DOI:10.12379/j.issn.2096-1057.2022.03.03

联邦学习安全威胁综述

王坤庆[1] 刘 婧[2] 李 晨[3] 赵语杭[4] 吕浩然[4] 李 鹏[1] 刘炳莹[1]
[1](中国人民武装警察部队 北京 100089)
[2](齐鲁师范学院生命科学学院 济南 250200)
[3](生态环境部信息中心 北京 100029)
[4](北京理工大学网络空间安全学院 北京 100081)
(282522085@qq.com)

A Survey on Threats to Federated Learning

Wang Kunqing[1], Liu Jing[2], Li Chen[3], Zhao Yuhang[4], Lü Haoran[4], Li Peng[1], and Liu Bingying[1]
[1](*Chinese People's Armed Police Force*, *Beijing* 100089)
[2](*School of Life Sciences*, *QiluNormal University*, *Jinan* 250200)
[3](*Information Center of Ministry of Ecology and Environment*, *Beijing* 100029)
[4](*School of Cyberspace Science and Technology*, *Beijing Institute of Technology*, *Beijing* 100081)

Abstract At present, federated learning has been considered as an effective solution to solve data island and privacy protection. Its own security and privacy protection issues have attracted widespread attentions from industry and academia. The existing federated learning systems have been proven to have vulnerabilities. These vulnerabilities can be exploited by adversaries, whether within or without the system, to destroy data security. Firstly, this paper introduces the concept, classification and threat models of federated learning in specific scenarios. Secondly, it introduces the confidentiality, integrity, and availability (CIA) model of federated learning. Then, it carries out a classification study on the attack methods that destroy the federated learning CIA model. Finally, it explores the current challenges and future research directions of federated learning CIA model.

Key words federated learning; privacy leakage; confidentiality integrity and availability (CIA) model; membership attack; generative adversarial network (GAN) attack

摘 要 当前,联邦学习已被认为是解决数据孤岛和隐私保护的有效解决方案,其自身安全性和隐私保护问题一直备受工业界和学术界关注.现有的联邦学习系统已被证明存在诸多漏洞,这些漏洞可被联邦学习系统内部或外部的攻击者所利用,破坏联邦学习数据的安全性.首先对特定场景下联邦学习的概念、分类和威胁模型进行介绍;其次介绍联邦学习的机密性、完整性、可用性(CIA)模型;然后对破坏联邦学习 CIA 模型的攻击方法进行分类研究;最后对 CIA 模型当前面临的问题挑战和未来研究方向进行分析和总结.

收稿日期:2021-12-21
基金项目:国家自然科学基金项目(61876019)
引用格式:王坤庆, 刘婧, 李晨, 等. 联邦学习安全威胁综述[J]. 信息安全研究, 2022, 8(3): 223-234

关键词　联邦学习；隐私泄露；机密性、完整性、可用性模型；成员攻击；生成对抗网络攻击

中图法分类号　TP181；TP183

随着人们对数据安全和隐私保护的意识逐渐增强，数据安全和隐私保护的重要性开始受到广泛关注.大多数企业出于各种目的不允许原始数据共享交换，不愿意贡献数据价值，导致出现大量数据孤岛和隐私保护问题.联邦学习旨在解决上述数据孤岛和数据安全问题.

联邦学习是一种分布式机器学习技术，允许多方通过本地训练集按照指定算法构建模型.具体来说，联邦学习过程就是联邦学习参与者对本地数据进行训练后，将训练得到的参数上传服务器，服务器聚合得到整体参数[1].根据参与者之间的数据特征和数据样本分布，联邦学习通常分为横向联邦学习(HFL)、纵向联邦学习(VFL)和联邦迁移学习(FTL).横向联邦学习也称为基于样本的联邦学习，是指不同的参与者之间共享数据集[2]的特征空间；纵向联邦学习也称为基于特征的联邦学习，用在参与者数据集的样本空间或特征空间有着明显重叠但又不同的场景中，即不同的参与者对同一条记录数据[3]有相互独立的属性；联邦迁移学习是指参与者之间几乎没有样本空间或者特征空间的重叠[4].近年来，联邦学习系统允许参与者在不暴露本地训练数据的情况下建立一个联合的机器学习模型，联邦学习已被广泛应用在词组预测[5]和视觉目标检测[6]等场景中.

隐私保护是研究联邦学习的一个主要切入点.当前研究主要集中在使用安全多方计算或差分隐私来增强联邦学习隐私安全，这些方法通常以降低模型性能或系统效率为代价来确保隐私安全.诸多研究表明，在训练过程中，联邦学习仍然存在模型更新过程中向第三方或中央服务器透露敏感信息的情况.Yang等人[4]已经证明联邦学习协议设计存在漏洞，如任何一个参与者可能获得全局参数并能控制这些参数的上传.随着训练的推进，服务器不断记录并维护参与者对全局参数的更新[7]，在模型训练和预测阶段，恶意参与者可通过对模型输入输出值的恶意修改来窃取模型参数.Zhu等人[8]提出攻击者仅在几个梯度迭代环节通过短短20多行代码就能窃取训练数据，文献[9]则证明了恶意第三方可以从服务器的共享数据更新中恢复参与者的部分数据.

1　联邦学习的CIA模型

联邦学习安全性要求可以概括为3点：机密性(confidentiality)、完整性(integrity)和可用性(availability)，这就是联邦学习的CIA模型.其中，机密性是指联邦学习系统保证模型不会泄露相关敏感信息；完整性是指联邦学习系统在模型学习和推理预测过程中完全不受干扰，输出结果符合模型的典型性能；可用性是指联邦学习系统可以被普遍使用.对联邦学习系统的攻击将会影响联邦学习数据以及模型的机密性、完整性和可用性，对应的攻击方式有机密性攻击、完整性攻击和可用性攻击.

1) 机密性攻击：联邦学习系统在模型学习阶段或预测阶段泄露敏感和相关信息，包括模型本身的信息(如模型参数、模型结构、训练方法等)以及模型使用的训练数据.

2) 完整性攻击：对模型学习过程和推理预测过程的干扰，使模型的输出结果不符合预期性能.完整性是研究人员相信人工智能模型的基础，也是人工智能模型最容易受到攻击的地方[10].

3) 可用性攻击：攻击者主要利用联邦学习系统中的软件漏洞对训练数据进行恶意操作，使模型无法得到正确更新，导致模型无法被正常使用.

2　对抗联邦学习的攻击模型

表1对破坏CIA模型的不同攻击方法进行了直观总结和比较.下面按照可用性攻击、机密性攻击和完整性攻击的顺序进行详细阐述.

2.1　可用性攻击

联邦学习模型的可用性意味着模型可以被普遍使用.攻击者主要利用联邦学习系统的漏洞，采用数据投毒攻击和拜占庭攻击破坏联邦学习模型的可用性.

表 1　不同攻击方法的总结比较

攻击方法	攻击目标		黑/白盒	主动/被动攻击	攻击意图	造成危害		
	数据	模型				完整性	机密性	可用性
数据投毒攻击	√		黑盒和白盒	主动	数据破坏			√
拜占庭攻击		√	—	主动	破坏可用性	√		√
模型提取攻击		√	黑盒	主动	构造替代模型		√	
成员推理攻击	√		黑盒和白盒	被动	推理获取信息		√	
属性推理攻击	√		白盒	主动和被动	重构特征		√	
生成对抗网络攻击		√	白盒	主动	推理获取信息		√	
梯度信息泄露攻击	√		白盒	被动	重构样本		√	
对抗攻击	√	√	白盒	被动	实现模型分类错误	√		√
模型投毒攻击		√	白盒	被动	推理获取信息	√	√	
后门攻击	√		黑盒和白盒	主动	模型中毒	√		

2.1.1　数据投毒攻击

数据投毒攻击是指攻击者污染了训练集中的样本.数据投毒攻击是攻击者在不改变目标机器学习系统的情况下,构造特定的输入样本来欺骗系统完成攻击.实际应用场景中,常用 3 种攻击方式:直接攻击、间接攻击和混合攻击.直接攻击是指攻击者通过提供假传感器数据的方式锁定目标节点,直接向目标节点注入经恶意修改的有毒数据;间接攻击是指攻击者在不能向任意目标节点注入有毒数据的条件下,利用设备间的通信协议缺陷等漏洞,通过向其他节点注入有毒数据来间接影响目标节点;混合攻击是直接攻击和间接攻击的结合,即攻击者既可以直接向目标节点也可以间接向目标节点注入有毒数据.例如,文献[11]利用直接攻击方法使被学习的模型参数无限接近恶意期望值,从而实现模型在测试样本上输出错误.此外,将少量有毒样本注入训练集的间接方法也获得了超过 90%的成功率.

特洛伊神经网络攻击也属于数据投毒攻击,这种攻击将特洛伊神经网络和目标模型网络打包在一起,将数据同时输入特洛伊网络和目标模型网络,并对它们的输出进行整合,这样就实现了将特洛伊网络分发.

总的来说,数据投毒攻击比较危险,给联邦学习带来一定挑战.其特点如下:一是在分发投毒模型之前,攻击者必须能够访问目标机器学习模型的训练管道;二是数据投毒往往会降低目标学习模型在主要任务上的准确率;三是数据投毒攻击对拜占庭鲁棒联邦学习无效;四是检测二进制文件的反恶意软件工具无法检测经数据投毒攻击的联邦学习算法中的后门[12],这往往使传统的安全检测方法失效.

2.1.2　拜占庭攻击

拜占庭攻击[13]是指攻击者控制多个授权节点,任意干扰或破坏网络.主要实现方法是使数据包延迟或无法送达而导致系统错误.例如,通过篡改数据包使得节点无法按照既定协议处理数据包,这样系统就遭到有目的性的破坏.

目前,联邦学习中对拜占庭攻击的研究主要集中在系统鲁棒性[14]以及拜占庭攻击和其他攻击方法的结合,如和数据投毒攻击方法的结合.文献[15]首次提出通过模型局部投毒的拜占庭攻击来攻击联邦学习的方法,攻击者的目标是在训练阶段破坏学习过程的完整性,通过多次迭代积累的偏差,使学习到的全局模型与攻击前的模型之间的差异变得明显,从而使模型无法正常使用.

实施拜占庭攻击的恶意联邦学习参与者的行为可能是完全任意的,但这些参与者的输出被调整为和正常模型更新相似的分布,这使得恶意联邦学习参与者的行为很难被检测到.归根到底,攻击者的目的是使用拜占庭攻击来控制和改变整个联邦学习模型的局部模型参数,提高全局模型的测试错误率,最终破坏联邦学习系统的可用性.

2.2　机密性攻击

机密性攻击主要破坏模型的机密性,通常通过特定的方法窃取模型信息或通过某种手段恢复

部分训练模型的数据，从而推断出用户的敏感数据.一般来说，破坏联邦学习机密性的攻击方法主要分为3类：模型提取攻击、模型反转攻击和重构攻击.

2.2.1 模型提取攻击

模型提取攻击是一种通过重建相同或相似模型，将替代模型作为目标的攻击方法.具体来说，在黑盒攻击条件下，攻击者试图窃取联邦学习模型的参数或超参数[16]，尽可能完整地重建模型，或构建与目标模型相似的替代模型.模型提取攻击一般从2个角度进行：一是构建替代模型；二是从目标模型中恢复信息.

在构建替代模型时，重点是创建与测试集的精度相匹配的模型[17-20].同时需要攻击者在训练替代模型时，尽可能少地进行查询操作，并尽量去复制决策函数 f 的决策边界.

Tramèr 等人[21]专注于从目标模型中恢复信息，首次提出窃取机器学习分类器参数的攻击，具体就是攻击 BigML 和 Amazon 机器在线学习模型(即功能映射模型和决策树模型)，并提取出和在线学习模型几乎相同的模型.文献[22]进一步开展了超参数窃取和架构提取等工作，黑盒攻击条件下成功推断出神经网络的隐藏模型结构 MLaaS 及其优化过程.

2.2.2 模型反转攻击

模型反转攻击是利用机器学习系统提供的API获取模型的初步信息，并利用这些初步信息对模型进行逆向分析.当成功实施模型反转攻击时，模型反转攻击生成的类成员类似于被攻击模型训练时的输入类.

Fredrikson 等人[23]根据模型的输出推断模型的输入，利用机器学习平台的预测接口通过查询实现对目标模型的推断.例如，对目标模型 f 给定一个输入样本 x，得到相应的输出 $f(x)$.攻击者首先训练一个与目标模型 f 无限接近的替代模型 f'，在 f' 的基础上反向恢复 f 的输入 x'.这样，攻击者就可以恢复 f 的训练集.文献[24]以网络数据流为数据集 x'，训练后得到网络流量分类器 f'，由此获得决定特定数据流的网络流量分类的目标模型 f 以及特定数据流的来源信息 x.

为实现模型反转攻击，攻击者也可以通过从已完成的模型中获取训练集信息来推测目标模型的某些数据.从逆向攻击中推断出的训练集信息可以是训练集的成员信息，也可以是训练集的某些统计特征.因此，模型反转攻击又可以进一步分为成员推理攻击和属性推理攻击.

1) 成员推理攻击

机器学习模型之所以在同一训练数据上对训练参数影响有所不同，过拟合是一个常见的原因，但不是唯一的原因.攻击者的目的是构建一个攻击模型，该模型可以识别目标模型行为中的这些差异，并利用它们来区分目标模型的成员和非成员.在联邦学习中，成员推理的目的是推断特定样本是否属于参与学习的一方或两方的私有训练数据.在成员推理攻击中，攻击者利用目标数据点 d^* 和模型 $h_\theta(x)$ 的访问权限，试图推断 $d^* \in X$，X 为私有训练数据.攻击者可以进行主动和被动的成员推理攻击.在被动情况下，攻击者观察更新后的模型参数，并在不改变任何本地或全局协作训练过程的情况下推断 d^* 是否为私有训练数据.在主动情况下，攻击者可以篡改联邦学习模型训练协议，实现对其他学习参与者的攻击，如利用随机梯度上升的攻击方法获取学习参与者的本地数据信息.

近年来，关于成员推理攻击的研究比较多[25-26]，包括影子模型的成员推理攻击、独立于数据模型的成员推理攻击、仅带标签的成员推理攻击和针对生成网络的成员推理攻击等.下面详细介绍这些成员推理攻击方法.

(1) 影子模型的成员推理攻击

Shokri 等人[26]提出一种成员推理攻击，使用多个影子模型来识别目标模型的行为，影子模型与目标模型的训练过程必须相似.如图1所示，训练影子模型必须使用和训练目标模型同一个平台提供的机器学习 API.目标模型和影子模型虽然具有相同的格式，但相互并不关联.影子模型的训练集和目标模型的训练集可能有所重叠，但必须独立训练得到所有的内部参数.因此，更多的影子模型提供了更多的训练素材，影子模型越多就越精确.

Salem 等人[27]提出的方法放宽了攻击者对目标模型训练集的限制，认为攻击者可以对目标模型训练集一无所知.影子模型的训练集使用了与目标模型不同的数据集，利用目标模型在训练数据上输出向量熵值低而在非训练数据上输出向量熵值高的特点，直接将输出向量的熵值与阈值进行

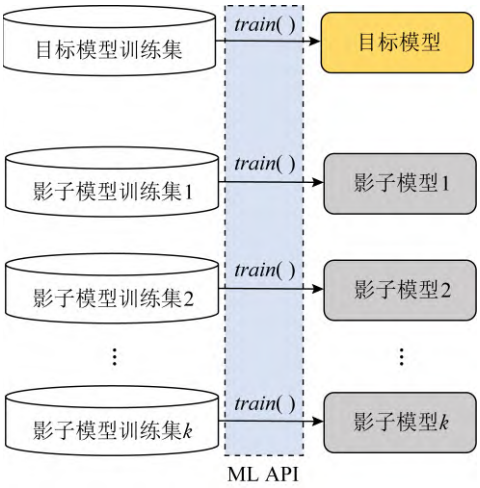

图 1　影子模型训练

比较.如果超过某个阈值,则认为是训练数据;否则,就是非训练数据.

(2) 独立于数据模型的成员推理攻击

介绍此类攻击的 2 种攻击方法.

① 独立于数据的成员推理攻击:攻击者利用与目标模型训练数据不同分布的现有数据集训练其影子模型,这意味着影子模型不是用来模拟目标模型的行为,而只是用来捕捉机器学习训练集中数据点的成员状态.

② 独立于模型和数据的成员推理攻击:攻击者不使用任何影子模型,仅使用目标数据点查询时目标模型返回的结果进行推理,不需要任何训练过程.具体就是利用目标模型的后验统计量,如最大值和熵,来区分成员数据点和非成员数据点.

(3) 仅带标签的成员推理攻击

文献[28]提出仅带标签的成员推理攻击,即在只提供预测标签的情况下,实现对目标模型的成员推理攻击.该文献给出了 2 种攻击方式:基于转移的攻击和基于持续时间的攻击.基于转移的攻击规定,如果本地建立的影子模型与目标模型足够相似,则攻击者可以利用影子模型信息预测目标样本成员.文献[29]仅对标签成员引入推理攻击,该攻击方法不依赖于置信度分数,而是评估模型在干扰条件下预测标签的鲁棒性,以获得细粒度的成员信号.

(4) 针对生成网络的成员推理攻击

Hayes 等人[30]较早提出针对生成模型的成员推理攻击.攻击思路是:利用生成对抗网络检测过拟合并识别训练集,利用鉴别器学习数据分布的统计差异.文献[31]对深度生成模型的成员推理攻击做了详细阐述,揭示了被攻击模型的训练数据信息.Hilprecht 等人[32]提出了针对变分自动编码的成员推理攻击和基于蒙特卡罗积分的成员推理攻击.基于蒙特卡罗积分的成员推理攻击只考虑来自模型的小距离样本,将这些从模型测试集或训练集中抽取的样本进行比较,以实现高精度的成员推理.

2) 属性推理攻击

属性推理的目的是从模型中提取无意中学到的信息,与训练任务无关[33].属性推理攻击可以是主动攻击也可以是被动攻击.攻击者试图提取未明确编码为特征或与学习任务无关的数据集属性,这些属性独立于联邦学习模型特征.该类攻击可以获得更多关于训练数据的信息,攻击者基于这些信息构建与联邦学习模型类似的模型.与成员推理攻击不同,攻击者可以推理出训练输入的子集,但不能推理出训练输入所属整个类的子集.攻击者的攻击内容包括类属性[34-35]和数据属性,并通过分类器观察数据做出推断,达到欺骗联邦学习模型的目的.

最新研究表明,属性推理攻击[36-39]可以有效用于联邦学习,导致信息泄露和模型中毒.文献[35]提出一种针对全连接且相对浅层神经网络的属性推理攻击,主要研究对象是经过训练后发布的白盒模型.由于这些模型是利用敏感数据进行训练的,因此模型属性可能被第三方通过与白盒任务相关或不相关的内容推断出来.Melis 等人[39]研究认为,联邦学习的参与者对每个协作学习迭代的贡献基于他们的训练数据批次,攻击者可以通过属性推理攻击推断出单个批次的特征以及特征出现的时间.

属性推理攻击用于联邦学习系统的限制条件是:

(1) 属性推理攻击需要辅助数据,这些辅助数据有时是难以获取的;

(2) 对于联邦学习参与者的数量,实验一般是在 20～30 个成员条件下进行,实际应用中可能会达到几百个;

(3) 部分属性信息本身不可分离,此时属性推理攻击会失效;

(4) 无法推理出属性信息的来源;

(5) 属性推理攻击依赖上下文环境;

(6) 攻击者需要持有推理出的部分数据来实施攻击.

2.2.3 重构攻击

在重构攻击中,攻击者的目标是重构部分或全部训练样本及标签.模型反转攻击的目标是恢复敏感特征或完整的数据样本,重构攻击侧重于重构出实际数据[40-43]以及可能属于训练集敏感特征的典型类[44-47].重构攻击包括生成对抗网络攻击(GAN)、变分自动编码攻击(VAE)和梯度泄露攻击(DLG).

1) 生成对抗网络攻击

生成对抗网络在生成具有与训练集相同的统计特征的新数据[48-52]方面已经被广泛使用.生成对抗网络原理主要来自博弈论中的零和博弈思想.生成对抗网络用于深度学习神经网络时,通过生成器 G 和鉴别器 D 的博弈使 G 学习数据的分布.G 是一个接受随机噪声并且应用这些噪声产生图像的生成网络.D 用来判断图像是否为"真",其输入是图像 X,输出是 $D(X)$.如果 $D(X)$ 为 1,则输入为 100%的自然图像;如果 $D(X)$ 为 0,则输入的图像不是真实图像.生成对抗网络的基本结构如图 2 所示,其中 Z 为原始输入图片,P_{noise} 为加入的噪声.

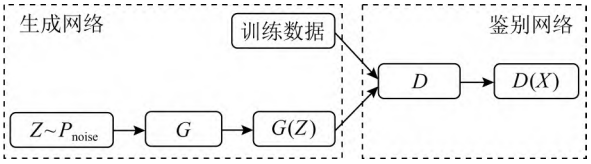

图 2 生成对抗网络基本结构

图 2 左边是生成网络,右边是鉴别网络.在训练过程中,生成网络尽可能生成图像去欺骗鉴别网络,鉴别网络的作用是区分生成网络生成的图像.生成网络和鉴别网络构成了动态"博弈过程",最终的均衡点是 Nash 均衡点.当鉴别器很强大时,生成器的数据仍然会对其造成混淆,使其无法做出正确判断,此时通常认为生成器已经学会了实际数据的分布[53].生成对抗网络中 G 的梯度信息更新必须来自 D,不能是原始样本.G 只与 D 的深度审计网络交互并学习到数据的分布,生成与训练集图像具有相同分布的相似样本.

生成对抗网络攻击构建的只是类的个例,而不是构建实际的训练输入数据.当且仅当特殊情况下所有类成员都相似时,生成对抗网络攻击构建的个例才和训练集数据相似.生成对抗网络攻击也有一些使用限制条件:

(1) 被攻击方参与训练的每轮数据必须有相似的数据分布;

(2) 实际攻击过程中,模型更新必须加入随机噪声;

(3) 并不十分适合处理类似文本数据的以离散形式存在的数据;

(4) 相比训练变分自动编码和像素循环神经网络模型,训练一个生成对抗网络并不稳定;

(5) 需要一定规模的计算资源.

2) 变分自动编码攻击

变分自动编码[54]也是生成模型的一种,该方法的实质是结合深度模型和静态推理将高阶数据映射到低维空间.实现原理为通过变分自动编码中的编码器生成置信度的分布区间,为每个隐藏变量生成一个确定值,并通过抽样得到全新的数据.变分自动编码模型如图 3 所示:

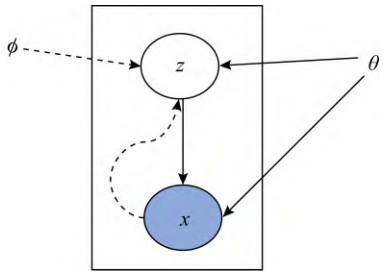

图 3 变分自动编码模型

图 3 中,x 代表数据(如能观测的图片),z 为隐藏变量(往往为多维数据),θ 和 ϕ 为初始化参数.生成模型 $p_\theta(x|z)$ 实现 $z\to x$ 的过程,从变分自动编码的角度看,它起到解码器的作用,重构得到的输出接近于数据 x.识别模型 $q_\phi(z|x)$ 实现 $x\to z$ 的过程,作用类似于编码器.假设 x 为参与联邦学习的数据,它具有多个分类属性,那么 z 就是决定 x 属性的隐藏因子.当攻击者的攻击发起要求 z 满足后,攻击者通过变分自动编码的解码作用就可以得到 x 的相关属性数据,从而达到攻击目的.

尽管生成对抗网络和变分自动编码都提供了

生成模型，但生成对抗网络无法通过编码和解码过程实现对重构图片和原始图片差异的直接比较，变分自动编码则可以实现对攻击获得的重构图片和原始图片的直接比较，这是变分自动编码的重要优势.变分自动编码的缺点是它不使用对抗网络，产生的图片会比较模糊.为了实现高清图像生成，研究人员通过改进变分自动编码方法设计了自省变分自动编码方法(IntroVAE)，较好地解决了这一问题.变分自动编码虽然提供了一个将概率图与深度学习相结合的案例，但并不意味着它提供了一个出色的生成模型，在实际应用中也有很多条件限制.

3) 梯度信息泄露攻击

梯度信息泄露是信息泄露研究的一个重要方面，该方法通过多次迭代获得训练输入数据和标签.以一个基于 PyTorch 实现的 20 行核心代码为例，攻击者利用生成对抗网络思想，在分布式训练条件下，在模型更新梯度过程中，不断生成与其他学习参与者相同的中间梯度数据，通过对这些梯度数据的正向和反向计算，成功“窃取”参与者参与学习的真实数据[8].

梯度信息泄露攻击在实际应用中效果较好，主要是因为它可以恢复像素级精度的原始图像和匹配符号级的原始文本[55].但梯度信息泄露攻击存在以下问题：

(1) 在收敛方面有诸多困难；

(2) 难以连续识别出基本的正确标签；

(3) 模型中的池化层会显著降低效果；

(4) 需要已知模型中参数的分布，且仅适用于具有平均分布初始化的模型，不能恢复正态分布和已经训练好的模型参数.

针对上述问题，文献[56]提出一种基于梯度深度泄露的可微交叉熵损失训练模型，但仅适用于独热标签.文献[57]提出一种基于梯度深度泄露的梯度自适应攻击，在学习梯度分布时使用高斯核来衡量梯度差异.

2.3 完整性攻击

模型在推理阶段和训练阶段最容易受到完整性攻击，模型的完整性一旦被破坏，模型的预测结果就会发生偏离.模型推理阶段最常见的完整性攻击是对抗攻击，模型训练阶段最常见的完整性攻击是模型投毒攻击[58].

2.3.1 对抗攻击

对抗攻击是利用深度学习的缺点破坏识别系统的方法，即通过对识别对象进行肉眼不可见的特殊改动来使模型识别出错.在模型的推理和预测阶段，通过在原始数据中加入精心设计的微扰，攻击者可以获得对抗样本，从而欺骗深度学习模型使其给出高信度的误判.近年来，针对机器学习模型的主要攻击方法就是构造对抗样本，如文献[58-62].逃逸攻击是对抗攻击的一个分支，即攻击者可以在不改变目标机器学习系统的情况下，通过构造特定的输入样本来欺骗目标系统.然而，目前关于联邦学习中实施逃逸攻击的研究还比较少.

在联邦学习中，攻击者在实施对抗攻击时最常见的问题是模型在预测阶段做出了错误的判断，虽然错误分类不会直接侵犯联邦学习参与者的隐私数据[63-64]，但会导致模型的准确性和可用性受到影响.

2.3.2 模型投毒攻击

模型投毒攻击主要指攻击者在全局聚合过程中通过发送错误的参数或破坏模型来扰乱联邦学习过程.例如，通过控制学习参与者传递给服务器的更新参数，影响整个学习过程模型参数走向和降低模型的收敛速度，甚至破坏训练模型的正确性，影响联邦学习模型的性能.

在联邦学习中，由于每个学习参与者都可以直接影响共享模型的权重，攻击者可以轻松地实现对共享模型的投毒，特别是在联邦学习框架中引入安全聚合(SecAgg)机制时，无法保证每个学习参与者的更新数据都经过检查以及安全聚合后的参数都能够在本地得到更新，攻击者可以在本地模型更新数据发送到服务器之前对更新数据进行投毒，也可以在全局模型中植入可供攻击者使用的后门[65-67].

模型投毒攻击分为有目标(针对某一分类的)攻击和非目标(泛化)攻击.所谓非目标攻击实现的是对所有样本的错误分类，并不只针对某一分类样本.文献[67]基于女巫攻击方法[68]提出一种模型投毒攻击方法，在该方法中，攻击者不依赖于增加样本或梯度的数量就能实现在每次迭代中向服务器返回大量有毒模型.文献[69]率先研究了针对拜占庭鲁棒联邦学习的局部模型投毒攻击，其目

标是通过破坏训练阶段学习过程的完整性破坏模型的完整性和机密性.

值得注意的是,在联邦学习中,数据投毒攻击没有将数据发送到服务器,模型投毒攻击则因将数据发送到服务器而需要复杂的技术和较高的计算资源,其综合效果比数据投毒攻击更有效[7].

2.3.3 后门攻击

后门攻击危害性较大,攻击者能够在模型中插入隐藏的后门,并在预测阶段通过触发简单的后门触发器完成恶意攻击.带有隐藏后门的深度神经网络对于非攻击样本表现良好,不易被察觉,但对特定的带有后门触发器的输入样本将实现特定的错误预测.由于后门可以被无限期地隐藏,直到被带有特定后门触发器的样本激活,因此给联邦学习系统带来严重的安全风险.

现有联邦学习框架无法判断本地学习模型的正确性,客户端可以随意提交恶意模型,如带有后门功能的模型[66,70-72],这些模型很难被识别出来.由于模型平均作用,联邦学习参与者能够对最终的全局学习模型权重产生直接影响,带有攻击目的的联邦学习参与者可以故意在全局模型中插入后门.利用联邦学习的这种漏洞,文献[66]提出一种基于模型替换的后门攻击方法,该方法将后门模型 X 的权重缩放为 $\gamma=n/\eta$,确保后门均值能够被保留下来,最终导致全局模型被替换.文献[72]提出一种动态后门攻击,该攻击方法可以实现将后门触发器以多种模式布置在不同位置.

近年来,变分自动编码和生成对抗网络被广泛用于欺诈检测和数据生成等关键领域.文献[72]探索了一种针对变分自动编码和生成对抗网络的后门攻击,其适用性扩展到了基于自动编码器和生成对抗网络的模型.

对抗攻击是利用不同类各个实例之间的边界生成使模型错误分类的输入样本,后门攻击则是通过故意改变上述决策边界使某些输入被错误分类.

3 攻击方法问题分析

表2对上述主要攻击方法的缺点和不足进行了汇总.通过汇总分析发现,破坏联邦学习CIA模型的攻击方法仍然存在以下问题:

表2 联邦学习主要攻击方法缺点和不足

攻击方法	缺点和不足
成员推理攻击	利用模型的过拟合特性,攻击效果会随着模型泛化程度的提高而下降;只适合小数据集
属性推理攻击	有效性取决于上下文,攻击者需要有利于推理的初始数据
生成对抗网络攻击	被攻击方参与每轮的训练数据分布需相似;模型更新时加入随机噪声对方法影响较大;仅适用于稀疏客户集
梯度信息泄露攻击	要求激活函数是二次可微的,与模型大多使用Relu函数作为激活函数相矛盾;模型中的池化层会显著降低方法的效果;需要知道模型中参数的分布情况;只适用于平均分布来初始化的模型,不能还原正态分布和已训练好的模型
对抗攻击	实际应用有限
模型投毒攻击	以牺牲整个系统的鲁棒性为代价,代价较高

1) 现有攻击方法是研究人员在不同条件下提出的,研究中数据集、目标模型和威胁模型差异较大,虽然攻击方法可行,但其有效性仍取决于实际应用场景.例如,攻击者对数据集和模型参数的理解不同,效果可能也截然不同;又如,攻击方法对某些特定用户的影响可能大于其他用户.目前攻击方法的通用性并不强,针对横向联邦学习模型的攻击方法不一定适合纵向联邦学习.

2) 目前大多数攻击方法都是基于联邦学习而不是协作学习,大多数研究人员只考虑了联邦平均算法(FedAvg),忽略了联邦学习中添加的其他保护机制.

3) 现有大部分攻击方法都是被移植到联邦学习中的,没有针对联邦学习框架的特定聚合算法或特征攻击.除非在特定场景中,否则现有攻击方法不一定在联邦学习环境中成功.例如,无论是纵向还是横向联邦学习,都限制了攻击者要使用2个客户端,一个客户端为正常合法用户,主要为攻击者提供联邦学习的全局参数,另一个客户端为攻击发起端,用来推理其他联邦学习参与者的数据.

4 未来研究方向

通过以上梳理发现,无论是从提高攻击效果还是从提高联邦学习安全性,都有很大的研究空间,具体可从以下几个方面入手.

1) 纵向和迁移联邦学习隐私保护:现有攻击方法都是针对横向联邦学习场景,对纵向联邦学

习和迁移联邦学习的研究相对较少，这可能是联邦学习遭受攻击的一个重点方面，也是未来联邦学习研究的一个重要方向.

2）通用攻击方法：无论是横向还是纵向联邦学习场景，都应致力于研究一种通用的攻击方法. 目前的攻击方法使用条件相对苛刻，难以满足需求.

3）联邦学习中的投毒攻击：由于投毒攻击的实施方式比较灵活，且不容易被传统安全检测方法检测到，因而可以将投毒攻击和其他攻击组成混合攻击方法，以达到更好的攻击效果.

4）联邦学习中的后门检测：现有针对联邦学习的攻击方法都是移植实现的攻击方法，未来可以针对这些方法研究联邦学习后门的检测方法.

5）联邦学习的鲁棒性：通过对各种攻击方法原理的掌握，未来可以在提高联邦学习鲁棒性和有效对抗攻击方面做一些深入的研究工作.

5 结束语

本文通过对联邦学习基本概念、安全性要求等内容的阐述，提出了联邦学习 CIA 模型，并根据攻击影响对攻击方法进行了分类；针对每种攻击类型常见攻击方法的攻击原理、攻击效果进行了详细阐述，并对相关重要文献进行了说明；对联邦学习主要攻击方法的缺点和不足进行了汇总，并对未来联邦学习安全性的主要研究方向进行了说明，为相关人员在联邦学习安全性方面的研究工作提供了参考.

参 考 文 献

[1] Yang Q, Liu Y, Chen T, et al. Federated machine learning: Concept and applications [J]. ACM Trans on Intelligent Systems and Technology, 2019, 10(2): 1-19

[2] Kantarcioglu M, Clifton C. Privacy-preserving distributed mining of association rules on horizontally partitioned data [J]. IEEE Trans on Knowledge & Data Engineering, 2004, 16(9): 1026-1037

[3] Vaidya J, Clifton C W. Privacy-preserving kth element score over vertically partitioned data [J]. IEEE Trans on Knowledge & Data Engineering, 2008, 21(2): 253-258

[4] Yang Qiang, Liu Yang, Cheng Yong, et al. Federated Learning [M]. Williston: Morgan & Claypool Publishers, 2019

[5] Mcmahan H B, Ramage D, Talwar K, et al. Learning differentially private recurrent language models [EB/OL]. (2017-10-18) [2021-12-01]. https://arxiv.org/abs/1710.06963v1

[6] Liu Yang, Huang Anbu, Luo Yun, et al. Fedvision: An online visual object detection platform powered by federated learning [EB/OL]. (2020-01-17) [2021-12-01]. https://arxiv.org/abs/2001.06202

[7] Lyu L, Yu H, Yang Q. Threats to federated learning: A survey [EB/OL]. (2020-03-04) [2021-11-12]. https://arxiv.org/abs/2003.02133

[8] Zhu L, Han S. Deep leakage from gradients [EB/OL]. (2019-06-21) [2021-11-12]. https://arxiv.org/abs/1906.08935v2

[9] Aono Y, Hayashi T, Wang L, et al. Privacy preserving deep learning: Revisited and enhanced [EB/OL]. (2017-06-23) [2021-12-02]. https://link.springer.com/chapter/10.1007/978-981-10-5421-1_9

[10] 宋蕾，马春光，段广晗，等. 基于数据纵向分布的隐私保护逻辑回归[J]. 计算机研究与发展，2019，56(10)：2243-2249

[11] Jiang W, Li H, Liu S, et al. A flexible poisoning attack against machinelearning [C] //Proc of 2019 IEEE Int Conf on Communications (ICC). Piscataway, NJ: IEEE, 2019: 1-6

[12] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines [EB/OL]. (2013-03-25) [2021-12-02]. https://arxiv.org/abs/1206.6389

[13] 王健宗，孔令炜，黄章成，等. 联邦学习隐私保护研究进展[J]. 大数据，2021，7(3)：130-149

[14] 李丽萍. 基于模型聚合的分布式拜占庭鲁棒优化算法研究[D]. 合肥：中国科学技术大学，2020

[15] Fang M, Cao X, Jia J, et al. Local model poisoning attacks to byzantine-robust federated learning [EB/OL]. (2021-09-21) [2021-12-02]. https://arxiv.org/abs/1911.11815

[16] Rigaki M, Garcia S. A survey of privacy attacks in machine learning [EB/OL]. (2021-07-15) [2021-12-02]. https://arxiv.org/abs/1910.12366v2

[17] Kalpesh K, Gaurav S T, Ankur P, et al. Thieves on sesame street! model extraction of BERT-based APIs [EB/OL]. (2020-10-12) [2021-12-05]. https://arxiv.org/abs/2007.07646v2

[18] Milli S, Schmidt L, Dragan A D, et al. Model reconstruction from model explanations [C] //Proc of 2018 Conf on Fairness, Accountability, and Transparency. New York: ACM, 2018: 1-9

[19] Orekondy T, Schiele B, Fritz M. Knockoff nets: Stealing functionality of black-box models [C] //Proc of 2019 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4954-4963

[20] Kariyappa S, Prakash A, Qureshi M. MAZE: Data-free model stealing attack using zeroth-order gradient estimation [EB/OL]. (2020-05-06) [2021-12-09]. https://arxiv.org/abs/2005.03161v1

[21] Tramèr F, Zhang Fan, Ari J, et al. Stealing machine learning models via prediction APIs [C] //Proc of the 25th USENIX Security Symp. Berkeley, CA: USENIX Association, 2016: 601-618

[22] Oh S J, Augustin M, Schiele B, et al. Towards reverse engineering black-box neural networks [EB/OL]. (2018-01-14) [2021-12-09]. https://arxiv.org/abs/1711.01768

[23] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures [C] //Proc of the 22nd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1322-1333

[24] Ateniese G, Mancini L V, Spognardi A, et al. Hacking smart machines with smarter ones [J]. International Journal of Security & Networks, 2015, 10(3): 137-137

[25] Backes M, Berrang P, Humbert M, et al. Membership privacy in MicroRNA-based studies [C] //Proc of ACM SIGSAC Conf on Computer & Communications Security. New York: ACM, 2016: 319-330

[26] Shokri R, Song L, Mittal P. Membership inference attacks against adversarially robust deep learning models [C] // Proc of 2019 IEEE Security and Privacy Workshops (SPW). Piscataway, NJ: IEEE, 2019: 50-56

[27] Salem A, Zhang Y, Humbert M, et al. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learningmodels [EB/OL]. (2018-10-14) [2021-12-11]. https://arxiv.org/abs/1806.01246v2

[28] Li Z, Zhang Y. Label-Leaks: Membership inference attack with label [EB/OL]. (2021-09-21) [2021-12-13]. https://arxiv.org/abs/2007.15528v1

[29] Choo C, Tramer F, Carlini N, et al. Label-only membership inference attacks [EB/OL]. (2021-10-05) [2021-12-13]. https://arxiv.org/abs/2007.14321

[30] Hayes J, Melis L, Danezis G, et al. LOGAN: Membership inference attacks against generative models [J]. Proceedings on Privacy Enhancing Technologies, 2019, 2019(1): 133-152

[31] Chen Dingfan, Yu Ning, Zhang Yang, et al. GAN-leaks: A taxonomy of membership inference attacks against GANs [EB/OL]. (2020-09-23) [2021-12-13]. https://arxiv.org/abs/1909.03935v1

[32] Hilprecht B, Hrterich M, Bernau D, et al. Monte carlo and reconstruction membership inference attacks against generative models [J]. Proceedings on Privacy Enhancing Technologies, 2019, 2019(4): 232-249

[33] He Yingzhe, Hu Xingbo, He Jinwen, et al. Overview of privacy and security issues in machine learning systems [J]. Computer Research and Development, 2019, 56(10): 2049-2070

[34] Giuseppe A, Luigi V M, Angelo S, et al. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers [EB/OL]. (2020-09-23) [2021-12-13]. https://arxiv.org/abs/1306.4447v1

[35] Ganju K, Wang Q, Yang W, et al. Property inference attacks on fully connected neural networks using permutation invariant representations [C] //Proc of 2018 ACM SIGSAC Conf. New York: ACM, 2018: 619-633

[36] Wang Zhibo, Song Mengkai, Zhang Zhifei, et al. Beyond inferring class representatives: User-level privacy leakage from federated learning [C] //Proc of 2019 IEEE Conf on Computer Communications (INFOCOM 2019). Piscataway, NJ: IEEE, 2019: 2512-2520

[37] Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: Information leakage from collaborative deep learning [C] //Proc of 2017 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 603-618

[38] Aono Y, Hayashi T, Wang Lihua, et al. Privacy preserving deep learning: Revisited and enhanced [EB/OL]. (2020-09-23) [2021-12-13]. https://link.springer.com/chapter/10.1007/978-981-10-5421-1_9

[39] Melis L, Song Congzheng, Cristofaro E D, et al. Exploiting unintended feature leakage in collaborative learning [C] //Proc of the 40th IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2019: 691-706

[40] Alufaisan Y, Kantarcioglu M, Zhou Y. Robust transparency against model inversion attacks [J]. IEEE Trans on Dependable and Secure Computing, 2021, 18(5): 2061-2073

[41] Wang Zhibo, Song Mengkai, Zhang Zhifei, et al. Beyond inferring class representatives: User-level privacy leakage from federated learning [C] //Proc of IEEE Conf on Computer Communications. Piscataway, NJ: IEEE, 2019: 2512-2520

[42] Yang Ziqi, Zhang Jiyi, Chang E, et al. Neural network inversion in adversarial setting via background knowledge alignment [C] //Proc of 2019 ACM SIGSAC Conf on Computer and Communications Security (CCS 2019). New York: ACM, 2019: 225-240

[43] Zhang Yuheng, Jia Ruoxi, Pei Hengzhi, et al. The secret revealer: Generative model-inversion attacks against deep neural networks [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 253-261

[44] Fredrikson M, Lantz E, Jha S, et al. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing [C] //Proc of the 23rd USENIX Security Symp (USENIX Security 14). Berkeley, CA: USENIX Association, 2014: 17-32

[45] Hidano S, Murakami T, Katsumata S, et al. Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes [C] //Proc of the 15th Annual Conf on Privacy, Security and Trust (PST). Piscataway, NJ: IEEE, 2018: 2665-2676

[46] Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: Information leakage from collaborative deep learning [C] //Proc of 2017 ACM SIGSAC Conf on Computer and Communications Security (CCS). New York: ACM, 2017: 603-618

[47] Salem A, Bhattacharya A, Backes M, et al. Updates-leak: Data set inference and reconstruction attacks in online learning [C] //Proc of the 29th USENIX Security Symp (USENIX Security). Berkeley, CA: USENIX Association, 2020: 1-13

[48] Arjovsky M, BottouL. Towards principled methods for training generative adversarial networks [C] //Proc of the 5th Int Conf on Learning Representations (ICLR). Toulon, France: ICLR, 2017: 1-17

[49] Goodfellow I J. On distinguishability criteria for estimating generative models [EB/OL]. (2015-05-21) [2021-12-15]. https://arxiv.org/abs/1412.6515

[50] Mescheder L, Nowozin S, Geiger A. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks [EB/OL]. (2018-06-11) [2021-12-15]. https://arxiv.org/abs/1701.04722v4

[51] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [EB/OL]. (2016-01-07) [2021-12-10]. https://arxiv.org/abs/1511.06434v1

[52] Salimans T, Goodfellow I J, Zaremba W, et al. Improved techniques for training GANs [C] //Proc of the 10th Advances in Neural Information Processing Systems. Barcelona: NIPS, 2016: 2226-2234

[53] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets [M]. Cambridge, MA: MIT Press, 2014

[54] Doersch C. Tutorial on variational autoencoders [EB/OL]. (2021-01-03) [2021-12-08]. https://arxiv. org/abs/1606.05908

[55] Zhao B, Mopuri K R, Bilen H. iDLG: Improved deep leakage from gradients [EB/OL]. (2020-01-08) [2021-12-08]. https://arxiv.org/abs/2001.02610

[56] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time [C] //Proc of the European Conf on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD). Berlin: Springer, 2013: 387-402

[57] Wang Y, Deng J, Guo D, et al. SAPAG: A self-adaptive privacy attack from gradients [EB/OL]. (2020-09-14) [2021-12-08]. https://arxiv.org/abs/2009.06228v1

[58] Kwon H, Yoon H, Park K W. Selective poisoning attack on deep neural networks [J]. Symmetry, 2019, 11(7): 892-892

[59] Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] //Proc of the 38th IEEE Symp on Security and Privacy (S&P). Piscataway, NJ: IEEE, 2017: 39-57

[60] Papernot N, McDaniel P D, Goodfellow I J, et al. Practical black-box attacks against machine learning [C] //Proc of the ACM Asia Conf on Computer and Communications Security (ASIACCS). New York: ACM, 2017: 506-519

[61] Shafahi A, Huang W R, Najibi M, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks [C] //Proc of the 32nd Int Conf on Neural Information Processing Systems (NeurIPS). New York: ACM, 2018: 6106-6116

[62] Tramer F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses [C] //Proc of the Int Conf on Learning Representations (ICLR). Toulon, France: ICLR, 2017: 1-20

[63] Papernot N, McDaniel P, Sinha A, et al. Wellman. Towards the science of security and privacy in machine learning [EB/OL]. (2016-11-11) [2021-12-08]. https://arxiv.org/abs/1611.03814

[64] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey [J]. IEEE Access, 2018, 6(2): 14410-14430

[65] Chen Xinyun, Liu Chang, Li Bo, et al. Targeted backdoor attacks on deep learning systems using data poisoning [EB/OL]. (2017-11-15) [2021-12-04]. https://arxiv.org/abs/1712.05526v1

[66] Bagdasaryan E, Veit A, Hua Yiqing, et al. How to backdoor federated learning [EB/OL]. (2019-08-06) [2021-12-04]. https://arxiv.org/abs/1807.00459

[67] Fung C, Yoon C J M, Beschastnikh I. Mitigating sybils in federated learning poisoning [EB/OL]. (2020-07-15) [2021-12-03]. https://arxiv.org/abs/1808.04866v5

[68] Douceur J R. The sybil attack [C] //Proc of the 1st Int Workshop on Peer-to-Peer Systems. Berlin: Springer, 2002: 251-260

[69] Yin Dong, Chen Yudong, Ramchandran K, et al. Byzantine-robust distributed learning: Towards optimal statistical rates [EB/OL]. (2021-02-25) [2021-12-05]. https://arxiv.org/abs/1803.01498

[70] Chen Xiaoyi, Salem A, BackesM, et al. BadNL: Backdoor attacks against NLP models [EB/OL]. (2021-10-04) [2021-12-06]. https://arxiv.org/abs/2006.01043v1

[71] Salem A, Sautter Y, Backes M, et al. Dynamic backdoor attacks against machine learning models [EB/OL]. (2020-03-07) [2021-12-06]. https://arxiv.org/abs/2003.03675

[72] Sun Z, Kairouz P, Suresh A T, et al. Can you really backdoor federated learning? [EB/OL]. (2019-11-02) [2021-12-07]. https://arxiv.org/abs/1911.07963

王坤庆

硕士,工程师.主要研究方向为网络与系统安全和智能对抗.

282522085@qq.com

刘　婧

博士,副教授.主要研究方向为生物信息学.

liujing_1205@163.com

李　晨

主要研究方向为信息安全.

li.chen@mee.gov.cn

赵语杭

博士研究生.主要研究方向为人工智能安全.

zhaoyuhang@bit.edu.cn

吕浩然

硕士.主要研究方向为机器学习和智能对抗攻击.

lyuhaoran@bit.edu.cn

李　鹏

主要研究方向为网络安全、信息管理与信息系统应用.

723352284@qq.com

刘炳莹

主要研究方向为信息安全和信息系统应用.

174432256@qq.com