GAN), 设计了一种联邦学习场景下针对深度神经网络的隐私攻击方法, 该文假设用户间的数据拥有互不相同的标签类, 敌手角色可以是任何一个客户端, 从系统内部攻击其他客户端, 推断其本地数据某一类的隐私信息. 具体地, 敌手在训练的每一轮下载服务器的全局模型, 将其作为判别器并在本地训练生成器, 生成与目标客户端本地数据相似的样本. 敌手将这些生成样本打上错误的标签, 混入本地数据参与模型训练, 从而某种程度上影响全局模型对特定类的识别能力. 为了正确分类这些包含错误标签的样本, 目标客户端下一轮训练中上传的梯度会包含更多与本地数据相关的信息, 据此敌手可进一步优化生成器, 生成更相似的样本. 该文强调记录级 (record-level) 差分隐私并不能有效抵抗此类攻击, 因为基于 GAN 的攻击方法旨在推断目标类的典型样本, 而非真实的训练样本.

Wang 等人[11]指出文献 [12] 的攻击方法存在 3 个缺陷: (1) 假设中客户端能改变共享模型架构, 权限过高, 不符合现实应用场景, 且攻击影响了正常的模型训练; (2) FedAvg 算法平均更新值的方式会减小恶意客户端造成的影响, 从而降低攻击性能; (3) 只能推断类的总体信息, 无法推断特定客户端隐私. 对此, 作者提出了一种包含多任务判别器的对抗生成网络 mGAN-AI, 同时对数据的真伪、类别和归属用户进行判别, 利用更新值还原每个客户端的典型数据, 并以此监督 GAN 训练. 对比文献 [12], 该方法可以恢复特定用户的典型数据, 造成用户级隐私泄露.

表 2 联邦学习中的典型攻击

文献	攻击阶段	敌手目标	敌手类型	敌手角色	敌手知识	敌手能力	攻击策略	目标模型
[12]	训练过程	类代表	恶意	客户端	白盒	强敌手	模型逆向攻击	神经网络
[11]	训练过程	类代表	半诚实/恶意	服务器	白盒	强敌手	模型逆向攻击	神经网络
[6]	推理过程	成员信息	半诚实	用户	黑盒	弱敌手	成员推断攻击	任意模型
[7]	推理过程	成员信息	半诚实	用户	黑盒	弱敌手	成员推断攻击	神经网络
[40]	推理过程	成员信息	半诚实	工程师/用户	白盒	弱敌手	成员推断攻击	任意二元分类器
[8]	训练过程	成员信息	半诚实/恶意 半诚实/恶意	服务器 客户端	白盒	强敌手	成员推断攻击	神经网络
[10]	训练过程	成员信息 属性信息	半诚实	客户端	白盒	强敌手	成员推断攻击 属性推断攻击	神经网络
[9]	推理过程	属性信息	半诚实	工程师/用户	白盒	弱敌手	属性推断攻击	全连接神经网络
[4]	训练过程	训练数据	半诚实	服务器	白盒	强敌手	重构攻击	神经网络
[13]	推理过程	训练数据	半诚实	工程师 用户	白盒 黑盒	弱敌手	模型逆向攻击	决策树/神经网络
[5]	训练过程	训练数据	半诚实	服务器/客户端	白盒	强敌手	重构攻击	任意二次可微模型
[3]	训练过程	训练数据	半诚实	服务器/客户端	白盒	强敌手	重构攻击	神经网络
[41]	训练过程	训练数据	半诚实	服务器/客户端	白盒	强敌手	重构攻击	神经网络

当数据集中每个类内部成员相似时, 获取类代表与获取训练数据将取得相近的攻击效果, 然而, GAN 只是生成了类的典型样本, 而非训练数据本身, 判别器无法有效区分训练样本和随机典型样本, 因此这两类攻击有着本质的区别. 例如, 目标类中是某一用户的照片, 敌手利用生成器输出的照片会呈现相似的脸, 从而判别目标用户的大致样貌, 然而给定一张真实照片, 同时按该类的分布随机生成一张照片, GAN 无法辨别哪一张是真实的.

2.2.2 获取成员信息

敌手可以在推理阶段获取成员信息. 通过访问输出模型和部署模型的接口, 尝试确定某样本是否属于训练集, 从而危害用户隐私, 此时攻击方式与传统机器学习场景相似.

Shokri 等[6]对黑盒模型下的成员推断攻击进行研究, 通过构建攻击模型来识别一条记录是否属于训练数据集. 为训练攻击模型, 作者提出了一种影子训练技术, 通过 3 类数据: (1) 基于模型的合成数据; (2) 基于统计信息的合成数据; (3) 含噪声的真实数据, 生成多个模仿目标模型行为的影子模型, 由于影子模型的训练数据集是确定的, 因此可根据其输入输出进行监督训练, 使攻击模型能分辨某记录是否属于影子模型的训练数据集. 实验表明训练好的攻击模型对 Google 和 Amazon 的机器学习服务平台 (MLaaS) 的成员推断攻击准确率可达 94% 和 74%.

上述攻击基于两个假设: 每个影子模型与目标模型具有相同结构; 用于训练影子模型和目标模型的训练数据具有相同分布. Salem 等人[7]认为该假设要求过高, 限制了实际情况中的攻击范围, 因此设计了 3 种敌手, 逐步放宽假设, 证明了成员推断攻击可能发生于更广泛的场景. 其中第 3 种敌手不使用任何影子模型, 无需任何训练流程, 仅依赖目标模型的预测结果. 该文表明目标模型的后验统计信息, 如熵和最大值, 可以有效区分成员和非成员数据点. 在实验中, 作者提出了一种阈值选择的方法, 在多个数据集上进行了有效的推理攻击.

Yaghini 等人[40]指出在面对成员推断攻击时, 训练数据的不同子集会表现出不一样的脆弱性, 而以往的工作只关注整个数据集的平均隐私损失. 对此提出了一个量化隐私泄露的框架, 无需对模型的重复训练和测试, 即可计算每个数据子集的隐私泄露程度, 而该框架也需要敌手掌握目标数据集的一些背景知识. 作者对 ADULT 数据集训练出的分类器进行测试, 实验表明规模小、代表性不足的子集更容易受到成员推断攻击, 而规模大的子集不易受攻击, 这一规律与分类器的结构无关. 作者进一步指出满足差分隐私的训练算法并不能完全消除这种差异性.

敌手也可以在训练阶段获取成员信息. 一些学者根据联邦学习的特点设计了训练过程的攻击方法, 研究客户端上传更新值引发的信息泄露问题. 此类攻击中敌手通过观察客户端的上传数据推断特定样本的成员信息, 由于敌手知识更多、能力更强, 攻击效果也更为显著.

Melis 等人[10]表明训练中恶意参与者可以推断其他参与方的特定样本信息, 该文假设存在$K(K\geqslant 2)$个节点的学习场景, 敌手作为参与节点之一观察不同轮次的全局模型, 计算得到每一轮的聚合梯度, 对于自然语言处理等任务, 嵌入层中的非零梯度揭示了哪些词汇在训练批次中, 从而进一步帮助敌手确定某段文本是否属于训练集.

Nasr 等人[8]对深度神经网络的白盒成员推断攻击进行研究, 分析表明当模型泛化能力强时, 针对激活函数等中间变量的攻击效果不好. 对此作者利用 SGD 算法中反向传播的梯度展开攻击, 由于深度神经网络中梯度的规模远大于训练数据本身, 泛化能力不强, 在面对训练和非训练数据时, 梯度的分布会产生较大差异. 文中分析评估了该方法的攻击效果, 结果表明: (1) 半诚实服务器的攻击成功率高于半诚实客户端; (2) 随着训练轮数增多, 攻击准确率更高; (3) 随着参与者增多, 攻击效果下降. 另外, 考虑恶意敌手, 作者设计了一种“梯度上升”攻击, 在训练中增大目标数据所产生的梯度, 若该数据在训练集中, 由 SGD 算法的特性后续训练中对应的梯度会明显减小. 进一步地, 恶意服务器可以针对目标客户端发起隔离攻击, 不向其传输其他客户端的更新, 从而获取该客户端的局部视图, 这种攻击能显著增加信息泄露概率.

2.2.3 获取属性信息

Melis 等人[10]对属性推断攻击进行了研究, 主要关注“非预期”特征, 即只对数据集的某一子集成立的特征, 如在用于训练性别分类器的照片数据集中, 某特定人物何时第一次出现, 又或是照片中人是否戴眼镜, 这些“非预期”特征与分类器的目标不相关, 也只对一小部分数据成立. 作者认为这些特征更有效地反映出用户隐私的泄露, 因为敌手在参与者不经意的情况下获取了额外信息. 同时该文表明, 记录级的差分隐私会限制成员推断攻击的成功率, 但无法阻止属性推断攻击; 而用户级 (participant-level) 差分隐私虽能抵抗属性推断, 但在参与方较少时严重影响模型精度. 该攻击方法假设敌手掌握带正确标签的额外训练数据, 其标签是敌手的目标属性, 如敌手目标是推断年龄, 则需预先掌握标签为年龄的照片数据集. 相似的研究还有 Ganju 等人[9]提出的针对全连接神经网络的属性攻击, 利用文献 [6] 中的影子训练技术来训练目标属性的元分类器 (meta classifier). 与文献 [10] 不同, 该方法作用于推理过程, 攻击的是训练后的输出模型或部署模型, 且假设敌手掌握模型的架构和参数.

目前的属性攻击方法都有一定程度的局限性, 如, 需要额外信息的支持; 攻击的属性与训练数据本身的特征和标签相关. 前者限制了攻击方法的实用性, 后者让防御机制的设计者有迹可循, 限制了敌手的攻击效果.

2.2.4 获取训练数据

目前学者们在设计联邦学习系统时, 普遍通过共享模型参数或梯度等更新值来训练模型[1,16,42–45], 既避免了暴露本地数据, 又能达到较好的训练效果. 然而越来越多研究表明, 若不设计特别的隐私保护机制, 敌手能根据这些更新值重构出用户训练数据.

Phong 等人[4]指出文献 [44] 中的学习算法存在隐私泄露问题, 在神经网络训练过程中, 一小部分的梯度即可泄露训练数据的相关信息. 以单个神经元为例, 损失函数定义为预测值$h_{W,b}(x)\triangleq f\left(\sum_{i=1}^{d}W_ix_i+b\right)$和真实值$y$之间的距离:

$$J(W,b,x,y)\triangleq (h_{W,b}(x)-y)^2 \tag{3}$$

因此对应梯度为:

$$\eta_k\triangleq\frac{\delta J(W,b,x,y)}{\delta W_k}=2(h_{W,b}(x)-y)f'\left(\sum_{i=1}^{d}W_ix_i+b\right)\cdot x_k \tag{4}$$

$$\eta\triangleq\frac{\delta J(W,b,x,y)}{\delta b}=2(h_{W,b}(x)-y)f'\left(\sum_{i=1}^{d}W_ix_i+b\right)\cdot 1 \tag{5}$$

中央服务器通过计算$\eta_k/\eta=x_k$即可获得用户输入数据, 同时观察可得梯度η_k和输入x_k成固定比例, 若$x=(x_1,\ldots,x_k)$为一幅图像, 敌手可利用梯度生成一幅同样“成比例”的相关图像, 并据此猜测真实值y. 作者同样对一般神经网络进行了分析, 实验表明对于手写体数字, 仅 3.89% 的梯度足以恢复用户原始数据.

Zhu 等人[5]提出一种利用梯度重构训练数据的攻击方法 DLG (deep leakage from gradients). 具体地, 敌手生成一对随机的“虚拟”数据和标签(x',y'), 然后对模型F进行前向及后向的计算, 在获得对应的虚拟梯度$\nabla W'$后, 并不像普通学习流程一样对模型参数进行优化, 而是优化虚拟输入和标签, 使得$\nabla W'$和真实梯度∇W的距离最小化, 当

两者距离很小时, (x',y') 和原数据 (x,y) 高度匹配. 总之, DLG 的本质是解决如下优化问题:

$$x'^*,y'^* \triangleq \arg\min_{x',y'} \|\nabla W' - \nabla W\|^2 = \arg\min_{x',y'} \left\| \frac{\partial L(F(x',W),y')}{\partial W} - \nabla W \right\|^2 \tag{6}$$

作者分别在计算机视觉和自然语言处理两种任务上验证了攻击的有效性, 结果表明对原数据达到了像素级和句柄级的还原. 另外, 作者对差分隐私的保护效果进行实验, 对于输入扰动, 当方差为 10^{-4} 或 10^{-3} 量级时无法抵抗攻击, 当方差大于 10^{-2} 时虽能有效抵抗攻击, 但噪声已开始影响模型精度. 同时, 作者对两类半精度梯度扰动方法进行实验, 结果表明其不能有效保护数据隐私.

基于上述工作, 一些学者对 DLG 进行了改进. Zhao 等人[3]观察到 DLG 进行数据重构时经常会生成错误的标签, 对此提出了改进算法 iDLG (improved DLG), 考虑使用交叉熵损失和独热码标签的神经网络模型, 作者利用输出层各标签概率及其上一层输出值梯度间的关系, 准确找出真实标签 y, 在解决优化问题 (6) 时只需对 x' 进行更新, 实验表明该方法对 x 的重构也具备更高准确率. Geiping 等人[41]为进一步优化攻击准确率, 将攻击算法的损失函数改为余弦相似度 $\ell(x,y) = \frac{\langle x,y\rangle}{\|x\|\|y\|}$ 并添加 TV (total variation) 正则项, 同时将数据空间限制在 $[0,1]$, 且基于 iDLG 假设已获取真实标签, 从而得到优化问题 (7). 进一步地, 该文研究了模型架构及参数对攻击效果造成的影响, 以及针对 FedAvg 的攻击方式.

$$x'^* \triangleq \arg\min_{x'\in[0,1]^n} 1 - \frac{\langle \nabla_W L_W(x',y), \nabla_W L_W(x,y)\rangle}{\|\nabla_W L_W(x',y)\|\|\nabla_W L_W(x,y)\|} + \alpha TV(x) \tag{7}$$

对于无法参与训练的敌手, 梯度、模型参数等中间参数是不可见的, 然而研究表明, 仅通过学习系统提供的预测接口, 敌手仍能对训练数据展开攻击. Fredrikson 等人[13]提出了针对决策树和人脸识别模型的逆向攻击方法, 以攻击人脸识别系统为例, 假设目标 MLaaS 系统会返回不同类标签及其置信度, 作者利用梯度下降搜寻假想输入, 使返回的置信度最大化且被归为相同类, 从而逼近和还原真实输入数据. 实验表明, 该方法能有效还原出初始图片的相关特征, 再由测试人员进行人工鉴别, 能以 60%~80% 的成功率匹配真实图片.

3 联邦学习中的隐私保护技术

机器学习领域内部的技术不能很好地抵抗上述各类隐私攻击, 对此研究人员将密码学和可信硬件等技术引入联邦学习, 通过密码技术的理论安全性以及可信硬件的物理层面安全来保障用户隐私. 目前面向联邦学习的隐私保护涉及的技术主要分为 3 类.

(1) 加密方法: 参与方在不交换明文的情况下, 进行安全的分布式计算. 相关技术包括安全多方计算、同态加密、函数加密等. 此类方法有效隐藏了计算输入和一些中间变量, 限制了敌手获取额外知识的能力, 从而影响敌手攻击成功率甚至直接使其攻击策略失效.

(2) 扰动方法: 参与方通过对数据添加噪声等方法获取可量化的隐私保证. 典型技术为差分隐私, 保证不同训练样本对最终模型的影响一定程度上不可区分, 从而抵抗敌手获取特定数据的隐私信息.

(3) 可信硬件: 参与方将数据加密, 在可信执行环境下执行数据解密及指定计算, 通过物理层面的安全性保证敌手无法接触原数据或推理相关信息. 典型架构有 intel SGX[46]、Sanctum[47]等.

本节将对上述 3 类中的典型技术进行介绍, 包括其定义和特点, 以及应用于联邦学习的关键问题.

3.1 安全多方计算

在一个安全多方计算协议中, n 个各自持有隐私数据 $d_1,\ldots,d_n$ 的参与者 $P_1,\ldots,P_n$ 可以计算一个公开函数 $F(d_1,\ldots,d_n)$, 同时保证隐私数据的机密性.

安全多方计算起源于 Yao[48,49]提出的百万富翁问题, 而后 Goldreich 等人提出了 GMW 协议[50], 证明即便存在恶意敌手, 任意函数都可以进行安全计算, Yao 方案的核心技术是混淆电路 (garbled circuit, GC) 和不经意传输 (oblivious transfer, OT), 而 GMW 利用秘密共享 (secret sharing, SS) 将两方计算 (2PC) 自然拓展到了多方计算. 另外一些基石性的工作有 BGW[51]、BMR[52]等. Ben-OR 和 Goldwasser 等利用 Shamir 秘密共享构建了 BGW 协议,

可在域$\mathbb{F}$上对运算电路 (arithmetic circuit) 进行计算, 包含加法、乘法、常数乘 3 种基础运算. 上述协议皆需要正比于电路规模的通信轮次, 而 BMR 协议通过一种分布式混淆电路生成方法, 将通信轮次降至常数.

根据函数 F 的表示方法和数据 $d_1,\ldots,d_n$ 的共享形式, 安全多方计算的后续研究主要分为两类.

(1) 基于秘密共享的运算电路. 用户数据以加法共享 (additively sharing) 的方式分散到参与节点. 此类协议进行加法、矩阵乘等线性代数运算时十分高效, 而进行比较等运算时开销较大. 代表性工作有 BDOZ[53]和 SPDZ[54], 利用加法秘密共享, Beaver 三元组[55]技术, 以及消息认证码实现了可抵抗恶意敌手的安全多方计算协议.

(2) 基于混淆电路的布尔电路. 用户数据以布尔共享 (boolean-sharing) 的方式分散到参与节点. 此类协议进行除法、比较、比特移位和 sign() 等易表示为布尔电路的运算时十分高效, 而对于加法、乘法等运算需要额外开销. 代表性工作有 WRK[56,57], 作者提出了一种可验证混淆 (authenticated garbling) 的技术, 将可验证秘密共享、Beaver 三元组、混淆电路、BMR 电路生成等技术相结合, 该协议同样可抵抗任意数量的恶意敌手, 实验表明该协议具备极高的效率.

除了上述通用协议, 安全多方计算也衍生出另一分支, 针对具体问题构造专用方案, 如集合求交[58], 电子投票[59], 不经意多项式计算 (OPE)[60]等. 此类协议在特定问题上往往比通用协议更高效, 且具备更简洁的安全性证明.

安全多方计算应用于联邦学习的关键问题主要在于: (1) 需针对计算类型, 选取合适的密码学工具. 安全多方计算是由同态加密、秘密共享、不经意传输、混淆电路等多种基础技术组成的综合密码学技术. 面对如全连接层等线性运算时, 可使用同态加密加速计算, 面对如激活函数等非线性运算时, 则使用混淆电路技术对布尔电路进行隐私计算. (2) 优化学习模型和计算协议, 使其适应密码技术, 从而提高协议效率. 例如将浮点数据进行截断并表示为定点整型, 使用多项式近似激活函数等. (3) 扩展参与方, 目前学界对安全两方计算的研究较为成熟, 而参与方增多会导致协议通信复杂度显著提升, 在跨设备联邦学习等场景下, 参与方可能是数百台终端甚至更多, 所有节点间直接进行安全多方计算是不可行的.

3.2 同态加密

令消息空间 $(M,\circ)$ 为一个有限 (半) 群 σ 为安全参数. M 上的一个同态加密方案[61]是由多项式时间算法组成的四元组 (K,E,D,A), 其中:

- 密钥生成函数 K. 输入 1^σ, 输出加密和解密密钥 $(k_e,k_d)=k\in\mathcal{K}$, 其中 $\mathcal{K}$ 为密钥空间.
- 加密函数 E. 输入 $1^\sigma,k_e$ 和明文 m, 输出密文 $c\in\mathcal{C}$, 其中 $\mathcal{C}$ 为密文空间.
- 解密函数 D. 输入 $1^\sigma,k$ 和密文 $c\in\mathcal{C}$, 输出 $m\in M$. 该过程满足: 若 $c=E(1^\sigma,k_e,m)$, 则 $\Pr[D(1^\sigma,k,c)\neq m]$ 可忽略, 也即 $\Pr[D(1^\sigma,k,c)\neq m]\leqslant 2^{-\sigma}$.
- 同态性. 算法 A 接收 $1^\sigma,k$ 和 $c_1,c_2\in\mathcal{C}$ 作为输入, 输出 $c_3\in\mathcal{C}$, 且满足对所有 $m_1,m_2\in M$, 若 $m_3=m_1\circ m_2$, $c_1=E(1^\sigma,k_e,m_1)$, $c_2=E(1^\sigma,k_e,m_2)$, 则 $\Pr[D(A(1^\sigma,k_e,c_1,c_2))]\neq m_3$ 可忽略.

对于同态性, 若 M 是加法 (半) 群, 则称该加密方案是加法同态的, 此时算法 A 中算符 $\circ$ 表示加法; 若 M 是乘法 (半) 群, 则称该加密方案是乘法同态的, 此时算法 A 中算符 $\circ$ 表示乘法.

同态加密方案主要分为 3 类[62,63]: 部分同态加密 (PHE), 类同态加密 (SHE), 全同态加密 (FHE). 在密文域中, PHE 支持加法或乘法其中一种的无限次同态运算; SHE 支持有限次的加法和乘法同态运算; FHE 支持无限次的加法和乘法同态运算. 3 种方案中全同态加密适用面最广, 然而计算开销也最大. 从定义上看, 全同态加密是一种适用于安全计算的理想方案, 可实现机器学习过程中端到端的隐私保护, Gentry[64]基于理想格 (ideal lattices) 提出的方案首次从理论上实现了全同态加密, 引入 bootstrapping 技术解决噪声增长的问题, 然而该过程计算开销很大, 导致方案并不实用, 对此学者们后续展开许多相关研究, 但目前仍未能将全同态加密投入实际大规模应用.

同态加密应用于联邦学习的关键问题在于: (1) 不能进行比较、比特位移等计算, 无法支持激活函数等复杂计算. (2) FHE 计算量大, 目前的硬件难以支持. (3) 很多同态加密方案是一对一的, 无法自然地应用于联邦学习的训练过程. 一些隐私保护方案让所有参与客户端共享密钥[4], 虽然抵抗了恶意服务器的攻击, 但需保证客户端间不会互相窃取密文. 加强多密钥同态加密[65,66]的研究有希望解决该问题.

3.3 函数加密

一个基于函数 f 的函数加密方案[67]包含 4 个算法.

- $(pk, msk) \leftarrow \text{Setup}\left(1^{\lambda}\right)$. 初始化算法创建公钥 pk 和主密钥 msk.
- $sk \leftarrow \text{Keygen}(msk, f)$. 密钥生成算法使用主密钥为函数 f 生成一个新的私钥.
- $c \leftarrow \text{Enc}(pk, x)$. 加密算法使用公钥加密消息 x.
- $y \leftarrow \text{Dec}(sk, c)$. 解密算法使用私钥计算 $y = f(x)$, 其中 x 是 c 对应的明文.

函数加密是公钥加密的推广, 拥有私钥的人能在只接触密文 c 的情况下获取函数 f 在明文 m 上的函数值. 函数加密的安全性要求敌手从密文 c 获取的任何信息只能来自 $f(x)$. Abdalla 等人[68]针对内积的高效计算问题, 基于 DDH 假设提出了一种多输入函数加密方案 (MIFE), 对应的内积函数形如:

$$f((x_1, x_2, \ldots, x_n), y) = \sum_{i=1}^{n}\sum_{j=1}^{\eta_i}\left(x_{ij} y_{\sum_{k=1}^{i-1}\eta_k + j}\right) \tag{8}$$

其中, $|y| = \sum_{i=1}^{n}\eta_i$, n 表示输入个数, η_i 是每个输入向量的长度, 且满足 $\dim(y) = \sum_{i=1}^{n}\dim(x_i)$.

函数加密应用于联邦学习的关键问题是无法高效计算复杂函数. 目前不存在实用函数加密方案能支持高于 2 次的多项式[69], 因此函数加密常被用于一些简单函数的隐私计算, 如聚合操作中服务器对客户端上传参数求和[70], 从而获得比安全多方计算和同态加密更高的效率, 但却无法用于神经网络等复杂模型的计算.

3.4 差分隐私

一个随机算法 $\mathcal{M}$ 具备 (ε, δ)-差分隐私[71], 若对相邻集合 D 和 D', 以及所有 $S \subseteq Range(\mathcal{M})$, 满足

$$\Pr[\mathcal{M}(D) \in S] \leqslant \exp(\varepsilon) \cdot \Pr[\mathcal{M}(D') \in S] + \delta \tag{9}$$

其中, 概率取自对 $\mathcal{M}$ 的随机掷币.

满足差分隐私的算法的输出对数据集中任何特定记录都不敏感, 敌手无法通过输出分布的差异推断一条数据的敏感信息, 因此可用于抵抗成员推理攻击. 差分隐私属于扰动技术, 即在模型训练中的某阶段添加一定的随机噪声, 常见的方法包括高斯机制 (Gaussian mechanism)、拉普拉斯机制 (Laplace mechanism)、二项式机制 (binomial mechanism)、指数机制 (exponential machanism). 根据添加噪声的位置可分为以下 4 类.

(1) 输入扰动: 对训练数据添加噪声.

(2) 算法扰动: 对算法的中间参数添加噪声.

(3) 目标扰动: 对学习算法的目标函数加噪声.

(4) 输出扰动: 对训练结果的输出参数加噪声.

与安全多方计算和同态加密等技术相比, 差分隐私机制的优点是计算复杂度低, 算法实现简单, 便于实际应用: 缺点是输出结果的偏差可能导致模型不收敛, 影响可用性, 特别是对深度学习等复杂模型, 更难平衡模型的可用性和隐私保护. 其次, 引入噪声会破坏模型本身的稀疏性, 影响模型剪枝等技术的应用. 另外, 参数 (ε, δ) 量化了隐私保护的程度, 然而在现实中不具备可解释性, 在现实任务中, (ε, δ) 如何设定并没有客观的指导方法.

差分隐私应用于联邦学习的关键问题是平衡隐私性和可用性, 由于计算高效、部署简单等优势, 近几年差分隐私被广泛用于联邦学习的隐私保护, 然而添加噪声不可避免会影响训练的准确性, 导致模型精度降低甚至不收敛. 而在横向联邦学习中, 若中央服务器是恶意的, 差分隐私也不能完全保护训练过程, 因为当噪声较小时, 用户的训练数据仍然暴露给敌手; 当噪声较大时, 会严重影响模型收敛性[5], 而联邦学习处理的往往是非独立同分布的数据, 本身就面临收敛性问题[72], 差分隐私的引入无疑加剧了这一现象.

3.5 可信执行环境

可信执行环境 (TEE) 是 CPU 中的一块区域, 提供安全隔离执行环境 (secure enclave), 能保证其中数据和代码的机密性、完整性等性质. TEE 是和操作系统并行运行的独立执行环境, 并为其提供安全服务, 其中包含了一组 API 来满足操作系统和 TEE 之间的通讯. 运行在 TEE 中的应用可以访问主处理器和内存的全部功能, 且被保护不遭受来自操作系统的恶意攻击, TEE 中运行的代码具有如下性质[73].

- 机密性. 除非代码本身公布一些消息, 否则其执行状态是秘密的.
- 完整性. 除非代码接受显式输入, 否则其执行过程不受影响.
- 可验证性. TEE 可以向远程用户证明一段特定二进制代码正在运行, 并处于何种状态.

相比于密码学技术, TEE 的效率更高. 然而, 目前将 TEE 技术应用于联邦学习存在一些挑战: (1) 技术本身存在缺陷, 易遭受侧信道攻击和微架构瞬态执行攻击, 使得可信环境内数据机密性受到影响[74]; (2) 受内存限制, 当程序数量和规模增大时, 为保证页面换进换出时的安全性, 系统开销明显增大[75], 利用 TEE 执行全流程的计算是不现实的; (3) 只能访问 CPU 资源, 无法保证 GPU 计算的安全性[76], 影响了 GPU 在学习任务中加速计算的应用.

3.6 技术对比与分析

上述技术具有不同的特点, 在联邦学习中的应用也有各自的优劣势, 适用于不同的隐私保护场景. 例如, 安全多方计算、同态加密、函数加密 3 类加密技术通过隐藏节点间的传输数据, 阻止敌手窃取其他节点的通讯消息, 限制了敌手获取额外信息的能力, 遏制了敌手知识的增长途径, 然而无法阻止一些合法信息的暴露, 如每轮聚合结果或最终模型, 敌手仍能从这些数据推断信息, 因此这些技术常用于防范恶意客户端的攻击; 差分隐私技术通过对传输数据添加扰动, 使得敌手无法通过分析中间结果或最终模型判断特定样本是否属于训练集, 然而无法阻止敌手窃取用户通讯内容获取额外信息, 因此该技术常用于防范恶意用户和分析者的攻击; 可信执行环境保证了运行代码的机密性、完整性和可验证性, 可以防止服务器在聚合数据时篡改数据或计算逻辑, 然而本身空间受限, 适用于数据聚合等相对简单的计算, 常被用于防范恶意服务器的攻击, 表 3 中对这些技术的特点进行了对比总结.

表 3 联邦学习中的典型隐私保护技术对比

名称	技术核心	优点	缺点
安全多方计算	隐藏输入的合作计算	隐私性好, 适用面广	效率较低, 通信开销大
同态加密	密文计算	隐私性好, 通信开销小	效率较低, 计算存储开销大
函数加密	隐藏输入的特定函数计算	效率较高, 通信开销小	支持的函数复杂度受限
差分隐私	添加噪声, 随机应答	效率高, 部署便捷	影响模型精度和收敛性
可信执行环境	硬件保护代码安全执行	效率高, 无需诚实节点假设	空间受限, 易受侧信道攻击

上述隐私保护技术的技术核心及优缺点各不相同, 研究者应根据实际场景选用合适的技术. 事实上, 隐私保护技术并非彼此独立水火不容, 可将一种技术用于其他技术的优化, 例如, 用 PHE 帮助 MPC 在无需第三方的情况下生成乘法三元组[54,77]; 也可将不同技术相结合, 取长补短, 来设计理想的隐私保护方案, 例如, 将加密技术和差分隐私相结合尝试同时解决效率和精度的问题[70,78], 虽然这一类方案并未完全成熟, 但多技术融合是隐私保护未来的一种发展趋势.

4 联邦学习中的隐私保护方案

目前, 许多学者基于上述技术探索了联邦学习中的隐私保护方案, 可按联邦学习类型、隐私保护技术、参与节点架构、学习模型等进行划分, 如表 4 所示. 为保证调研的全面性, 本文选取了 60 篇左右涵盖上述各种分类的代表性文献[79−126], 主要包括: 发表于著名期刊和会议的研究; 在 arXiv 等平台发布的近期热点文献; 受人工智能或信息安全社区广泛认可的报告或手稿; 应用于医疗、金融等领域等现实场景的研究.

本文依据作用阶段、防护策略及所用技术, 将这些隐私保护方案分为 6 大类: 安全聚合机制、安全多方机制、同态加密机制、可信硬件机制、安全预测机制、模型泛化机制. 首先, 前 4 种机制作用于训练过程, 安全预测机制作用于推理过程, 模型泛化机制可令两个阶段同时受益. 其次, 训练过程的 4 种机制主要区别在于对数据的保护策略, 安全聚合机制遵循用户数据不出本地的核心思想, 通过交换中间参数进行训练. 其余 3 种机

制则允许数据加密后传出本地, 而其特点和应用场景又因采用的技术产生区分: 安全多方机制允许数据通过安全的方式共享以进行模型训练, 同时保证其在计算过程中的隐私性, 直至计算结果公布; 同态加密机制利用密文计算技术, 保证数据加密后的隐私性和计算正确性; 可信硬件机制则通过硬件层面的安全保证计算时数据不被破解.

表 4 联邦学习中的典型隐私保护方案分类

分类			典型方案			
按联邦学习类型划分	参与节点类型	跨设备	[26]	[70]	[88]	[96]
		跨筒仓	[95]	[96]	[110]	[111]
	数据分布形式	横向联邦学习	[78]	[99]	[100]	[105]
		纵向联邦学习	[20]	[98]	[108]	[109]
		横/纵向联邦学习	[65]	[66]	[101]	
		联邦迁移学习	[104]	[110]		
按防护过程划分	训练过程防护		[4] [81] [96]	[26] [84] [103]	[70] [86] [112]	[78] [88] [113]
	推理过程防护		[94] [119]	[97] [121]	[100] [124]	[115] [125]
按学习模型划分	贝叶斯		[100]	[116]		
	支持向量机		[85]	[111]		
	Logistic回归		[85]	[88]	[109]	
	决策树		[78]	[111]	[116]	[121]
	k-means聚类		[98]	[99]	[111]	
	神经网络		[4]	[81]	[82]	[87]
	卷积网络		[70]	[90]	[94]	[96]
按防护机制划分	安全聚合机制	基于数据加密	[4]	[26]	[80]	[81]
		基于数据扰动	[82]	[84]	[85]	[86]
		结合加密与扰动	[70]	[78]	[87]	
	安全多方机制	外包计算架构	[77] [91]	[88] [92]	[89] [94]	[90] [95]
		去中心化架构	[81]	[103]	[105]	
	同态加密机制		[20]	[108]	[109]	[110]
	可信硬件机制		[111]	[112]	[113]	
	安全预测机制	基于MPC	[94]	[114]	[115]	
		基于HE	[93]	[100]	[116]	[118]
		结合MPC和HE	[90]	[97]	[121]	[122]
		基于TEE	[124]	[125]	[126]	
	模型泛化机制		[6]	[7]		

4.1 安全聚合机制

安全聚合机制是由第 1.2.2 节典型框架衍生的, 进一步加强隐私保护的模型训练方法, 也是目前横向联邦学习主流的隐私保护机制. 其典型架构为一个中央服务器和多个客户端, 服务器负责调度整个训练流程并维护全局模型, 期间每个客户端利用本地数据集对全局模型进行训练, 通过梯度下降等优化算法得到新的梯度或模型参数, 然后由服务器执行数据的安全聚合, 如图 2 所示.

由第 2.2 节可知, 数据聚合需要节点间的参数传递, 往往成为敌手的突破口, 因此安全聚合机制基于目前联邦学习的典型模式, 对聚合过程进行安全加固, 通过数据加密和扰动等手段防止中间参数泄露隐私, 相关方案的总结如表 5 所示.

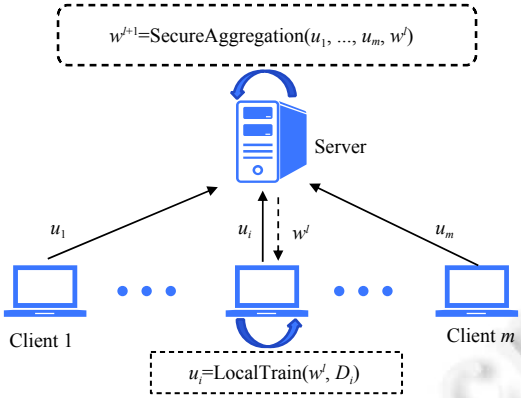


图 2 安全聚合机制

表 5 安全聚合机制的典型方案

文献	方案类型	加密			扰动			聚合对象	学习模型
		核心技术	加密角色	解密角色	核心机制	DP类型	作用位置		
[26]	加密	OTP, SS	客户端	服务器	N/A			权重	N/A
[79]		SS, PKI	客户端	服务器				权重	N/A
[4]		同态加密	客户端	客户端				梯度	神经网络
[80]		HE, CRT	客户端	客户端				梯度	MLP
[81]		对称密码	客户端	客户端				权重	神经网络
[82]	扰动	N/A			高斯机制	中心化	输出扰动	权重	神经网络
[84]					二项式机制	本地化	输入扰动	梯度	神经网络
[85]					拉普拉斯机制	本地化	目标扰动	梯度	感知机, SVM, LR
[86]					高斯机制	本地化	输入扰动	权重	MLP
[78]	混合	同态加密	客户端	客户端	高斯机制	本地化	输入扰动	质询数据	DTs, CNN, SVM
[87]		同态加密	客户端	客户端	拉普拉斯机制	本地化	输入扰动	梯度	神经网络
[70]		函数加密	客户端	服务器	高斯机制	本地化	输入扰动	权重	CNN

4.1.1 基于数据加密的安全聚合

安全聚合的一种方式是加密客户端上传的数据, 服务器对密文进行聚合, 只向服务器暴露聚合的结果, 从而减小暴露个体隐私的风险. 相关技术有安全多方计算、同态加密、函数加密、公钥加密等.

Bonawitz 等人[26]基于半诚实服务器的假设构造了一种安全聚合协议 SecAgg, 用以计算多个更新值之和. 其核心思想是使用一次一密 (one time pad) 为每个输入加上混淆值. 任意两个客户端 $(C_u, C_v), u < v$ 间协商随机向量 $s_{u,v}$, 记 C_u 的输入向量为 x_u, 则其计算:

$$y_u = x_u + \sum_{C_v \in \mathcal{C}, v>u} s_{u,v} - \sum_{C_v \in \mathcal{C}, v<u} s_{v,u} (\mathrm{mod}\, R) \tag{10}$$

并将 y_u 发送至服务器 S, S 计算:

$$z = \sum_{C_u \in \mathcal{C}} y_u = \sum_{C_u \in \mathcal{C}} \left(x_u + \sum_{C_v \in \mathcal{C}, v>u} s_{u,v} - \sum_{C_v \in \mathcal{C}, v<u} s_{v,u} \right) = \sum_{C_u \in \mathcal{C}} x_u (\mathrm{mod}\, R) \tag{11}$$

从而求得正确的聚合结果, 并且 S 无法由 y_u 推知 x_u. 该方法会导致两个问题: (1) 客户端间协商 $s_{u,v}$ 需要的通信复杂度为 $O(|\mathcal{C}| \times |x|)$; (2) 任一终端 C_u 交换完 $s_{u,v}$, 但在向中心提交 y_u 前离线, 都会导致聚合结果出错. 对此, 作者引入了伪随机数生成器、Shamir 秘密共享和双重混淆等方法来解决衍生的问题. 该协议具有容忍节点掉线, 计算复杂度低, 以及 RO 模型下抵抗恶意敌手等优点. 然而, 该方案中密钥协商和秘密共享及恢复会带来巨大开销,

Mandal 等人[79]对此进行了优化, 引入非共谋的密钥提供者实现非交互密钥生成, 引入正则图和相邻用户的概念, 只在相邻用户间协商掩码, 从而实现了高效聚合.

Phong 等人[4]基于 AHE 提出了一种隐私保护的聚合方法, 所有客户端掌握一对 AHE 方案的公私钥, 并对服务器保密, 同时每个客户端与服务器创建 TLS/SSL 安全信道, 用以保证密文完整性和隐私性. 训练开始前, 由一个客户端将初始化模型参数加密后发送给服务器, 训练开始后, 每个客户端下载加密权重参数并解密, 然后用本地数据进行模型更新, 并将得到的梯度加密发送给服务器, 服务器直接用密文梯度对全局模型进行更新. 该方案在半诚实服务器的假设下是安全的, 同时不会导致模型精度下降. 类似的, Zhang 等人[80]基于 HE 和中国剩余定理等技术设计了抵抗恶意服务器的隐私保护方案, 在保证聚合正确性和隐私性的同时, 利用双线性聚合签名技术 (bilinear aggregate signature) 提供数据可验证性, 可纠察并阻止服务器伪造聚合结果从而影响模型正常更新.

Phuong 等人[81]提出了一种基于对称密钥的隐私保护方案, 在服务器的协调下, 客户端间安全地传输权重并更新模型. 该方案中无需使用近似函数替代激活函数, 并且在半诚实敌手假设下能抵抗合谋攻击, 只要一个客户端是诚实的, 即使服务器与其他客户端共谋, 也无法恢复该客户端的数据. 其核心思想是利用对称密码加密权重并传输, L 个客户端 $\{C_i\}_{i\in[L]}$ 共享一个密钥 K, 并对服务器 S 保密. 当 C_i 收到 S 发送的加密权重 $Enc_K(W)$, 利用本地数据 (X,Y) 更新权重 $W'\leftarrow W-\alpha\cdot\delta J(W,X,Y)/\delta W$, 并上传 $Enc_K(W')$ 至 S; S 收到后随机或按某既定规则将其发送给另一个客户端 $C_j(j\neq i)$. 可见该方案中各节点的训练是串行的, 一定程度上影响整体效率.

4.1.2 基于数据扰动的安全聚合

安全聚合的另一种方式是利用差分隐私, 对客户端的数据添加扰动, 从而使敌手无法识别特定客户端贡献的数据, 根据添加噪声的位置一般分为中心化模型和本地化模型两种.

中心化模型中每个客户端将它们未受保护的数据发送给一个可信的中央服务器, 服务器在聚合这些数据时添加噪声. Geyer 等人[82]提出一种保护客户端级别隐私的聚合方案, 利用高斯机制在中心平均客户端的上传参数时添加噪声, 同时利用时刻累计技术 (moments accoutant)[83]保证当个体贡献过高时及时停止训练, 从而保护个体隐私. 中心化模型可有效防止客户端和用户等角色的推断攻击, 然而由于服务器能看到客户端上传的准确数据, 这类方案无法抵抗恶意服务器的攻击.

本地化模型中每个客户端先对数据添加噪声, 再将其发送给一个不可信的中央服务器进行聚合. Agarwal 等[84]提出了一种通信高效且满足差分隐私的分布式 SGD 算法 cpSGD, 考虑客户端不信任服务器的场景, 利用二项式机制对上传梯度进行扰动, 从而使服务器的输出模型满足差分隐私. Choudhury 等人[85]考虑联邦学习在医药领域的应用, 利用差分隐私完成含隐私数据的二分类任务, 在客户端本地训练时, 先对目标函数加上扰动然后优化模型. 作者将该方法用于感知机、支持向量机和逻辑回归 (logistic regression, LR) 这 3 种模型. Wei 等人[86]同样提出了一种基于本地化差分隐私的联邦学习框架 NbAFL, 同时分析了模型的收敛性并得出了以下结论: (1) 模型收敛性与隐私保护强度存在矛盾, 成负相关; (2) 固定隐私保护强度, 增加参与方的数目可以提高模型收敛表现; (3) 给定隐私保护强度, 对于模型收敛性存在最优的训练轮数.

4.1.3 结合加密与扰动的安全聚合

上述两类方案存在各自的缺陷, 基于数据加密的方案效率较低, 且无法有效抵抗模型 API 处发起的推断攻击, 而基于数据扰动的方案当噪声方差较小时仍会暴露原数据的信息, 方差较大时导致模型可用性丧失, 特别是参与方数目多, 而数据量小时, 精度下降明显. 对此, 一些研究人员提出了结合加密与扰动的安全聚合方案.

Truex 等人[78]利用 DP 和 AHE 提出一种联邦学习方案, 为决策树 (decision trees, DTs), 卷积神经网络和支持向量机 3 种模型设计了安全聚合算法, 服务器根据学习模型向客户端质询相关的数据形式, 如对于决策树, 服务器请求满足特定条件的样本个数, 对于神经网络, 服务器请求当前模型权重. 客户端在本地扰动数据, 再通过门限版本的 (n,t)-Paillier 加密来聚合扰动后的数据. 门限加密允许不少于 t 个客户端进行密文解密, 因此对相同的隐私预算 ε, 每个客户端添加噪声的方差可降为原来的 $1/(t-1)$, 从而提高了模型准确性. 类似的, Hao 等[87]利用 DP 和 AHE 设计了神经网络的安全聚合方案, 在半诚实服务器与多个客户端合谋时, 仍能保护训练数据隐私.

Xu 等人[70]结合 MIFE 和 DP, 提出了一种高效联邦学习框架 HybridAlpha, 主要包括 5 种算法: Setup、PKDistri-

bute、SKGenerate、Encrypt、Decrypt; 和 3 种角色: 可信第三方 (trusted third party, TTP)、客户端、聚合服务器. 协议开始时, TTP 运行前 3 个算法进行初始化和函数密钥的分发, 然后每个客户端利用 *Encrypt* 加密本地的模型权重, 最后聚合器运行 *Decrypt* 解密得到所有加密权重的均值. 为了抵抗推理攻击, 作者为 TTP 添加了一个抗推理模组, 同时客户端加密本地数据前需添加噪声. 相较于文献 [78], 该方案在不影响模型表现的情况下, 训练时间平均减少了 68%, 数据传输量平均减少了 92%. 文献中使用的 MIFE 技术只支持线性函数运算, 因此该方案只能进行诸如求和等线性聚合运算.

4.2 安全多方机制

安全多方机制是指参与方通过安全多方计算、同态加密等技术直接构建一个多方计算协议, 共同训练机器学习模型的方法. 其关键在于为学习算法中每个底层算子选取合适的密码学工具, 并针对性地进行优化. 根据参与节点架构区分, 本文将目前基于安全多方机制的联邦学习训练方案分为两类: 外包计算架构和去中心化架构. 其中, 外包计算架构中客户端作为数据拥有者, 将学习任务外包给服务器, 服务器作为计算节点进行模型训练; 去中心化架构中, 参与方既是数据拥有者也是计算执行者, 在无可信第三方协助的情况下完成训练任务.

4.2.1 外包计算架构

外包计算架构中数据拥有者将训练集通过秘密共享技术安全地发布至多个计算节点, 由计算节点共同完成训练任务. 其典型架构为 N ($N \geqslant 2$) 个服务器和 m 个客户端, 如图 3 所示, 训练开始前, 客户端将本地数据集秘密共享至 N 台服务器, 然后服务器间基于共享份额执行 MPC 协议进行训练. 整个流程中, 客户端完成数据共享后无需参与训练, 由服务器完成主体计算任务, 在服务器不共谋和半诚实敌手的假设下, 单一服务器无法由本地份额获取训练数据的相关信息.

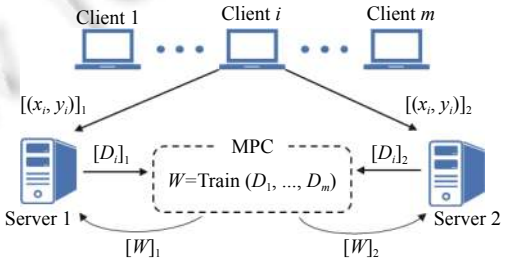


图 3 外包训练机制

由第 3.1 节可知, 对于不同类型的计算, 应选取合适的电路表示和数据共享方法来减少额外开销. Demmler 等人[77]提出了一种两方安全计算框架 ABY, 同时支持运算共享, 布尔共享以及 Yao 共享 3 种数据共享方式. 文中利用扩展不经意传输 (OT extension) 进行密码学操作的高效预计算, 并设计了 3 种共享份额的相互转换方法, 显著提高了计算效率. 作者在隐私集合求交、生物特征匹配和模幂运算 3 种应用上证明了该混合协议的高效性. Mohassel 等人[88]在 ABY 的基础上, 设计了隐私保护机器学习系统 SecureML, 该系统基于 2-服务器模型, 包含两阶段协议, 在线阶段服务器间依据训练算法对份额进行计算, 离线阶段通过 OT、LHE (linearly HE) 等技术生成 Beaver 三元组. 为对系统进行优化, 作者在乘法中对浮点数进行截断并表示为有限域上的整数; 设计了新的线性激活函数; 将训练数据向量化, 从而降低计算复杂度. 该系统在线性回归、logistic 回归以及局域网下的神经网络训练中具备较高效率. 然而, 由于服务器间交互频繁、通信量大, 广域网下神经网络的训练暂未达到实用标准.

ABY 提供了 3 种数据共享方式以提高面对不同计算时的执行效率, 但没有友好的编程接口, 造成了编码人员和密码学者间的鸿沟. Chandran 等人[89]针对这一问题提出了一种高效易编程的 2PC 框架 EzPC, 用户无需关心密码学层面的细节, EzPC 编译器会根据高层运算符的运算代价, 为不同的子运算自动选择合适的电路表示. 同时, 作者使用安全代码划分 (secure code partitioning) 的技术解决面对复杂函数时内存容量不够的问题.

模型精度是衡量学习算法的重要指标, SecureML 为了计算效率, 在训练神经网络时使用一种线性分段函数替

代原有的激活函数, 这种方法会导致一定的精度损失. Liu 等人[90]基于 2PC、HE 和 SIMD 等技术设计了一个 2PC 框架 MiniONN, 提出了一种茫然神经网络 (oblivious neural network) 技术, 不改变神经网络的结构, 而是为一些基本运算设计了安全协议, 包括线性变换、常见的激活函数、池化操作等. Rouhani 等人[91]提出了一个可扩展及可证明安全的深度学习系统 DeepSecure, 主要基于 GC 执行深度学习中的计算, 包括各种非线性函数, 减少精度损失. 并针对深度学习的特点用一个预处理步骤对 GC 进行优化, 避免不必要的计算和通信开销.

此类工作的另一个研究重点是提高训练效率. SecureML 中为提高离线阶段的效率, 引入了可信第三方帮助生成 Beaver 三元组. 借鉴此思想, Riazi 等人[92]基于 OT、GC、GMW、SS 等技术提出一个混合安全计算框架 Chameleon, 借助可信第三方进行 OT 预计算, 生成乘法三元组, 以及优化向量点乘运算. 实验表明 Chameleon 的运行效率比 CryptoNets[93]提升了 133 倍, 比 MiniONN 提升了 4.2 倍. Agrawal 等人[94]则从机器学习和安全计算两个角度对训练流程进行优化, 基于 GC、COT (correlated OT) 等技术提出了一个两方计算框架 QUOTIENT, 用以训练包含全连接层、卷积层和残差层的深度神经网络, 作者在安全计算框架中实现了正则化及动态步长, 进一步提升了模型精度. 对比 SecureML, QUOTIENT 的模型精度提升了约 6%, WAN 模式下的训练效率提升了超过 50 倍. 然而相较于实际应用需求, 该方法中 CNN 训练仍然较慢, 且会产生较大的通信负担.

一些研究也对参与服务器的数目进行了拓展. Mohassel 等人[95]提出了三服务器计算协议 ABY^3, 在 3PC 场景下实现了共享十进制数乘法和共享份额转换, 并进行相关优化. 例如, 提出延迟重共享技术减小通信复杂度, 基于广义三方 OT 计算分段多项式函数等. Wagh 等人[96]针对神经网络训练提出了一种三方安全计算框架 SecureNN, 文中为 CNN 中的常见计算分别设计了安全计算协议, 包括线性运算、卷积、ReLU、最大池化层、正则化等, 这些运算可被高效组合形成复杂网络. 作者设计了新协议解决 Yao 共享和 GC 计算带来的高额通信开销, 实验表明, 相比于 SecureML、MiniONN、Gazelle[97]、Chameleon 等系统, SecureNN 的运行效率提升了 6–113 倍.

4.2.2 去中心化架构

去中心化架构中所有参与方既是数据拥有者也是计算执行者, 且无需可信第三方, 其架构如图 4. 外包计算方案中往往需要引入第三方帮助加速计算[88,92], 故需添加第三方的可靠性假设, 而现实应用中其合法性和可靠性难以保证, 可能带来额外风险, 因此设计仅依赖参与方自身的去中心化学习方法, 是联邦学习的一个重要研究方向.

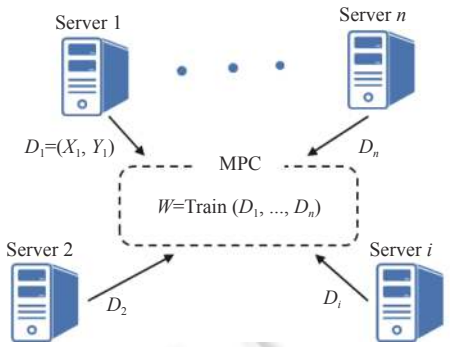


图 4 去中心化训练机制

数据挖掘领域中已有学者研究去中心化架构的模型学习方法, 参与方各自拥有隐私数据, 利用安全多方计算合作进行聚类、分类等任务. 针对纵向划分的数据集. Vaidya 等人[98]提出了一种保护隐私的 k-means 聚类算法, 基于 MPC、HE、安全置换算法 (secure permutation algorithm) 设计了一系列子算法, 如门限检查、计算最近簇等, 参与方在不暴露隐私的情况下, 能对数据进行距离比较等运算, 该方案在半诚实敌手模型下抵抗隐私泄露. 但该方案需要较高的计算开销且在大数据集的情况下可扩展性不强, 文中无实验证明其高效性. 针对横向划分的数据集, Gheid 等人[99]认为引入密码学原语会降低聚类算法的性能, 对此设计了一种专用 MPC 方案, 用一个多方加和协议

来安全地求均值, 有效提高了协议执行效率. Prasad 等人[100]针对数据挖掘任务中用户因担心隐私泄露不提供正确数据的问题, 提出了一种保护隐私的朴素贝叶斯分类器, 解决完全分布式环境下对离散和连续数据的分类问题, 并通过 AHE 保护数据传输过程中的隐私安全. 针对横向和纵向两类联邦学习任务, Samet 等人[101]为后向传播 (BP) 算法和极限学习机 (ELM) 设计了安全协议, 根据不同的参与方数目, 作者在两方计算中使用 HE, 而为多方计算设计了专用子协议. 作者基于文献 [102] 中的安全标量积协议, 移除了对可信第三方的需求, 并将熵函数替换为分段线性函数, 以便在 2PC 协议中高效计算. 这一类方案在目标场景下能获得较高的效率和精度, 然而处理的模型复杂性不高, 通用性较差, 无法使用与神经网络等复杂模型.

加强敌手假设会导致额外的计算和通信开销, 因此目前大多基于 MPC 的隐私保护方案只能抵抗半诚实敌手, 而 Zheng 等人[103]针对线性模型提出了一种抵抗恶意敌手的学习系统 Helen, 作者将该系统适用的场景称为竞争合作学习 (coopetitive learning), 多个团体合作训练一个模型, 但不披露自己的数据, 同时每一方都可能偏离协议侵害他人隐私. Helen 使用交替方向乘子法 (ADMM) 替代 SGD, 显著减小了 MPC 的同步操作次数; 使用零知识证明 (zero-knowledge proof, ZKP) 保证训练过程中参与方始终使用同一个数据集未进行篡改; 利用奇异值分解将数据降维, 从而避免 MPC 中矩阵求逆等昂贵的运算. 文中用基于 SGD 和 SPDZ 的训练方法作为基线, 在每方拥有 10K 个数据点, 90 个特征的四方计算中, 对照方法耗时约 3 个月, 而 Helen 耗时不到 3 小时. Sharma 等人[104]利用 SPDZ 提出了一套安全且高效的联邦迁移学习方案, 在两方间进行知识迁移, 该方案可扩展至多方情形, 且保证即便存在大多数恶意敌手的情况, 仍能保护诚实节点的数据隐私.

还有一类方案并不直接应用密码工具, 而是通过顺序计算并由安全信道传输权重的方式达到相似效果, 敌手恢复客户端梯度或样本的难度相当于求解 NPC 问题. Phuong 等人[81]针对 MLP、CNN 等模型提出了一种去中心化联邦学习方法, 该方案中假设所有参与方间有 TLS 安全信道, 一个参与方用本地数据训练全局模型, 并将模型权重安全传输至下一节点, 按此方式不断更新全局模型, 参数的传递顺序可以事先约定, 也可以随机选取. 由于每个参与方本地训练需要多个 mini-batch, 因此极端合谋情况下, 窃取唯一诚实节点梯度的问题相当于求解子集和问题 (subset sum problem). 相似的, Chang 等人[105]将顺序训练并传递权重的方式用于医疗领域的图像识别任务.

4.3 同态加密机制

同态加密机制是指利用同态加密技术保证参与方间只进行密文传输的隐私保护方法. 由第 3.2 节可知, 现有的同态加密方案不易直接应用于诸如跨设备横向联邦学习等涉及大规模节点的场景. 目前同态加密机制主要用于两方的纵向联邦学习和联邦迁移学习.

对于纵向联邦学习, 一般假设参与方是半诚实的, 同态加密机制一般分为两个步骤.

(1) 隐私实体匹配[106,107]. 参与双方 A 和 B 先找出具有相同 id 的样本对象, 以确保训练开始前数据集包含的样本对象全部匹配. 该过程除了双方匹配的数据集外, 不应泄露其他信息

(2) 加密模型训练. 参与双方通过同态加密技术加密和交换中间结果, 用于计算梯度, 具有标签的一方还需要计算损失. 显然, 为保护各自数据隐私, 双方不能共享同一密钥对, 因此该过程一般需要引入一个可信第三方 C 创建和分发密钥, 并协助中间结果的交换. 本地加密梯度和损失计算完成后, A, B 双方加上一个加密的随机掩码再上传给 C, 防止其解密结果并窃取信息. 最后 C 进行解密将混淆后的梯度明文发回, A 和 B 去除掩码得到真实梯度, 据此更新模型.

根据上述流程, Yang 等人[20]提出了一种安全联邦线性回归算法, Cheng 等人[108]提出了安全联邦提升树算法, 两种方法训练出的模型都是无损的, 与集中学习场景下的算法具有相同准确度.

由于同态加密本身的性质, 面对非线性模型时, 一般需要对计算的函数进行多项式近似. Hardy 等人[109]基于 AHE 提出了一种纵向联邦学习算法, 两个数据持有者 A, B 在服务器 C 的协调下, 对数训练数据进行实体匹配 (entity resolution) 并训练 logistic 回归二分类模型 $\theta \in \mathbb{R}^d$, 该算法抵抗半诚实敌手, 且精度与明文训练相同. 下面对训练过程的核心——梯度计算进行说明, 假设 A 和 B 已完成实体匹配, 得到纵向划分的共有数据集 $X = [X^A|X^B] \in \mathbb{R}^{n\times d}$, A 掌握标签向量 y. 令 x 为 X 的一行, x_A 中下标表示取 x 中只包含 A 特征的部分, 对 x_B, θ_A, θ_B 同理.

模型在训练集 S 上的平均损失为 $\ell_S(\theta)=\frac{1}{n}\sum_{i\in S}\log(1+\mathrm{e}^{-y_i\theta^{\mathrm{T}}x_i})$, 则对大小为 s' 的 batch $S'\subset S$ 有:

$$\nabla\ell_{S'}(\theta)=\frac{1}{s'}\sum_{i\in S'}\left(\frac{1}{1+\mathrm{e}^{-y\theta^{\mathrm{T}}x}}-1\right)y_ix_i \tag{12}$$

由于 AHE 无法直接计算公式 (12), 作者利用其二次 Taylor 展开作为近似 $\nabla\ell_{S'}(\theta)\approx\frac{1}{s'}\sum_{i\in S'}\left(\frac{1}{4}\theta^{\mathrm{T}}x_i-\frac{1}{2}y_i\right)x_i$, 同时为防止 C 获取梯度信息, A、B 需为加密梯度乘上一个掩码 m_i 如下:

$$[\![\nabla\ell_{S'}(\theta)]\!]\approx\frac{1}{s'}\sum_{i\in S'}[\![m_i]\!]\left(\frac{1}{4}\theta^{\mathrm{T}}x_i-\frac{1}{2}y_i\right)x_i \tag{13}$$

于是, 梯度的安全计算方法如表 6 所示, 该过程中 A, B 之间发送的明文只有模型 θ 和 batch 标号 S', C 能获取加掩码后的梯度 $\nabla\ell_{S'}(\theta)$.

表 6 文献 [109] 中的安全梯度计算算法

阶段	执行方	执行内容
1	C	发送模型参数 θ 给 A
2	A	选择batch $S'\subset S$, 计算 $u=\frac{1}{4}X_{AS'}\theta_A$, $[\![u']\!]=[\![m]\!]'_S\circ\left(u-\frac{1}{2}y_{S'}\right)$, 发送 $\theta, S', [\![u']\!]$ 给 B
3	B	计算 $v=\frac{1}{4}X_{BS'}\theta_B$, $[\![w]\!]=[\![u']\!]+[\![m]\!]_{S'}\circ v$, $[\![z]\!]=X_{BS'}[\![w]\!]$, 发送 $[\![w]\!], [\![z]\!]$ 给 A
4	A	计算 $[\![z']\!]=X_{AS'}[\![w]\!]$, 发送 $[\![z']\!], [\![z]\!]$ 给 C
5	C	拼接 $[\![z']\!]$ 和 $[\![z]\!]$ 得到 $[\![\nabla\ell_{S'}(\theta)]\!]$, 用私钥解密得到 $\nabla\ell_{S'}(\theta)$

对于联邦迁移学习, 参与双方面临样本对象和特征重叠较少的问题, 该场景下, 学习算法的目的是从信息丰富的源域 A 向信息缺乏的目标域 B 迁移知识, 共同建立有效模型并为 B 的样本提供预测标签. 其核心在于: (1) 选取合适的模型生成源域和目标域的隐式表征; (2) 利用合适的预测函数为目标域预测标签. Liu 等人[110]基于 AHE 提出了一种安全的联邦迁移学习框架, 用神经网络生成数据域的隐式表征, 用二阶泰勒多项式近似计算损失函数. 该方法无需第三方参与, A 和 B 各自创建一对 AHE 密钥加密传输的中间结果. 由于最后计算得到的加密梯度仍需对方解密, 同样需要添加一个掩码进行混淆, 防止梯度攻击. 训练完成后, 以同样的技术路线利用选定的预测函数对 B 的样本进行安全标签预测. 该方案利用 HE 的特性在两方间进行迁移学习, 当涉及多方学习任务时, 则需要进一步探索其他方法.

4.4 可信硬件机制

可信硬件机制是指利用 TEE 保证学习算法在不可信环境下安全运行的隐私保护方法. 其架构为一台带 TEE 的中央服务器和多个客户端, 敌手可能控制服务器和客户端, 但无法观察和篡改可信环境的内部状态. TEE 的空间受限, 只能执行有限的代码段, 而在外部执行的代码仍可能受到敌手的监控、推断和篡改, 所以此类方法的关键在于对学习算法进行精心设计和改造, 保证可信硬件能容纳核心代码, 且与内存、硬盘等外部环境的交互不会泄露隐私信息.

Ohrimenko 等人[111]基于 SGX 提出一种允许多方进行联合学习的隐私保护方案, 其基本架构如图 5 所示, 多个数据拥有者各自用不同的密钥加密隐私数据, 然后上传至云端数据中心并共享密钥, SGX 在内部解密且合并数据集, 并执行各方约定的学习算法, 最后输出加密模型. 该文设计了比较赋值、排序等一系列基础模糊算子, 进一步为 SVM、k-means 聚类、矩阵分解、神经网络、决策树 5 种模型提出了模糊化算法, 这些算法执行过程中的内存引用、磁盘访问和网络传输的顺序与隐私数据无关. 因此即使敌手控制了数据中心除 SGX 外的所有硬件, 观察到算法与外部环境的交互内容, 仍然无法推断客户端的输入. 文中作者假设拒绝服务攻击和侧信道攻击不会发生, 从而规避 TEE 本身的脆弱性问题.

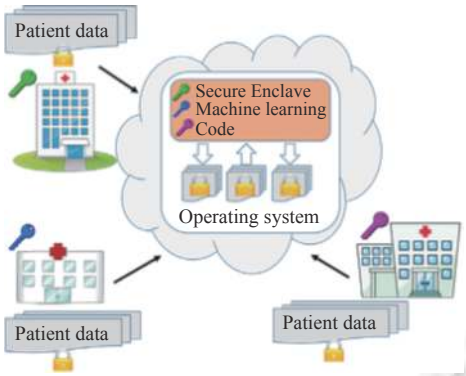


图 5 基于可信执行环境的隐私保护合作学习方案[111]

上述方案本质上是利用加密技术和可信执行环境, 将数据集中到服务器并进行训练, 同时保证全流程的数据隐私不外泄. 而 Lin 等人[112]则基于参数聚合的训练方法做出了改进, 作者认为目前联邦学习中的隐私保护方案普遍存在两个问题: (1) 引入加密技术会显著增加计算和通信复杂度; (2) 引入扰动技术会影响模型精度, 同时妨碍模型剪枝等技术的应用. 对此提出一种基于 SGX 的隐私保护框架 ESMFL, 开始时服务器初始化一个受信任的执行空间, 每个客户端向服务器发送远程证明请求 (remote attestation request), 当服务器证明了本地的软硬件环境后, 两者协商对称密钥. 每个客户端利用本地数据集训练模型得到参数更新, 加密后发送至服务器, 在可信空间内解密并进行聚合, 训练过程中客户端的参数更新仅自身和经验证的 SGX 飞地可见, 有效阻止了敌手窃取隐私. 同时为提高训练效率和减少通信开销, 作者针对本地训练过程提出了一种基于 ADMM 优化算法的剪枝技术, 实验表明在 MNIST 和 CIFAR-10 数据集上, ESMFL 相比 FedAvg 通信开销分别减少了 34.85% 和 15.68%.

由于 TEE 空间受限, 应用于执行更容易受隐私攻击的计算步骤, 以深度神经网络为例, Mo 等人[113]提出一种隐私暴露的量化方法, 分析网络每一层包含多少隐私信息, 其中隐私信息是指某样本集合是否属于训练集. 该文表明最初几层网络往往只记住了样本的总体特征, 而最后几层记住了特定图像的关键特征, 另外, 最后几个卷积层的神经元能泄露更多关于训练数据的信息, 因此作者使用 TEE 保护这些层的计算, 从而抵抗白盒成员推理攻击.

4.5 安全预测机制

安全预测机制是指利用加密和可信硬件等技术隐藏推理过程中用户输入数据, 从而保护用户隐私的方法. 当训练结束得到可用模型后, 一个重要应用场景是将模型部署至云端, 由服务提供商向用户提供预测服务 (prediction-as-a-service, PaaS). 该场景下, 既要保护推理过程的计算正确性, 又要防止用户数据被服务提供商窃取.

4.5.1 基于 MPC 的安全预测方法

具备隐私保护的 PaaS 可自然地看作一个安全两方计算的过程, 如图 6. 服务商和用户分别提供模型和待预测数据作为输入, 用户得到预测标签作为输出.

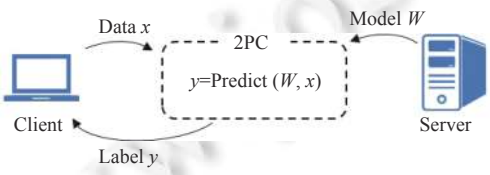


图 6 基于两方计算的安全预测

由于预测的底层算子集合往往是训练算子集合的一个子集, 很多研究在设计基于 MPC 的安全训练协议时, 也实现了安全预测过程. 如前文中的 SecureML[88]支持线性回归、logistic 回归和神经网络的安全预测; QUOTIENT[94]支持深度神经网络的安全预测; Barni 等人[114]基于 HE、OT 和 GC 实现了线性分支程序 (linear branching program, LBP) 的安全计算. Chaudhari 等提出了一种高效 3PC 框架 ASTRA[115], 允许服务商对用户提供预测服务时, 将推理

计算外包给 3 个非共谋服务器, 并可同时抵抗半诚实和恶意敌手. 该方案沿用线下-线上的两阶段 MPC 方法, 且线上阶段具有极高的效率, 在半诚实和恶意敌手假设下, 每个乘法门分别只需传输 2 个和 4 个元素, 具有比 ABY^3 更高的吞吐量.

4.5.2 基于 HE 的安全预测方法

同态加密提供密文计算的特性天然切合安全预测机制的需求. 用户加密待预测数据并上传, 服务商对密数据进行运算并返回加密结果, 用户对结果进行解密从而获取预测标签, 如图 7 所示.

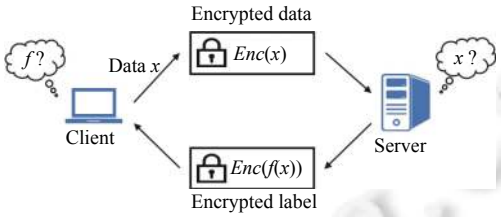


图 7 基于同态加密的安全预测

Bost 等人[116]为超平面决策、朴素贝叶斯和决策树 3 种分类算法设计了安全计算协议, 并可通过 AdaBoost 结合这 3 种分类器, 从而提高预测效果. Dowlin 等人[93]基于 leveled HE 方案 YASHE[117]提出了一种神经网络的密文预测系统 CryptoNets, 在 MNIST 数据集上达到 99% 的准确率. Sanyal 等人[118]基于 FHE 提出了一种加速密文预测的方法 TAPAS, 借鉴神经网络二值化和稀疏化的思想, 同时并行化密文计算, 从而提高推理速度. 类似的, Bourse 等人[119]基于 FHE 提出了一种神经网络密文计算框架 FHE-DiNN, 其计算复杂度与网络深度成线性关系. 该文同样使用了二值神经网络 (binarized neural networks), 并利用 Chillotti 等人[120]提出的 FHE 构造方法, 在 bootstrapping 阶段扩大消息空间, 并使用 sign 函数激活神经元. 相比于 CryptoNets, 该方法由于将网络离散化, 在 MNIST 数据集上的准确率降低了 2.6%, 但大幅提升了数据预测效率.

4.5.3 结合 MPC 和 HE 的安全预测方法

为充分提高协议效率, 学者们往往将 GC、SS、OT、HE 等技术结合起来. 前文中的 MiniONN[90]和 Gazelle[97]都实现了神经网络的安全预测, 对比 CryptoNets, MiniONN 的通信延迟降至 1/230, 数据传输量降至 1/8; Gazelle 的通信延迟降至 1/10000, 数据传输量降至 1/7440. Wu 等人[121]基于 HE 和 OT 实现了决策树和随机森林的安全计算, 可抵抗半诚实和恶意敌手. 该文基于 GC 提出一种决策树的专用计算协议, 对客户端隐藏决策树结构, 计算程序从服务端接收决策树的一个“描述”τ, 从客户端接收一个特征向量 x, 输出 $\tau(x)$. 该方法比通用 2PC 协议高效. 并且, 由于针对决策树计算做了专项优化, 该方法也比解决一般化 LBP 问题的方法[116]更高效. Chen 等人[122]实现了一个抵抗半诚实敌手的两方 k 近邻搜索协议 SANNS, 利用 AHE 计算数据点间距离, 利用 DORAM[123]安全地取回数据点, 利用 GC 实现 top-k 选择算法, 该协议能在包含千万条目的数据集上高效运行.

4.5.4 基于 TEE 的安全预测方法

具备可信执行环境的云端服务器向用户提供远程证明后, 可提供受信任的安全预测服务. Hunt 等人[124]基于 SGX 设计了一套 MLaaS 系统 Chiron, 允许用户向服务商提供数据进行模型训练, 训练过程中, 用户不暴露训练数据, 服务商不暴露训练算法和模型结构. 同时, 为用户提供了模型的黑盒访问权限, 可保证推理过程中用户隐私不泄露. Acs 等人[125]提出了一种安全预测方法, 将机器学习模型部署到客户端, 而非服务端, 从而避免了用户频繁查询产生的网络通信, 同时利用 SGX 保证终端用户不能审查部署模型的具体细节.

Grover 等人[126]为深度学习设计了一套实用的安全预测系统 Privado, 与上述工作不同, 该文假设模型由一个模型所有者提供, 如图 8 所示. 模型所有者将模型对应的二进制代码发送至支持 SGX 的云服务器, 双方进行远程证明并创建安全信道, 模型所有者将加密的权重发送至安全飞地. 然后, 用户同样和云服务器间进行远程证明和信道创建, 用户将加密后的输入发送至服务器进行推理计算, 随后收到加密输出. Privado 在保证用户输入隐私的同时也保证了模型权重不被泄露.

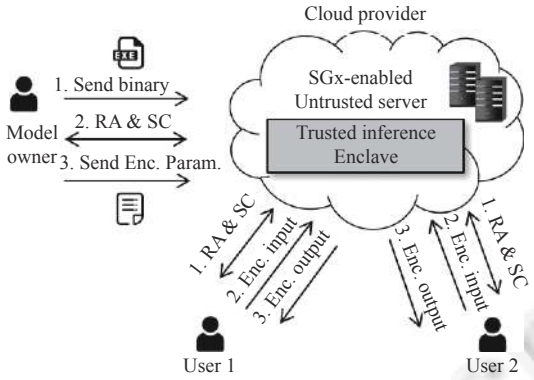


图 8 基于可信执行环境的安全预测[126]

4.6 模型泛化机制

学习模型易受隐私攻击的根本原因是泛化性不强, 训练过程是信息从训练数据向模型转化的过程, 因此模型在某种程度上“记住”了原数据中的相关信息, 特别当过拟合时, 其面对训练数据和非训练数据表现出明显差异[6−8]. 对此, 从模型本身入手, 防止过拟合能有效抵抗推理攻击, 常见的方法有以下 5 种.

(1) L1&L2 正则化[6]: 在损失函数加上惩罚项, L1 正则化添加权重的绝对值之和, L2 正则化添加权重的平方和, 其中正则化参数 λ 是超参数.

$$L_1(x,y) \triangleq \sum_{i=1}^{n}(y_i - h_w(x_i))^2 + \lambda \sum |w| \tag{14}$$

$$L_2(x,y) \triangleq \sum_{i=1}^{n}(y_i - h_w(x_i))^2 + \lambda \sum w^2 \tag{15}$$

然后针对新的损失函数进行参数优化.

(2) dropout[127]: 在神经网络的每轮模型迭代中, 随机丢弃某些神经元及其连接边, 其中每个节点的移除概率 p 是超参数. 由于每轮的模型都包含不同的神经元组合, dropout 可看作一种模型集成方法.

(3) 早停 (early stopping)[128]: 将部分训练集作为验证集, 在训练的同时通过验证集观察模型表现, 当模型表现有下降趋势时立即停止训练.

(4) 数据扩增 (data augmentation)[129]: 通过旋转、平移、放缩、添加噪声等方法增加训练样本, 更大的数据集规模往往意味着具备更强泛化能力的模型.

(5) 模型堆叠 (model stacking)[7]: 将多个不同类型的学习器分层堆叠, 且分别在互不相交的数据集上进行训练. 由于各学习器见过的数据不同, 最终模型在保证预测准确率的同时, 有更低的过拟合倾向.

模型泛化机制在黑盒场景下, 面对半诚实敌手时具有较好的表现. 然而当面临恶意敌手, 或敌手具有白盒攻击权限时, 若不对模型本身或交互中的信息进行保护, 依然容易发生隐私泄露.

4.7 总 结

根据隐私性、高效性、可扩展性、适用场景等方面, 对本节隐私方案的横向对比总结如表 7 所示.

(1) 安全聚合机制允许客户端在本地进行训练, 实现了数据并行, 将算力和存储空间的需求分摊到了各计算节点, 在深度学习等任务中极为高效, 且具有可扩展性高、容忍节点掉线等优点. 然而, 每轮训练中客户端与服务器间的交互更易引起隐私泄露, 因此对中间变量的保护提出了更高的要求. 另外, 要求客户端本身具有一定的计算和存储能力. 此类方案适合包含大规模节点的跨设备横向联邦学习场景.

(2) 安全多方机制通过加密手段隐藏了计算的中间变量, 只暴露最终输出, 具有极强的理论安全性. 外包计算架构的方案由于客户端无需参与训练过程, 因此同样可容忍用户掉线, 且具备高可扩展性. 此类方案中, 秘密共享

的数据可通过 MPC 协议执行任意计算, 适用于任意场景. 然而由于计算压力全部集中于服务器集群, 当训练数据总体规模很大时, 因未能充分利用客户端的计算能力以及密码协议本身的复杂性, 训练效率会显著降低. 而去中心化架构的方案中, 所有参与节点也是计算节点, 计算和通信负担较重, 对节点要求很高, 适合参与节点为大型机构的跨筒仓联邦学习场景.

(3) 同态加密机制应用于两方的纵向和迁移学习场景, 加密参与方之间所有通讯内容, 具有很强的隐私性. 由于参与方的限制以及同态加密技术本身的计算开销, 适合应用于跨筒仓联邦学习, 可扩展性不强.

(4) 可信硬件机制利用 TEE 保证了在不可信服务器上计算的隐私性, 相较于密码学方案效率较高, 由于 TEE 本身空间受限, 目前此类方案不支持多方参与且数据集规模较大的场景, 可扩展性较弱. 且 TEE 本身易受侧信道攻击, 不具备密码学协议的理论安全性.

(5) 安全预测机制可看作一个隐私保护的两方计算场景, 第 3 节中的多数技术都可以应用于此, 其特点主要随应用的技术而产生差异. 此类方案适用于任何场景, 如跨设备联邦学习中输出模型部署到云端向用户提供服务, 或跨筒仓联邦学习中一个机构向其他机构提供预测服务.

(6) 模型泛化机制通过对模型本身或训练方法的改造增强其隐私性, 相比于上述机制, 总体复杂度较低, 效率高, 在实际应用中往往能取得较好的效果, 且泛用性强, 与其他隐私保护技术相兼容.

表 7　隐私保护方案总结

方案类型		隐私保护	方案效率	模型精度	可扩展性	典型适用场景	主要保护对象
安全聚合机制	基于加密	较强	较高	高	较强	跨设备HFL	训练数据
	基于扰动	较强	高	较低	强		成员信息/属性信息
	混合	强	较高	较高	较强		训练数据/成员信息
安全多方机制	外包计算	强	较低	高	强	跨筒仓HFL/VFL/FTL	训练数据/类代表/成员信息/属性信息
	去中心化	较强	较低	较高	较强	跨设备HFL/VFL/FTL	训练数据
同态加密机制		强	较低	较高	弱	跨筒仓VFL/FTL	训练数据/成员信息
可信硬件机制		较强	高	高	较弱	跨筒仓HFL	训练数据
安全预测机制		强	较高	N/A		任意两方推理场景	训练数据
模型泛化机制		较强	高			任意训练场景	成员信息

5　未来挑战及展望

目前联邦学习尚处于研究起步阶段, 不同于传统机器学习中的隐私问题, 新的攻击形式和场景需求对隐私保护提出了更严苛的挑战. 本文结合现有工作中的问题, 指出联邦学习中隐私保护面临的挑战, 并提出未来值得研究的方向.

5.1　平衡隐私保护、模型精度、算法效率的矛盾

保护用户隐私是联邦学习的核心, 然而随着隐私保护程度的增强, 会不可避免地提高学习算法的复杂性, 并引入额外的计算和通信开销, 从而降低模型精度和算法效率. 因此, 如何加强隐私性、可用性和高效性, 同时平衡好三者间的关系, 成为联邦学习隐私保护的一大挑战. 未来可从如下几个方面开展工作.

(1) 从隐私保护技术入手, 解决其内部短板, 并进行针对性优化. 以加密技术为例, 其瓶颈在于计算和通信开销过大影响可用性, 如 MPC 中 OT 和 SS 技术通信复杂度较高, FHE 技术计算复杂度极高, 这些都影响了隐私保护方案的整体效率. 对此可以展开两类研究: 一方面是根据具体的机器学习任务进行技术优化, 如 QUOTIENT[94]利用文献 [130] 中基于整数的训练和推理方法设计对应的 2PC 协议, 从而无需对训练数据进行截断, 同时提高算法效率. 另一方面是对工具本身的优化, 如在 MPC 中设计混合电路从而提高综合计算效率[131]. 事实上, 目前的 MPC 协议未在真正的大规模数据集上运行, 例如, WRK[57]是目前基于混淆电路最高效的协议之一, 然而当电路规模增大时, 该协议的内存消耗会显著提升[132]. 因此密码学工具要真正面向大数据应用, 需要优化和平衡自身的时空复杂度.

(2) 从系统设计入手, 结合多种技术, 弥补技术短板. 隐私保护技术中, 加密技术可以有效保护算法的中间变量, 但不能掩盖数据本身的统计特征; 差分隐私可以抵抗敌手对特定样本的识别, 但作为一种有损运算, 会造成精度损失; TEE 在保证代码执行安全的同时, 具备较高的效率, 然而本身易受各类侧信道攻击. 因此, 如何根据给定场景和具体任务, 有针对性地选用并结合这些技术, 形成一个完备且实用的隐私保护联邦学习系统, 值得进一步研究.

(3) 从模型入手, 从本质上加强模型的隐私保护能力. 在关注训练和推理过程中数据和通信的隐私保护时, 如何提高模型本身的泛化能力, 也是一个重要的研究点. 利用正则化技术防止过拟合, 可以有效减小模型在成员和非成员数据集上表现的差异性, 这一类方法可以兼容第 3 节中任意一种隐私保护技术, 因此值得开展广泛研究.

(4) 从机器学习理论入手, 提高算法效率. 联邦学习对于网络带宽、计算参与方的内存和算力等都提出了较高的要求, 尤其是面临深度神经网络等复杂模型. 目前有一些工作使用模型压缩技术[133−135], 在保证模型精度的同时, 有效减少训练和推理过程的计算开销. 如何将这些技术与隐私保护方案相结合, 提联邦学习系统的整体性能, 也是一个值得研究的方向.

(5) 从应用场景入手, 针对实际需求选取合适的安全假设及对应隐私保护方案. 目前大多隐私保护方案只能抵抗半诚实敌手, 而抵恶意敌手的方案往往需要负担额外的计算和通信代价. 设计系统前须明确应用场景和需求, 合理降低安全假设, 从而减小方案复杂度.

5.2 建立隐私泄露和隐私保护程度的度量标准

联邦学习的隐私攻击和隐私保护方法相互对抗、相互促进, 成螺旋上升的发展趋势, 然而仍未建立起统一的隐私度量标准.

从整体来看, 缺乏对联邦学习系统隐私保护的评估标准, 研究人员无法准确评判设计方案的效果, 用户也无法获知自身在系统内的受保护程度. 目前已有学者展开隐私量化问题的研究[136], 尝试系统化地衡量用户在系统中享有的隐私保护程度, 以及不同技术提供的保护量. 遵循此类工作的思路, 针对联邦学习系统构建统一完善的隐私保护度量标准, 不仅有利于完善系统的评价指标, 也有利于隐私攻击和隐私保护方案的迭代研究.

从局部来看, 缺乏系统内各环节隐私泄露风险的评估体系, 例如, 安全聚合机制需要服务器聚合客户端的上传参数, 添加输入混淆等方法固然能隐藏用户的上传数据, 但服务器仍能观察到每轮的聚合结果, 并据此发掘用户上传参数的统计特征, 甚至发起白盒推断攻击. 如何评估暴露此类中间参数带来的隐患, 需要进一步研究. 另外, 研究人员无法量化引入特定技术对隐私保护的增强程度, 建立完善的隐私度量体系有助于指导隐私保护技术的选择和局部优化.

5.3 研究去中心化架构的联邦学习隐私保护方案

目前的联邦学习算法大多依赖可信或半诚实的第三方, 例如, 安全聚合机制需要中央服务器进行参数聚合; 基于外包计算架构的安全多方机制需要多个服务器运行安全多方计算协议; 同态加密机制需要可信第三方协助加密训练; 可信硬件机制需要支持 TEE 的服务器执行可信计算.

然而这种架构在实际应用中可能出现各种问题: (1) 不存在满足安全假设的可信第三方, 如安全聚合机制中服务器被敌手侵入, 外包架构的安全多方机制中服务器合谋等; (2) 第三方节点故障, 如安全聚合机制中服务器失效, 不正确的全局模型进一步损坏各客户端的本地模型. 因此, 如何在参与方互不信任, 且无第三方协助的情况下完成联邦学习, 是未来研究中的一个挑战.

目前有一些去中心化架构的隐私保护方案已在第 4.2.2 节中进行讨论, 然而, 其中大都只面向小规模参与节点. 文献 [81] 和文献 [103] 理论上允许大规模节点参与学习, 但在实际应用中都面临可扩展性问题, 前者需要安全多方计算和零知识证明, 计算复杂度和通信复杂度高, 当参与方增多时, 效率明显降低; 后者是串行的顺序学习流程, 无法利用并行化提升算法效率, 也无法很好处理大规模节点的场景. 因此, 沿着这些工作进一步增强方案可扩展性, 降低方案复杂度, 是一个可行的研究方向. 另一种思路是引入区块链实现学习流程的去中心化[137], 利用区块链的权威性和防篡改性去除对可信第三方的需求.

5.4 研究面向移动边缘设备的联邦学习隐私保护方案

各类移动边缘设备存储着海量数据, 将这些数据利用起来挖掘有价值的信息, 是一个很有意义的课题, 然而这

些设备大多面临在线时间不稳定、计算和存储能力受限、通讯状况不佳等问题. 如何在这些限制下完成联邦学习, 同时提供隐私保护是个不小的难题. 事实上, 目前学者已对此开展了一些研究, 然而并不足以全面的解决这些问题. 例如, 文献 [26] 设计了容忍用户掉线的隐私保护方案, 只需保证中央服务器的稳定性, 然而仍需要客户端存储完整模型并执行优化算法, 且参与节点间需进行多轮通信, 这对参与节点设定了门槛, 部分不满足条件的终端无法参与到联邦学习中.

目前, 基于外包计算架构的安全多方机制和可信硬件机制有希望解决上述问题, 边缘设备完成隐私数据分享后, 无需承担计算任务, 极大减小了负担. 然而这两种机制都需要保证计算服务器的诚实性和可靠性, 一旦计算服务器腐化或故障, 会导致数据失窃或模型错误. 相较而言, 安全聚合机制和基于去中心化架构的安全多方机制中, 终端真正享有数据自治权, 不将原始数据以任何形式送出本地, 但终端仍需负担计算任务, 因此需进一步研究减小计算和通信复杂度的方法.

总的来说, 在设备能力受限的前提下, 隐私保护技术为联邦学习系统引入了额外的计算和通信开销, 进一步加重了节点负担, 因此, 设计实用化的面向移动边缘设备的联邦学习隐私保护方案, 是未来的一个挑战.

5.5 加强纵向联邦学习和迁移学习的隐私问题研究

目前隐私攻击和隐私保护方案大都针对横向联邦学习, 而纵向联邦学习和联邦迁移学习的相关文献较少. 例如, 安全聚合机制就是横向联邦学习的隐私保护方案, 该场景下所有用户的数据特征都是相同的, 具有一定对称性. 而在纵向联邦学习中, 用户数据形式不对称, 可能只有一方拥有数据标签, 目前并不清楚该用户在隐私攻击中是否具有更强的攻击能力, 或是在隐私保护中是否应受到更强的保护, 迁移学习场景更加剧了这种不对称性. 诸如此类的问题还有很多, 故未来需要加强这两种场景下隐私攻击和隐私保护方案的研究.

5.6 加强图像数据的隐私保护

与数值型的数据集不同, 联邦学习中图像类数据更容易受到隐私攻击, 由第 2.2.1 节可知, 敌手试图获取类代表时, 可利用 GAN 生成与原数据具有相似分布的数据, 当目标数据是用户照片时, 敌手可生成极其相似的图片, 从而识别目标人物, 而对于数值型数据, 复现相似分布的类代表无法达到相同的攻击效果. 因此一些传统的隐私保护技术不能直接用于保护图像数据的隐私, 例如, 由于敌手的目标不是识别和恢复原始数据, 差分隐私等技术无法抵抗此类攻击. 研究新的方法和技术来保护图像数据的隐私, 是很有意义的方向.

5.7 解决参与方的激励问题和建立公平性准则

数据是有价值的, 隐私保护机制保证了参与方贡献数据过程的私密性, 却没有提供相应激励, 特别是在跨设备联邦学习场景中, 参与方不能因贡献数据获得直接的回报, 从而丧失参与的动力. 例如 Google 等公司希望收集用户手机的文本记录用于训练词预测模型, 从长远来看有助于所有用户获得更好的输入体验, 而由于缺乏直接的激励机制, 且参与学习过程本身存在计算、通信和存储开销, 即使数据隐私得以保证, 很多用户仍会拒绝参与联邦学习. 进一步的, 参与节点间的公平性准则有待建立, 在联合学习过程中, 需要准确衡量每个参与方的贡献, 如本地数据的数量和质量, 以及对全局模型精度的贡献度, 并根据参与方贡献给予等比例的回报, 这也有助于促进参与方持续提供高质量的数据. 因此, 在保护用户隐私的前提下, 建立行之有效的激励机制和公平性准则, 是保证用户积极参与联邦学习的关键.

6 总 结

联邦学习的出现有效解决了数据孤岛的问题, 充分挖掘了边缘设备、移动设备中存储数据的价值, 然而敌手可通过隐私攻击获取训练数据的相关信息, 严重威胁了正常的训练和推理过程, 危害参与方的隐私权益, 为联邦学习的系统设计及相关标准的制定带来了巨大挑战.

本文深入分析了联邦学习的定义、特点和分类, 描述了联邦学习系统可能面临隐私攻击的敌手模型和攻击类型, 总结并分析了隐私攻击和隐私保护的最新研究, 对联邦学习中的隐私保护方法进行归纳和抽象, 并指出了现有方案中存在的问题, 探讨了未来的挑战和值得研究的方向. 总之, 在平衡好隐私保护、模型精度和算法效率的前提

下, 如何根据特定应用场景设计有针对性的隐私保护方案, 最小化用户隐私泄露风险, 是一个长期的挑战, 需要持续跟进与研究.

References:

[1] McMahan HB, Moore E, Ramage D, Hampson S, Areas BA. Communication-efficient learning of deep networks from decentralized data. In: Proc. of the 20th Int'l Conf. on Artificial Intelligence and Statistics. Fort Lauderdale: PMLR, 2017. 1273–1282.

[2] Bonawitz KA, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, Kiddon C, Konečný J, Mazzocchi S, McMahan B, Van Overveldt T, Petrou D, Ramage D, Roselander J. Towards federated learning at scale: System design. In: Proc. of the Machine Learning and Systems 2019. Stanford: MLSys.org, 2019.

[3] Zhao B, Mopuri KR, Bilen H. iDLG: Improved deep leakage from gradients. arXiv:2001.02610, 2020.

[4] Phong LT, Aono Y, Hayashi T, Wang LH, Moriai S. Privacy-preserving deep learning via additively homomorphic encryption. IEEE Trans. on Information Forensics and Security, 2018, 13(5): 1333–1345. [doi: 10.1109/TIFS.2017.2787987]

[5] Zhu LG, Liu ZJ, Han S. Deep leakage from gradients. In: Proc. of the Advances in Neural Information Processing Systems. Vancouver, 2019. 14747–14756.

[6] Shokri R, Stronati M, Song CZ, Shmatikov V. Membership inference attacks against machine learning models. In: Proc. of the 2017 IEEE Symp. on Security and Privacy. San Jose: IEEE, 2017. 3–18. [doi: 10.1109/SP.2017.41]

[7] Salem A, Zhang Y, Humbert M, Berrang P, Fritz M, Backes M. ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In: Proc. of the 26th Annual Network and Distributed System Security Symp. San Diego: The Internet Society, 2019.

[8] Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: Proc. of the 2019 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2019. 739–753. [doi: 10.1109/SP.2019.00065]

[9] Ganju K, Wang Q, Yang W, Gunter CA, Borisov N. Property inference attacks on fully connected neural networks using permutation invariant representations. In: Proc. of the 2018 ACM SIGSAC Conf. on Computer and Communications Security. Toronto: Association for Computing Machinery, 2018. 619–633. [doi: 10.1145/3243734.3243834]

[10] Melis L, Song CZ, De Cristofaro E, Shmatikov V. Exploiting unintended feature leakage in collaborative learning. In: Proc. of the 2019 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2019. 691–706. [doi: 10.1109/SP.2019.00029]

[11] Wang ZB, Song MK, Zhang ZF, Song Y, Wang Q, Qi HR. Beyond inferring class representatives: User-level privacy leakage from federated learning. In: Proc. of the IEEE INFOCOM 2019 IEEE Conf. on Computer Communications. Paris: IEEE, 2019. 2512–2520. [doi: 10.1109/INFOCOM.2019.8737416]

[12] Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: Information leakage from collaborative deep learning. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. Dallas: Association for Computing Machinery, 2017. 603–618. [doi: 10.1145/3133956.3134012]

[13] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. Denver: Association for Computing Machinery, 2015. 1322–1333. [doi: 10.1145/2810103.2813677]

[14] Kairouz P, McMahan HB, Avent B, Bellet A, *et al*. Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 2021, 14(1–2): 1–210. [doi: 10.1561/2200000083]

[15] Konečný J, McMahan HB, Ramage D, Richtarik P. Federated optimization: Distributed machine learning for on-device intelligence. arXiv:1610.02527, 2016.

[16] Konečný J, McMahan HB, Yu FX, Suresh AT, Bacon D. Federated learning: Strategies for improving communication efficiency. arXiv:1610.05492, 2016.

[17] Sahu AK, Li T, Sanjabi M, Zaheer M, Talwalkar A, Smith V. On the convergence of federated optimization in heterogeneous networks. arXiv:1812.06127, 2018.

[18] Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. In: Proc. of the Machine Learning and Systems 2020. Austin: MLSys.org, 2020.

[19] Yu H, Yang S, Zhu SH. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 5693–5700. [doi: 10.1609/aaai.v33i01.33015693]

[20] Yang Q, Liu Y, Chen TJ, Tong YX. Federated machine learning: Concept and applications. ACM Trans. on Intelligent Systems and Technology, 2019, 10(2): 12. [doi: 10.1145/3298981]

[21] Tan ZW, Zhang LF. Survey on privacy preserving techniques for machine learning. Ruan Jian Xue Bao/Journal of Software, 2020, 31(7): 2127–2156 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6052.htm [doi: 10.13328/j.cnki.jos.006052]

[22] Dean J, Corrado GS, Monga R, Chen K, Devin M, Le QV, Mao MZ, Ranzato MA, Senior A, Tucker P, Yang K, Ng AY. Large scale distributed deep networks. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1223–1231.

[23] Lin YJ, Han S, Mao HZ, Wang Y, Dally W. Deep gradient compression: Reducing the communication bandwidth for distributed training. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[24] Xing EP, Ho QR, Dai W, Kim JK, Wei JL, Lee S, Zheng X, Xie PT, Kumar A, Yu YL. Petuum: A new platform for distributed machine learning on big data. IEEE Trans. on Big Data, 2015, 1(2): 49–67. [doi: 10.1109/TBDATA.2015.2472014]

[25] Zinkevich MA, Weimer M, Smola A, Li LH. Parallelized stochastic gradient descent. In: Proc. of the 23rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2010. 2595–2603.

[26] Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, Ramage D, Segal A, Seth K. Practical secure aggregation for privacy-preserving machine learning. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. Dallas: Association for Computing Machinery, 2017. 1175–1191. [doi: 10.1145/3133956.3133982]

[27] Leontiadis I, Elkhiyaoui K, Önen M, Molva R. PUDA–privacy and unforgeability for data aggregation. In: Proc. of the 14th Int'l Conf. on Cryptology and Network Security. Marrakesh: Springer, 2015. 3–18. [doi: 10.1007/978-3-319-26823-1_1]

[28] Ghazi B, Manurangsi P, Pagh R, Velingker A. Private aggregation from fewer anonymous messages. In: Proc. of the 39th Annual Int'l Conf. on the Theory and Applications of Cryptographic Techniques. Zagreb: Springer, 2020. 798–827. [doi: 10.1007/978-3-030-45724-2_27]

[29] Bonawitz K, Salehi F, Konečný J, McMahan B, Gruteser M. Federated learning with autotuned communication-efficient secure aggregation. In: Proc. of the 2019 Asilomar Conf. on Signals, Systems, and Computers. Pacific Grove: IEEE, 2019. 1222–1226. [doi: 10.1109/IEEECONF44664.2019.9049066]

[30] Goryczka S, Xiong L, Sunderam V. Secure multiparty aggregation with differential privacy: A comparative study. In: Proc. of the 2013 Joint EDBT/ICDT Workshops. Genoa: Association for Computing Machinery, 2013. 155–163. [doi: 10.1145/2457317.2457343]

[31] Bagdasaryan E, Veit A, Hua YQ, Estrin D, Shmatikov V. How to backdoor federated learning. In: Proc. of the Int'l Conf. on Artificial Intelligence and Statistics. Sicily: PMLR, 2020. 2938–2948.

[32] Bhagoji AN, Chakraborty S, Mittal P, Calo S. Analyzing federated learning through an adversarial lens. In: Proc. of the Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 634–643.

[33] Fung C, Yoon CJM, Beschastnikh I. Mitigating sybils in federated learning poisoning. arXiv:1808.04866, 2018.

[34] Sun ZT, Kairouz P, Suresh AT, McMahan HB. Can you really backdoor federated learning? arXiv:1911.07963, 2019.

[35] Lyu LJ, Yu H, Yang Q. Threats to federated learning: A survey. arXiv:2003.02133, 2020.

[36] Hazay C, Venkitasubramaniam M, Weiss M. The price of active security in cryptographic protocols. In: Proc. of the 39th Annual Int'l Conf. on the Theory and Applications of Cryptographic Techniques. Zagreb: Springer, 2020. 184–215. [doi: 10.1007/978-3-030-45724-2_7]

[37] He YZ, Hu XB, He JW, Meng GZ, Chen K. Privacy and security issues in machine learning systems: A survey. Journal of Computer Research and Development, 2019, 56(10): 2049 –2070 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2019.20190437]

[38] Liu JX, Meng XF. Survey on privacy-preserving machine learning. Journal of Computer Research and Development, 2020, 57(2): 346–362 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2020.20190455]

[39] Liu RX, Chen H, Guo RY, Zhao D, Liang WJ, Li CP. Survey on privacy attacks and defenses in machine learning. Ruan Jian Xue Bao/Journal of Software, 2020, 31(3): 866–892 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5904.htm [doi: 10.13328/j.cnki.jos.005904]

[40] Yaghini M, Kulynych B, Troncoso C. Disparate vulnerability: On the unfairness of privacy attacks against machine learning. arXiv: 1906.00389, 2019.

[41] Geiping J, Bauermeister H, Dröge H, Moeller M. Inverting gradients—How easy is it to break privacy in federated learning? In: Proc. of the 33rd Advances in Neural Information Processing Systems. 2020.

[42] Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, Jolly S, Matuszak M, Ten Haken R, Van Soest J, Oberije C, Faivre-Finn C, Price G, De Ruysscher D, Lambin P, Dekker A. Developing and validating a survival prediction model for NSCLC patients

through distributed learning across 3 countries. Int'l Journal of Radiation Oncology, Biology, Physics, 2017, 99(2): 344–352. [doi: 10.1016/j.ijrobp.2017.04.021]

[43] Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, Dries W, Lambin P, Dekker A. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital–A real life proof of concept. Radiotherapy and Oncology, 2016, 121(3): 459–467. [doi: 10.1016/j.radonc.2016.10.002]

[44] Shokri R, Shmatikov V. Privacy-preserving deep learning. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. Denver: Association for Computing Machinery, 2015. 1310–1321. [doi: 10.1145/2810103.2813687]

[45] Tran NH, Bao W, Zomaya A, Nguyen MNH, Hong CS. Federated learning over wireless networks: Optimization model design and analysis. In: Proc. of the IEEE Conf. on Computer Communications. Paris: IEEE, 2019. 1387–1395. [doi: 10.1109/INFOCOM.2019.8737464]

[46] Costan V, Devadas S. Intel SGX explained. IACR Cryptology ePrint Archive, 2016: 86.

[47] Costan V, Lebedev IA, Devadas S. Sanctum: Minimal hardware extensions for strong software isolation. In: Proc. of the 25th USENIX Security Symp. Austin: USENIX Association, 2016. 857–874.

[48] Yao AC. Protocols for secure computations. In: Proc. of the 23rd Annual Symp. on Foundations of Computer Science (SFCS 1982). Chicago: IEEE, 1982. 160–164. [doi: 10.1109/SFCS.1982.38]

[49] Yao AC. How to generate and exchange secrets. In: Proc. of the 27th Annual Symp. on Foundations of Computer Science (SFCS 1986). Toronto: IEEE, 1986. 162–167. [doi: 10.1109/SFCS.1986.25]

[50] Goldreich O, Micali S, Wigderson A. How to play ANY mental game. In: Proc. of the Nineteenth ACM Symp. on Theory of Computing, STOC. New York: Association for Computing Machinery, 1987. 218–229. [doi: 10.1145/28395.28420]

[51] Ben-Or M, Goldwasser S, Wigderson A. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In: Proc. of the 20th Annual ACM Symp. on Theory of Computing (STOC). Chicago: Association for Computing Machinery, 1988. 1–10. [doi: 10.1145/62212.62213]

[52] Beaver D, Micali S, Rogaway P. The round complexity of secure protocols. In: Proc. of the 22nd Annual ACM Symp. on Theory of Computing. Baltimore: Association for Computing Machinery, 1990. 503–513. [doi: 10.1145/100216.100287]

[53] Bendlin R, Damgård I, Orlandi C, Zakarias S. Semi-homomorphic encryption and multiparty computation. In: Proc. of the 30th Annual Int'l Conf. on the Theory and Applications of Cryptographic Techniques. Tallinn: Springer, 2011. 169–188. [doi: 10.1007/978-3-642-20465-4_11]

[54] Damgård I, Pastro V, Smart N, Zakarias S. Multiparty computation from somewhat homomorphic encryption. In: Proc. of the 32nd Annual Cryptology Conf. on Advances in Cryptology. Santa Barbara: Springer, 2012. 643–662. [doi: 10.1007/978-3-642-32009-5_38]

[55] Beaver D. Efficient multiparty protocols using circuit randomization. In: Proc. of the 11th Annual Int'l Cryptology Conf. on Advances in Cryptology. Santa Barbara: Springer, 1991. 420–432. [doi: 10.1007/3-540-46766-1_34]

[56] Wang X, Ranellucci S, Katz J. Authenticated garbling and efficient maliciously secure two-party computation. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. Dallas: Association for Computing Machinery, 2017. 21–37. [doi: 10.1145/3133956.3134053]

[57] Wang X, Ranellucci S, Katz J. Global-scale secure multiparty computation. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. Dallas: Association for Computing Machinery, 2017. 39–56. [doi: 10.1145/3133956.3133979]

[58] Pinkas B, Rosulek M, Trieu N, Yanai A. PSI from PaXoS: Fast, malicious private set intersection. In: Proc. of the 39th Annual Int'l Conf. on the Theory and Applications of Cryptographic Techniques. Zagreb: Springer, 2020. 739–767. [doi: 10.1007/978-3-030-45724-2_25]

[59] Nair DG, Binu VP, Kumar GS. An improved e-voting scheme using secret sharing based secure multi-party computation. arXiv: 1502.07469, 2015.

[60] Naor M, Pinkas B. Oblivious polynomial evaluation. SIAM Journal on Computing, 2006, 35(5): 1254 –1281. [doi: 10.1137/S0097539704383633]

[61] Sen J. Homomorphic encryption —Theory and application. In: Sen J. Theory and Practice of Cryptography and Network Security Protocols and Technologies. London: IntechOpen, 2013. 1–21. [doi: 10.5772/56687]

[62] Acar A, Aksu H, Uluagac AS, Conti M. A survey on homomorphic encryption schemes: Theory and implementation. ACM Computing Surveys, 2019, 51(4): 79. [doi: 10.1145/3214303]

[63] Li ZY, Gui XL, Gu YJ, Li XS, Dai HJ, Zhang XJ. Survey on homomorphic encryption algorithm and its application in the privacy-preserving for cloud computing. Ruan Jian Xue Bao/Journal of Software, 2018, 29(7): 1830−1851 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5354.htm [doi: 10.13328/j.cnki.jos.005354]

[64] Gentry C. Fully homomorphic encryption using ideal lattices. In: Proc. of the 41st Annual ACM Symp. on Theory of Computing. Bethesda: Association for Computing Machinery, 2009. 169–178. [doi: 10.1145/1536414.1536440]

[65] López-Alt A, Tromer E, Vaikuntanathan V. On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption. In: Proc. of the 44th Annual ACM Symp. on Theory of Computing. New York: Association for Computing Machinery, 2012. 1219–1234. [doi: 10.1145/2213977.2214086]

[66] Mukherjee P, Wichs D. Two round multiparty computation via multi-key FHE. In: Proc. of the 35th Annual Int'l Conf. on the Theory and Applications of Cryptographic Techniques. Vienna: Springer, 2016. 735–763. [doi: 10.1007/978-3-662-49896-5_26]

[67] Boneh D, Sahai A, Waters B. Functional encryption: Definitions and challenges. In: Proc. of the 8th Theory of Cryptography Conf. on Theory of Cryptography. Providence: Springer, 2011. 253–273. [doi: 10.1007/978-3-642-19571-6_16]

[68] Abdalla M, Catalano D, Fiore D, Gay R, Ursu B. Multi-input functional encryption for inner products: Function-hiding realizations and constructions without pairings. In: Proc. of the 38th Annual Int'l Cryptology Conf. on Advances in Cryptology. Santa Barbara: Springer, 2018. 597–627. [doi: 10.1007/978-3-319-96884-1_20]

[69] Marc T, Stopar M, Hartman J, Bizjak M, Modic J. Privacy-enhanced machine learning with functional encryption. In: Proc. of the 24th European Symp. on Research in Computer Security. Luxembourg: Springer, 2019. 3–21. [doi: 10.1007/978-3-030-29959-0_1]

[70] Xu RH, Baracaldo N, Zhou Y, Anwar A, Ludwig H. HybridAlpha: An efficient approach for privacy-preserving federated learning. In: Proc. of the 12th ACM Workshop on Artificial Intelligence and Security. London: Association for Computing Machinery, 2019. 13–23. [doi: 10.1145/3338501.3357371]

[71] Dwork C, Lei J. Differential privacy and robust statistics. In: Proc. of the 41st Annual ACM Symp. on Theory of Computing. Bethesda: Association for Computing Machinery, 2009. 371–380. [doi: 10.1145/1536414.1536466]

[72] Li X, Huang KX, Yang WH, Wang SS, Zhang ZH. On the convergence of FedAvg on Non-IID data. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[73] Subramanyan P, Sinha R, Lebedev I, Devadas S, Seshia SA. A formal foundation for secure remote execution of enclaves. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. Dallas: Association for Computing Machinery, 2017. 2435–2450. [doi: 10.1145/3133956.3134098]

[74] Wu XH, He YP, Ma HT, Zhou QM, Lin SF. Microarchitectural transient execution attacks and defense methods. Ruan Jian Xue Bao/Journal of Software, 2020, 31(2): 544–563 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5979.htm [doi: 10.13328/j.cnki.jos.005979]

[75] Brasser F, Müller U, Dmitrienko A, Kostiainen K, Capkun S, Sadeghi AR. Software grand exposure: SGX cache attacks are practical. In: Proc. of the 11th USENIX Conf. on Offensive Technologies. Vancouver: USENIX Association, 2017.

[76] Wang J, Fan CY, Cheng YQ, Zhao B, Wei T, Yan F, Zhang HG, Ma J. Analysis and research on SGX technology. Ruan Jian Xue Bao/Journal of Software, 2018, 29(9): 2778–2798 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5594.htm [doi: 10.13328/j.cnki.jos.005594]

[77] Demmler D, Schneider T, Zohner M. ABY-A framework for efficient mixed-protocol secure two-party computation. In: Proc. of the 22nd Annual Network and Distributed System Security Symp. San Diego: The Internet Society, 2015.

[78] Truex S, Baracaldo N, Anwar A, Steinke T, Ludwig H, Zhang R, Zhou Y. A hybrid approach to privacy-preserving federated learning. In: Proc. of the 12th ACM Workshop on Artificial Intelligence and Security. London: Association for Computing Machinery, 2019. 1–11. [doi: 10.1145/3338501.3357370]

[79] Mandal K, Gong G, Liu CY. NIKE-based fast privacy-preserving high-dimensional data aggregation for mobile devices. CACR Technical Report, CACR2018-10, Waterloo: University of Waterloo, 2018.

[80] Zhang XL, Fu AM, Wang HQ, Zhou CY, Chen ZZ. A privacy-preserving and verifiable federated learning scheme. In: Proc. of the IEEE Int'l Conf. on Communications (ICC). Dublin: IEEE, 2020. 1–6. [doi: 10.1109/ICC40277.2020.9148628]

[81] Phong LT, Phuong TT. Privacy-preserving deep learning via weight transmission. IEEE Trans. on Information Forensics and Security, 2019, 14(11): 3003–3015. [doi: 10.1109/TIFS.2019.2911169]

[82] Geyer RC, Klein T, Nabi M. Differentially private federated learning: A client level perspective. arXiv:1712.07557, 2017.

[83] Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In: Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security. Vienna: Association for Computing Machinery, 2016. 308–318. [doi: 10.1145/2976749.2978318]

[84] Agarwal N, Suresh AT, Yu F, Kumar S, McMahan HB. cpSGD: Communication-efficient and differentially-private distributed SGD. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 7575–7586.

[85] Choudhury O, Gkoulalas-Divanis A, Salonidis T, Sylla I, Park Y, Hsu G, Das A. Differential privacy-enabled federated learning for

sensitive health data. arXiv:1910.02578, 2019.

[86] Wei K, Li J, Ding M, Ma C, Yang HH, Farokhi F, Jin S, Quek TQS, Poor HV. Federated learning with differential privacy: Algorithms and performance analysis. IEEE Trans. on Information Forensics and Security, 2020, 15: 3454–3469. [doi: 10.1109/TIFS.2020.2988575]

[87] Hao M, Li HW, Xu GW, Liu S, Yang HM. Towards efficient and privacy-preserving federated deep learning. In: Proc. of the IEEE Int'l Conf. on Communications (ICC). Shanghai: IEEE, 2019. 1–6. [doi: 10.1109/ICC.2019.8761267]

[88] Mohassel P, Zhang YP. SecureML: A system for scalable privacy-preserving machine learning. In: Proc. of the 2017 IEEE Symp. on Security and Privacy. San Jose: IEEE, 2017. 19–38. [doi: 10.1109/SP.2017.12]

[89] Chandran N, Gupta D, Rastogi A, Sharma R, Tripathi S. EzPC: Programmable and efficient secure two-party computation for machine learning. In: Proc. of the 4th IEEE European Symp. on Security and Privacy. Stockholm: IEEE, 2019. 496–511. [doi: 10.1109/EuroSP.2019.00043]

[90] Liu J, Juuti M, Lu Y, Asokan N. Oblivious neural network predictions via minionn transformations. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. Dallas: Association for Computing Machinery, 2017. 619–631. [doi: 10.1145/3133956.3134056]

[91] Rouhani BD, Riazi MS, Koushanfar F. Deepsecure: Scalable provably-secure deep learning. In: Proc. of the 55th Annual Design Automation Conf. California: Association for Computing Machinery, 2018. 2. [doi: 10.1145/3195970.3196023]

[92] Riazi MS, Weinert C, Tkachenko O, Songhori EM, Schneider T, Koushanfar F. Chameleon: A hybrid secure computation framework for machine learning applications. In: Proc. of the Asia Conf. on Computer and Communications Security. Incheon: Association for Computing Machinery, 2018. 707–721. [doi: 10.1145/3196494.3196522]

[93] Dowlin N, Gilad-Bachrach R, Laine K, Lauter K, Naehrig M, Wernsing J. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: JMLR.org, 2016. 201–210.

[94] Agrawal N, Shahin Shamsabadi A, Kusner MJ, Gascón A. QUOTIENT: Two-party secure neural network training and prediction. In: Proc. of the 2019 ACM SIGSAC Conf. on Computer and Communications Security. London: Association for Computing Machinery, 2019. 1231–1247. [doi: 10.1145/3319535.3339819]

[95] Mohassel P, Rindal P. ABY^3: A mixed protocol framework for machine learning. In: Proc. of the 2018 ACM SIGSAC Conf. on Computer and Communications Security. Toronto: Association for Computing Machinery, 2018. 35–52. [doi: 10.1145/3243734.3243760]

[96] Wagh S, Gupta D, Chandran N. SecureNN: 3-party secure computation for neural network training. Proc. on Privacy Enhancing Technologies, 2019, 2019(3): 26–49. [doi: 10.2478/popets-2019-0035]

[97] Juvekar C, Vaikuntanathan V, Chandrakasan A. GAZELLE: A low latency framework for secure neural network inference. In: Proc. of the 27th USENIX Conf. on Security Symp. Baltimore: USENIX Association, 2018. 1651–1668.

[98] Vaidya J, Clifton C. Privacy-preserving *k*-means clustering over vertically partitioned data. In: Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington: Association for Computing Machinery, 2003. 206–215. [doi: 10.1145/956750.956776]

[99] Gheid Z, Challal Y. Efficient and privacy-preserving k-means clustering for big data mining. In: Proc. of the IEEE Trustcom/BigDataSE/ISPA. Tianjin: IEEE, 2016. 791–798. [doi: 10.1109/TrustCom.2016.0140]

[100] Prasad KD, Reddy KAN, Vasumathi D. Privacy-preserving naive bayesian classifier for continuous data and discrete data. In: Proc. of the 1st Int'l Conf. on Artificial Intelligence and Cognitive Computing. Singapore: Springer, 2019. 289–299. [doi: 10.1007/978-981-13-1580-0_28]

[101] Samet S, Miri A. Privacy-preserving back-propagation and extreme learning machine algorithms. Data & Knowledge Engineering, 2012, 79–80: 40–61. [doi: 10.1016/j.datak.2012.06.001]

[102] Goethals B, Laur S, Lipmaa H, Mielikainen T. On private scalar product computation for privacy-preserving data mining. In: Proc. of the 7th Int'l Conf. on Information Security and Cryptology. Seoul: Springer, 2005. 104–120. [doi: 10.1007/11496618_9]

[103] Zheng WT, Popa RA, Gonzalez JE, Stoica I. Helen: Maliciously secure coopetitive learning for linear models. In: Proc. of the 2019 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2019. 724–738. [doi: 10.1109/SP.2019.00045]

[104] Sharma S, Xing CP, Liu Y, Kang Y. Secure and efficient federated transfer learning. In: Proc. of the 2019 IEEE Int'l Conf. on Big Data. Los Angeles: IEEE, 2019. 2569–2576. [doi: 10.1109/BigData47090.2019.9006280]

[105] Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, Rosen B, Rubin DL, Kalpathy-Cramer J. Distributed deep learning networks among institutions for medical imaging. Journal of the American Medical Informatics Association, 2018, 25(8): 945–954. [doi: 10.1093/jamia/ocy017]

[106] Inan A, Kantarcioglu M, Bertino E, Scannapieco M. A hybrid approach to private record linkage. In: Proc. of the 24th IEEE Int'l Conf.

on Data Engineering. Cancun: IEEE, 2008. 496–505. [doi: 10.1109/ICDE.2008.4497458]

[107] Scannapieco M, Figotin I, Bertino E, Elmagarmid AK. Privacy preserving schema and data matching. In: Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of Data. Beijing: Association for Computing Machinery, 2007. 653–664. [doi: 10.1145/1247480.1247553]

[108] Cheng KW, Fan T, Jin YL, Liu Y, Chen TJ, Papadopoulos D, Yang Q. SecureBoost: A lossless federated learning framework. IEEE Intelligent Systems, 2021, 36(6): 87–98. [doi: 10.1109/MIS.2021.3082561]

[109] Hardy S, Henecka W, Ivey-Law H, Nock R, Patrini G, Smith G, Thorne B. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv:1711.10677, 2017.

[110] Liu Y, Kang Y, Xing CP, Chen TJ, Yang Q. Secure federated transfer learning. arXiv:1812.03337, 2018.

[111] Ohrimenko O, Schuster F, Fournet C, Mehta A, Nowozin S, Vaswani K, Costa M. Oblivious multi-party machine learning on trusted processors. In: Proc. of the 25th USENIX Conf. on Security Symp. Austin: USENIX Association, 2016. 619–636.

[112] Lin S, Wang CH, Li HJ, Deng JR, Wang YZ, Ding CW. ESMFL: Efficient and secure models for federated learning. arXiv:2009.01867, 2020.

[113] Mo F, Shamsabadi AS, Katevas K, Cavallaro A, Haddadi H. Towards characterizing and limiting information exposure in DNN layers. arXiv:1907.06034, 2019.

[114] Barni M, Failla P, Kolesnikov V, Lazzeretti R, Sadeghi AR, Schneider T. Secure evaluation of private linear branching programs with medical applications. In: Proc. of the 14th European Symp. on Research in Computer Security. Saint-Malo: Springer, 2009. 424–439. [doi: 10.1007/978-3-642-04444-1_26]

[115] Chaudhari H, Choudhury A, Patra A, Suresh A. ASTRA: High throughput 3PC over rings with application to secure prediction. In: Proc. of the 2019 ACM SIGSAC Conf. on Cloud Computing Security Workshop. London: Association for Computing Machinery, 2019. 81–92. [doi: 10.1145/3338466.3358922]

[116] Bost R, Popa RA, Tu S, Goldwasser S. Machine learning classification over encrypted data. In: Proc. of the 22nd Annual Network and Distributed System Security Symp. San Diego: The Internet Society, 2015. [doi: 10.14722/NDSS.2015.23241]

[117] Bos JW, Lauter K, Loftus J, Naehrig M. Improved security for a ring-based fully homomorphic encryption scheme. In: Proc. of the 14th IMA Int'l Conf. on Cryptography and Coding. Oxford: Springer, 2013. 45–64. [doi: 10.1007/978-3-642-45239-0_4]

[118] Sanyal A, Kusner MJ, Gascon A, Kanade V. Tapas: Tricks to accelerate (encrypted) prediction as a service. In: Proc. of the Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 4490–4499.

[119] Bourse F, Minelli M, Minihold M, Paillier P. Fast homomorphic evaluation of deep discretized neural networks. In: Proc. of the 38th Annual Int'l Cryptology Conf. on Advances in Cryptology. Santa Barbara: Springer, 2018. 483–512. [doi: 10.1007/978-3-319-96878-0_17]

[120] Chillotti I, Gama N, Georgieva M, Izabachene M. Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds. In: Proc. of the 22nd Int'l Conf. on the Theory and Application of Cryptology and Information Security. Hanoi: Springer, 2016. 3–33. [doi: 10.1007/978-3-662-53887-6_1]

[121] Wu DJ, Feng T, Naehrig M, Lauter K. Privately evaluating decision trees and random forests. Proc. on Privacy Enhancing Technologies, 2016, 2016(4): 335–355. [doi: 10.1515/popets-2016-0043]

[122] Chen H, Chillotti I, Dong YH, Poburinnaya O, Razenshteyn I, Riazi MS. SANNS: Scaling up secure approximate k-nearest neighbors search. In: Proc. of the 29th USENIX Conf. on Security Symp. USENIX Association, 2020. 119.

[123] Doerner J, Shelat A. Scaling ORAM for secure computation. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. Dallas: Association for Computing Machinery, 2017. 523–535. [doi: 10.1145/3133956.3133967]

[124] Hunt T, Song CZ, Shokri R, Shmatikov V, Witchel E. Chiron: Privacy-preserving machine learning as a service. arXiv:1803.05961, 2018.

[125] Ács D, Coleşa A. Securely exposing machine learning models to web clients using intel SGX. In: Proc. of the 15th IEEE Int'l Conf. on Intelligent Computer Communication and Processing (ICCP). Cluj-Napoca: IEEE, 2019. 161–168. [doi: 10.1109/ICCP48234.2019.8959635]

[126] Grover K, Tople S, Shinde S, Bhagwan R, Ramjee R. Privado: Practical and secure DNN inference with enclaves. arXiv:1810.00602, 2018.

[127] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958.

[128] Yao Y, Rosasco L, Caponnetto A. On early stopping in gradient descent learning. Constructive Approximation, 2007, 26(2): 289–315. [doi: 10.1007/s00365-006-0663-2]

[129] Wong SC, Gatt A, Stamatescu V, McDonnell MD. Understanding data augmentation for classification: When to warp? In: Proc. of the 2016 Int'l Conf. on Digital Image Computing: Techniques and Applications (DICTA). Gold Coast: IEEE, 2016. 1–6. [doi: 10.1109/DICTA.2016.7797091]

[130] Wu S, Li GQ, Chen F, Shi LP. Training and inference with integers in deep neural networks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[131] Rotaru D, Wood T. MArBled circuits: Mixing arithmetic and boolean circuits with active security. In: Proc. of the 20th Int'l Conf. on Cryptology in India. Hyderabad: Springer, 2019. 227–249. [doi: 10.1007/978-3-030-35423-7_12]

[132] Zhu RY, Cassel D, Sabry A, Huang Y. NANOPI: Extreme-scale actively-secure multi-party computation. In: Proc. of the 2018 ACM SIGSAC Conf. on Computer and Communications Security. Toronto: Association for Computing Machinery, 2018. 862–879. [doi: 10.1145/3243734.3243850]

[133] He Y, Zhang X, Sun J. Channel pruning for accelerating very deep neural networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1389-1397. [doi: 10.1109/iccv.2017.155]

[134] Han S, Mao HZ, Dally WJ. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv:1510.00149, 2015.

[135] Louizos C, Ullrich K, Welling M. Bayesian compression for deep learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 3290–3300.

[136] Wagner I, Eckhoff D. Technical privacy metrics: A systematic survey. ACM Computing Surveys, 2019, 51(3): 57. [doi: 10.1145/3168389]

[137] Lyu LJ, Yu JS, Nandakumar K, Li YT, Ma XJ, Jin J. Towards fair and decentralized privacy-preserving deep learning. arXiv: 1906.01167, 2019.

附中文参考文献:

[21] 谭作文, 张连福. 机器学习隐私保护研究综述. 软件学报, 2020, 31(7): 2127–2156. http://www.jos.org.cn/1000-9825/6052.htm [doi: 10.13328/j.cnki.jos.006052]

[37] 何英哲, 胡兴波, 何锦雯, 孟国柱, 陈恺. 机器学习系统的隐私和安全问题综述. 计算机研究与发展, 2019, 56(10): 2049–2070. [doi: 10.7544/issn1000-1239.2019.20190437]

[38] 刘俊旭, 孟小峰. 机器学习的隐私保护研究综述. 计算机研究与发展, 2020, 57(2): 346–362. [doi: 10.7544/issn1000-1239.2020.20190455]

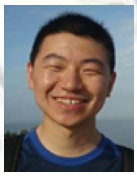
[39] 刘睿瑄, 陈红, 郭若杨, 赵丹, 梁文娟, 李翠平. 机器学习中的隐私攻击与防御. 软件学报, 2020, 31(3): 866–892. http://www.jos.org.cn/1000-9825/5904.htm [doi: 10.13328/j.cnki.jos.005904]

[63] 李宗育, 桂小林, 顾迎捷, 李雪松, 戴慧珺, 张学军. 同态加密技术及其在云计算隐私保护中的应用. 软件学报, 2018, 29(7): 1830–1851. http://www.jos.org.cn/1000-9825/5354.htm [doi: 10.13328/j.cnki.jos.005354]

[74] 吴晓慧, 贺也平, 马恒太, 周启明, 林少锋. 微架构瞬态执行攻击与防御方法. 软件学报, 2020, 31(2): 544–563. http://www.jos.org.cn/1000-9825/5979.htm [doi: 10.13328/j.cnki.jos.005979]

[76] 王鹃, 樊成阳, 程越强, 赵波, 韦韬, 严飞, 张焕国, 马婧. SGX技术的分析和研究. 软件学报, 2018, 29(9): 2778–2798. http://www.jos.org.cn/1000-9825/5594.htm [doi: 10.13328/j.cnki.jos.005594]

汤凌韬(1994－), 男, 博士生, CCF 学生会员, 主要研究领域为信息安全, 机器学习隐私保护.



张鲁飞(1986－), 男, 博士, 工程师, 主要研究领域为高性能计算, 操作系统, 机器学习.

陈左宁(1957－), 女, 博士, 博士生导师, 中国工程院院士, CCF 会士, 主要研究领域为软件理论, 操作系统, 信息安全.

吴东(1971－), 男, 博士, 研究员, 主要研究领域为人工智能, 密码学.