DOI:10.13878/j.cnki.jnuist.2022.02.001

石聪聪[1]　高先周[1]　黄秀丽[1]　毛云龙[2]

联邦学习隐私模型发布综述

摘要

联邦学习这一类分布式机器学习技术旨在保证使用大数据进行机器学习训练时保护本地数据不泄露.然而一系列机器学习隐私攻击表明,即使不直接暴露本地数据,仅仅通过获取机器学习模型的参数就可以进行数据隐私的窃取.从训练时参与者和聚合端之间传递的中间模型到最后发布的聚合模型,联邦学习的模型发布过程存在诸多隐私威胁.由此出现了大量相关的保护技术,包括基于差分隐私以及基于密码学的联邦学习隐私保护技术.本文针对联邦学习本地模型和聚合模型发布过程中可能出现的各种隐私威胁和敌手模型进行了简要介绍,并且对相关的防御技术和研究成果进行系统性综述.同时也对相关技术在联邦学习隐私保护中的发展趋势进行了展望.

关键词

联邦学习;隐私保护;差分隐私

中图分类号 TP309;TP181

文献标志码 A

收稿日期 2021-10-20

资助项目 国家电网有限公司总部管理科技项目(5700-202190184A-0-0-00)

作者简介

石聪聪,男,高级工程师,主要研究方向为电力数据安全.shicongcong@ geiri.sgcc.com.cn

毛云龙(通信作者),男,博士,助理研究员,主要研究方向为网络信息安全.maoyl@nju.edu.cn

1 全球能源互联网研究院有限公司南京分公司/信息网络安全国网重点实验室,南京,210094

2 南京大学 计算机科学与技术系,南京,210023

0 引言

由于在大数据分析处理方面出色的表现,人工智能技术得到了广泛的应用.机器学习以及深度学习模型凭借着性能优良、鲁棒性强、适应性高等优点,在多种需求场景中发挥了重要作用.现阶段,随着国内现代化传感器设备、智能手机、IoT 设备的普及,产生了海量个人信息数据,这些数据充分反映了不同用户的使用习惯和个人需求,能够对提供个性化服务的人工智能平台提供有力的数据支持(如输入法、语音助手、购物推荐等),导致用户个人数据分析的需求越来越迫切.但是这也带来了潜藏在传统中心化机器学习场景下的数据隐私安全风险.服务器直接收集客户端或者边缘端的数据用以模型训练虽然简单直接,然而由于越来越多的针对机器学习模型的隐私攻击的出现以及国家法律法规的要求,这种中心化训练的方式逐渐加剧人们的担忧,引起了人们对个人数据隐私安全的顾虑.在这种背景下,联邦学习顺势而生[1].数据拥有者训练本地模型随后上传更新的梯度至聚合服务端,聚合端将收集到的多份梯度聚合随后更新维护最终模型.之后将聚合更新过的模型再发送至各个训练参与方开始新一轮的训练.通过这种方式避免了数据直接暴露且依旧能够利用隐私数据进行机器学习训练.

尽管不会直接暴露用户的训练数据,联邦学习依旧存在诸多安全风险.已经有研究表明,联邦学习易于遭受多种隐私窃取攻击的威胁.恶意攻击者能够在获取模型梯度的情况下窃取数据拥有者的隐私信息[2].由此产生了一系列的防御解决方案,包括基于安全多方计算的防御方案[3]、基于差分隐私的隐私训练技术[4]等.安全多方计算通过一系列如同态加密、混淆电路以及秘密分享等技术[5-6],能够使得在联邦学习训练过程中,训练参与方的本地模型梯度对除了自身以外的其他方不可见,从而实现保护本地模型隐私的模型发布目标.然而,采用安全多方计算技术保护的联邦学习过程也仅仅局限于训练过程中模型的安全,无法对聚合之后的模型提供保护.为了解决聚合模型安全发布问题,有研究提出了基于差分隐私的联邦学习方案,用来保护聚合模型中的用户数据隐私.其中,差分隐私随机梯度下降方法(Differential Privacy-Stochastic Gradient Descent,DP-SGD)是最著名的隐私机器学习训练方法之一,通过应用它的矩计算技术可以使得采用差分隐私训练的联邦学习成为可行的方案[7].

为了给联邦学习提供全生命周期多角度的隐私保护,出现了一些结合差分隐私和安全多方计算技术的联邦学习模型发布方案.对于安全多方计算,往往是提供黑盒式的功能保护梯度不暴露,对于整体模型的效果并不会有影响,差距往往在于不同方案的通信和计算效率.对于分布式训练中的差分隐私,如何在理论正确的情况下寻求与安全多方计算相结合的最高效的方案是一个难题.本文针对隐私安全的联邦学习模型发布方法做了系统性的整理和介绍,包括:1)系统性地介绍联邦学习和涉及到的相关概念;2)对于当下联邦学习模型发布面临的隐私安全威胁和敌手模型的介绍;3)介绍主流的联邦学习模型发布的隐私保护方案;4)针对联邦学习模型发布场景中可行的保护方案做出展望.

1 相关概念

本节介绍联邦学习的概念,以及涉及到的隐私安全保护技术,包括差分隐私、安全多方计算等技术.

1.1 联邦学习

传统机器学习中,服务商不得不从各个客户端收集私人数据组成一个巨大的数据集供中心化训练.一方面,个人隐私敏感的数据将会暴露给服务商并且会加剧数据泄露带来的后果;另一方面,中心化的训练消耗的服务资源会更多一些.联邦学习由谷歌提出,主要是为了保护个人隐私,是支持大规模参与者的分布式机器学习,也包括深度学习的训练方式[1,8].它主要是在采用反向求导优化的机器学习算法中,在训练每一轮上传本地梯度更新值,再由中心服务端收集且加权求和共同更新维护中心模型,拓扑结构如图 1 所示.随后,根据参与训练的不同数据特征的差异,出现了包括横向联邦学习、纵向联邦学习和联邦迁移学习的概念.在横向联邦学习中,训练参与方的数据拥有一致的特征,而在纵向联邦学习中,训练参与方的数据来自一致的个体但是拥有不同的特征.

本文主要介绍经典的横向联邦学习.具体来说,每个参与方拥有相同特征的数据但是数据所有者不尽相同,每个训练参与方从服务端接收一个初始化模型 θ,随后使用本地的数据训练 E 轮得到更新梯度 G_i,再上传至服务端聚合.服务端收集到一定数量 k 的 G 后求和平均得到 $G=\sum_{i}^{K}\frac{n_k}{n}G_i$,其中 n_k 表示第 k 个参与者的本地数据样本大小,$n=\sum_{0}^{K}n_k$.接下来服务端使用梯度 G 更新中心模型然后下发至新的参与方开始新一轮的训练[1,9].

通常会选取参数 B,E 和学习率来控制联邦学习,B 代表本地训练的批大小(batch size),E 代表本地训练轮次(epoch).根据实验经验,B 和 E 需要被设定为较小的适合值来平衡模型可用性和联邦学习通信和计算的开销.同时,为了兼顾联邦学习训练效率以及稳定性,也有许多工作对联邦学习的优化目标、损失函数、超参数设置等做一些相关的改进.

在纵向联邦学习协议中,服务器和客户端各自会采取更为复杂一些的方法进行模型训练.比如 GELU-net 采用了子模型的方式令服务器和客户端分别训练不同的特征部分,随后组合两个子模型的输出再进行统一训练[10-11].

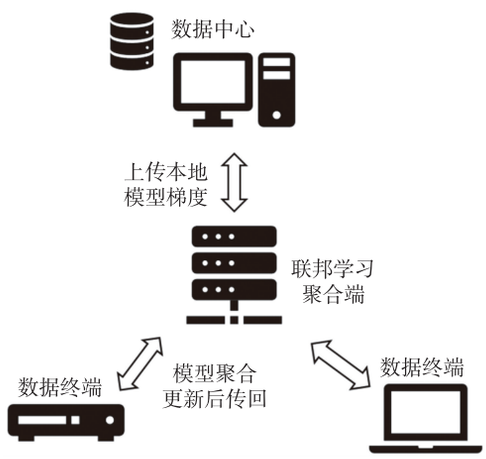

图 1 联邦学习聚合模型

Fig. 1 Topology of federated learning

1.2 差分隐私

差分隐私(Differential Privacy,DP)提供了一种可靠的中心化隐私保护属性.由于严谨的理论基础和实用性,差分隐私成为被广泛接受和采用的中心化数据隐私保护技术[12].

定义 1 对于两个相邻数据集 D,$D'\subset D$ 和任意值域的子集 $S\subset R$,如果一个随机机制 $M:D\to R$ 满足 (ϵ,δ)-DP,它需要满足如下约束条件:

$$P[M(D)\in S]\leqslant e^{\epsilon}P[M(D')\in S]+\delta.$$

如定义 1 所示,δ 表示 ϵ-DP 可能会不满足上述约束的概率.通过预设的噪声来使得某查询函数满足 (ϵ,δ)-DP 隐私要求,该噪声的级别除了上述参数 ϵ 外,还与敏感度 S_f 相关.S_f 可以被定义为

$$S_f = \max_{d,d' \in D} | f(d) - f(d') |.$$

那么,一个满足(ϵ,δ)-DP 的随机化查询函数能够通过如下方式获得:

$$M(d) \triangleq f(d) + N(0,S_f^2 \cdot \sigma^2),$$

其中,$N(0,S_f^2 \cdot \sigma^2)$ 是一个均值为0标准差为 $S_f\sigma$ 的高斯分布,并且 $S_f = \max_{d,d' \in D} | f(d) - f(d') |_2$,$\epsilon \in (0,1)$,被称之为高斯机制.根据高斯机制,在单次查询的情况下,对于固定的 σ 和 ϵ,则有

$$\delta \leqslant \frac{4}{5}\exp(-(\sigma\epsilon)^2/2).$$

1.2.1 组合定理

在应用差分隐私的情况下,若连续查询,则 (ϵ,δ)-DP 的隐私属性将会被破坏.因为可串通的连续查询将会使得预设0均值噪声由于真实值的增强失效,所以考虑此种情况下有效的组合定理对于隐私联邦学习至关重要.根据差分隐私串行组合的属性,基础的组合定理是:对(ϵ,δ)-DP 做 N 次查询,它会成为$(N\epsilon,N\delta)$-DP.但是显然,这种隐私边界太松以至于无法投入到实际生产使用中.所以,出现了强组合定理,使得上述情况下为$(\epsilon\sqrt{N\ln/\delta},N\delta)$-DP.最近,差分隐私随机梯度下降方法(DP-SGD)应用矩计算技术进一步改进了组合定理的隐私边界,使得差分隐私结合随机子采样和按层梯度裁剪技术之后能够满足实际应用,将 N 次连续查询的隐私开销压缩到$(\epsilon\sqrt{N},\delta)$-DP[7].

1.2.2 本地差分隐私

传统的中心化差分隐私往往用于中心化数据的发布,即可信的数据拥有方采用一些隐私机制进行数据发布,模糊被查询数据相对于整体数据集的存在感,满足中心化的差分隐私.然而,这种隐私化方法不适用于无可信第三方的应用场景.本地差分隐私(Local Differential Privacy,LDP)则将数据隐私化的任务分配给本地用户,即在没有可信第三方的情况下解决了本地数据隐私发布问题.通过本地差分隐私技术,对于任意本地数据,其相比于任意其他数据都是不可区分的.因此在这种需求下,数据的可用性显然会更低,往往只能满足一些特殊需求的数据采集,例如频率统计、均值统计等.

1.3 安全多方计算

安全多方计算(secure Multi-Party Computation,MPC)是一类基于密码学的安全与隐私保护技术.自从 Yao[13] 提出百万富翁问题以来,出现了许多安全多方计算的解决方案.安全多方计算所研究的问题是,各个参与方如何在不直接暴露自身数据的前提下交换信息合作共同完成某个计算任务.

安全多方计算最经典的应用场景是百万富翁问题,即两个富翁 Alice 和 Bob 希望在不暴露各自拥有财富的情况下知道谁更富有,即典型的无可信三方的多方计算场景.一般来说,当前学术界和工业界主流的实现方式是通过混淆电路(garbled circuit)、茫然传输(oblivious transfer)、秘密分享(secret sharing)、同态加密(homomorphic encryption)、零知识证明(zero-knowledge proof)等多种现代密码学技术组成的多方安全计算方法,目前诸多主流的计算框架包括 ABY、SPDZ、PICCO、Obliv-C 等[5,14-15].同时也有应用同态加密技术的三方方案,即需要一个密钥分发机构进行密码体系的构建[16-17].安全多方计算能够保障在分布式场景中不暴露参与者的隐私信息,对私有数据交互共享提供密码学级别的保护,因此在金融、征信等对隐私保护程度较高的场景有着广泛的应用.

2 隐私威胁和敌手模型

本节介绍联邦学习模型发布中大致的隐私概念和威胁,随后具体介绍可能的敌手模型.

2.1 隐私攻击与威胁

2.1.1 隐私攻击

隐私的传统定义是指用户数据本身以及可能泄露数据的相关信息.虽然联邦学习协议的初衷就是为了避免暴露用户本身的私有数据,事实上已经在某种程度上直接保护了数据隐私,但是对于机器学习模型来说,攻击者仅凭模型参数就能够窃取原始数据,构成严重的隐私威胁.根据已有的研究[2,18-21],许多经典的针对机器学习隐私攻击在联邦学习场景中大多能够奏效,包括标签复现、属性推断攻击、成员推断攻击等.这些攻击主要分为三种场景:1)攻击者是联邦学习聚合端,在这种情况下,聚合者甚至能够极大程度地复原用户的个人信息[19];2)攻击者是训练参与方,攻击者能够实施成员推断攻击和属性推断攻击,但是攻击效果会随着参与者数量的提升而降低[2];3)以模型观察者的姿态对联邦学习训练的模型实施一系列针对普通机器学习模型的攻击[18,20-21].

因此,本文将机器学习或深度学习模型的梯度或者模型参数视为需要保护的隐私数据.在联邦学

习场景中,各个参与方的本地模型会在模型聚合时暴露给服务端,同时最终聚合后的模型也包含参与者的隐私信息,所以,用户本地模型以及聚合模型的发布都是可能遭受隐私攻击的风险对象.在机器学习训练或者模型发布场景中,如何真正地保护用户隐私是重中之重.

2.1.2 隐私威胁

如上所述,联邦学习场景中的隐私威胁主要存在于三个方面:一是训练过程中,本地发布出的局部模型;二是训练阶段后期发布给各个参与方的聚合过的中间模型;三是整体训练完后发布的模型.其中:第一点的威胁最大,这是因为本地模型最易受隐私攻击且成功率最高;第二点和第三点本质上是相似的,聚合模型对于隐私攻击具有一定的天然抵御能力然而依旧需要采取防御措施.联邦学习的整个生命周期可以分为两个阶段,包括本地模型发布阶段和最终模型发布阶段,具体如下:

1)本地发布.在联邦学习的本地发布阶段,聚合端在每一轮都能够直接得到每个参与者上传的梯度,这对于参与者的隐私安全来说是非常危险的,所以当一个本地梯度被发送出去的时候,它不应该是明文.

2)聚合发布.在联邦学习中,模型训练后期分发的模型或者最终发布的整体模型都可以看作是直接暴露的.恶意攻击者能够作为参与者获得训练过程中的聚合模型,并且可以推断其他参与者的梯度.尤其是对于一些合谋的参与者,他们能够通过与他人合作获得其他参与者的本地模型,因此在这种情况下保证每个参与者的隐私也非常重要.

2.2 敌手模型

联邦学习的隐私威胁主要来自两种场景(表1):一类来自内部,主要包括联邦学习的参与者包括数据提供者和模型聚合服务端;另一类则来自外部,主要是模型的使用者[22].根据以往的诸多相关工作,普遍会将敌手分为两类:一种是半诚实者(semi-honest),会遵守协议但对隐私好奇;另一类是半恶意者(semi-malicious),为了达成目的可能会不遵守协议.我们的目标是保护联邦学习中训练参与方的数据隐私.根据现实情况,聚合端和除了自身以外的训练参与方一般都是诚实但好奇(honest but curious)的,即会遵守训练协议但是对别人的隐私好奇,并且参与方中的一部分可以串通.另外,训练过程中每一轮的聚合模型和最终发布的聚合模型的观察者都是好奇的.

很多已有的工作也是基于这样的敌手假设.本文总结为以下两类敌手威胁:

1)训练内部参与方,会遵守既定协议但是对其他人的隐私感兴趣,会寻求和其他方合作以完成隐私窃取攻击.这类威胁主要是来自联邦学习训练过程中的参与者和聚合者.

2)联邦学习完成后发布模型的观察者.这类敌手对训练参与者的隐私感兴趣,但是所能获取的信息有限,即只有最终发布的模型(最大限度为白盒模型).

另外,在一些联邦学习具体问题特别是安全聚合相关工作中,也会将联邦学习聚合者或参与者假设为半恶意者,以此来进一步扩展工作.

表1 敌手模型和角色

Table 1 The adversaries and their characteristics

敌手	角色	特征
半诚实者	模型观察者、参与者、聚合端	遵守协议,只对隐私好奇
主动恶意者	参与者、聚合端	可以不遵守协议

3 隐私联邦学习模型发布

根据上述的联邦学习现实场景中会面临的诸多隐私威胁,研究者们提出了一系列相应的保护措施.这其中主要包括采用密码学技术对训练过程中本地模型发布的保护、采用差分隐私技术对聚合模型的隐私保护.总体来说,隐私联邦学习模型发布包含基于密码学的本地模型发布和基于差分隐私的聚合模型发布两部分.其中,基于密码学的本地模型发布技术将安全计算技术融入联邦学习中,使得每个参与方本地模型梯度并不直接暴露,极大程度降低了数据拥有者本地面临的隐私威胁.而聚合模型发布技术则聚焦于训练阶段下发的聚合模型和最终发布模型的隐私保护,通常采用隐私机器学习训练技术保护模型以应对模型观察者的隐私窃取.还有很多工作致力于探索结合两者进行全方位从联邦学习训练到最终模型发布全生命周期的隐私保护.

3.1 聚合模型发布

联邦学习训练后期和训练完成之后聚合模型的发布是联邦学习模型发布面临的主要威胁场景之一.由于发布的聚合模型除了聚合端以外还需要面对诸多的模型观察者,如何保证聚合模型中整体训练数据的隐私安全是这一类工作的重点.当前,学术

界往往采用差分隐私的保护方案,包括隐私机器学习训练方法、基于教师-学生模型的隐私聚合方法以及一系列差分隐私增强技术和弱差分隐私方案.

3.1.1 基于差分隐私组合定理的隐私训练方法

最直接的隐私训练方案就是直接应用高斯机制进行隐私训练.文献[23]将高斯机制的简单组合定理应用在联邦学习中,提出了 NbAFL 方案,最终取得的实验结果的隐私开销远超可以接受的范围.它将联邦学习的本地模型上传和模型分发分为两个都需要保护的独立阶段,即在参与者本地进行一轮保护之后,聚合完成在聚合端再进行一轮保护.一方面这种应用场景不具有很大的实际价值,另一方面这种方案下的隐私开销极大.当然文献[23]也提出了一些有价值的见解,比如联邦学习参与者的随机选取数量在平衡隐私保护程度和模型可用性之间存在一个最优值,这是一个值得深入探索的问题.

文献[4,24]尝试将文献[7]提出的矩计算技术分别应用在图像识别任务和自然语言处理任务的联邦学习中,站在训练参与者的视角保护数据隐私.在此方案的隐私训练中,使用隐私参数、高斯分布的标准差 σ 以及每一轮选取的参与者数量 m 来控制隐私开销;给定 σ,m 和 ϵ,应用矩计算技术可以计算 δ,只要 δ 达到预设阈值训练过程就会停止;同时也采用了基础的梯度裁剪和随机选取技术来进一步优化隐私开销.其中,进行梯度裁剪的敏感度则是来自于本地训练阶段各参与方上传至服务端的最大值,随后进行梯度加噪,使得方案理论上符合对各客户端数据的差分隐私保护.另外也有学者尝试将矩计算技术应用在多层结构的特异联邦学习中[25],矩计算技术已经成为事实上的隐私机器学习训练标准.

在当前中心化差分隐私理论体系下,矩计算技术似乎已经是高斯机制组合定理的理论极限,然而采用隐私训练的机器学习包括联邦学习的模型效果依旧不足以满足生产环境的需要,模型的可用性依旧不够高.

3.1.2 基于差分隐私技术的集成方法

除了机器学习的隐私训练外还有一系列其他的基于差分隐私的模型发布方案.

在基于教师-学生模型的隐私聚合方法(Private Aggregation of Teacher Ensembles,PATE)[26]中(图2),本地数据将被随机划分后送入多个教师模型训练,随后再取各个教师模型的输出的最大值当作教师模型聚合的输出,同时统计其出现的频率来计算隐私的开销.同一个结果在多个教师模型中出现说明不同的训练数据集对结果的影响很小,即表明其暴露的数据隐私越少.在教师模型的输出结果上加噪声则可以定制一定的隐私保护.再采用知识迁移的方式用公共数据集通过有限次的查询获取教师模型中蕴含的信息得到公共数据集的标签,最后送至学生模型统一学习得到最终学习结果[26].随后也出现了改进方案,基于 Renyi 差分隐私的计算后施加更松弛的高斯噪声以及交互式学生模型查询方案,进一步提高了模型的可用性[27].作为一类基于差分隐私技术的本地隐私模型的发布方案,PATE 具有可用性高的特点,但是需要更复杂的训练和发布过程.文献[28]将模型参数独立处理,提供另类的隐私模型的发布方案,相比于隐私训练方法提升了模型可用性.

3.1.3 近似差分隐私及隐私增强技术

也有许多学者提出了一些近似差分隐私技术如贝叶斯差分隐私、集中式差分隐私(Concentrated Differential Privacy,CDP)等,在一定程度的隐私和模型

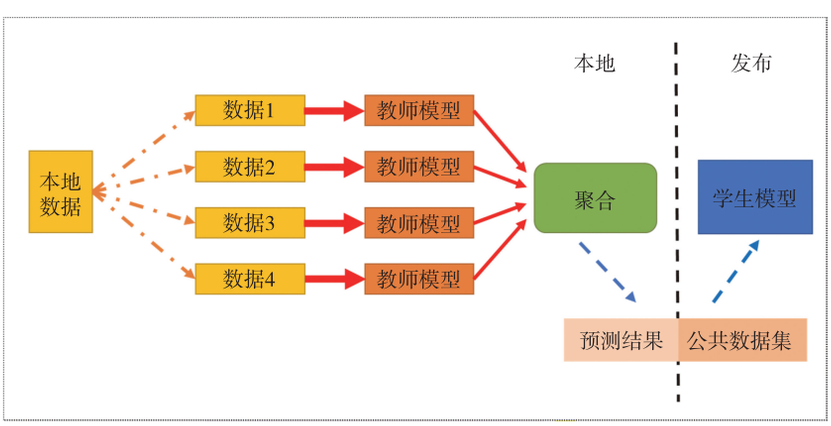

图2 教师-学生模型的隐私聚合方法[26]

Fig. 2 Private aggregation of teacher ensembles[26]

可用性之间做进一步的取舍[29-31].当然也有一些采用安全洗牌、子采样、指数机制等技术进行的无噪声方案[28,32-34].

总体来说,差分隐私技术在联邦学习中的应用除了基本的差分隐私机制之外还包含以下几类增强技术,如梯度裁剪、参与者随机选取、本地数据随机选取以及安全洗牌等[35-37].当然匿名洗牌往往需要一个可信的洗牌者或是可以匿名传输的强假设,需要更成熟的联邦学习的通信机制,相对来说应用门槛更高.目前来看差分隐私在机器学习隐私保护中的应用依旧需要进一步地工作来推进和优化,幸运的是在联邦学习这一复杂的系统中,如安全洗牌、子采样技术的加入使得其越来越令人期待.

3.2 本地模型发布

在联邦学习训练阶段,数据拥有者即参与方将本地模型发布给聚合端参与联邦学习.由于本地模型对于隐私攻击最为敏感,通常倾向于采用本地差分隐私技术提供隐私保护或者结合基于密码学技术的安全多方计算方案,在本地资源受限或者特征不完全匹配的场景则采用模型分割的方式进行联邦学习和本地模型发布.表 2 列出了本节将要讨论的主要发布方法.

表 2 隐私发布方案对比

Table 2 Comparison of private model publishing approaches

分类	方案(文献)	应用场景	优点
隐私训练	[4,7,24-26]	聚合发布	严格隐私定义
隐私集成方法	[27-29]	聚合和本地发布	模型可用性高
本地差分隐私	[31,38-40]	本地发布	保护程度高
模型分割	[10,41-42]	本地发布	适用本地资源受限
可信执行环境	[43]	本地发布	可信硬件隔离

3.2.1 基于本地差分隐私的方法

差分隐私的保护方案根据添加噪声的主体可以分为本地差分隐私和中心化差分隐私两大类.中心化差分隐私方案采用中心化差分隐私的定义进行隐私保护,通常针对的是模型中整体数据的隐私.本地差分隐私则结合符合本地差分隐私定义的机制进行隐私模型发布,针对本地数据进行隐私保护.当然也有一些基于知识蒸馏的非典型隐私联邦学习方案[44]不在本文的讨论范围内.

由于提供更高要求的隐私保护,本地差分隐私相比于中心化差分隐私需要提供更高的数据扰动或者噪声扰动,使得其获得的最终数据或者模型的可用性都较低.

在联邦学习训练中,本地差分隐私也是被广泛考虑的一类技术.最知名的当属苹果公司和斯坦福大学合作的这篇文章,为了大规模可实用的联邦学习提出了新的本地差分隐私机制并且在对某条记录感兴趣且无先验知识的敌手假设下,采用大隐私参数的严格定义的本地差分隐私保护本地数据隐私[38].由于本地差分隐私技术本身的特点,隐私开销往往巨大.为了降低隐私开销,有学者提出将模型的每个参数独立处理,随后采用洗牌和随机技术来完成对每个参数独立的本地差分隐私方案.虽然此方案采用洗牌技术有效进行隐私增强,然而将参数独立开来已经违背了差分隐私保护数据的理论出发点,只能算得上是对模型本身的保护[39],当然这也不失为另一种隐私联邦学习思路.同时也有一些采用隐私增强技术和弱差分隐私技术的方案[30].

3.2.2 基于差分隐私和密码学技术的方法

通过密码学使得本地模型不暴露,直接杜绝了联邦学习中风险最大的内部聚合者的隐私窃取行为.一般来说,密码学技术的目的主要是为了安全多方计算,经典联邦学习往往是一个星形拓扑结构,构成多个数据拥有方与一个模型聚合端的多方计算.当然也存在一些针对两方的安全机器学习框架,但需要与联邦学习区分[40,45].

针对隐藏本地模型进行安全计算,同时也考虑到训练过程中参与方的掉线情况以及通信和计算开销的影响,由此谷歌公司提出了针对移动设备使用的联邦学习安全聚合方法[3].使用双掩码技术和秘密分享技术以一定代价实现了能够抵御主动恶意聚合者的实用的联邦学习本地模型发布技术.然而有系统实验表明此方案在现实场景中仅可以支持数百个参与者的规模,与联邦学习在实际应用场景中的规模还存在一定差距[46].

同时考虑到联邦学习训练中本地模型和聚合模型以及最终模型发布时的隐私威胁,研究者采用阈值同态技术[17]将密码学技术和隐私训练结合以防御这两种威胁情况[47].由于采用了阈值同态技术,聚合端的每一次模型更新都需要足够多的参与者参与,以防止少数参与者和聚合端串通攻击;同时考虑到只需要对最终模型加噪声进行发布模型的隐私保护,作者将高斯机制的噪声方差做了线性的衰减,提高了最终模型的可用性.然而考虑到每个参与者梯度更新的权重,最终的聚合梯度将会是 $\sum_{0}^{K}\frac{n_k}{n}\theta_k+$

$\sum_{0}^{K}\frac{n_k}{n}N(0,S_f^2\cdot\sigma_k^2)$,最终的噪声会进一步衰减,由此带来的差分隐私保护级别还有待商榷.文献[48]将函数加密[49]和类似的隐私机制结合.同样地,也不断有人尝试将同态加密或是安全计算技术结合差分隐私技术实现隐私的联邦学习模型发布,同态加密技术从纯理论的通信复杂度方面也确实优于双掩码的安全聚合技术[50].

3.2.3 基于模型分割的发布方法

为了进一步提升联邦学习的容错能力以及应对更多的数据特征不匹配的情况,同时考虑到一些参与方设备的资源有限,推出了一系列基于模型分割的模型发布方法.与传统联邦学习不同的是,基于模型分割的方法引入了另一种分布式模型训练和发布机制.

文献[51]为了应对本地资源受限的参与方场景,提出将模型进行分割,本地留存小部分模型,大部分模型留存在服务端,同时对传输的数据进行差分隐私的保护,对参与方小部分模型采用密码学方案进行安全发布.在模型分割中,被分割的模型切面层(cut layer)传输的梯度数据称为碎片数据(smashed data),参与联邦学习的各方通过碎片数据流通完成跨切面层的梯度传播进行模型训练,这样可以保证既不泄露原始训练样本,参与训练的各方也无法获取完整的模型,进而提高了本地模型保护程度[41].对于一些数据特征不一致的联邦学习场景,往往会采用模型分割的方式来进行特征的分别提取.

另外,基于模型分割的发布方法也可以与前文所述的基于差分隐私、密码学技术的方法进行结合用以保障本地模型发布的隐私性和安全性.例如:在纵向联邦学习中,不同的数据拥有方持有部分模型,在有一个可信执行者的场景下,各方采用基于同态加密的方案进行模型发布[10,42].现阶段针对基于模型分割的发布方法研究尚不充分,关于其隐私保护方面的分析工作仍需要进一步探索[52].

3.2.4 基于可信执行环境的方法

为了隐匿本地模型,防止参与联邦学习的其他用户或者服务端窃取本地模型的发布结果,每个用户可以在一个提供密码学保护功能的可信执行环境(Trusted Execution Environment,TEE)中训练其本地模型.通过可信硬件隔离,TEE 能够保障对模型的安全性和完整性的证明.服务端将验证用户的本地联邦学习模型是否在 TEE 中运行,然后将加密的模型梯度更新传输给设备,模型的加密和解密均只能在 TEE 中执行,不安全的行为将被隔离在外部环境.

虽然可信执行环境能够为本地模型提供强力的保护,但是这种方法并不具备普适性,特别是当终端设备运算能力不足时(如智能手机、普通 IOT 设备等).为了降低计算开销,有研究工作提出可以将部分计算移出可信执行环境,同时保持整个计算过程的完整性和保密性需求[43].然而,基于可信执行环境的联邦学习的计算效率和通信开销依然是需要后续研究工作解决的问题.

4 展望

联邦学习隐私模型发布的重点是进一步提升以差分隐私为主流的隐私模型发布方案的效果,提高模型的可用性,降低方案的计算和通信开销.其中在联邦学习中一些非典型的无噪声方案将会带来新的可能性[35-37].

除了单独改进差分隐私系列技术,如何结合密码学和差分隐私技术,保证联邦学习全生命周期的模型发布的隐私安全并且提供可行高效的方案将会是这一方向工作的重点.

隐私安全的联邦学习依旧处在研究探索之中,一方面相关的理论技术不够完善,如能够最大程度保证模型效果的差分隐私理论技术的欠缺使得相关技术无法被真正应用,又如同态加密或是基于混淆电路的安全多方计算系列技术在效率上依旧有很大的进步空间.另一方面,依旧缺乏切实可行的系统性方案的尝试,比如大规模的实验对于很多研究者来说很难达成,需要相关企业的介入[46],以及实用的相关工具链的完善也是推动相关方案发展的要素.

5 总结

本文对联邦学习本地模型到聚合模型和最终模型发布整个生命周期潜在的隐私威胁做了系统性分析和介绍,同时也对已有的一系列防御方案包括基于差分隐私的聚合模型发布方法和基于本地差分隐私以及安全多方计算技术的本地模型发布方案进行了系统介绍.

两类方案分别针对不同的隐私威胁和场景并且具有不同的发展方向和趋势.其中对于基于差分隐私的聚合隐私模型发布技术,在隐私训练技术目前的理论极限下,如何寻求更低的隐私开销或者寻求

一些新颖的隐私机制将会给这一技术带来更多可能性.而对于本地模型发布,针对联邦学习本地模型发布可能遇到的严峻的隐私威胁,一方面采用本地差分隐私技术进行高度的隐私保护,另一方面采用密码学方案,旨在于隐匿参与者本地更新梯度的情况下参与联邦学习共同完成梯度求和的安全计算任务,不同的方案主要差异在于通信和计算效率.同时对于一些新颖的如近似差分隐私方案、模型分割方案等,都是为了进一步提升隐私发布方案的可用性.如何权衡模型可用性和隐私性是未来联邦学习研究的一个重要方向.可以预见的是联邦学习的大规模应用前景和国家社会对个人信息安全的重视必然会推进联邦学习隐私保护相关问题的发展.

参考文献

References

[1] McMahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data [J]. arXiv e-print, 2016, arXiv:1602.05629

[2] Melis L,Song C Z,De Cristofaro E,et al.Exploiting unintended feature leakage in collaborative learning[C]//2019 IEEE Symposium on Security and Privacy (SP). May 19-23,2019,San Francisco,CA,USA.IEEE,2019:691-706

[3] Bonawitz K,Ivanov V,Kreuter B,et al.Practical secure aggregation for privacy-preserving machine learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas, Texas, USA.New York,NY,USA:ACM,2017:1175-1191

[4] Geyer R C, Klein T, Nabi M. Differentially private federated learning: a client level perspective[J]. arXiv e-print,2017,arXiv:1712.07557

[5] Keller M, Pastro V, Rotaru D. Overdrive: making SPDZ great again[M]//Advances in Cryptology:EUROCRYPT 2018. Cham: Springer International Publishing, 2018:158-189

[6] Damgård I,Jurik M,Nielsen J B.A generalization of Paillier's public-key system with applications to electronic voting[J].International Journal of Information Security, 2010,9(6):371-385

[7] 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. Vienna, Austria. New York, NY, USA: ACM, 2016:308-318

[8] Shokri R,Shmatikov V.Privacy-preserving deep learning[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security.Denver Colorado USA. New York, NY, USA: ACM, 2015. DOI: 10.1109/ALLERTON.2015.7447103

[9] McMahan H B, Moore E, Ramage D, et al. Federated learning of deep networks using model averaging[J]. arXiv e-print,2016,arXiv:1602.05629

[10] Zhang Q,Wang C,Wu H Y,et al.GELU-net:a globally encrypted,locally unencrypted deep neural network for privacy-preserved learning[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence.July 13-19,2018,Stockholm,Sweden. California: International Joint Conferences on Artificial Intelligence Organization,2018:3933-3939

[11] Zhang Y F,Zhu H.Additively homomorphical encryption based deep neural network for asymmetrically collaborative machine learning[J].arXiv e-print,2020, arXiv:2007.06849

[12] Dwork C, Roth A. The algorithmic foundations of differential privacy[J].Foundations and Trends® in Theoretical Computer Science,2013,9(3/4):211-407

[13] Yao A C.Protocols for secure computations[C]//23rd Annual Symposium on Foundations of Computer Science. November 3-5, 1982, Chicago, IL, USA. IEEE, 1982:160-164

[14] Demmler D, Schneider T, Zohner M. ABY: a framework for efficient mixed-protocol secure two-party computation[C]//Proceedings 2015 Network and Distributed System Security Symposium.San Diego, CA.Reston, VA: Internet Society,2015.DOI:10.14722/ndss.2015.23113

[15] Hastings M, Hemenway B, Noble D, et al. SoK: general purpose compilers for secure multi-party computation[C]//2019 IEEE Symposium on Security and Privacy (SP). May 19-23,2019,San Francisco,CA,USA.IEEE,2019:1220-1237

[16] Cramer R, Damgård I, Nielsen J B. Multiparty computation from threshold homomorphic encryption[M]// Lecture Notes in Computer Science. Berlin, Heidelberg:Springer Berlin Heidelberg,2001:280-300

[17] Damgård I,Jurik M.A generalisation,a simplification and some applications of Paillier's probabilistic public-key system[C]//4th International Workshop on Practice and Theory in Public Key Cryptography,2001:119-136

[18] Shokri R, Stronati M, Song C Z, et al. Membership inference attacks against machine learning models[C]//2017 IEEE Symposium on Security and Privacy (SP). May 22-26,2017,San Jose,CA,USA.IEEE,2017:3-18

[19] Wang Z B,Song M K,Zhang Z F,et al.Beyond inferring class representatives:user-level privacy leakage from federated learning[C]//IEEE INFOCOM 2019-IEEE Conference on Computer Communications.April 29-May 2, 2019,Paris,France.IEEE,2019:2512-2520

[20] Ganju K R,Wang Q,Yang W,et al.Property inference attacks on fully connected neural networks using permutation invariant representations[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. Toronto, Canada. New York, NY,USA:ACM,2018:619-633

[21] Leino K,Fredrikson M.Stolen memories:leveraging model memorization for calibrated white-box membership inference[J].arXiv e-print,2019,arXiv:1906.11798

[22] Yin X F,Zhu Y M,Hu J K.A comprehensive survey of privacy-preserving federated learning [J]. ACM

Computing Surveys,2021,54(6):1-36
[23] Wei K,Li J,Ding M,et al.Federated learning with differential privacy:algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security,2020,15:3454-3469
[24] McMahan H B,Ramage D,Talwar K,et al.Learning differentially private recurrent language models[J].arXiv e-print,2017,arXiv:1710.06963
[25] Yang H.H-FL:a hierarchical communication-efficient and privacy-protected architecture for federated learning[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. August 19-27, 2021, Montreal, Canada. California: International Joint Conferences on Artificial Intelligence Organization,2021: 479-485
[26] Papernot N, Abadi M, Erlingsson Ú, et al. Semi-supervised knowledge transfer for deep learning from private training data [J]. arXiv e-print, 2016, arXiv:1610.05755
[27] Papernot N, Song S, Mironov I, et al. Scalable private learning with pate [J]. arXiv e-print, 2018, arXiv:1802.08908
[28] Mao Y L,Zhu B Y,Hong W B,et al.Private deep neural network models publishing for machine learning as a service[C]//2020 IEEE/ACM 28th International Symposium on Quality of Service (IWQoS). June 15-17, 2020,Hang Zhou,China.IEEE,2020:1-10
[29] Triastcyn A,Faltings B.Bayesian differential privacy for machine learning [J]. arXiv e-print, 2019, arXiv:1901.09697
[30] Truex S, Liu L, Chow K H, et al. LDP-Fed: federated learning with local differential privacy[C]//Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking. Heraklion, Greece. New York,NY,USA:ACM,2020:61-66
[31] Yu L,Liu L,Pu C,et al.Differentially private model publishing for deep learning[C]//2019 IEEE Symposium on Security and Privacy (SP).May 19-23,2019,San Francisco,CA,USA.IEEE,2019:332-349
[32] Cheu A,Smith A,Ullman J,et al.Distributed differential privacy via shuffling[M]//Advances in Cryptology:EUROCRYPT 2019. Cham: Springer International Publishing,2019:375-403
[33] Bittau A,Erlingsson Ú,Maniatis P,et al.Prochlo:strong privacy for analytics in the crowd[C]//Proceedings of the 26th Symposium on Operating Systems Principles. Shanghai,China. New York,NY,USA:ACM,2017.DOI: 10.1145/3132747.3132769
[34] Balle B,Barthe G,Gaboardi M.Privacy amplification by subsampling:tight analyses via couplings and divergences [J].arXiv e-print,2018,arXiv:1807.01647
[35] Girgis A M,Data D,Diggavi S,et al.Shuffled model of federated learning:privacy,accuracy and communication trade-offs[J].IEEE Journal on Selected Areas in Information Theory,2021,2(1):464-478
[36] Liu R X,Cao Y,Chen H,et al.FLAME:differentially private federated learning in the shuffle model[J].arXiv e-print,2020,arXiv:2009.08063
[37] Bell J H,Bonawitz K A,Gascón A,et al.Secure single-server aggregation with (poly) logarithmic overhead[C]//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security.New York,NY, USA:ACM,2020:1253-1269
[38] Bhowmick A, Duchi J, Freudiger J, et al. Protection against reconstruction and its applications in private federated learning [J]. arXiv e-print, 2018, arXiv:1812.00984
[39] Sun L C,Qian J W,Chen X.LDP-FL:practical private aggregation in federated learning with local differential privacy[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence.August 19-27, 2021, Montreal, Canada. California: International Joint Conferences on Artificial Intelligence Organization,2021: 1571-1578
[40] Mohassel P,Zhang Y P.SecureML:a system for scalable privacy-preserving machine learning[C]//2017 IEEE Symposium on Security and Privacy (SP).May 22-26, 2017,San Jose,CA,USA.IEEE,2017:19-38
[41] Thapa C,Chamikara M A P,Camtepe S,et al.SplitFed: when federated learning meets split learning[J].arXiv e-print,2014,arXiv:2004.12088
[42] 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 e-print,2017,arXiv:1711.10677
[43] Tramèr F,Boneh D.Slalom:fast,verifiable and private execution of neural networks in trusted hardware[J].arXiv e-print,2018,arXiv:1806.03287
[44] Sun L C,Lyu L J.Federated model distillation with noise-free differential privacy [C] // Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence.August 19-27,2021,Montreal,Canada.California:International Joint Conferences on Artificial Intelligence Organization,2021:1563-1570
[45] Liu Y, Kang Y, Xing C P, et al. A secure federated transfer learning framework [J]. IEEE Intelligent Systems,2020,35(4):70-82
[46] Bonawitz K,Eichner H,Grieskamp W,et al.Towards federated learning at scale:system design[J].arXiv e-print, 2019,arXiv:1902.01046
[47] Truex S,Baracaldo N,Anwar A,et al.A hybrid approach to privacy-preserving federated learning [C] // Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security(AISec'19).November 15,2019, London,United Kingdom.New York:ACM Press,2019: 1-11
[48] Xu R H,Baracaldo N,Zhou Y,et al.HybridAlpha:an efficient approach for privacy-preserving federated learning [C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security(AISec'19).November 15, 2019,London,United Kingdom.New York:ACM Press, 2019:13-23
[49] Boneh D,Sahai A,Waters B.Functional encryption:definitions and challenges[M]//Theory of Cryptography. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011:

253-273
[50] Fereidooni H,Marchal S,Miettinen M,et al.SAFELearn:secure aggregation for private federated learning[C]// 2021 IEEE Security and Privacy Workshops (SPW).May 27,2021,San Francisco,CA,USA.IEEE,2021:56-62
[51] Mao Y L,Hong W B,Wang H,et al.Privacy-preserving computation offloading for parallel deep neural networks training [J]. IEEE Transactions on Parallel and Distributed Systems,2021,32(7):1777-1788
[52] Kairouz E B P,McMahan H B.Advances and open problems in federated learning[J].Foundations and Trends ® in Machine Learning,2021,14(1):1-122

Survey on private model publishing for federated learning

SHI Congcong[1] GAO Xianzhou[1] HUANG Xiuli[1] MAO Yunlong[2]
1 Global Energy Interconnection Research Institute Nanjing Branch/State Grid Key Laboratory of Information & Network Security,Nanjing 210094
2 Department of Computer Science and Technology,Nanjing University,Nanjing 210023

Abstract Federated learning is a kind of distributed machine learning technology to ensure that local data is not compromised when training with big data for machine learning models.However,a series of attacks shows that the adversary can steal private information from machine learning model parameters even if local data is inaccessible.Thus,many privacy threats must be mitigated,since they can arise from the intermediate model parameters transmitted between participants and the aggregator in the training phase or from the finally released aggregated model.Therefore,various privacy-preserving federated learning approaches have emerged,primarily based on cryptography and differential privacy technology.This paper surveys the privacy threats and adversary models that may appear when we publish local models and aggregated model of federated learning.Furthermore,we systematically summarize the related defense technologies and research advances.Finally,we also presents a prospect for the development trend of privacy-preserving federated learning.

Key words federated learning;privacy-preserving;differential privacy