计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20200966
58(5):927-943,2021

分布式深度学习隐私与安全攻击研究进展与挑战

周纯毅[1,2] 陈大卫[1] 王 尚[1] 付安民[1,2] 高艳松[1]
[1](南京理工大学计算机科学与工程学院 南京 210094)
[2](桂林电子科技大学广西可信软件重点实验室 广西桂林 541004)
(zhouchunyi@njust.edu.cn)

Research and Challenge of Distributed Deep Learning Privacy and Security Attack

Zhou Chunyi[1,2], Chen Dawei[1], Wang Shang[1], Fu Anmin[1,2], and Gao Yansong[1]
[1](*School of Computer Science and Engineering, Nanjing University of Science & Technology, Nanjing* 210094)
[2](*Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi* 541004)

Abstract Different from the centralized deep learning mode, distributed deep learning gets rid of the limitation that the data must be centralized during the model training process, which realizes the local operation of the data, and allows all participants to collaborate without exchanging data. It significantly reduces the risk of user privacy leakage, breaks the data island from the technical level, and improves the efficiency of deep learning. Distributed deep learning can be widely used in smart medical care, smart finance, smart retail and smart transportation. However, typical attacks such as generative adversarial network attacks, membership inference attacks and backdoor attacks, have revealed that distributed deep learning still has serious privacy vulnerabilities and security threats. This paper first compares and analyzes the characteristics of the three distributed deep learning modes and their core problems, including collaborative learning, federated learning and split learning. Secondly, from the perspective of privacy attacks, it comprehensively expounds various types of privacy attacks faced by distributed deep learning, and summarizes the existing privacy attack defense methods. At the same time, from the perspective of security attacks, the paper analyzes the attack process and inherent security threats of the three security attacks: data poisoning attacks, adversarial sample attacks, and backdoor attacks, and analyzes the existing security attack defense technology from the perspectives of defense principles, adversary capabilities, and defense effects. Finally, from the perspective of privacy and security attacks, the future research directions of distributed deep learning are discussed and prospected.

Key words deep learning; distributed deep learning; privacy attack; privacy protection; backdoor attack

摘 要 不同于集中式深度学习模式,分布式深度学习摆脱了模型训练过程中数据必须中心化的限制,实现了数据的本地操作,允许各方参与者在不交换数据的情况下进行协作,显著降低了用户隐私泄露风险,从技术层面可以打破数据孤岛,显著提升深度学习的效果,能够广泛应用于智慧医疗、智慧金融、

收稿日期:2020-11-25;**修回日期**:2021-02-09
基金项目:国家自然科学基金项目(62072239,62002167);广西可信软件重点实验室研究课题(KX202029);中央高校基本科研业务费专项资金(30920021129)
This work was supported by the National Natural Science Foundation of China (62072239, 62002167), the Guangxi Key Laboratory of Trusted Software (KX202029), and the Fundamental Research Funds for the Central Universities (30920021129).
通信作者:付安民(fuam@njust.edu.cn)

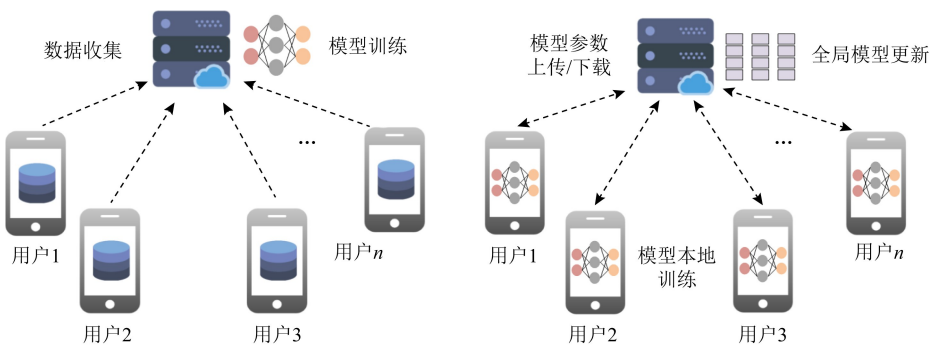
智慧零售和智慧交通等领域.但生成对抗式网络攻击、成员推理攻击和后门攻击等典型攻击揭露了分布式深度学习依然存在严重隐私漏洞和安全威胁.首先对比分析了联合学习、联邦学习和分割学习3种主流的分布式深度学习模式特征及其存在的核心问题.其次,从隐私攻击角度,全面阐述了分布式深度学习所面临的各类隐私攻击,并归纳和分析了现有隐私攻击防御手段.同时,从安全攻击角度,深入剖析了数据投毒攻击、对抗样本攻击和后门攻击3种安全攻击方法的攻击过程和内在安全威胁,并从敌手能力、防御原理和防御效果等方面对现有安全攻击防御技术进行了度量.最后,从隐私与安全攻击角度,对分布式深度学习未来的研究方向进行了讨论和展望.

关键词 深度学习;分布式深度学习;隐私攻击;隐私保护;后门攻击

中图法分类号 TP391

近年来全球掀起人工智能研发浪潮,美国、日本、英国、德国等世界科技强国纷纷将人工智能上升为国家战略,力图在新一轮国际科技竞争中掌握主导权.2017年我国发布了《新一代人工智能发展规划》,明确提出要抢抓人工智能发展的重大战略机遇,构筑我国人工智能发展的先发优势,加快建设创新型国家.深度学习作为实现人工智能的一种重要方法,通过海量训练数据构建具有很多隐层的深度学习模型,获得强大的数据特征学习能力.在深度学习过程中,普遍认为训练数据量越大,训练得到的模型的鲁棒性和准确性越高[1].因此,深度学习通常需要着重考虑数据的多源性,即通过汇聚各个机构或者用户数据完成整体计算任务,以提高训练模型的准确性.但在深度学习模型训练过程中,运营商可能会窃取用户的隐私信息.同时,公司之间的数据共享需要用户的授权,而许多用户出于隐私泄露的顾虑而拒绝数据共享.这些因素会导致"数据孤岛",难以创造出"1+1>2"的数据价值[2].因此,随着各国法律法规对于隐私信息使用的严格限制和公众隐私保护意识的加强,如何在保护数据隐私的前提下实现行业协作与协同治理,如何破解数据隐私保护与数据孤岛的两难困境,成为当下深度学习应用中亟待解决的技术难题.

不同于传统的集中式深度学习,分布式深度学习通过将深度学习与协作性模型相结合,使各个机构或者用户在不交换数据的情况下进行协作训练并获得更加精准的深度学习模型[3],以便在满足隐私保护和数据安全的前提下实现数据的有效利用[4].分布式深度学习模型将模型训练过程从云端转移至用户端,允许各方参与者在不暴露数据的情况下完成训练,降低了用户隐私泄露风险[5-6]和通信开销[7],从技术层面可以打破数据孤岛,明显提高深度学习的性能,能够实现多个领域的落地应用,比如智慧医疗、智慧金融、智慧零售和智慧交通等[8].分布式深度学习作为大数据使用的新范式,是破解数据隐私保护与数据孤岛难题的新思路,一经提出就成为国际学术界和产业界关注的焦点.图1展示了集中式深度学习和分布式深度学习训练模式的区别.



(a) 集中式深度学习 (b) 分布式深度学习

Fig. 1 Centralized deep learning and distributed deep learning

图1 集中式深度学习与分布式深度学习

海量的用户数据、丰富的应用场景促进了分布式深度学习技术的蓬勃发展,但生成对抗式网络攻击、成员推理攻击和属性推理攻击等隐私攻击揭露了分布式深度学习依然存在严重隐私漏洞[5].在2017年

的计算机和通信安全(ACM Conference on Computer and Communications Security)会议上,Hitaj 等人[9]设计了一种基于生成式对抗网络攻击的新型隐私攻击方式.攻击者伪装成正常用户加入模型训练后,可以基于生成式对抗网络的方法获得其他参与训练者的数据仿真集,极大地威胁到用户的数据安全.同时,Shokri 等人[10]提出了一种称为成员推理攻击的隐私攻击方法.攻击者通过训练多个影子分类器为攻击模型提供训练集,训练好的攻击模型可以输出某条记录是否在目标训练集中.最近,文献[11]又提出了一种属性推理攻击,可以在攻击者仅拥有训练集子集的情况下获取用户数据集的各类属性信息,如性别分布、年龄分布、收入分布等.可见,在分布式深度学习模式下,攻击者依然可以通过生成式对抗网络攻击等一系列典型隐私攻击方式获取用户的隐私数据信息.另一方面,在现有的分布式深度学习框架中,参与者在本地训练模型并向云服务器共享明文梯度.由于无法监管参与者在本地的训练行为,使得分布式学习容易遭受数据投毒、对抗样本和后门攻击的威胁.怀有恶意的参与者,在训练过程中可以利用数据投毒攻击,有目的性地更改本地数据的标签,上传"有毒"的梯度,"污染"协作训练的模型,使得模型预测结果准确率下降.此外攻击者可以上传后门本地模型,替换全局模型,以便模型有更高的概率输出攻击者的目标标签[12].

综上所述,分布式深度学习是破解数据隐私保护与数据孤岛难题的新思路[13],但其依然面临严重隐私漏洞和安全威胁.本文系统研究和分析了分布式深度学习面临的隐私与安全攻击问题,主要包含 4 个方面内容:

1) 对比分析了联合学习、联邦学习和分割学习 3 种主流的分布式深度学习模式,归纳总结了它们各自特征及其存在的核心问题;

2) 从隐私攻击角度,全面阐述了分布式深度学习所面临的各类隐私攻击,并归纳和分析了差分隐私、同态加密和安全多方计算等隐私攻击防御手段;

3) 从安全攻击角度,深入剖析了数据投毒攻击、对抗样本攻击和后门攻击 3 种安全攻击方法的攻击过程和内在安全威胁,并从数据集、模型输入和模型训练角度对现有的安全防御技术进行了归纳与总结;

4) 针对现有的隐私和安全攻击与防护研究中存在的主要问题,讨论和指出了分布式深度学习领域下一步可能的研究方向.

1 分布式深度学习概述

分布式深度学习无需用户上传本地数据就可以协作完成模型训练,消除了用户关于数据云端存储不可控的担忧,缓解了传统集中式深度学习收集用户数据所带来的隐私泄露问题.从训练模式上来看,分布式深度学习目前主要有联合学习、联邦学习和分割学习 3 种.

1.1 联合学习

联合学习首次由 Shokri 等人[14]于 2015 年提出,它打破了集中式深度学习的固有模式.如图 2 所示,在这种训练模式下,云服务器首先收集一批用户的数据集训练初始的全局模型,然后参与联合学习的第 1 个用户下载初始模型并基于自己的数据集使用随机梯度下降法(stochastic gradient descent, SGD)在本地训练模型.训练结束后,该用户按照一定比例随机选择部分模型参数上传到云服务器完成全局模型的更新.当第 1 个用户上传完毕后,下一个用户下载新的全局模型,并重复上述的训练和上传操作,这个过程将持续到模型收敛或达到预先设定的迭代次数.

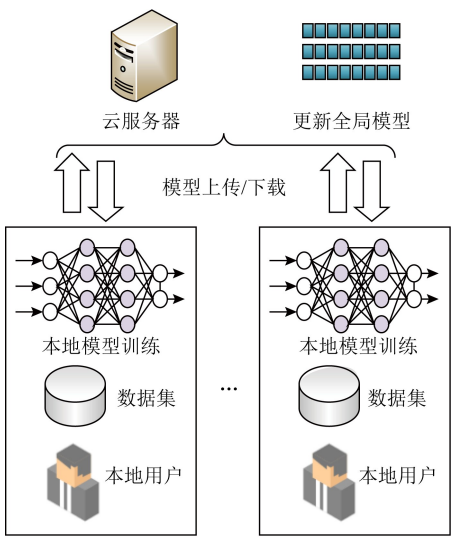


Fig. 2 Training mode of collaborative learning

图 2 联合学习训练模式

然而,由于联合学习的训练流程是每个用户异步进行的,因此当网络拥塞或用户离线时会导致全局参数无法顺利更新,其余用户会进入等待状态,训练效率可能会受到严重的影响.

1.2 联邦学习

联邦学习[15]在 2017 年由 Google 提出,它改进

了联合学习的异步训练模式,以并行训练的方式进行分布式深度学习任务,提高了训练效率.根据数据分布类型联邦学习可分为横向联邦、纵向联邦与迁移学习[8].与联合学习类似,联邦学习的用户下载云服务器提供的初始全局模型,然后根据自己的数据在本地训练并上传模型,服务器以加权平均的方式更新全局模型,满足了更复杂的用户数据分布并行场景.如图 3 所示,用户在本地训练模型后上传至云服务器,云服务器执行模型参数聚合算法得到全局模型,然后用户可以下载新的全局模型用于新一轮训练.在该训练模式下,所有用户不需要依次上传模型参数,而是同步向云服务器进行模型传输,即使出现某个用户离线或传输延时的情况,云服务器仍可聚合得到新的全局模型.联邦学习从根本上避免了联合学习的异步性限制,提高了系统的鲁棒性和训练效率.

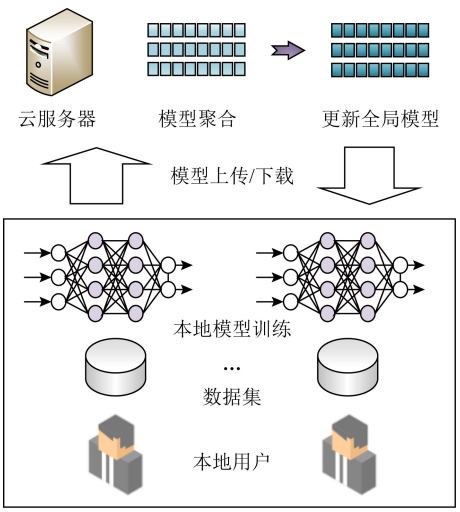


Fig. 3 Training mode of federated learning

图 3 联邦学习训练模式

虽然联邦学习在模型更新模式上对联合学习的异步性进行了改进,但仍然存在着效率问题.由于分布式深度学习的应用场景包含大量离散用户,在实际情况下他们拥有的计算能力和数据量是不均匀的,这会导致整体训练进度不一致,云服务器无法按时完成模型的聚合从而影响整体训练任务的进度.

1.3 分割学习

分割学习[16]是分布式深度学习中的一种最新训练模式.如图 4 所示,它考虑了用户计算资源的局限性,认为某些用户可能并不拥有足够的计算资源,因此将训练任务分成 2 部分,由用户和云服务器共同执行.在本地训练阶段,第 1 个用户首先训练神经网络的前一部分直至分割层,然后将分割层的输出发送到云服务器.云服务器进一步训练模型的后半部分,这也是训练中最繁重的计算任务.云服务器在获得输出后,将其反向传播回第 1 个用户.最后,用户完成前一部分模型的反向传播,并将用户端模型交给参与分割学习的下个用户.

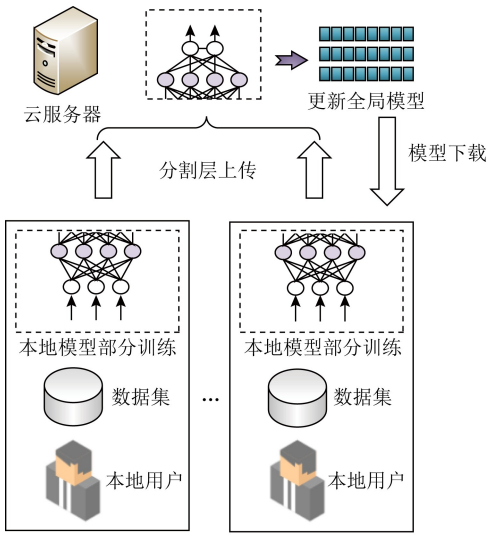


Fig. 4 Training mode of split learning

图 4 分割学习训练模式

由于分割学习将训练任务分割成 2 部分分别交由用户与服务器完成,所以它降低了用户的计算压力.但因为分割学习的训练过程是异步的,全局模型的更新要求在用户之间进行传递,这无疑增加了用户在通信上的负担,对于一些通信条件差的用户以及环境来说,分割学习的效率可能会有所影响.

1.4 小 结

联合学习、联邦学习和分割学习 3 种分布式深度学习模式在训练形式上都实现了各方参与者在不交换训练数据的情况下进行协作,降低了用户隐私泄露风险,从技术层面打破了数据孤岛,同时消除了用户关于数据云端存储不可控的担忧.但这 3 种训练模式也存在一些明显异同,表 1 从多方面对比分析了它们的特征.

由表 1 可以看出,虽然 3 种分布式深度学习模式都是以用户数据保留在本地为核心,但它们之间仍然存在一些显著区别.

1) 训练模式.联邦学习采用了并行化的训练,这种训练模式可以使数百万计的移动终端同时训练 1 个深度学习模型,更适用于大规模的用户量以及数据总量场景;而联合学习采用异步的模型参数部分上传训练模式,更适用于几个拥有庞大数据量的

大型企业或者机构之间相互协作完成 1 个合作模型的场景.分割学习由于其将训练任务分割成 2 部分交由用户与服务器分开执行的特性,用户的计算开销可以明显降低,但同时对于用户与服务器之间的通信开销增加,因此比较适用于小规模的用户群.

2) 模型参数更新方式.异步性更新的联合学习和分割学习依靠着多个实体间的传递,尤其是分割学习,1 次全局模型参数更新需要 3 次实体间的传递(用户—服务器,服务器—用户,用户—用户),因此分割学习对于通信网络传输效率要求更高;而同步性的联邦学习采用的是加权平均方式,虽然模型参数平均的方式可以保证准确率,减少过拟合,但是在恶意参数检测时很难辨别恶意参数来自于哪个实体,因此联邦学习对于模型参数筛选要求更加严格.

Table 1 Analysis of Distributed Deep Learning Feature

表 1 分布式深度学习特征分析

分布式框架	联合学习	联邦学习	分割学习
训练模式	异步	同步	异步
模型更新方式	选择性上传	联邦平均	用户间传递
训练位置	用户	用户	用户+服务器
用户计算开销	高	高	低
应用场景	小规模用户	大规模用户	小规模用户
缺点	用户离线导致的训练中止	用户不平均的计算能力	高通信开销

2 分布式深度学习隐私攻击

分布式深度学习模式摆脱了模型训练过程中数据必须中心化的制约,允许各方参与者在不交换数据的情况下进行协作,降低了用户隐私泄露风险[8].但生成对抗式网络攻击等新型隐私攻击方式表明分布式深度学习依然存在严重隐私漏洞[5].

2.1 隐私攻击方法

目前分布式深度学习中存在的隐私攻击主要包括生成对抗式网络攻击、成员推理攻击、属性推理攻击和模型反演攻击 4 种.

2.1.1 生成对抗式网络攻击

生成对抗式网络(generative adversarial networks,GAN)攻击[9]中,攻击者通常伪装成正常用户加入模型训练后,基于 GAN 获得其他参与训练者的数据仿真集,从而威胁用户的数据隐私[17],攻击发生在模型训练阶段,如图 5 所示.

首先,攻击者在本地训练一个生成器网络,用来生成目标类的训练样本,利用从服务器获取的最新

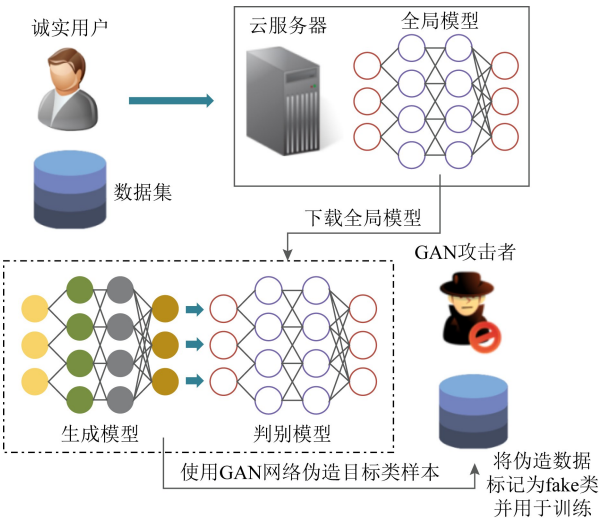


Fig. 5 GAN attack process

图 5 GAN 攻击过程

系统模型构建判别器.然后,攻击者将本地生成器生成的样本作为训练数据,并标记为 fake 类用于参与联合训练过程,上传并更新系统模型.持有目标类训练样本的一般训练者被迫暴露关于该类数据更详细的特征信息,用来提高模型区分目标类和 fake 类的能力,而这些信息将有助于提高攻击者本地生成器的伪造能力.通过在本地训练生成对抗式网络,可以有效获取目标类的隐私训练数据的信息.

为了将攻击发起者由用户端转移至云端,Wang 等人[18]进一步提出了一种基于云端的生成对抗式网络攻击,恶意云服务器采用一种多任务的生成对抗式网络攻击模型能够成功还原出参与训练用户的数据集.

2.1.2 成员推理攻击

成员推理攻击[10]是一种攻击者试图推断目标模型的训练集中是否包含特定信息的隐私攻击方法.分布式深度学习下,攻击者可以作为本地用户在模型训练阶段进行攻击,因为用户可以事先知晓目标模型的类型与结构,便于训练攻击模型.如图 6 所示,攻击者可以使用与目标模型相同分布的数据集,通过训练多个影子分类器模仿目标模型预测行为.然后根据训练集以及非训练集的 2 种输出分别进行标记,并以此为训练集训练二元攻击模型.训练好的攻击模型可以输出某条记录是否在目标训练集中.

Salem 等人[19]对成员推理攻击进行了进一步优化,减轻了攻击所需要的前置条件,在没有目标模型的知识结构和训练数据集分布的情况下使影子分类器的数量由 N 降低到 1.成员推断攻击可以确定

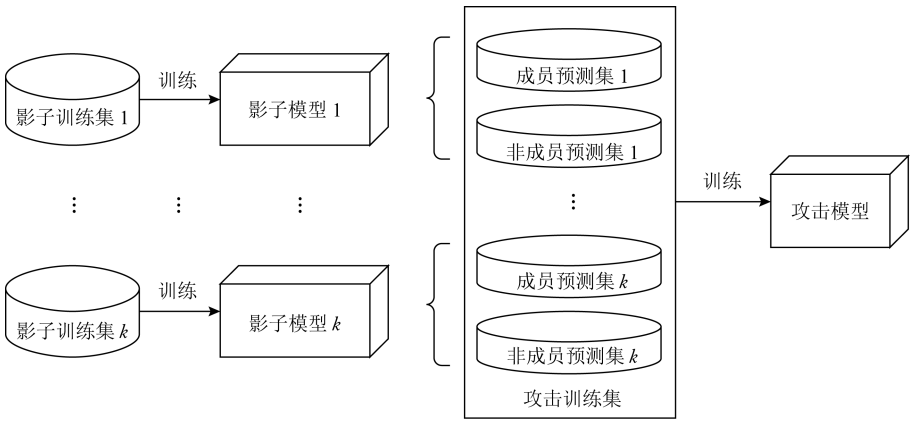


Fig. 6 Membership Inference Attack

图 6 成员推理攻击过程

数据记录是否在目标模型的训练数据中,并且在白盒和黑盒设置中都可以执行有效攻击.攻击者可以利用影子模型和少量确定的训练样本来实现成员关系推理.

2.1.3 属性推理攻击

属性推理攻击[11,20]指攻击者在仅拥有训练集子集的情况下可以推断用户数据集的各类属性信息,如性别分布、年龄分布、收入分布等,如图 7 所示:

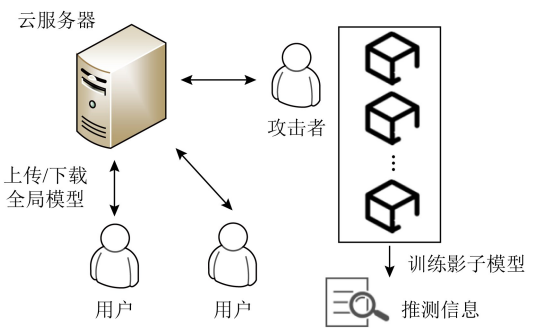


Fig. 7 Property inference attack process

图 7 属性推理攻击过程

属性推理攻击是基于全连接神经网络各层节点的排列不变性展开的一种隐私攻击.在分布式深度学习中,攻击者可以伪装成普通用户,训练攻击模型在模型训练阶段对良性用户训练数据中的属性进行推断.通过这种攻击,攻击者可以推断出训练者不想发布的训练数据的属性,例如判断包含敏感属性的某类数据是否存在训练集中.对于要推断的属性 P,攻击者通过调用模型生成 k 个数据集,其中一半将属性 P 的标签设置为 true,另一半设置为 false.然后基于 k 个数据集分别训练 k 个影子分类器.再以这 k 个模型参数作为元数据集,将每个样本对应地标记为 P_{true} 或 P_{false},进行再次训练构造元分类器.攻击者可以利用元分类器从目标模型中预测目标属性的存在性.

2.1.4 模型反演攻击

模型反演攻击[21]指攻击者利用深度学习系统提供的应用程序接口向模型发送大量的预测数据,然后对目标模型返回的类标签和置信度系数进行重新建模,通过得到的模型还原出目标模型的训练数据集.由于分布式深度学习并未限制模型的使用,任何参与者都可以无条件地访问系统模型,这无疑为模型反演攻击提供了有利的攻击条件.甚至攻击者事先不知道模型信息的情况下(训练数据、模型参数、模型类型等)仍然可以使用应用程序接口(application programming interface, API)在黑盒设置中发起此攻击.此类攻击中,攻击者大多是在预测阶段通过滥用对模型 API 的访问来检索特定训练数据.即使攻击者事先对训练集信息一无所知,该方法仍然有很高的攻击效率.这种攻击利用了“置信信息”的价值,它代表特征向量与类之间关联的可能性估计.攻击者通过输入特征向量来计算一个或多个类的分类或回归,然后选择置信度最高的类作为结果.Tramèr 等人[22]提出了一种模型反演攻击可以利用开放的深度学习模型 API 来“仿造”出原学习模型,从而造成模型提供商的损失.

2.1.5 小结

表 2 给出了分布式深度学习中的隐私攻击特征对比分析.从表 2 中可以看出,分布式深度学习隐私攻击大多是基于白盒展开的,因为目前分布式深度学习框架并没有对用户的模型使用权加以约束,任何参与模型训练的用户都可以轻易获取完整的模型,这给攻击者提供了绝佳的攻击条件.分布式深度

学习中的隐私攻击者既可能是恶意用户,也可能是云服务器,他们都可以在模型训练和测试阶段对用户的隐私进行攻击.从攻击手段方面来看,大多攻击方式都是通过建立攻击模型来推断出用户隐私信息,并且不同攻击者的目标也不相同,包括但不限于目标用户的数据、数据分布、模型参数等.这些攻击大多难以察觉,因为攻击者是在本地进行攻击模型的训练,而且目前缺乏对于此类攻击检测方面的研究.因此,如何在用户训练阶段实现隐私攻击的检测与防御是保证分布式深度学习用户隐私的关键.

Table 2 Summary and Comparison of Distributed Deep Learning Privacy Attacks

表 2 分布式深度学习隐私攻击总结对比

攻击名称	攻击模式	攻击过程	白/黑盒	攻击目标
GAN 攻击	使用生成器和判别器生成诚实用户信息	模型训练阶段	白盒	生成与受害者用户相似数据
成员推理攻击	训练多个影子分类器构造攻击模型	模型训练阶段	黑盒	判别特定数据记录是否存在于训练集
属性推理攻击	训练目标属性二元分类器	模型训练阶段	黑盒	判别特定属性是否存在于训练集
模型反演攻击	滥用模型 API 推测模型信息	模型测试阶段	黑盒	重构训练数据或窃取模型

2.2 隐私攻击防御技术

针对分布式深度学习面临的各类隐私攻击问题,目前已有一些研究分别从差分隐私、同态加密、安全多方计算等不同角度对分布式深度学习中的隐私保护方法进行了探讨.

2.2.1 基于差分隐私的防御方法

差分隐私是一种专门针对数据集统计特性设计的新型隐私保护技术,主要通过限制查询统计数据库时识别具体数据实例的机会而实现用户数据的隐私保护.在深度学习领域,差分隐私通常用于在模型梯度中添加噪声,从而实现隐私保护.针对 GAN 攻击问题,Xu 等人[23]提出基于差分隐私的 GANobfuscator 方案,通过在梯度中添加噪声实现差分隐私,并开发完整的梯度裁剪策略以提高数据训练的可扩展性和稳定性,使用户能够在不泄露自身隐私的前提下使用 GAN 生成大量合成数据.此外,针对联邦学习面临数据分布不均匀和计算能力差距大而导致训练效率低的问题,Zhou 等人[24]首次提出了一种雾计算环境下的基于本地差分隐私的联邦学习新方法,有效实现了对参数更新的保护,并提高了模型训练效率.Li 等人[25]基于多数据源广泛分布的特性,设计了一个基于差分隐私的分布式在线学习框架,使用最小批次稀疏学习方式进行训练保证了相同隐私预算下的更低噪声.

2.2.2 基于同态加密的防御方法

同态加密技术允许数据在密文条件下进行安全计算,实现密文状态下对明文的操作,从而保证数据的隐私性.在分布式深度学习领域,同态加密技术经常被用于参数加密,以密文状态下实现全局参数聚合更新,使服务器无法获取明文.Phong 等人[26]认为分布式深度学习中用户的梯度会泄露给半诚实的服务器而导致泄露训练集信息,从而提出了一种同态加密下的隐私保护联邦学习方案,使服务器无法接触用户的梯度明文,保证了模型安全.Li 等人[27]建立了一种非交互式联合学习框架,有效避免了数据拥有者与训练者之间数据的多轮交互,降低了通信开销,并通过掩码技术与 Paillier 同态加密技术保护了训练集以及训练模型的隐私.在模型鲁棒性研究中,Bonawitz 等人[28]设计了一种使用随机子集的用户更新矢量的加权平均值的安全聚合方案,在聚合过程中用户间的盲化因子相互抵消,可以在用户退出的情况下恢复其盲化因子,使方案可以适应用户离线的情况.Zhang 等人[29]使用 Paillier 加密实现梯度的安全聚合,并利用同态散列改进双线性聚合签名来验证聚合结果的正确性.

2.2.3 基于安全多方计算的防御方法

安全多方计算技术是指在无可信第三方的条件下多方参与者安全地计算一个约定函数问题,主要研究参与方在协作计算时如何对各方隐私数据进行保护,重点关注各参与方之间的隐私安全性问题,即在安全多方计算过程中必须保证各方私密输入独立,计算时不泄露任何本地数据.Chen 等人[17]应用安全多方计算的 Improved Du-Atallah 协议将参与者与模型参数分离开,通过交互模式实现参与者的本地模型训练,保证参与者的数据安全和服务器的模型安全.最近,Chen 等人[30]提出了一个基于可信执行环境的联邦学习机制,首次实现了用户本地训练结果的完整性验证,但依然没有解决用户数据可用性验证问题.Esposito 等人[31]采用博弈论思想将多用户和云中心模式的联合学习抽象为单个用户和

云服务器之间的交互模式,实现了用户和云之间的安全交互.

2.2.4 小结

表 3 对分布式深度学习隐私防御技术进行了对比分析.从表 3 中我们可以看出,目前的隐私保护研究主要围绕差分隐私、同态加密和安全多方计算 3 项关键技术展开,一些其他的技术,如盲化[32]等也在分布式深度学习隐私保护领域有所研究.目前的防御方法虽然实现了隐私保护,但无论是差分隐私、同态加密还是安全多方计算,都存在着一些应用上的弊端:

1) 差分隐私技术大多是在模型参数上进行加噪,但是与此同时模型精度会降低,一旦加噪的程度降低就达不到隐私保护的标准.

2) 同态加密技术计算开销巨大,全同态技术目前并不适用于具体应用场景,而且目前使用同态加密技术的方案对于多个实体间共谋的情况抵御能力差,容易造成密钥的泄露.

3) 安全多方计算技术可以实现无可信方场景下的安全计算,但是为了实现隐私保护,需要执行多次不经意传输协议,通信开销较大.该方法比较适用于小规模实体间的安全计算,对于具有大量用户的分布式深度学习场景的适应性差.

此外,由于分布式深度学习中隐私攻击很难被察觉,目前的研究针对成员推理攻击、属性推理攻击等隐私攻击并没有很好的检测和抵御手段.

Table 3 Summary and Comparison of Deep Learning Privacy Defense Techniques

表 3 分布式深度学习隐私防御总结对比

相关工作所载文献	技术方法	分布式训练模式	优势
文献[21]	差分隐私	联合学习	抵御 GAN 攻击
文献[29]	博弈论	联合学习	服务器与用户之间的安全交互
文献[24]	加性同态加密	联邦学习	模型参数安全聚合
文献[25]	盲化技术和 Paillier 同态加密	联合学习	降低通信开销和模型参数隐私保护
文献[26]	盲化技术	联邦学习	模型参数安全聚合与用户离线时参数恢复
文献[30]	同态散列函数和盲化技术	联邦学习	模型参数验证机制
文献[28]	可信执行环境	联邦学习	用户本地训练模型的完整性认证
文献[22]	盲化技术和 Paillier 同态加密	联邦学习	提高训练效率
文献[23]	差分隐私	联合学习	以低噪声实现模型参数保护
文献[14]	改进 Du-Atallah 协议	联邦学习	将用户与模型参数分离实现参数保护
文献[27]	Paillier 同态加密和同态散列函数	联邦学习	验证模型聚合结果正确性

3 分布式深度学习安全攻击

隐私攻击的对象一般是用户的数据集、训练模型的参数、模型的预测标签以及模型返回的结果.与此同时,一些攻击者可以通过一些恶意样本,对深度学习模型的结果产生负面影响,从而对整个分布式深度学习系统造成严重的安全隐患,例如数据投毒攻击、对抗样本攻击和后门攻击等.因此,分布式深度学习领域中的安全攻击与防御也是近年来研究者关注的热点.

3.1 安全攻击方法

目前的安全攻击方法大致可以分为 3 类:数据投毒攻击、对抗样本攻击和后门攻击.

3.1.1 数据投毒攻击

数据投毒攻击[33]通常也称为模型失效中毒攻击,攻击者通过修改、删除或者注入精心挑选的数据内容,来影响模型的训练结果.攻击者的主要目的是降低模型的准确性.通过有害数据样本,使分类器的决策边界发生变化,导致数据样本被错误分类.数据投毒攻击的本质在于改变训练数据上的局部分布.为了便于理解,我们在图 8 中将数据投毒攻击的整个过程进行了抽象化表示.如图 8 所示样本空间中存在 2 类样本数据,基于样本的特征 A 和 B 对数据进行建模、训练,实线表示原本分类器决策边界,虚线表示遭受投毒攻击后的决策边界.遭受攻击后,决策边界发生倾斜,原本属于 A 的矩形测试样本被错误分类,导致模型准确度降低.Xiao 等人[33]通过植入反转标签来攻击向量机模型,将训练数据中 (x,y) 标签替换为 (x,y'),使模型在训练过程中学习错误的特征.作者通过设计最大化支持向量机(support vector machines, SVM)损失函数的优化

框架,进行随机标签翻转,使分类误差最大化,从而降低模型准确性.而分布式中毒攻击通过多个攻击者相互勾结,将恶意训练数据注入到各自的局部模型中,会给分布式深度学习带来更大的灾难.Hayes等人[34]首先在多方深度学习上实现了单用户数据中毒,攻击者仅在本地数据集中反转5%的数据标签,攻击成功率就达到38%.Cao等人[35]研究联邦学习中的分布式中毒攻击,多个攻击者使用与参与者相同的损失函数和超参数训练自己的局部模型,利用标签反转污染数据集.实验表明攻击成功率、效率随中毒样本和攻击者数量呈线性增加.

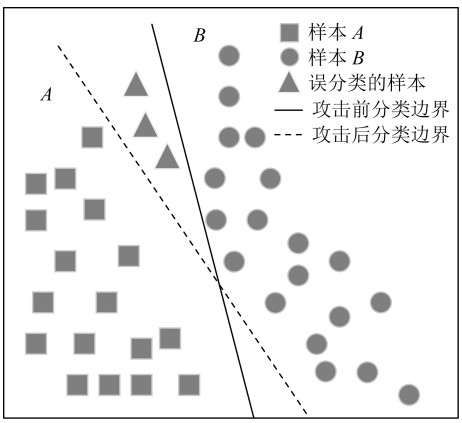


Fig. 8 Poisoning attack process

图8 投毒攻击过程

3.1.2 对抗样本攻击

对抗样本攻击[36]是一种能够让模型产生错误分类的攻击方式,仅作用于模型预测阶段.攻击者利用模型对输入样本的中间参数构造相应的对抗扰动,并将其注入输入样本即得到对抗样本.其中构造对抗扰动的实质是搜索使输入样本跨越决策边界的微小扰动,因此对抗样本可以导致模型错误分类.同时,由于对抗扰动比较微小,对抗样本与输入样本肉眼难以区分,增加了攻击的实用性.图9将对抗样本攻击的整个过程进行了抽象,样本空间中存在2类样本数据,基于样本的特征A和B对数据进行建模、训练,训练后的模型能够区分2种样本数据.采用对抗样本生成算法,矩形样本添加扰动后,该样本移动至圆形样本空间.与数据投毒攻击相比,该攻击下模型的决策边界未发生变化.

白盒攻击下攻击者已知目标模型的参数和结构,在集中深度学习中不易实现,但在分布式深度学习中,本地用户可以修改数据集和本地模型参数,对全局模型造成极大威胁.Szegedy等人[36]首先发现

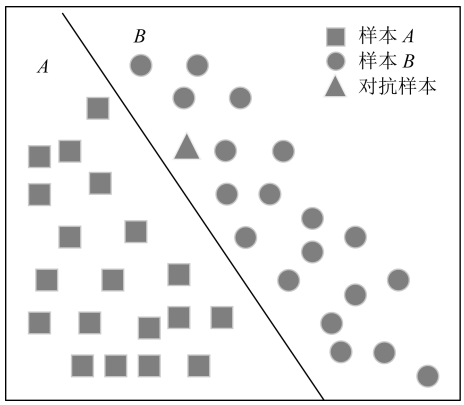


Fig. 9 Adversarial sample attack process

图9 对抗样本攻击过程

对图片添加轻微扰动可以欺骗深度学习模型,他们尝试寻找最小的损失函数添加项,使深度学习模型误分类.Goodfellow等人[37]提出FGSM(fast gradient sign method)扰动算法,利用模型对输入的导数与符号函数得到其具体的梯度方向,接着乘以1个步长即得到的恶意扰动.Kurakin等人[38]以FGSM攻击为基础提出了BIM(basic iterative method)快速生成对抗样本方法,他们进行了多次小步的迭代,并在每一步之后都修剪迭代结果的像素值.Carlini等人[39]提出了C&W(Carlini&Wagner)基于优化的攻击,通过限制L_∞,L_2,L_0范数使得扰动无法被察觉,将对抗样本的生成视为优化问题.Papernot等人[40]提出JSMA(jacobian-based saliency map attack)扰动生成算法,该方法限制了扰动的L_0范数,只需要在图片中修改几个像素点就可以使模型错误分类.

3.1.3 后门攻击

后门攻击[41-42]又被称为木马攻击,是一种针对深度学习最新的攻击方式.攻击者在模型的训练过程中隐藏后门,遭受后门的模型在不包含触发器的情况下通常作为其干净的对等模型,当后门遇到攻击者预先准备的触发器时,后门被激发.模型的输出变为攻击者预先指定的标签以达到攻击者的意图.后门攻击主要发生在数据收集与模型训练2个阶段,导致一些特殊的深度学习模式较容易遭受此类攻击,如联邦学习与迁移学习.因此,我们探索了这2种特殊的训练模式下的后门攻击.

1) 针对联邦学习

联邦学习训练模式下,恶意参与者很容易完成后门攻击.如图10所示,攻击者可控制部分参与者,利用后门数据进行模型训练,并提交本地后门模型.

然而,联邦聚合后,本地后门模型的贡献被抵消,所以全局模型在多轮聚合后可以缓解后门攻击.于是,攻击者在训练本地后门模型时应用约束与放缩技术,使得联邦聚合后全局模型被本地后门模型取代,实现了有效的后门攻击.

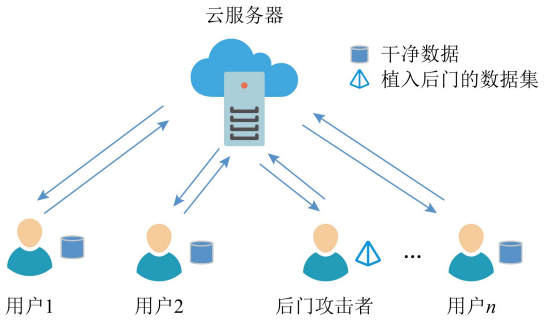


Fig. 10 Federated learning backdoor attack process

图 10 联邦学习后门攻击过程

Bagdasaryan 等人[43]采用模型替代植入关键后门,通过优化增加本地后门模型的权重以确保后门能够在多轮聚合后存活,使全局模型被后门模型取代.当全局模型收敛时,后门模型由于与全局模型权重近似,因此后门模型不会被大量良性用户模型所抵消.这种攻击也称为模型中毒.即使仅有 1 个恶意参与者在 1 个回合中被选中参与联邦模型聚合,当本地模型被植入后门情况下,联邦模型也会立即具有 100%的攻击成功率.但该方案随着联邦模型的多轮迭代聚合,攻击成功率会被削弱.

Xie 等人[44]在 Bagdasaryan 等人[43]研究的基础上提出了分布式后门攻击,多个攻击者只训练本地局部触发器,每个触发器的目标与原本集中式后门攻击一致,多个本地触发器共同形成全局触发器.文献[44]作者的实验针对 RFA[45]和 FoolsGold[46] 2 种模型聚合鲁棒性算法.结果表明,与以往的联邦学习后门攻击相比分布式后门攻击具有更好的攻击成功率.

2) 针对迁移学习

迁移学习训练模式下,攻击者通过向上游模型植入后门来感染用户训练的定制下游模型.图 11 形象地描述了迁移学习下后门攻击的过程.可见,对于上游模型,攻击者使用带有触发器的数据训练模型,调整前 $N-1$ 层的权重,使受攻击模型的前 N 层输出与任意干净模型的前 N 层输出非常相似.由于这些权重不会随着用户训练而更改,因此植入的潜在后门会成功传播到下游模型.因此,带触发器的输入将激活下游模型的后门,从而被分类为攻击者的目标类.

Liu 等人[47]在网络上发布嵌入后门的预训练模型.攻击者不需要访问原始的数据集,使用触发器和逆向生成的数据集对模型的一部分进行再训练,训练好的模型能够将带有触发器的图像分类为攻击者的目标.Yao 等人[48]提出了潜在的后门攻击来感染预训练模型,他们设计了一种嵌入到预训练模型中的不完整后门,并通过迁移学习感染多个下游模型.攻击者将受感染的预训练模型的最后 1 层替换为原始预训练模型的最后 1 层,以此嵌入高隐蔽性后门并保证模型正常的分类精度,提高了后门攻击的威胁.

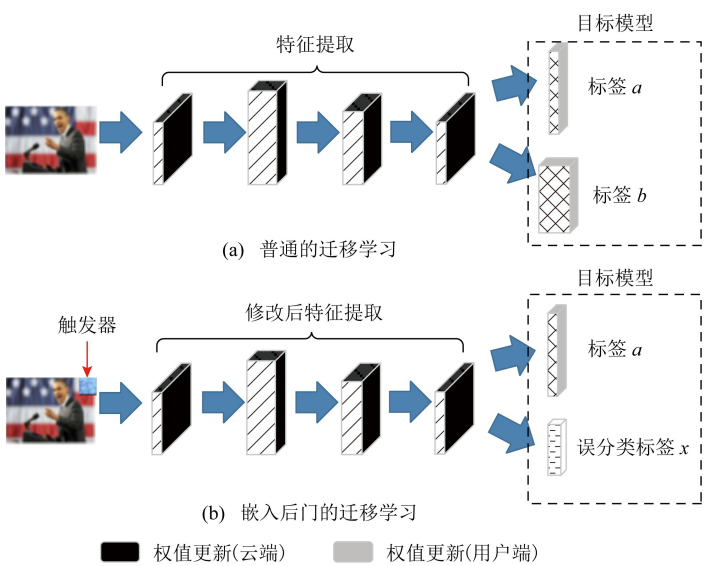


Fig. 11 Transfer learning backdoor attack process

图 11 迁移学习后门攻击过程

3.1.4 小结

为了清楚地显示3种攻击方式之间的差异,我们从敌手知识与攻击效果2个角度对3种安全攻击进行了总结,如表4所示.其中敌手知识包括数据集访问度与模型访问度,表示攻击方案成功的假设条件.完整的分布式深度学习过程大致分为5个步骤:数据样本收集、模型选择、模型训练、模型部署和更新模型.数据投毒攻击只能在数据样本收集阶段进行,对抗样本攻击只能发生在模型部署后的推理阶段,而后门攻击可在每个步骤中攻击.此外,从表4中可以看出3种安全攻击威胁程度不同,我们从攻击结果的角度分别对其进行分析:

1) 投毒攻击能够导致模型错误分类、准确性降低,但这种攻击方式通常不具有针对性,相当于幼年期的"病毒",简单地破坏模型的可用性;

2) 对抗样本攻击通常是对每个输入进行精心设计对抗扰动,绑架模型来输出攻击者设置的分类目标,相当于成熟期的"病毒",懂得适当隐藏自己并达到自己攻击目的;

3) 后门攻击可以使用相同的触发器(模式补丁)进行错误分类任何输入,并且植入后门的模型对无触发器的输入样本表现与干净对等模型并无区别,相当于高隐蔽性的"病毒",将自己伪装成正常模样并进行极大程度地破坏.

Table 4 Summary and Comparison of Distributed Deep Learning Security Attacks

表4 分布深度学习安全攻击总结对比

攻击方式	攻击阶段	相关工作所载文献	黑/白盒	数据访问权限	模型访问权限	攻击能力
数据投毒攻击	数据收集阶段	文献[31]	黑盒	全部	部分	▲▲
		文献[32]	白盒	全部	全部	▲▲▲
		文献[33]	白盒	全部	全部	▲▲▲△
对抗样本攻击	模型部署阶段	文献[34]	白盒	全部	全部	▲▲
		文献[35]				▲▲▲
		文献[36]				▲▲▲△
		文献[37]				▲▲▲▲
		文献[38]				▲▲▲△
后门攻击	全阶段	文献[41]	白盒	全部	全部	▲▲▲
		文献[42]	白盒	全部	全部	▲▲▲▲△
		文献[45]	黑盒	部分	部分	▲▲▲
		文献[46]	黑盒	部分	部分	▲▲▲▲

注:▲个数的多少代表攻击能力的强弱;△表示半个▲的攻击能力强度.

3.2 安全攻击防御技术

针对3类安全攻击,目前的防御手段主要从数据集、输入样本与模型3个方面进行考虑.

3.2.1 基于数据集的防御方法

数据集是投毒攻击与后门攻击的主要攻击对象.因此,通过一些手段来识别数据集中的恶意样本是抵御2种攻击的一条主要思路.例如防御者可以计算数据对模型的贡献从而剔除具有消极贡献的数据.文献[49]将训练数据按照来源信息划分为若干段,测试每段对模型的贡献.该方案可用于部分可信和不可信数据集的检测.此外,考虑到良性样本与恶意样本的深层特征差别很大,防御者对其使用鲁棒统计技术来检测恶意样本.Tran等人[50]利用深层特征对待检测数据集进行划分.对于每个输出类,他们搜集该类所有输入数据的模型中间表示的协方差矩阵,通过奇异值分解计算每条输入数据的异常值分数.同时,一些防御者利用恶意样本的异常行为过滤数据集.Chan等人[51]发现输入层中触发图像的触发位置有较大的梯度绝对值,他们利用聚类算法识别触发样本并修正其标签.

3.2.2 基于输入样本的防御方法

对抗样本攻击与后门攻击均需要构造恶意输入样本以实现攻击者目标.因此,针对输入数据进行防御是抵御2种攻击的另一条主要思路.目前,此类防御手段主要包括检测恶意样本与修补输入样本.

1) 检测恶意样本

防御者通过输入数据的异常特性或异常行为来检测恶意样本.McCoyd等人[52]用背景类填充关键类之间的区域,以检测恶意样本.经验证,该方案可缓解FGSM[37]攻击,但对C&W[39]攻击无效.此外,文献[53]提出一种特征压缩技术,可抵御多种对抗样本攻击.它通过空间平滑与减少每个像素颜色位

的深度来压缩特征,然后观察输入数据与压缩样本预测结果的一致性.文献[54]中利用触发器与攻击目标类的强关联性提出一种检测方案.他们发现触发输入添加强烈扰动后的预测结果具有弱随机性,该方案适用于复杂触发器的检测,尤其对触发器大小不敏感.

2) 修补输入样本

该策略通过修补输入样本的对抗性或触发器来缓解攻击.Meng 等人[55]训练降噪器(编码器与解码器)MagNet 来修补输入,突出图像中的主要成分,进而校正分类结果.实验证明:降噪器能够有效缓解 FGSM 与 C&W 攻击.Doan 等人[56]提出一种 2 阶段图像修补方法,首先删除最影响预测的区域,然后利用 GAN 对该区域进行修补.该方法涉及图像的移除和恢复,所以它对触发器规模很敏感.

3.2.3 基于模型的防御方法

安全攻击本质上是对模型可用性的破坏,因此针对模型的防御策略亦是安全防御的一种主要思路.根据防御策略,可以将其分为提升模型鲁棒性与模型诊断 2 类方案.

1) 提升模型鲁棒性

该策略主要根据安全攻击的思路调整模型参数或改进训练过程,从而使模型难以实现攻击者的目标.Feng 等人[57]提出一种鲁棒逻辑回归模型,他们删除幅度较大的样本并最大化其余样本与估计的逻辑回归模型之间的相关性.而 Biggio 等人[58]利用多分类器系统构造鲁棒分类器.他们利用引导聚集算法与随机子空间方法使特征权值分布更均匀,以此降低恶意数据的影响.这些手段主要用于投毒攻击,不适用于对抗样本攻击.对抗防御最直接的手段是利用对抗样本进行对抗训练,使得模型可以正确预测类似的对抗样本.此外,文献[59]中利用输入梯度正则化训练可微模型并惩罚输入中的微小变化,降低了对抗扰动的影响.Papernot 等人[60]用防御蒸馏技术降低深度神经网络结构的计算复杂度.它可以防御 FGSM 与基于雅可比矩阵(Jacobian Matrix)的迭代攻击.针对后门攻击,防御者主要从消除或抑制模型中的后门隐患来考虑.文献[61]通过修剪与反向触发器相关性较高的神经元来移除潜在的后门.Du 等人[62]采用加噪的 SGD 训练满足差分隐私机制的模型.由于训练过程中的随机性,触发样本的贡献将因随机噪声而减少,从而导致后门植入失败.实际应用中,加噪的 SGD 训练过程会降低模型的精度,所以此方案需要权衡模型的精度与鲁棒性.

2) 模型诊断

模型诊断主要针对后门攻击,它可以判断模型是否嵌入了后门,进而重构受损模型的后门触发器.

首先,防御者可以通过分析模型中神经元的行为进行检测.Liu 等人[63]利用人工脑刺激扫描模型中的神经元进而识别后门模型.它需要对每个神经元完成扫描,其中受损神经元对特定类的输出激活异常高.虽然方案具有很高的检测效率,但是它的有效性靠一个强假设保证,即后门攻击仅由 1 个神经元执行,这在实际中不完全适用.受编程中使用的控制流的启发,Ma 等人[64]从模型结构出发,利用神经网络进行源不变性与激活值分布不变性的检测.源不变性针对 2 个连续的隐藏层,而激活值分布不变性针对给定层神经元的激活分布.触发输入会违背这 2 种不变性,所以作者通过检测输入样本的 2 种不变性来识别触发样本并判断受感染的标签.

其次,一些防御者利用神经网络模型的参数、中间状态或输出训练元分类器以判断模型是否被感染.Xu 等人[65]将良性影子模型与后门影子模型作为训练样本训练元分类器.他们对每个影子模型进行若干次输入查询.对应的若干输出结果(如置信度评分)拼接后即为该影子模型的特征表示.训练过程中,防御者需要充分了解模型结构并且拥有强大的算力,为元分类器构造高质量训练集.

识别后门模型后,防御者需要还原其后门触发器,进而抵御后门攻击.Wang 等人[61]首先提出一种通用的重构触发器的技术,称为 Neural cleanse.对于每个输出类,他们利用逆向工程重建 1 个触发器,然后利用触发器识别恶意样本并修剪与触发器相关的神经元.该方法需要一部分保留的良性数据,而 Chen 等人[66]提出一种利用最少信息完成检测的方法.通过模型逆向技术与条件生成对抗式网络(conditional GAN, cGAN)为每个输出类生成 1 个触发器.

3.2.4 小结

安全攻击防御技术可以部署于数据集、输入样本和训练模型.为了清晰地展示 3 种防御策略的差异,表 5 从敌手知识与防御效果对目前已存在的分布式深度学习中的安全防御方法进行对比.其中敌手知识包括数据集访问度和模型访问度,表示防御方案成功的假设条件.而防御效果表示抵御安全攻击的能力.通过分析,我们可以看出2种防御策略各有优劣.

Table 5 Summary and Comparison of Distributed Deep Learning Security Defense Technology

表 5 分布深度学习安全防御技术总结对比

防御实体	抵御攻击	防御思路	相关工作所载文献	数据访问权限	模型访问权限	防御能力
训练数据集	数据投毒攻击	检测源	文献[47]	全部	部分	▲▲▲▲△
	后门攻击	统计	文献[48]	全部	部分	▲▲▲▲△
			文献[49]	全部	部分	▲▲▲▲
输入数据	对抗样本攻击	检测	文献[50]	全部	部分	▲▲▲
			文献[51]	部分	部分	▲▲▲▲
		修复	文献[53]	部分	部分	▲▲▲▲△
	后门攻击	检测	文献[52]	部分	部分	▲▲▲▲△
		修复	文献[54]	无	部分	▲▲▲▲
模型	数据投毒攻击	提高鲁棒性	文献[55]	部分	全部	▲▲▲△
			文献[56]	部分	全部	▲▲▲△
			文献[63]	全部	无	▲▲▲△
	对抗样本攻击		文献[57]	部分	全部	▲▲▲▲△
			文献[58]	无	全部	▲▲▲▲
	后门攻击		文献[63]	部分	全部	▲▲▲△
			文献[59]	部分	部分	▲▲▲△
	后门攻击	检测神经元	文献[60]	部分	部分	▲▲▲▲
			文献[61]	全部	全部	▲▲▲▲▲
		元分类器	文献[62]	部分	全部	▲▲▲▲△
		触发器	文献[63]	部分	全部	▲▲▲▲
			文献[64]	少量	全部	▲▲▲△

注:▲个数的多少代表防御能力的强弱;△表示半个▲的防御能力强度.

1) 基于数据集的检测方案需要访问整个数据集且检测效果较好.然而出于隐私考虑,防御者很难接触到数据集,所以此类方案的应用场景比较有限.

2) 基于输入样本的检测与修补只需要少量数据,敌手知识弱于基于数据集的检测方案.另外,它只需要访问部分良性数据即可完成高质量检测.所以,此类方案应用场景更广泛.

3) 针对模型的防御策略包括提升模型鲁棒性以及模型诊断.前者旨在利用一些已知的知识或假设提高模型抵抗安全攻击的能力;后者通过分析模型的参数与内部行为检查其完整性,甚至还原攻击行为.尽管这 2 种针对模型的防御策略仅需要访问部分良性数据且实验结果较好,但是它们大多存在局限性,比如,较大的计算开销、较强的假设、有限的适用领域、受限制的触发器性质等.

4 未来挑战及研究方向

分布式深度学习由于其白盒的特性,用户可以直接获取完整的全局模型参数,但与此同时恶意攻击者也可以利用这个特性伪装成诚实用户来展开攻击窃取隐私或破坏模型.这给分布式深度学习的隐私保护和安全防御带来极大挑战.目前,分布式深度学习的研究仍处于起步阶段,并没有一套完整成熟的恶意攻击检测和防御机制,极易受到各种隐私攻击和安全攻击的影响.通过对分布式深度学习隐私与安全攻击研究现状的深入分析,我们认为未来分布式深度学习中隐私与安全攻击研究可以重点从 4 个方面展开:

1) 构建新型分布式深度学习隐私保护框架.随着生成对抗式攻击、成员推理攻击等隐私攻击的不断提出与发展,目前提出的分布式深度学习框架已经不能满足用户训练数据和模型参数的隐私保护要求.而且分布式深度学习隐私攻击形式隐蔽,攻击者以诚实用户的身份混入训练队列,在本地即可发起攻击,使得服务器更加难以检测攻击行为.Awan 等人[67]将联邦学习部署于区块链,利用其匿名性与可追溯性保证参与者的隐私与全局模型的准确性.

虽然该方案并未具体探究隐私性与模型精度,但仍给了我们一条构建新型分布式深度学习隐私保护框架的思路.因此,如何给出基于多维度的隐私量化、刻画及演化方法,揭示数据隐私、模型准确性以及训练效率等关键要素之间的内在联系,构建安全、高效的新型分布式深度学习隐私保护框架是亟待解决的关键问题.

2) 推进轻量级隐私保护方法研究.现有分布式深度学习的隐私保护方法大多是利用安全多方计算、同态加密和差分隐私等经典密码学方法来保护用户数据和模型参数,但此类方法会带来沉重的计算开销,降低了模型训练效率和准确率.Cao 等人[68]设计了一种低开销的联邦学习框架.他们按照本地数据对全局模型的贡献选定一个参与者上传本地模型,其余参与者上传本轮模型的更新方向.该方案计算开销与通信开销较低,模型精度降低不显著.虽然该方案未进行严格的隐私分析,但给我们一条设计轻量级隐私保护方案的思路.一方面使用同态加密、安全多方计算等开销巨大的安全方法仅针对模型中的关键部分进行加密处理,减少计算开销,另一方面设计出更加轻量级的隐私保护算法,优化现有隐私保护框架,在保证隐私的同时降低计算开销.因此,新型轻量级的高质量的隐私保护方法是未来分布式深度学习隐私保护另一个新的研究方向.

3) 提升深度学习模型鲁棒性.随着数据投毒攻击、对抗样本攻击、后门攻击的不断发展,现有模型已经不足以抵御这些不断增强的攻击方法.并且因为分布式深度学习系统旨在利用参与者非独立同分布的本地训练数据,在数据收集过程中会收集到攻击者发布的"有毒"数据,因此模型必须对行为不当的参与者都具有较强的鲁棒性.一方面可以从数据收集环节出发,对数据进行清洗,提高数据的质量,提升模型的安全性.另一方面从模型更新环节出发,剔除异常更新.Pillutla 等人[45]提出了一种鲁棒的聚合方法,将加权算术平均值替换为近似的几何中位数,以最大程度减少异常值更新的影响.尽管该框架还没有证明其抵御攻击的实用性,但仍是一次成功的尝试.研究能够抵抗更强攻击手段的高鲁棒性深度学习模型是今后比较重要的研究方向.

4) 探索后门攻击防御策略.后门攻击作为高隐蔽性"病毒"引起广泛重视,因此相应的防御手段相继被提出.出于隐私考虑,目前主流的防御手段从检测训练数据与输入数据转移至模型检测.受感染的模型内部神经元或参数受后门影响,对其检测涉及到后门攻击的原理和后门触发器的激活机制.例如,利用反向工程重构后门触发器并分析单个神经元的行为[61]等.尽管这些方案存在问题,但它们为后门防御提供一些思路,比如可以利用后门模型尽可能还原高质量的后门触发器,或者检测多个神经元联合产生的异常行为.这些思路涉及后门攻击的原理,将是消除后门攻击威胁的关键,因此值得不断探索.

5 结束语

分布式深度学习摆脱了集中式深度学习数据中心化的限制,实现了模型本地训练,使多方参与者在不交换数据的情况下协作完成深度学习任务,降低了数据隐私泄露风险.但目前越来越多的隐私攻击和安全攻击时刻威胁着分布式深度学习的隐私与安全.因此,本文在国内外分布式深度学习研究调研和分析的技术上,对分布式深度学习模式进行梳理并归纳总结了不同训练模式的特征与问题.我们对分布式深度学习面临的隐私威胁与安全问题按照攻击类型进行详细的分类,并从敌手能力、防御原理和防御效果等方面分析归纳了现有的防御措施.最后阐述了目前研究中存在的问题与挑战,并展望了其未来的发展方向.

参 考 文 献

[1] Zhang Lei, Cui Yong, Liu Jing, et al. Application of machine learning in cyberspace security research [J]. Chinese Journal of Computers, 2018, 41(9): 1943-1975 (in Chinese)
(张蕾, 崔勇, 刘静, 等. 机器学习在网络空间安全研究中的应用[J]. 计算机学报, 2018, 41(9): 1943-1975)

[2] Yang Qiang. AI and data privacy protection: The way to federated learning [J]. Journal of Information Security Research, 2019, 5(11): 961-965 (in Chinese)
(杨强. AI 与数据隐私保护: 联邦学习的破解之道[J]. 信息安全研究, 2019, 5(11): 961-965)

[3] Zhang Yuqing, Dong Ying, Liu Caiyun, et al. The status, trends and prospects of deep learning in cyberspace security [J]. Journal of Computer Research and Development, 2018, 55(6): 1117-1142 (in Chinese)
(张玉清, 董颖, 柳彩云, 等. 深度学习应用于网络空间安全的现状、趋势与展望[J]. 计算机研究与发展, 2018, 55(6): 1117-1142)

[4] Qi Jia, Lin Keguo, Zhan Pengjin, et al. Preserving model privacy for machine learning in distributed systems [J]. IEEE Transactions on Parallel and Distributed Systems, 2018, 29(8): 1808-1822

[5] Chen Yufei, Shen Chao, Wang Qian, et al. Security and privacy risks in artificial intelligence systems [J]. Journal of Computer Research and Development, 2019, 56(10): 2135-2150 (in Chinese)
(陈宇飞, 沈超, 王骞, 等. 人工智能系统安全与隐私风险[J]. 计算机研究与发展, 2019, 56(10): 2135-2150)

[6] Fu Anmin, Zhang Xianglong, Xiong Naixue, et al. VFL: A verifiable federated learning with privacy-preserving for big data in industrial IoT [J]. IEEE Transactions on Industrial Informatics, arXiv preprint, arXiv:2007.13585, 2020

[7] Zhou Lei, Fu Anmin, Yang Guomin, et al. Efficient certificateless multi-copy integrity auditing scheme supporting data dSVMynamics [J/OL] IEEE Transactions on Dependable and Secure Computing, 2020 [2020-08-04]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9158490

[8] Yang Qiang, Liu Yang, Chen Tianjian, et al. Federated machine learning: Concept and applications [J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19

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

[10] Shokri R, Stronati M, Song C, et al. Membership inference attacks against machine learning models [C] //Proc of the IEEE Symp Security and Privacy. Piscataway, NJ: IEEE, 2017: 3-18

[11] Ganju K, Wang Qi, Yang Wei, et al. Property inference attacks on fully connected neural networks using permutation invariant representations [C]//Proc of the 2018 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2018: 619-633

[12] Chen Yuefeng, Mao Xiaofeng, Li Yuhong, et al. AI security—Research and application on adversarial example [J]. Journal of Information Security Research, 2019, 5(11): 1000-1007 (in Chinese)
(陈岳峰, 毛潇锋, 李裕宏, 等. AI安全—对抗样本技术综述与应用[J]. 信息安全研究, 2019, 5(11): 1000-1007)

[13] Chen Si, Fu Anmin, Shen Jian, et al. RNN-DP: A new differential privacy scheme base on recurrent neural network for dynamic trajectory privacy protection [J]. Journal of Network and Computer Applications, 2020, 168: No.102736

[14] Shokri R, Shmatikov V. Privacy-preserving deep learning [C] //Proc of the 2015 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1310-1321

[15] Mcmahan H, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data [C] //Proc of the 20th Int Conf on Artificial Intelligence and Statistics. Brooklyn, MA: Microtome Publishing, 2016: 1273-1282

[16] Vepakomma P, Gupta O, Swedish T, et al. Split learning for health: Distributed deep learning without sharing raw patient data [J] //arXiv preprint, arXiv:1812.00564, 2018

[17] Chen Zhenzhu, Fu Anmin, Zhang Yinghui, et al. Secure collaborative deep learning against GAN attacks in theInternet of things [J/OL] IEEE Internet of Things Journal, 2020 [2021-01-14]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9235504

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

[19] Salem A, Zhang Yang, Humbert M, et al. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models [C/OL] //Proc of the 2019 Network and Distributed System Security Symp. Reston: The Internet Society, 2019 [2021-01-15]. https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_03A-1_Salem_paper.pdf

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

[21] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures [C] //Proc of the 2018 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1322-1333

[22] Tramèr F, Fan Zhang, Juels A, et al. Stealing machine learning models via prediction APIs [C] //Proc of the 29th USENIX Security Symp. Berkeley, CA: USENIX Assocaiation, 2016: 601-618

[23] Xu Chugui, Ren Ju, Zhang Deyu, et al. GANobfuscator: Mitigating information leakage under GAN via differential privacy [J]. IEEE Transactions on Information Forensics and Security, 2019, 14(9): 2358-2371

[24] Zhou Chunyi, Fu Anmin, Yu Shui, et al. Privacy-preserving federated learning in fog computing [J]. IEEE Internet of Things Journal, 2020, 7(11): 10782-10793

[25] Li Chencheng, Zhou Pan, Xiong Li, et al. Differentially private distributed online learning [J]. IEEE Transactions on Knowledgeand Data Engineering, 2018, 30(8): 1440-1453

[26] Phong L, 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

[27] Li Tong, Li Jin, Chen Xiaofeng, et al. NPMML: A framework for non-interactive privacy-preserving multi-party machine learning[J/OL]. IEEE Transactions on Dependable and Secure Computing, 2020 [2021-01-15]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8981947

[28] Bonawitz K, Ivanov V, Kreuter B, et al. Practical secure aggregation for privacy-preserving machine learning [C] // Proc of the 2017 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 1175-1191

[29] Zhang Xianglong, Fu Anmin, Wang Huaqun, et al. A privacy-preserving and verifiable federated learning scheme [C/OL] //Proc of the 2020 IEEE Int Conf on Communications. Piscataway, NJ: IEEE, 2020 [2021-01-15]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9148628

[30] Chen Yu, Luo Fang, Li Tong, et al. A training-integrity privacy-preserving federated learning scheme with trusted execution environment [J]. Information Sciences, 2020, 522: 69-79

[31] Esposito C, Xin Su, Shadi A, et al. Securing collaborative deep learning in industrial applications within adversarial scenarios [J]. IEEE Transactions on Industrial Informatics, 2018, 14(11): 4972-4981

[32] Xu Guowen, Li Hongwei, Liu Sen, et al. VerifyNet: Secure and verifiable federated learning [J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 911-926

[33] Han Xiao, Huang Xiao, Eckert C. Adversarial label flips attack on support vector machines [C] //Proc of the 20th EUROPEAN Conf on Artificial Intelligence. Amsterdam: IOS Press, 2012: 870-875

[34] Hayes J, Ohrimenko O. Contamination attacks and mitigation in multi-party machine learning [C] //Proc of the Annual Conf on Neural Information Processing Systems. Amsterdam: Elsevier, 2018: 6604-6615

[35] Cao Di, Chang Shan, Lin Zhijia, et al. Understanding distributed poisoning attack in federated learning [C] //Proc of the 25th IEEE Int Conf on Parallel and Distributed Systems (ICPADS). Piscataway, NJ: IEEE, 2019: 233-239

[36] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [C/OL] //Proc of the 2nd Int Conf on Learning Representations. 2014 [2021-01-16]. https://iclr.cc/archive/2014/conference-proceedings

[37] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [J]. arXiv preprint, arXiv: 1412.6572, 2014

[38] Kurakin A, Goodfellow I J, Bengio S, et al. Adversarial examples in the physical worlds [J]. arXiv preprint, arXiv: 1804.00097, 2018

[39] Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] //Proc of the IEEE Symp Security and Privacy. Piscataway, NJ: IEEE, 2017: 39-57

[40] Papernot N, Mcdaniel P, Jha S, et al. The limitations of deep learning in adversarial settings [C] //Proc of the IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 372-387

[41] Rakin A, He Zhezhi, Fan Deliang. TBT: Targeted neural network attack with bit trojan [C] //Proc of the 2020 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 13195-13204

[42] Gu Tianyu, Dolan-Gavitt B, Garg S. BadNets: Identifying vulnerabilities in the machine learning model supply chain [J]. arXiv preprint, arXiv:1708.06733, 2019

[43] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning [C] //Proc of Int Conf on Artificial Intelligence and Statistics. Cambridge, MA: MIT Press, 2020: 2938-2948

[44] Xie Chulin, Huang Keli, Chen Pinyu, et al. DBA: Distributed backdoor attacks against federated learning [C/OL] //Proc of the 7th Int Conf on Learning Representations. 2019 [2021-01-16]. https://openreview.net/group?id=ICLR.cc/2019/Conference

[45] Pillutla K, Kakade S M, Harchaoui Z. Robust aggregation for federated learning [J]. arXiv preprint, arXiv: 1912.13445, 2019

[46] Fung C, Yoon C J M, Beschastnikh I. Mitigating sybils in federated learning poisoning [J]. arXiv preprint, arXiv:1808.04866, 2018

[47] Liu Yingqi, Ma Shiqing, Aafer Y, et al. Trojaning attack on neural networks [C/OL] //Proc of the 2018 Network and Distributed System Security Symp. Reston, VA: The Internet Society, 2018 [2021-01-16]. https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf

[48] Yao Yuanshun, Li Huiying, Zheng Haitao, et al. Latent backdoor attacks on deep neural networks [C] //Proc of the 2019 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 2041-2055

[49] Baracaldo N, Chen B, Ludwig H, et al. Mitigating poisoning attacks on machine learning models: A data provenance based approach [C] //Proc of the 10th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2017: 103-110

[50] Tran B, Li J, Madry A. Spectral signatures in backdoor attacks [C] //Proc of the Annual Conf on Neural Information Processing Systems. Amsterdam: Elsevier, 2018: 8000-8010

[51] Chan A, Ong Y S. Poison as a cure: Detecting & neutralizing variable-sized backdoor attacks in deep neural networks [J]. arXiv preprint, arXiv:1911.08040, 2019

[52] McCoyd M, Wagner D. Background class defense against adversarial examples [C] //Proc of the IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018: 96-102

[53] Xu Weilin, Evans D, Qi Yanjun. Feature squeezing: Detecting adversarial examples in deep neural networks [J]. arXiv preprint, arXiv:1704.01155, 2017

[54] Gao Yansong, Xu Change, Wang Derui, et al. Strip: A defence against trojan attacks on deep neural networks [C] // Proc of the 35th Annual Computer Security Applications Conf. New York: ACM, 2019: 113-125

[55] Meng Dongyu, Chen Hao. Magnet: A two-pronged defense against adversarial examples [C] //Proc of the 2017 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 135-147

[56] Doan B G, Abbasnejad E, Ranasinghe D C. Februus: Input purification defense against trojan attacks on deep neural network systems [J]. arXiv preprint, arXiv: 1908.03369, 2019

[57] Feng Jiashi, Xu Huan, Mannor S, et al. Robust logistic regression and classification [C] //Proc of the Annual Conf on Neural Information Processing Systems. Amsterdam: Elsevier, 2014: 253-261

[58] Biggio B, Fumera G, Roli F. Multiple classifier systems for robust classifier design in adversarial environments [J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1/2/3/4): 27-41

[59] Ross A S, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients [C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 1660-1669

[60] Papernot N, McDaniel P. On the effectiveness of defensive distillation [J]. arXiv preprint, arXiv:1607.05113, 2016

[61] Wang Bolun, Yao Yuanshun, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks [C] //Proc of the IEEE Symp Security and Privacy. Piscataway, NJ: IEEE, 2019: 707-723

[62] Du Min, Jia Ruoxi, Song D. Robust anomaly detection and backdoor attack detection via differential privacy [J]. arXiv preprint, arXiv:1911.07116, 2019

[63] Liu Yingqi, Lee W C, Tao Guanhong, et al. ABS: Scanning neural networks for back-doors by artificial brain stimulation [C] //Proc of the 2019 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 1265-1282

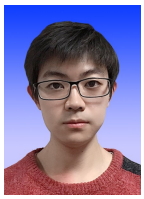
[64] Ma Shiqing, Liu Yingqi, Tao Guanhong, et al. Nic: Detecting adversarial samples with neural network invariant checking [C/OL] //Proc of the 2019 Network and Distributed System Security Symp. Reston: The Internet Society, 2018 [2021-01-16]. https://par.nsf.gov/servlets/purl/10139597

[65] Xu Xiaojun, Wang Qi, Li Huichen, et al. Detecting AI trojans using meta neural analysis [J]. arXiv preprint, arXiv: 1910.03137, 2019

[66] Chen Huili, Fu Cheng, Zhao Jishen, et al. DeepInspect: A black-box trojan detection and mitigation framework for deep neural networks [C] //Proc of the 28th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2019: 4658-4664

[67] Awan S, Li Fengjun, Luo Bo, et al. Poster: A Reliable and Accountable Privacy-Preserving Federated Learning Framework using the Blockchain [C] //Proc of the 2019 ACM SIGSAC Conf on Computer and Communications. New York: ACM, 2019: 2561-2563

[68] Cao T, Tram T, Hien D, et al. A federated learning framework for privacy-preserving and parallel training [J]. arXiv preprint, arXiv:2001.09782, 2020

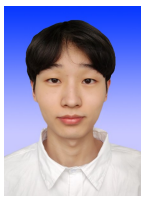


Zhou Chunyi, born in 1995. PhD candidate. His main research interests include machine learning security and privacy preserving.

周纯毅,1995 年生.博士研究生.主要研究方向为机器学习安全与隐私保护.

Chen Dawei, born in 1997. Master candidate. His main research interests include machine learning security and privacy preserving.

陈大卫,1997 年生.硕士研究生.主要研究方向为机器学习安全与隐私保护.



Wang Shang, born in 1998. Master candidate. His main research interests include machine learning security and privacy preserving.

王　尚,1998 年生.硕士研究生.主要研究方向为机器学习安全与隐私保护.

Fu Anmin, born in 1981. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include cryptography and privacy preserving.

付安民,1981 年生.博士,教授,博士生导师,CCF 高级会员.主要研究方向为密码学以及隐私保护.

Gao Yansong, born in 1986. PhD, associate professor. His main research interests include hardware security, AI security and privacy, and system security.

高艳松,1986 年生.博士,副教授.主要研究方向为硬件安全、人工智能安全和隐私、系统安全.