联邦学习中隐私攻击与防御综述

王恺楠[1,2], 张玉会[1,2], 侯 锐[1]

[1] 中国科学院信息工程研究所信息安全国家重点实验室 北京 中国 100093
[2] 中国科学院大学 网络空间安全学院 北京 中国 101408

摘要 联邦学习作为一种新兴的分布式机器学习框架。在保护用户隐私的同时实现数据共享与模型训练, 已逐渐成为人工智能领域的重要研究方向。该方法通过多个数据提供方共同训练机器学习模型, 能够在不泄露原始数据的前提下完成模型更新和优化。近年来, 联邦学习因其在医疗、金融等领域的广泛应用而备受关注。然而, 随着技术的不断发展, 学术界也提出了多种针对联邦学习框架的攻击手段。本文对联邦学习领域中常见的攻击方法进行了系统性分析与分类。通过对现有攻击方法的不同属性进行深入研究, 本文提出了基于攻击特性的分类策略, 并基于这一分类策略对已有攻击方法进行了全面的总结、归纳和介绍。例如, 根据攻击方法的目标性质, 本文将其划分为模型污染, 数据污染攻击、成员推理, 重建推理攻击等类别。此外, 为了解决这些漏洞, 学术界在联邦学习框架内提出了多种防御策略。针对现有的多种攻击模型, 本文还总结了一系列防御策略。这些防御策略主要基于防御原理, 包括鲁棒聚合、模型对抗, 差分隐私方法以及同态加密、多方计算等技术。通过系统地总结和分析现有的防御模型, 本文不仅为理解现有防护机制提供了清晰的框架, 也为未来研究方向提供了新的思路。例如, 如何在有限资源条件下实现不同粒度的防御策略, 如何在压缩通信量的同时保持模型训练效果的提升, 以及将非监督学习应用到联邦学习等问题均成为未来值得深入探索的研究方向。

关键词 联邦学习; 隐私攻击; 机器学习
中图法分类号 TP309.02 DOI 号 10.19363/J.cnki.cn10-1380/tn.2025.03.14

Survey of Privacy Attack and Defense in Federated Learning

WANG Kainan[1,2], ZHANG Yuhui[1,2], HOU Rui[1]

[1]State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 101408, China

Abstract Federated learning, as a new distributed machine learning framework, realizes data sharing and model training while protecting user privacy, and has gradually become an important research direction in the field of artificial intelligence. This method trains the machine learning model through multiple data providers, which can update and optimize the model without sharing the raw data. In recent years, federated learning has attracted much attention due to its wide application in medical, financial and other fields. However, with the continuous development of technology, the academic community has also proposed a variety of attacks against the federated learning framework. This paper systematically analyzes and classifies the common attack methods in the field of federated learning. By deeply studying the different attributes of existing attack methods, this paper proposes a classification strategy based on attack characteristics, and summarizes, generalizes and introduces the existing attack methods based on this classification strategy. For example, according to the target attributes of the attack method, this paper divides it into model contamination, data contamination attack, member inference, reconstruction inference attack and so on. In addition, to address these vulnerabilities, multiple defense strategies have been developed within the federated learning framework. This paper summarizes a series of defense strategies against various attack models. These defense strategies are mainly based on defense principles, including robust aggregation, model antagonism, differential privacy methods, homomorphic encryption, multi-party computing and other technologies. By systematically summarizing and analyzing the existing defense models, this paper not only provides a clear framework for understanding the existing defense mechanisms, but also provides new ideas for future research. For example, how to implement different granularity defense strategies under limited resources, how to compress the communication while maintaining the improvement of model training effect, and how to apply unsupervised learning to federated learning have become the future research directions worthy of further exploration.

Key words federated learning; privacy attack; machine learning

---

通讯作者: 侯锐, 研究员, 主要研究方向包括处理器芯片设计与安全、AI 芯片安全, Email: hourui@iie.ac.cn。
本课题得到中国科学院战略性先导科技专项(No. XDC02010200)资助。
收稿日期: 2021-02-10; 修改日期: 2021-03-09; 定稿日期: 2023-08-11

1 介绍

人工智能在近十年来成为最为重要的研究问题之一, 而人工智能中最受欢迎的则是机器学习算法。机器学习算法关键在于训练过程, 使用标记好的数据集对算法模型中的参数进行更新, 当达到一定的条件后可以认为这个模型被训练完毕。随后训练好的算法模型会被应用在对应的需求场景中, 例如图像分类, 自然语言处理等。机器学习算法训练过程通常需要多次迭代, 算法模型才能更好的收敛。大量的训练数据有助于得到一个更加准确的算法模型。如今, 大量移动、可穿戴设备逐渐进入日常生活。不只是手机, 智能手表, 智能眼镜都开始渐渐成为一些用户的常用设备。为了更好的训练机器学习算法模型, 许多互联网公司采用各种手段收集来自用户的数据, 甚至出现了利用用户身边的设备进行监视的现象。在这些被收集到的数据中, 不可避免地存在用户的隐私数据, 例如个人信息, 每日行程等。随着大量用户数据被应用在机器学习算法训练中, 用户的隐私安全成为了 AI 落地面临和亟需解决的问题。例如, 2018 年, Facebook 被曝光其用户信息泄露给剑桥分析公司。该公司通过个人信息与好友关系等数据, 创建用户画像, 向潜在支持者定向推送竞选广告, 被指操纵美国大选。引发了人们对 Facebook 的信任危机和竞选公平性的质疑。就在同一年, 中国消费者协会展开问卷调查, 回收有效问卷 5458 份, 调查结果显示, 遇到过信息泄露的人高达 85.2%, 他们接收到了推销电话、诈骗电话、垃圾邮件, 甚至出现账号密码被盗的情况。2018 年 3 月, 欧盟出台的《通用数据保护条例》(General Data Protection Regulation, GDPR)也正式生效[1], 该条例对企业处理用户数据的行为提出了明确的要求。如何保护用户的隐私数据逐渐成为机器学习算法中的关键问题。

为了更好保护用户的隐私数据, 许多机器学习的隐私保护框架被提出。其中, 联邦学习[2]是应用最广, 讨论最多的一种隐私保护框架。联邦学习通常的应用场景是多个用户协作训练一个算法模型。在联邦学习中, 用户的原始数据并不会离开本地。与之对应的, 被各方共享的是机器学习算法中的梯度值。也就是说, 算法的训练被搬运到了用户本地。在这样的机制下, 用户数据可以实现物理隔离, 保证了隐私安全性。

1.1 联邦学习中的隐私问题

在机器学习发展早期, 就有很多攻击机器学习算法的方法与防御框架被提出。对机器学习算法的攻击可以分为三类。第一类叫做模型提取(Model extraction), 攻击者通过循环发送数据并查看对应的响应结果来推测机器学习的参数和功能, 从而复制出一个功能相似甚至相同的机器学习模型。这种攻击方法由 Tramèr 提出, 并分别针对逻辑回归, 决策树和神经网络进行了有效攻击[3]。第二类叫做模型逆向(Model inversion)[4], 攻击者将训练好的线性模型以黑匣子的方式提供给受害者。同时, 攻击者根据此黑匣子提供的接口来获取模型的一些初步信息, 并通过初步信息对模型进行逆向分析, 最终获取模型信息。第三类叫做成员推断(Membership inference)[5], 攻击者通常训练多个模仿正常行为的算法模型, 随后据此训练攻击模型[5]。攻击通过推断的手法决定某一数据样本属于训练集与否。与此对应的, 许多防御模型也被建立起来。常见防御模型有差分隐私(Differential privacy)[6-7], 模型折叠(Model stacking)[8], 安全多方计算(Secure multiparty computation)[9]。这些都是针对机器学习算法的安全问题提出的模型, 旨在解决算法训练时存在的安全问题。

从算法训练的过程来讲, 联邦学习依然与传统的机器学习算法训练保持一致。尽管联邦学习通过物理隔离将用户隐私数据保护了起来。但是联邦学习依然面临着传统机器学习中的某些攻击方法的威胁。例如机器学习攻击中的推断攻击, 就可以在不读取隐私数据的情况下, 根据算法迭代中的参数, 对隐私数据的具体内容做出推断。

在联邦学习的概念被提出后, 许多学者致力于研究攻击联邦学习框架的方法和防御模型。对这些研究成果的结构化的分析和总结, 有助于研究人员更好的理解联邦学习中的隐私安全问题。虽然, 已有一些工作对这些研究成果进行了概述, 但是现有的概述中存在: (1) 重点不突出的问题[10], (2) 总结不够具体的问题[11]。因此, 本文致力于深入分析现有攻击和防御模型, 对主要的研究工作做出总结分类, 并依照类别对常见攻击和防御模型进行介绍。

这篇综述将会按照如下组织来展开, 第二节介绍联邦学习的背景知识, 第三节介绍联邦学习中的攻击分类框架。接着第四节会介绍当前学术界针对联邦学习的各种攻击细节。针对不同攻击方法而提出的防御模型将会在第五节进行分类和细节说明, 第六节则会根据当前研究现状给出一些可能的研究方向。这篇文章的结论会在第七节给出。

2 背景

2.1 联邦学习起源与概念

联邦学习的概念在 2017 年被 McMahan[2]在谷歌

研究中心提出。联邦学习是由多个参与者(Client, 数据提供者), 和一个服务中心(Server)组成, 并且, 参与者与服务中心共同合作训练某一机器学习模型[10]。每一个参与者(Client)的隐私数据都被存储在本地并且不会被交换或转移。联邦学习主要解决了机器学习中隐私安全的问题。在联邦学习模型中, 算法训练不再是中心化的, 而是分散在每个参与者(Client)中。如图 1.1 所示, 全局模型, 也称作服务中心(Server)会将算法模型参数发送给每一个参与者(Client), 随后每个节点都对自己的私有训练数据进行更新。每一轮更新后, 全局模型采集部分或所有参与者的更新参数, 然后根据不同的聚合(Aggregation)策略更新全局参数。大量数据被分布式存储可以减轻服务中心的压力, 同时每个节点的隐私数据并不会被服务中心或其他参与者读取, 也保证了一定的隐私安全性。

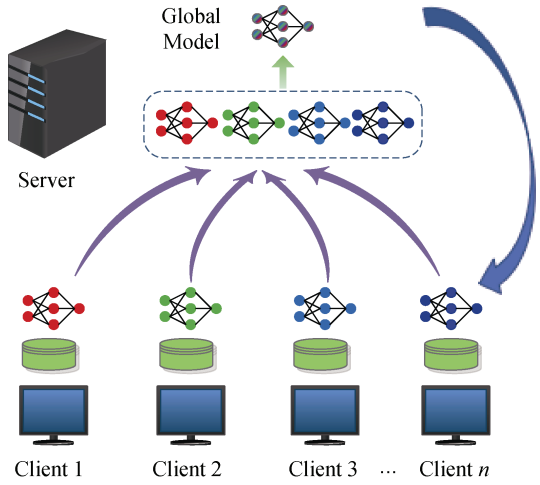


图 1　联邦学习框架

Figure 1　Federated learning framework

联邦学习借鉴了数据最小化的原则。训练数据不必和训练模型存储在一起。同时, 利用机器学习算法中更新公式只需要梯度向量这一特性, 将训练模型和更新参数巧妙地分离开来。分布式的思想既保证了数据分布化也保证了隐私加固。联邦学习还有一个重要特征是数据未加密。在常见的数据保护方法中, 加密都是一个不可或缺的环节。但是在联邦学习这种分布式学习的框架下, 加密过程会大规模加剧边缘节点的资源耗费。同时加密过后的数据可能会对全局模型的聚合产生消极的影响。这一点后续章节会具体分析。最早的分布式加密思想来自于 20 世纪 80 年代[12-13]。Srikant 等人[14]也曾提出利用分布式思想保证隐私的具体方法。Jaideep 等人[15]也探索过对于机器学习中隐私保护的问题。在机器学习算法大力发展的近十年, 卷积神经网络, 循环神经网络等新模型的出现, 一定程度上将隐私保护复杂化了。因为研究人员必须针对每一个复杂的算法模型做出对应分析。在这样复杂的背景下, 联邦学习这样的设置更好的结合了大部分机器学习算法的特征, 更好的协调了隐私保护和性能维持之间的平衡。这也是联邦学习成为目前学术界最受欢迎的安全框架以及研究热点的原因[10]。

2.2　联邦学习算法模型

在联邦学习的设置中, 通常分为如下步骤。首先算法模型会被服务中心(Server)初始化并发送给每一位参与者(Client); 随后参与者根据本地隐私数据对模型进行迭代训练, 每一轮迭代都会产生一个梯度向量; 服务中心也在每一次迭代后都收集参与者输出的更新向量, 并进行汇总。这一步叫作梯度聚合, 这也是联邦学习中最为关键的一步。在经过大量迭代计算后, 全局模型的准确度已经达到期望水平, 这时算法训练就可以结束。随后算法模型会被传输给应用方, 在对应的应用场景中发挥作用。

可以看到, 联邦学习框架中通常有两个参与角色。如图 1.1 所示, 第一个是服务中心(Server), 通常是现实世界中的服务器端。服务中心负责全局模型的更新和整体把控, 联邦学习流程中最重要的聚合就是由服务中心来完成。服务中心拥有着待训练算法的模型信息, 并将算法模型分发给每一个参与者(Client), 随后负责收集所有或其中一部分的参与者(Client)传送回来的梯度向量。第二个是参与者(Client), 通常是现实世界中的终端设备, 例如手机, 物联网设备。参与者负责接收来自服务中心的数据并依据不同的更新公式输出对应的梯度更新值。算法随后被传送给应用方, 在现实生活中可以是手机端以及各种物联网设备。应用方只需要接收训练好的全局模型并应用在当前智能化场景中即可。

因为机器学习中算法模型数量过于庞大, 这里就不展开论证。只讨论通用的参数更新公式。如公式(1)所示

$$\omega(t+1)=\omega(t)-\eta\sum_{k=1}^{K_t}\frac{n_t}{n}\nabla\omega_k(t) \tag{1}$$

其中 t 代表当前迭代轮数, 则 t+1 代表下一轮。ω 代表机器学习算法中的参数, 常数 η 代表学习率。后面的参数 k 代表参与者(Client)的标号, 即共有 K_t 个参与者, $\frac{n_t}{n}$ 代表每一个参与者的权重。对每一个参与者的梯度更新值 $\nabla\omega$ 进行加权平均, 这一步加权平均也就是关键的梯度聚合。随后将聚合得到的最终梯度更新值乘以学习率 η, 来对全局模型的参数 ω 进

行更新。

当然, 加权平均只是梯度聚合中最常见的一种方法。在综合衡量效率, 安全和准确度后, 学术界还提出了许多其他的聚合方法, 分为如下四类。首先做出如下假设, 当前环境中有 m 个参与者; 这 m 个参与者中最多有 c 个恶意的攻击者; 服务中心进行梯度聚合时挑选 j 个参与者的数据。

第一类方法被称作切尾均值(Trimmed mean)。在每一轮更新中, 将挑选出的 j 个模型进行排序, 然后去掉这 j 个模型中更新参数最大和最小的两个[16]。随后对剩下的 j-2 个模型进行加权平均, 也就是做聚合。

第二类方法被称作中位数(Median)聚合。与第一类相似, 在每一轮更新中将挑选出的 j 个模型进行排序, 随后采用这个序列中的中位数作为聚合值[16]。

第三类方法被称作 KRUM 聚合。Peva 等人提出, 为了削弱来自恶意模型的影响, 在选择参与者时需要有一定的限制[17]。首先每一轮迭代结束后, 找到 m 个模型中和大部分其他模型相似的一个模型作为中心模型。这么做的目的是即使这个中心模型被恶意的控制, 也能根据它相似度的变化幅度迅速识别出来, 这样可以过滤掉一些恶意污染攻击。随后在这个中心模型周围找到 m-c-2 个和它欧氏距离最近的模型, 随后将挑选出来的所有模型进行聚合。数据分析表明, 当 c 小于 m-2 的二分之一时, 这个聚合方法可以在理论上保证安全的聚合[17]。

第四类方法被称作 Bulyan 聚合。在 KRUM 聚合中, 欧氏距离的使用会被模型参数影响[18], 因此某些异常参数就会大大影响聚合效率。Bulyan 聚合的方法因此被提出。每一轮聚合都从 m 个模型中挑选 j 个, 并保证 j 小于 m-2c。然后使用切尾均值(Trimmed mean)的方法对这 j 个模型进行聚合。数据分析表明, 当 c 小于 m-3 的四分之一时, 这个聚合方法可以在理论上保证安全的聚合[18]。

联邦学习的框架吸取了以往框架的优点, 又在探索性能方面提出了更高的要求。如何解决学习框架中的沟通效率和安全问题, 将一直是联邦学习社区的研究热点[10]。

3 攻击分类框架

尽管联邦学习的模型在设计之初就考虑了隐私数据的安全性, 安全问题依然没有得到彻底解决。分布式学习的思想对算法做了分离保护处理, 但是某种程度上也会将攻击者更好的隐藏起来。攻击者可能恶意的控制联邦学习框架中的任意一方, 然后使用既定的方法对隐私数据进行窃取攻击。

在真实的攻击场景中, 不同的攻击模型的操控对象, 攻击方法, 攻击目标等存在差异。因此, 针对这些攻击模型的不同特征对其进行细致化分类。

3.1 黑盒子攻击和白盒子攻击

黑盒子攻击模型指代攻击者无法掌握数据内部信息, 只能依靠观测模型输出做出攻击; 而白盒子攻击模型则指攻击者可以掌握模型内部信息, 并对模型本身做出一定攻击。

在联邦学习框架中, 因为隐私数据存在物理隔离, 因此白盒子攻击模型中的攻击者掌握的信息只包括完整的算法模型, 梯度更新以及模型参数等信息, 并不包括隐私数据。框架中的参与者(Client)因为拥有完整的算法模型信息和自己本地的隐私数据, 因此更容易受到白盒子攻击。另一方面, 黑盒子攻击的目标通常是服务中心(Server)。服务中心拥有的信息通常过少, 会受到黑盒子攻击。

3.2 目标化攻击和非目标化攻击

非目标化的攻击一般指攻击采用的方法会使算法不能完成训练, 导致算法最终无法成功。换言之, 攻击者的首要任务就是使得算法模型不能很好的收敛。这样的攻击下, 不仅会浪费训练者大量的训练时间和空间, 还可能导致算法模型崩溃, 无法达到预期目标。在现实生活中, 这样的攻击产生的效果可能会更大。而目标化的攻击则通常不会影响算法模型的整体准确度, 攻击者的首要任务是让算法模型针对某些特定的数据产生预期的恶意变化而不对其他分类数据产生影响。这种攻击也被称作后门(Backdoor)攻击。一个常见例子是, 在医学图像识别中, 攻击者攻击某一类别, 例如肿瘤, 的训练数据。最终算法模型不能准确检测对应的医学图像, 同时还有可能产生误判。这样的攻击在真实世界中会产生很大的危害性。

这两种攻击不同之处在于是否对算法收敛产生影响。从隐蔽性的角度, 非目标化的攻击可以提前被受害者发现, 而目标化的攻击更具有隐蔽性, 很可能在产生大量错误后才会被察觉。这是一种通过判断是否对算法收敛产生影响而做出的攻击分类。

3.3 攻击场景分类

首先需要明确的是攻击者所在的场景。可以简单地将攻击场景分为三类, 三种攻击场景中攻击者的最终目标和衡量攻击是否成功的标准都是不同的。

第一种攻击场景是数据推断(Data inference), 攻击者的目标是更加具体化的学习隐私数据。这些数据首先可能是训练集的具体包含对象, 也就是训练

者隐私数据中存在哪些具体的分类对象。攻击者试图重建的是包含在已有类别中的数据, 并不属于不存在的新类别, 这种情况下攻击者窃取到的信息会包含大量隐私数据。其次, 这些数据可能是训练集的数据属性, 也就是训练者隐私数据中不同数据的特征, 例如某一个训练集中黑色的数字一最多。再其次, 这些数据可能是训练集的类别, 也就是训练者的数据中包含哪些标签, 例如本次分类训练的是不同种类的动物, 可能包括狗, 猫, 鸟等标签。最后这种数据可能仅仅是训练数据是否被使用过, 攻击者想要窃取的数据是某一个数据点是否被此次数据迭代使用过。

第二种攻击场景是样本重建(Sample reconstruction), 攻击者的目标在于重构参与者的训练数据。因此可以将重构的训练数据和原始训练数据的相似度作为衡量攻击是否成功的标准。在这样的场景中, 攻击者往往窃取的是训练数据的全部内容, 因此难度往往较大。从隐私保护的角度来看, 当一个攻击者可以操控参与者或者服务中心来窃取完整的训练数据, 这个平台框架可以被认为存在隐私安全问题。

第三个攻击场景是模型污染(Model poisoning), 攻击者的目标是向有利于攻击者的方向改变算法模型。在这样的攻击背景下, 攻击者可以通过植入恶意训练数据的方式改变模型的分类属性。结果是算法模型可能错误的做出预测或者算法模型无法达到收敛。模型污染的攻击场景并没有对参与者的隐私数据进行窃取, 某种程度上并没有过多破坏学习框架中的隐私性。不同的攻击场景在攻击方法和攻击目标上都有所不同, 这是一种依据应用场景而做出的分类方式。

3.4 攻击方向分类

在联邦学习的框架中, 存在两个框架对象, 分别是服务中心(Server)和参与者(Client)。从攻击的角度来看, 攻击者可以分别控制服务中心和参与者。因此根据攻击者利用的对象, 可以把联邦学习中的攻击分为两个方向。

第一类是从服务中心(Server)端向参与者(Client)发起攻击。在这样的攻击背景下, 攻击者可以观测到的要素有算法模型的参数, 每次更新迭代中模型参数的更新向量。攻击者可以窃取的要素有参与者的本地数据类别, 参与者的数据内容等。

第二类是从参与者(Client)端向服务中心(Server)发起攻击。在这样的攻击方式下, 攻击者可以掌握到的信息有恶意参与者的隐私数据和算法模型参数。攻击者可以窃取的信息有未被恶意控制的参与者的隐私数据内容和类别, 全局模型的梯度更新向量等。

3.5 攻击方式分类

根据联邦学习中不同的攻击方式, 可以分为污染(Poisoning)攻击和推理(Inference)攻击。

首先解释污染攻击。顾名思义, 污染攻击就是攻击者通过恶意植入, 将现有正常模型或者是数据进行污染, 从而使得算法模型向着攻击者预想的方向进行更新, 最终产生严重的后果。在污染攻击中, 再根据攻击目标的不同, 分为模型污染(Model poisoning)攻击和数据污染(Data poisoning)攻击。

推理攻击也可以从字面上理解, 攻击者试图通过推断的方法, 来重构隐私数据。例如参与者的训练集类别, 数据内容等。而推断的方法, 既可以交给常见的分析工具, 也可以交给专用的恶意神经网络来做推断。我们继续将推理攻击分为两类, 第一类叫做成员推理(Membership inference), 攻击者猜想是否特定的数据记录被包含在了训练集中。第二类叫做重建推理(Reconstruction inference), 攻击者通过不同的猜想模型, 将想要窃取的隐私数据进行重建, 猜想训练集中的数据属性。

总体来说, 本文会根据攻击方式的类别对联邦学习攻击做介绍, 表格 1 详细介绍了每一种具体的攻击方法对应的类别信息, 攻击介绍分为四大类别

1) 模型污染(Model poisoning)攻击
2) 数据污染(Data poisoning)攻击
3) 成员推理(Membership inference)攻击
4) 重建推理(Reconstruction inference)攻击

4 联邦学习中的攻击

4.1 模型污染攻击

第一个攻击方法是 Bagdasaryan 等人[19]提出的后门攻击。该论文第一次提出比起数据污染攻击, 模型污染攻击会对联邦学习造成更严重的威胁。一个恶意的参与者可以通过模型替换给算法模型加入一个后门模型。例如, 强迫单词预测器使用攻击者要求的恶意词汇造句, 或者修改一个图像识别器在特征空间加入攻击者选择的恶意标签。实验表明, 控制少于 1%参与者的攻击者就可以有效的进行后门攻击, 同时还不会降低模型预测其他分类类别的准确度[19]。不仅如此, 这个后门攻击还可以在使用不常见的聚合方法 KRUM[17]时产生更大的破坏力。另外, 文中还提出了一种叫做 constrain and scale 的技巧来破解可能存在的恶意探测防御。

另一个经过改良的模型污染攻击是 Arjun 等人[20]提出的长度依赖攻击。当联邦学习框架中的参与者数量过多时, 单个参与者对全局模型产生的污染往

表 1 联邦学习攻击分类

Table 1 Federated learning attack classification

	后门攻击[19]	长度依赖攻击[20]	女巫攻击[21]	拜占庭污染[23]	ATTFL[25]	基于 GAN 攻击[26]	mGAN-AI[28]	白盒子推理[29]	DLG[30]	基于 ReLu 攻击[31]
目标或非目标	目标	目标	非目标	非目标	目标	目标	目标	目标	目标	目标
攻击方向	参与者发起	参与者发起	服务中心发起	参与者发起	参与者发起	参与者发起	参与者发起	参与者发起	都存在	服务中心发起
攻击方式	模型污染	模型污染	模型污染	模型污染	数据污染	成员推理	成员推理	成员推理	重建推理	重建推理
黑盒子或白盒子攻击	白盒子	白盒子	黑盒子	白盒子	白盒子	白盒子	白盒子	白盒子	黑盒子	黑盒子

往没有那么强烈。为了增强单个模型污染对训练过程的影响, 同时加强这个攻击的隐蔽性, 交替最小化的策略被提出。在此策略中, 攻击者只需控制少量参与者, 就可以影响模型对于某些特定分类类别的准确度。另一方面, 与以往的模型污染不同, 此攻击只污染模型中的一部分参数, 不对整个模型做替换。Arjun 还提出一定的防御策略。当一次更新传递过来后, 服务中心可以对两方面做检查。第一个是在已验证的数据集上, 这个更新参数是在增强还是减弱当前模型性能。第二个是这个更新参数是否和其他正常参数有很大不同。通过类似于这样的检查, 模型污染攻击的隐蔽性就会大大减少。最后, 为了继续削弱安全检查的威力。攻击中还可以保证恶意的训练与无害的训练之间的数学距离尽量接近于两个正常的训练之间的距离, 使得攻击更不容易被探测。

其次, 不同于常见污染攻击增大恶意参与者数量的方式, Fung 等人[21]提出新的方法。该论文指出, 攻击者应选择不同时刻控制不同的参与者, 也就是说尽管恶意参与者的数量很少, 但是累计产生恶意影响的参与者却很多。这种对于分布式系统的攻击其实可以归类为女巫(Sybil)攻击[22]。通过这样的方式, 攻击者最终可以在单次迭代中返回大量被污染的模型。

最后, Minghong 等人[23]提出一种模型污染攻击。与其他模型污染不同的是, 这个攻击首次将其应用在拜占庭鲁棒(Byzantine-Robust)系统上。拜占庭鲁棒系统是为了防御拜占庭攻击者(Byzantine attacker)的系统。拜占庭攻击者常被定义为训练过程中可能会传递任意错误信息的参与者, 起因有数据污染, 交流失败以及恶意攻击。攻击者在控制本地模型后, 恶意的篡改更新参数。在这样的情况下, 本地模型常常会有很大的出错率。这个攻击被应用在实际的数据集中, 包括 MNIST, Fashion-MNIST, CHMNIST 等, 都展现出了不错的效果。

4.2 数据污染攻击

数据污染攻击可以成功的被攻击者用来做标签翻转[24]。一个例子是 Sun 等人[25]提出的 ATTFL 攻击。在这个攻击中, 恶意参与者可以通过数据污染改变梯度更新方向。这篇攻击还提出了一个最优化计算模型, 用来计算联邦学习中最佳的污染攻击。具体的做法有根据更新规则修正注入数据的标签, 随后进行梯度计算, 最后输出已被污染的梯度和损失函数值。因为联邦学习的框架原因, 每个参与者的数据都是私有的, 所以很少有攻击专注于数据污染, 这篇文章也是为数不多的数据污染攻击案例。

从联邦学习配置的角度看, 因为参与者的训练数据是私有的, 攻击者在恶意控制参与者后也只能局限于当前计算窗口。同时, 攻击者不管是通过数据污染改变更新向量影响全局模型, 还是直接通过模型污染的方法去影响全局模型, 最终的攻击效果大同小异。因此从某种意义上来说, 数据污染攻击也可以被看作是在做模型污染攻击。

4.3 成员推理攻击

成员推理攻击中, 攻击者会利用各种推理方法来对私密信息进行推断。一个推断方法是使用 GAN 网络[26]。GAN 网络是一种通过对抗思想生成人类可以识别数据的一种神经网络[27]。在联邦学习框架中, 这个攻击也是第一次利用 GAN 网络。使用 GAN 网络来攻击联邦学习, 使得任意参与者都可以读取到受害者的隐私数据, 更严重的是, 攻击还可以影响训练过程, 使受害者释放更多信息。这种利用神经网络做攻击的方法更加的通用有效, 以至于可以攻破以往很难攻破的 CNN 网络。更进一步的是, 当学习框架采用差分隐私这样的防御方法后, 这个攻击依然有效[26]。与黑盒子攻击下攻击者只能看到模型输出的情况不同, 这个攻击方法归属于白盒子攻击。攻击者需要看到训练中的参数, 这一点不算是这个攻击的缺点。因为联邦学习的初衷就是参数共享, 即使共享的比例很小, 白盒子攻击依然可以开展。攻击最终的目标是通过 GAN 网络生成与受害者隐私数据相同的数据, 最终完成成员推理攻击。

另一个攻击在GAN网络攻击的基础上进行改良, 就是 Zhibo 等人[28]提出的 mGAN-AI 攻击。尽管利用 GAN 攻击的方法可以在全局数据上对类别做区分, 但是依然无法做到从具体的某一个参与者做攻击, 该论文第一次探索用户级别的隐私泄露, 从服务中心发起攻击, 可以恢复特定用户的隐私数据。之前的利用 GAN 攻击模型训练的方法存在三个局限性。首先攻击需要一个强力的攻击者, 同时保证它可以深刻的影响训练过程, 这在现实生活中不是很常见, 属于理想情况。其次这类攻击容易受到梯度下降影响, 因为攻击者的训练参数可能会因为联邦学习的特性, 也就是平均化参数传递, 而减少对训练模型的影响。最后这类攻击只能对通用的数据分类做攻击, 无法具化到某一类特定的分类上。针对这样的局限性, 改良版的攻击在不影响训练过程的情况下提出了一种更通用, 实际, 可视化强的攻击方法。从服务中心方向做出攻击, 探索参与者的隐私问题。有一个比较局限的点是这个攻击只能应用在存在大量数据的联邦学习框架下, 同时恶意控制的参与者必须有一个协助参与者才能完成攻击。

Milad 等人[29]提出一种白盒子推理攻击。该攻击利用了随机梯度下降(Stochastic gradient descent)算法的漏洞。在随机梯度下降算法中, 为了减小训练中的损失, 算法会向着逼近误差为零的方向上重复训练模型。在这样的背景下, 每一个训练数据都会在梯度更新中留下独一无二的痕迹。该论文还发现, 即使是在全局模型准确度很高的情况下, 恶意的参与者也可以利用这个攻击反抗其他参与者。实验数据表明, 即使是训练拟合度很好的神经网络也有可能泄露训练集的敏感信息, 因此也更容易受到白盒子推理攻击的影响。该论文还从另一个角度对推理攻击做了区分, 主动的攻击讲的是训练的参与者可以影响模型的最终训练结果, 而消极的攻击则是在不改变训练过程的前提下攻击者仅做一些观测。与此同时, 该论文还对联邦学习的隐私泄露做出了一定的量化分析, 算法训练中预测向量的不确定性被发现与隐私泄露存在一定的相关性[29]。

4.4 重建推理攻击

一个攻击是由 Ligeng 等人[30]提出的 DLG 攻击。在联邦学习的设置中, 参数共享是被允许并且认为是安全的, 但是这篇攻击证明了这种参数共享的机制会产生很大的安全问题。攻击者可以从分享的梯度数据上窃取敏感信息。作者将攻击方法应用在现实生活中的图像处理和自然语言处理网络上, 都取得了不错的效果。在攻击中, 只需要迭代中的梯度数据就能重建精确到像素点的图像和精确到token的语言序列。相对以往的攻击需要更多的信息, 这个攻击依据很少的信息就能重建出敏感信息, 可以说是向前迈了一大步。在具体的攻击中, 首先在一个独立的模型中初始化一些杂乱无章的输入输出与梯度值。在正常用户训练时使用独立模型进行恶意训练, 不断匹配正常用户分享的梯度值。最终逼近正常用户的训练数据, 最终完成重建推理攻击。

另一个重建推理攻击由 Sannai 等人[31]提出。该论文对激活函数使用ReLu深度神经网络的损失函数进行了量化分析, 并最终发现由输入计算出的损失函数值与输出存在一定的相关性。因为在神经网络的每一层都使用了 ReLu 激活函数, 攻击者可以通过反向计算找到被激活的节点。在这种输入输出关系可以中采用多项式进行拟合。最终, 这些多项式就可以被使用计算和观测到的损失函数匹配的私有训练数据了。这种重建攻击还存在一定的局限性, 只可以被使用在采用 ReLu 激活函数的线性模型中。另一方面, 因为攻击方法中严格的数学推导, 这个攻击对噪声敏感。

5 联邦学习隐私保护方法

这部分主要介绍在联邦学习框架中的防御方法。通常来说, 联邦学习的框架可以被看作服务中心通过请求其余节点的计算并收集结果的过程。那么每个节点, 也就是参与者, 都可以被抽象为一个函数, 服务中心只需要给出输入并收集输出即可。我们把这个函数称作参与者函数。在这个过程中, 有三个关于隐私的问题需要解决。

首先, 我们需要考虑参与者函数是如何计算的以及函数计算时数据流是怎样分布的。因为这会影响到包括服务中心, 参与者和其余算法参与方的敏感性。在早期设计算法框架中数据流分布时, 包括安全多方计算(Secure multiparty computation)[9], 可信执行环境(Trusted Execution Environments)等方法都被考虑解决隐私问题。其次, 我们需要考虑什么样的数据被计算了。换言之, 参与者函数在计算中暴漏了多少数据给其他算法参与方。针对这方面的考虑, 用来对抗隐私泄露的技术, 尤其是差分隐私(Differential privacy)[6-7], 被广泛使用。最后, 还需要考虑的问题是身份验证。身份验证指代参与者或者服务中心向系统或其他算法参与方证明可信身份的能力。还需要注意的是, 身份验证采用的数学计算不能用到超出框架限制的多余隐私数据, 否则身份验证本身就会存在安全风险。针对这方面的考虑, 包括

远程证明(Remote attestation), 零知识证明(Zero-knowledge proofs)在内的技术被广泛使用。

安全多方计算可以被看作加密学的子领域, 旨在解决多个节点在只输出约定俗成的数据的情况下如何共同完成一项任务的问题。这个领域最早被 Yao 等人[32]提出, 归功于大量突破性学术成果[33-36], 这个理论随后被广泛应用在学术界和工业界。通常密码学的解决方案是在有限的空间里完成计算, 这种方法在节点数量巨大时会存在很大困难。一个比较通用的方法是通过细致的和归一化的量化方法将参与者节点的计算过程控制在非溢出的状态。

归因于联邦学习的分布式框架, 可信执行环境可以提供一种平台, 将联邦学习中的部分算法迭代过程移动到云端安全环境上。在云上安全环境中, 参与者们共享的数据可以被安全的验证。可信执行环境已经被应用在多个方面, 英特尔的 SGX 技术[37], 来自 ARM 的 TrustZone[38], 多种应用都可提供包括保密性, 完整性和可信性的属性保证。尽管如此, 可信执行环境仍然存在局限性, 可信环境遭受各种各样的测信道攻击[39]。

差分隐私的核心思路是在不使用加密的情况下, 通过对为加密数据加入一定规律的噪声来影响攻击者的攻击模型。在联邦学习的背景下, 差分隐私需要根据分布式系统做出一定改变, 这一点已经有一些优秀的研究成果[40-41]。在这样的模型下, 参与者首先计算模型输出并加入一定的噪声, 随后输出给服务中心。服务中心的算法模型也要保证对于加入的噪声有鲁棒性。经这样的处理可能损失一些准确度, 但是在安全性上的提升可以盖过性能的损失。这样的背景下, 差分隐私就是一种有效的隐私保护方法。

远程证明在某种程度上可以看作是可信执行环境的一个具体应用。而零知识证明则是一种密码学的方法, 两个节点在不暴露自身隐私数据到验证节点的前提下进行双方互验。最早是由 Goldwasser 等人[42]提出。随后 Parno 等人[43]经过改良并简化做出了更具体的应用。发展到今天, 零知识证明已经可以实现百字节水平的校对和毫秒级别的验证速度。

常见的隐私保护方法在应用到具体的框架中时都需要做出一定改良。在联邦学习应用中也不例外, 分布式的学习系统对防御框架提出了更高的要求。本文会对一些具有代表性的防御框架做出陈述性介绍。不同防御模型的分类详见表 2。

表 2 防御模型分类

Table 2 Defense model classification

	RSA [44]	拜占庭鲁棒框架[16]	来源可靠性分析[45]	Cronus[46]	anti-GAN[47]	服务中心差分隐私[48]	贝叶斯差分隐私[49]	DSGD [51]	cryptonets[52]	spark [54]	Aby3 [58]
鲁棒聚合	√	√	√	√							
模型对抗					√						
差分隐私						√	√				
同态加密								√	√		
随机失活										√	
安全多方计算											√

5.1 鲁棒聚合

在联邦学习框架中, 参与者只需向服务中心输出梯度更新值即可。那么参与者就存在修改样本数据以及传递错误信息的情况, 这种情况下就需要一个更加鲁棒性的系统框架。

第一个框架是 RSA[44]。这个框架是用来防御拜占庭攻击者(Byzantine attacker)的, 拜占庭攻击者被视为在训练过程中可能会传递任意错误信息的节点。该论文提出了一种具有高鲁棒性的随机策略, 针对联邦学习的隐私安全问题, 进行了优化。具体的做法包括在算法更新公式中加入一个距离惩罚量, 这个距离惩罚量被定义为参与者与服务中心自变量的数学距离乘以一个系数。同时, 这个系数会根据算法计算的迭代过程做出相应调整。因此, 这个惩罚量的加入, 可以使得RSA对于拜占庭攻击者具有鲁棒性。在这样的情况下, 只有来自拜占庭攻击者的数量而不是攻击者的非法数据可以对更新产生消极影响。

其次, Dong等人[16]提出了一种成功的防御框架。这种防御模型专注于实现最佳化统计性能。曾经的防御方法有两种思路。首先是将模型简化只输出 0, 这会大大减少算法模型的精准度。其次是减少精度, 输出一些不容易被攻击的数据, 这种思路会对性能产生消极影响。该论文做出了两点优化, 分别针对性能和通信效率。并最终提出了两种具有鲁棒性的分布式梯度下降算法。两种算法在参数平均化上存在差异, 一个是正常加权平均, 另一个是切尾均值。这

一点在第二节中有具体介绍。

与之对应的, Li 等人[45]也曾提出了一种具有鲁棒性的防御框架, 可以被称作来源可靠性分析框架。与 RSA 相比, 这个防御模型更加轻量化。同时, 作者将数据可靠性进行了量化分析, 使用聚合后的梯度更新值与参与者的梯度更新值之间的距离作为衡量可靠性的标准。Chang 等人[46]也提出了自己的鲁棒防御框架。一种聪明的知识共享算法被作者提出, 这种算法依赖一个所有参与者可见的公共数据集。参与者不再向服务中心输出梯度更新值, 而是输出根据公共数据集推断出来本地数据的标签值。

5.2 模型对抗

前面提到, 利用 GAN 神经网络做推理攻击会产生巨大的破坏性。与之对应的, 一种防御基于 GAN 网络做推理攻击的防御框架被提出[47]。这个防御框架也被命名为 anti-GAN。该论文中指出, 防御基于 GAN 网络的攻击方法存在两个难点。第一个是不能采用加密方法, 因为加密计算对于边缘节点的资源耗费巨大, 同时对于全局模型的聚合会产生消极的影响。第二个是减少参数共享并不会阻止 GAN 网络对图像模式进行学习, 仍然会重建出私有图像数据。也就是说, 只要受害者的模型准确度不是很低, 那么 GAN 网络的攻击就会有效。基于以上分析, 该论文设计了一个中间网络, 使得输入到 GAN 网络里的重建图像与私有图像不仅有相似的分类特征, 还有一些不可分辨的可视化特征, 来抵御这种猜测攻击。这种中间网络就是受害者自己的 GAN 网络, 它会训练出与人类不可辨别的与原始图像相似特征的新图片。用这个新图片作为输出更新全局模型, 受害者的 GAN 网络一直在无监督学习, 作为对抗, 目标是使得攻击者还原出来的图像不能被人类识别。这样巧妙地防御方法正是使用了 GAN 网络来对抗自己的思想, 可以说是十分成功。

5.3 差分隐私

差分隐私[6]的主要原理是通过加入一定的噪声来缓解攻击算法的精度。尽管差分隐私凭借优秀的性能被应用在各种机器学习算法中, 他却不能直接应用在联邦学习的框架中。差分隐私要求算法参与者掌握大量数据集, 在联邦学习的框架中, 参与者往往只拥有自己的本地数据。Bonawitz 等人[48]提出了一种应用差分隐私的方法。作者将差分隐私的思想应用在联邦学习中的服务中心端。这样可以保证参与者不能攻击其他参与者, 因为服务中心做聚合的参数对于每一个参与者都是带噪声干扰的。这样做的局限性在于联邦学习框架更可能受到来自服务中心的攻击。

Triastcyn 等人[49]也提出了一种应用差分隐私的方法。作者提出一种贝叶斯差分隐私的防御框架。这种防御框架可以通过增强数据相似度保证更快的聚合度。贝叶斯差分隐私比正常的差分隐私会加入更少的噪音, 这样也会进一步提升框架的准确率。

5.4 同态加密

前面提到, 联邦学习中不采用加密的数据是因为边缘节点会耗费大量计算资源。同态加密则是一种具有优秀对称性的数学加密算法, 在同态加密中, 处理经过同态加密的数据处理并将输出进行解密, 最终结果与用同一方法处理未加密的原始数据得到的输出结果是一样的。已经有研究证明, 同态加密的方法在联邦学习中只会造成很小的性能损失[50]。

同态加密可以防御不安全的服务中心, 这样的工作已经被 Phong 等人[51]提出。作者还提出了一种安全的梯度下降算法 DSGD, 并将其成功应用在了联邦学习框架上。与之对应的, Dowlin 等人[52]也提出了一种叫做 cryptonets 的网络用来实现同态加密算法。这种方法要求参与者将共享的参数以加密形式输出, 同时将密钥保留在本地隐私数据中。这样来自服务中心的攻击者就很难使用已经加密的数据进行攻击。

5.5 随机失活

另一种防御思路是利用深度学习中一种优化方法增加防御性, 这种技术被称作随机失活(Dropout)[53]。尽管这种优化方法被广泛应用, 他仍然被证明会对梯度更新产生一些随机化的影响。随机失活的使用, 会在训练单个数据节点时对梯度更新产生不确定的影响, 这会降低攻击者可以利用的窗口, 从而做出防御。将随机失活应用在联邦学习上的方法也在不断地被研究人员提出[54-55]。

5.6 安全多方计算

常见的安全多方计算算法[9]通常针对两个算法参与方, 两方共享的数据只会被两位参与方知晓, 不会泄漏到第三方。从联邦学习的框架来看, 这样的安全多方计算可以应用在本地参与者(Client)之间的数据共享。这样做的好处是所有参与者都可以防御来自服务中心的攻击。为了将安全多方计算更好的应用在复杂的联邦学习框架上, 一些三方数据共享的防御框架也在被提出[56-57]。

Mohassel 等人[58]提出了一种更复杂的多方安全计算框架。在这个防御框架中, 最多存在三个服务中心。当共享数据的两方或者三方进行了安全验证计算后, 这些算法参与方中的服务中心就会被承认为

可信赖的。与此同时, 可信赖的服务中心对应的参与者(Client)也可以被进一步认为是可信赖的。

6 联邦学习未来存在的研究方向

联邦学习自提出后就凭借优秀的框架和强大的兼容性被广泛使用, 但是学术界不断提出的隐私攻击方法证明了这个框架存在一定隐私安全问题。本文会在现有学术工作的基础上, 对联邦学习未来研究方向提出一定展望。主要分为以下三点。

第一, 如何实现不同粒度的隐私保护。在隐私泄露的攻击中, 不同的方法窃取的隐私数据也不尽相同。某些信息可能只是当前样本点是否被使用过, 而某些信息可能是完整的训练样本。目前学术界存在的防御模型并没有很好的涵盖所有粒度数据。因此, 必须对不同粒度下的隐私泄露做分析, 最后完成不同粒度的隐私保护。

第二, 如何大量压缩通信量。在当前隐私攻击的方法中, 大部分的攻击目标都和通信数据有关。压缩通信数据在某种程度上可以减少信息泄露量。如何在更快速准确的压缩信息量, 以及如何在压缩信息的同时保证信息的完整性, 这都是研究工作需要重点关注的地方。同时, 压缩数据的算法不宜设计过于复杂。耗费大量资源对数据进行压缩会产生得不偿失的效果。

第三, 探索联邦学习中使用非监督学习的可能性。目前在联邦学习的框架中都是采用的监督学习对全局模型进行训练。在真实世界中, 还存在许多样本数据并没有明确的标签分类或者没有标签。同时, 学术界现存的隐私攻击方法和防御模型都建立在监督学习的前提下。不可排除联邦学习的算法数据必须采用非监督学习的情况, 例如算法参与方在不同时间段展现不同的行为。当新的特性被应用到联邦学习框架中时, 如何更好的利用这个特性做攻击以及对特定的攻击做出防御都将是一个值得研究的方向。

7 结论

本文详细的讨论了联邦学习中的研究热点, 隐私安全问题。尽管联邦学习通过物理隔离保护用户隐私数据保护了起来。但是部分传统机器学习中的攻击方法对联邦学习仍然有效。因此, 联邦学习中的隐私安全问题一直在对框架推出更高的要求。针对隐私安全不同的攻击和防御方法也在不断对抗中出现。攻击方法可以分为污染攻击和推理攻击, 分别聚焦于模型的破坏和隐私数据的重构。防御方法又可以分为差分隐私、同态加密、模型对抗和鲁棒聚合等类别。文章利用分类学的方法分别从攻击和防御角度对现有工作做出了系统化的讨论。同时, 论文背景部分也对联邦学习概念的起源和算法模型进行了详细的介绍。最后, 论文还给出了联邦学习可能存在的未来研究方向。

参考文献

[1] Council of European Union. General data protection regulation. https://gdpr-info.eu/. May 2018.

[2] McMahan B, Ramage D. Federated learning: Collaborative machine learning without centralized training data. https://www.googblogs.com/federated-learning-collaborative-machine-learning-without-centralized-training-data/. Mar. 2017.

[3] Tramèr F, Zhang F, Juels A, et al. Stealing Machine Learning Models *via* Prediction APIs[C]. *The 25th USENIX Conference on Security Symposium*, 2016: 601-618.

[4] Fredrikson M, Jha S, Ristenpart T, et al. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures[C]. *The 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015: 1322-1333.

[5] Shokri R, Stronati M, Song C Z, et al. Membership Inference Attacks Against Machine Learning Models[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 3-18.

[6] Zuba M. Cynthia Dwork on Differential Privacy[J]. *XRDS*: *Crossroads*, *the ACM Magazine for Students*, 2013, 20(1): 58-59.

[7] Shokri R, Shmatikov V. Privacy-Preserving Deep Learning[C]. *2015 53rd Annual Allerton Conference on Communication*, *Control*, *and Computing*, 2015: 909-910.

[8] Salem A, Zhang Y, Humbert M, et al. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models[EB/OL]. 2018: 1806.01246. https://arxiv.org/abs/1806.01246v2.

[9] Bonawitz K, Ivanov V, Kreuter B, et al. Practical Secure Aggregation for Privacy-Preserving Machine Learning[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 1175-1191.

[10] Kairouz P, McMahan H B, Avent B, et al. Advances and Open Problems in Federated Learning[J]. *Foundations and Trends® in Machine Learning*, 2021, 14(1/2): 1-210.

[11] Li T, Sahu A K, Talwalkar A, et al. Federated Learning: Challenges, Methods, and Future Directions[J]. *IEEE Signal Processing Magazine*, 2020, 37(3): 50-60.

[12] Rivest R L, Adleman L, Dertouzos M L. On data banks and privacy homomorphisms[J]. *Foundations of secure computation*, 1978, 4(11): 169-180.

[13] Yao A C. Protocols for Secure Computations[C]. *23rd Annual Symposium on Foundations of Computer Science*, 1982: 160-164.

[14] Agrawal R, Srikant R. Privacy-Preserving Data Mining[J]. *ACM SIGMOD Record*, 2000, 29(2): 439-450.

[15] Vaidya J, Yu H, Jiang X Q. Privacy-Preserving SVM Classifica-

tion[J]. *Knowledge and Information Systems*, 2008, 14(2): 161-178.

[16] Yin D, Chen Y D, Ramchandran K, et al. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates[EB/OL]. 2018: 1803.01498. https://arxiv.org/abs/1803.01498v2.

[17] Blanchard P, El Mhamdi E M, Guerraoui R, et al. Machine Learning with Adversaries[C]. *The 31st International Conference on Neural Information Processing Systems*, 2017: 118-128.

[18] El Mhamdi E M, Guerraoui R, Rouault S. The Hidden Vulnerability of Distributed Learning in Byzantium[EB/OL]. 2018: 1802.07927. https://arxiv.org/abs/1802.07927v2.

[19] Bagdasaryan E, Veit A, Hua Y Q, et al. How to Backdoor Federated Learning[EB/OL]. 2018: 1807.00459. https://arxiv.org/abs/1807.00459v3.

[20] Bhagoji A N, Chakraborty S, Mittal P, et al. Analyzing Federated Learning through an Adversarial Lens[EB/OL]. 2018: 1811.12470. https://arxiv.org/abs/1811.12470v4.

[21] Fung C, Yoon C J M, Beschastnikh I. Mitigating Sybils in Federated Learning Poisoning[EB/OL]. 2018: 1808.04866. https://arxiv.org/abs/1808.04866v5.

[22] Douceur J R. The Sybil Attack[M]. Peer-to-Peer Systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002: 251-260.

[23] Fang M H, Cao X Y, Jia J Y, et al. Local Model Poisoning Attacks to Byzantine-Robust Federated Learning[EB/OL]. 2019: 1911.11815. https://arxiv.org/abs/1911.11815v4.

[24] Biggio B, Nelson B, Laskov P. Poisoning Attacks Against Support Vector Machines[EB/OL]. 2012: 1206.6389. https://arxiv.org/abs/1206.6389v3.

[25] Sun G, Cong Y, Dong J H, et al. Data Poisoning Attacks on Federated Machine Learning[EB/OL]. 2020: 2004.10020. https://arxiv.org/abs/2004.10020v1.

[26] Hitaj B, Ateniese G, Perez-Cruz F. Deep Models under the GAN: Information Leakage from Collaborative Deep Learning[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 603-618.

[27] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks[J]. *Communications of the ACM*, 2020, 63(11): 139-144.

[28] 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*, 2019: 2512-2520.

[29] Nasr M, Shokri R, Houmansadr A. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-Box Inference Attacks Against Centralized and Federated Learning[EB/OL]. 2018: 1812.00910. https://arxiv.org/abs/1812.00910v2.

[30] Zhu L G, Han S. Deep Leakage from Gradients[M]. Federated Learning. Cham: Springer, 2020: 17-31.

[31] Sannai A. Reconstruction of Training Samples from Loss Functions[EB/OL]. 2018: 1805.07337. https://arxiv.org/abs/1805.07337v1.

[32] Yao A C. How to Generate and Exchange Secrets[C]. *27th Annual Symposium on Foundations of Computer Science*, 1986: 162-167.

[33] Bogdanov D, Talviste R, Willemson J. Deploying Secure Multi-Party Computation for Financial Data Analysis[C]. Financial Cryptography and Data Security, 2012: 57-64.

[34] Bogetoft P, Christensen D L, Damgård I, et al. Secure Multiparty Computation Goes Live[C]. Financial Cryptography and Data Security, 2009: 325-343.

[35] Ion M, Kreuter B, Nergiz E, et al. Private Intersection-Sum Protocol with Applications to Attributing Aggregate Ad Conversions[J]. *IACR Cryptol EPrint Arch*, 2017: 738.

[36] Ion M, Kreuter B, Nergiz A E, et al. On Deploying Secure Computing: Private Intersection-Sum-with-Cardinality[C]. *2020 IEEE European Symposium on Security and Privacy*, 2020: 370-389.

[37] Intel Corporation. Architecture instruction set extensions programming reference. https://software.intel.com/content/dam/develop/external/us/en/documents-tps/architecture-instruction-set-extensions-programming-reference.pdf. Feb. 2012.

[38] Pinto S, Santos N. Demystifying Arm TrustZone: A Comprehensive Survey[J]. *ACM Computing Surveys*, 51(6)Article No. 130.

[39] Van Bulck J, Minkin M, Weisse O, et al. Foreshadow: Extracting the Keys to the Intel SGX Kingdom with Transient Out-of-Order Execution[C]. *USENIX Security Symposium*, 2018.

[40] Dwork C, Kenthapadi K, McSherry F, et al. Our Data, Ourselves: Privacy via Distributed Noise Generation[M]. Advances in Cryptology - EUROCRYPT 2006. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 486-503.

[41] Shi E, Chan T H H, Rieffel E, et al. Privacy-preserving aggregation of time-series data[C]. *In Annual Network & Distributed System Security Symposium*. 2011, 2: 1-17.

[42] Goldwasser S, Micali S, Rackoff C. The Knowledge Complexity of Interactive Proof-Systems[J]. 2019: 203-225.

[43] Parno B, Howell J, Gentry C, et al. Pinocchio: Nearly Practical Verifiable Computation[C]. *2013 IEEE Symposium on Security and Privacy*, 2013: 238-252.

[44] Li L P, Xu W, Chen T Y, et al. RSA: Byzantine-Robust Stochastic Aggregation Methods for Distributed Learning from Heterogeneous Datasets[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 1544-1551.

[45] Li Q, Li Y L, Gao J, et al. Resolving Conflicts in Heterogeneous Data by Truth Discovery and Source Reliability Estimation[C]. *The 2014 ACM SIGMOD International Conference on Management of Data*, 2014: 1187-1198.

[46] Chang H Y, Shejwalkar V, Shokri R, et al. Cronus: Robust and Heterogeneous Collaborative Learning with Black-Box Knowledge Transfer[EB/OL]. 2019: 1912.11279. https://arxiv.org/abs/1912.11279v1.

[47] Luo X J, Zhang X L. Exploiting Defenses Against GAN-Based Feature Inference Attacks in Federated Learning[EB/OL]. 2020: 2004.12571. https://arxiv.org/abs/2004.12571v4.

[48] McMahan H B, Ramage D, Talwar K, et al. Learning Differentially Private Recurrent Language Models[EB/OL]. 2017: 1710.06963. https://arxiv.org/abs/1710.06963v3.

[49] Triastcyn A, Faltings B. Federated Learning with Bayesian Differ-

ential Privacy[C]. *2019 IEEE International Conference on Big Data*, 2019: 2587-2596.

[50] Phong L T, Aono Y, Hayashi T, et al. Privacy-Preserving Deep Learning *via* Additively Homomorphic Encryption[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(5): 1333-1345.

[51] Phong L T, Aono Y, Hayashi T, et al. Privacy-Preserving Deep Learning: Revisited and Enhanced[C]. *Applications and Techniques in Information Security*, 2017: 100-110.

[52] Dowlin N, Gilad-Bachrach R, Laine K, et al. CryptoNets[C]. *The 33rd International Conference on International Conference on Machine Learning - Volume 48*, 2016: 201-210.

[53] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors[EB/OL]. 2012: 1207.0580. https://arxiv.org/abs/1207.0580v1.

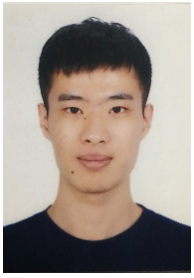
[54] Moritz P, Nishihara R, Stoica I, et al. SparkNet: Training Deep Networks in Spark[EB/OL]. 2015: 1511.06051. https://arxiv.org/abs/1511.06051v4.

[55] 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*, 2019: 691-706.

[56] Araki T, Furukawa J, Lindell Y, et al. High-Throughput Semi-Honest Secure Three-Party Computation with an Honest Majority[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 805-817.

[57] Mohassel P, Rosulek M, Zhang Y. Fast and Secure Three-Party Computation: The Garbled Circuit Approach[C]. *The 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015: 591-602.

[58] Mohassel P, Rindal P. ABY^3: A Mixed Protocol Framework for Machine Learning[C]. *The 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018: 35-52.



王恺楠 于 2018 年在北京航空航天大学飞行器设计(航天工程)专业获得学士学位。现在中国科学院大学网络安全学院攻读硕士学位。研究领域为计算机体系结构、系统安全。研究兴趣包括: 计算机体系结构、硬件安全。Email: wangkainan@iie.ac.cn

侯锐 于 2007 年在中国科学院计算技术研究所获得博士学位。现在中国科学院信息工程研究所国家重点实验室担任研究员。研究领域为处理器芯片设计与安全、AI 芯片安全与数据隐私, 以及数据中心服务器等。Email: hourui@iie.ac.cn

张玉会 于 2018 年在北京交通大学获得学士学位, 现在中国科学院信息工程研究所国家安全重点实验室攻读博士学位。研究领域为隐私安全。研究兴趣包括: 联邦学习, 机器学习, 可信执行环境。Email: zhangyuhui@iie.ac.cn