

## 分割学习数据隐私研究综述

秦轶群<sup>1,2</sup>, 马晓静<sup>1,2</sup>, 付佳韵<sup>1,2</sup>, 胡平一<sup>1,2</sup>, 徐鹏<sup>1,2</sup>, 金海<sup>1,3</sup>

(1. 大数据技术与系统国家地方联合工程研究中心, 服务计算技术与系统教育部重点实验室, 湖北 武汉 430074;  
2. 华中科技大学网络空间安全学院, 湖北 武汉 430074; 3. 华中科技大学计算机科学与技术学院, 湖北 武汉 430074)

**摘要:** 随着机器学习的快速发展, 以其为核心的人工智能技术已经应用于生活的各个领域, 但人们也日益担忧机器学习会泄露隐私信息。为了保护国家和公民的信息安全, 我国颁布了《中华人民共和国个人信息保护法》来规范隐私信息的收集、使用和传输。然而机器学习需要使用大量数据, 因此需要隐私保护技术在合法合规的情况下完成数据的收集和处理。分割学习可以在不共享原始数据的情况下在多个参与方之间训练分布式模型, 成为机器学习隐私保护领域的研究热点。分割学习面临着数据隐私遭受攻击的风险, 目前已有多种不同的攻击方法和对应的防御方法被提出, 但已有的综述未对分割学习训练阶段数据隐私的研究内容进行讨论和总结。因此对分割学习训练阶段的数据隐私攻防技术进行综述。归纳了分割学习定义, 给出了其原理和分类; 介绍了分割学习中两类常见的攻击方法, 原始数据重建攻击和标签推断攻击, 分析了分割学习训练阶段数据隐私受到数据隐私攻击的原因, 并总结了相应的防御方法; 最后, 讨论了分割学习数据隐私研究的未来发展方向。

**关键词:** 隐私保护; 人工智能安全; 分布式机器学习; 分割学习

**中图分类号:** TP309.2

**文献标志码:** A

**DOI:** 10.11959/j.issn.2096-109x.2024037

## Survey of split learning data privacy

QIN Yiqun<sup>1,2</sup>, MA Xiaojing<sup>1,2</sup>, FU Jiayun<sup>1,2</sup>, HU Pingyi<sup>1,2</sup>, XU Peng<sup>1,2</sup>, JIN Hai<sup>1,3</sup>

1. Services Computing Technology and System Lab, National Engineering Research  
Center for Big Data Technology, Wuhan 430074, China

2. School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

3. School of Computer Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

**Abstract:** With the rapid development of machine learning, artificial intelligence technology has been widely applied across various domains of life. However, concerns regarding the privacy risks associated with machine learning have increased. In response to these concerns, the Personal Information Protection Law of the People's Republic of China was promulgated to regulate the collection, use, and transmission of private information. Despite

收稿日期: 2023-03-02; 修回日期: 2024-03-04

通信作者: 马晓静, lindahust@hust.edu.cn

基金项目: 国家自然科学基金 (62272175)

**Foundation Item:** National Nature Science Foundation of China (62272175)

**引用格式:** 秦轶群, 马晓静, 付佳韵, 等. 分割学习数据隐私研究综述[J]. 网络与信息安全学报, 2024, 10(3): 20-37.

**Citation Format:** QIN Y Q, MA X J, FU J Y, et al. Survey of split learning data privacy[J]. Chinese Journal of Network and Information Security, 2024, 10(3): 20-37.

this, machine learning requires a large amount of data, necessitating the development of privacy protection technologies that allow for the collection and processing of data under legal and compliant conditions. Split learning, a privacy-preserving machine learning technique that enables the training of distributed models among multiple participants without sharing raw data, has emerged as a research focus. It has been recognized that split learning is vulnerable to data privacy attacks, and various attacks along with corresponding defenses have been proposed. However, existing surveys have not discussed and summarized research on data privacy during the training phase of split learning. The comprehensive overview of data privacy attack and defense techniques in the training phase of split learning was offered. Initially, the definition, principles, and classifications of split learning were summarized. Subsequently, two common attacks in split learning, namely the raw data reconstruction attack and the label leakage attack, were introduced. The causes of these attacks in the training phase of split learning were then analyzed, and corresponding defenses were presented. Finally, future research directions in the area of data privacy for split learning were discussed.

**Keywords:** privacy protection, artificial intelligence security, distributed machine learning, split learning

## 0 引言

以机器学习为核心的人工智能技术已经在生活的各个领域得到了广泛应用,但其也面临着挑战。一是数据隐私问题,在机器学习中,数据和模型都是高价值资产,需要受到保护,但不同公司和实体间存在数据壁垒,导致数据不能充分发挥其价值。以智慧医疗为例,虽然机器学习在医疗领域已经取得了显著的成果,但是,用户可能并不希望自己的隐私信息泄露给医疗公司,同时医院只能获取本院的数据信息,而某些病症需要联合多家医院的数据共同完成诊断,但医院之间不能直接以明文形式交换数据,因此难以实现联合诊断。对此,我国颁布了《中华人民共和国个人信息保护法》,数据隐私问题亟待解决。二是计算能力问题,由于机器学习需要大量的计算和存储资源,而低算力节点无法满足此类需求,因此需要借助高算力云设备协同完成整个计算过程。

为了打破数据壁垒,谷歌提出了联邦学习<sup>[1]</sup>(federated learning, FL)的概念,参与联邦学习的客户端在本地训练完整的模型,将每一轮训练结束后的模型参数发送给服务器,服务器负责聚合所有客户端的模型并生成全局模型,最后将全局模型发送给客户端,客户端以全局模型为基础,进行下一轮训练。在联邦学习中,客户端需要训练与服务器模型结构相同的完整模型,对客户端的计算能力有一定的要求,并且客户端和服务端需要传输完整的模型参数,会占用大量网络带宽。

麻省理工学院(Massachusetts Institute of Technology, MIT)提出的分割学习<sup>[2]</sup>(split learning, SL)通过分割模型结构和训练过程,将模型和计算分散到拥有数据的客户端和拥有高算力设备的服务器上,各方拥有部分模型并执行部分训练过程,在训练过程中仅传输部分中间结果,共同训练完整的分布式模型。

分割学习与联邦学习均为分布式隐私保护机器学习框架,二者存在一定的相似之处,即数据不直接交给服务器训练。但分割学习的客户端负责提取特征,服务器负责处理特征,联邦学习的客户端主要负责训练模型,服务器负责聚合模型,二者存在如下的联系和区别:①当数据纵向分布于客户端时,联邦学习经常与分割学习结合使用,其中分割学习负责提取特征部分,联邦学习负责特征聚合推断部分。②当数据横向分布于客户端时,联邦学习与分割学习的训练流程不同,在联邦学习中,多个客户端并行训练并传输完整模型,服务器负责聚合多个客户端的模型,模型训练较为稳定,但通信开销和客户端计算开销较大;在分割学习中,单个客户端和服务端串行训练所持有的部分模型,仅传输中间结果,通信开销和客户端计算开销较小,但客户端数量较多时训练效果较差。③联邦学习为多客户端训练而设计,分割学习同时支持多客户端和单客户端训练。

研究人员已经证明了分割学习在医学图像分类<sup>[3-4]</sup>、图像分割<sup>[5]</sup>、Transformer<sup>[6-7]</sup>、图计算<sup>[8]</sup>和长短期记忆(long short-term memory, LSTM)<sup>[9]</sup>等训练任务中的可行性,由分割学习训练出来的

模型拥有与本地训练出的模型相似的性能和准确率,并且在训练过程中,参与方并不解除自己以外的任何模型和数据细节。这种训练方式不仅可以保护数据持有方训练数据的隐私性,还可以利用高算力服务器加速训练过程,更可以与纵向联邦学习<sup>[10-11]</sup>相结合,从而打破数据壁垒,充分地发挥数据价值,因此成了人工智能领域的研究热点。

已有相关综述<sup>[12-18]</sup>对分割学习进行了详细的介绍,但是大部分综述<sup>[12-17]</sup>未对分割学习面临的隐私威胁和可行的防御方法进行细致的描述和分析。MIT 的 Vepakomma 等<sup>[18]</sup>在 2022 年发表的综述中,讨论了分割学习训练阶段完成后,推断阶段的隐私泄露问题和对应的防御方法。分割学习训练阶段传输的中间结果也可能会泄露隐私,本文重点针对分割学习训练阶段的数据隐私问题和可行的防御方法进行归纳和总结。

## 1 分割学习基本概念

### 1.1 分割学习

分割学习<sup>[2]</sup>是为了解决数据隐私和计算资源匮乏问题提出的分布式机器学习框架。在该框架中,拥有数据的实体称为客户端,拥有计算资源或者协调计算过程的实体称为服务器。分割学习分为两个阶段,训练阶段和推断阶段,本节主要描述训练阶段的细节流程。

传统分割学习的训练过程通过层级分割模型结构,以分割层为界限,客户端拥有底层模型,服务器拥有顶层模型,客户端和服务器通过训练本地模型并传输计算的中间结果协同训练整个分布式模型,最终得到收敛的模型。由于模型训练过程中,服务器和客户端之间传输的中间结果是经过模型计算得到的高维特征,所以隐藏了原始数据的信息。

### 1.2 分割学习分类

分割学习在发展的过程中,针对不同的应用场景,产生了许多变体。Vepakomma 等<sup>[2]</sup>在 SplitNN 中将分割学习分为传统分割学习、标签不共享的分割学习以及纵向数据切分的分割学习。作为一个快速发展的研究方向,分割学习的变体层出不穷,如结合联邦学习和分割学习以加速训练的

SplitFed<sup>[9]</sup>以及优化分割学习训练开销的 AdaSplit<sup>[20]</sup>等。在本文中,主要针对上述的 3 类分割学习框架中的数据隐私问题而展开。

#### 1.2.1 传统分割学习

传统分割学习是客户端与服务器双方协作的学习方法,由一对客户端与服务器共同完成模型训练,以解决客户端的原始数据泄露和算力不足的问题。在传统分割学习中,客户端和服务器分别拥有部分模型,客户端利用模型将原始数据映射为高维特征,并将其交给服务器训练,服务器计算出梯度后交还给客户端进行模型更新,通过这种方式,双方在不共享原始数据的情况下完成模型训练。

在传统分割学习中,客户端拥有原始数据和底层模型,服务器拥有标签和顶层模型。其中标签通常是由客户端共享给服务器的,传统分割学习框架如图 1 所示。

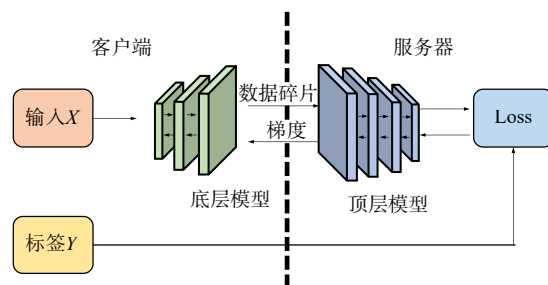


图 1 传统分割学习框架  
Figure 1 Vanilla split learning framework

轮训练流程如下。

(1) 客户端的本地使用数据  $x_0$  在底层模型  $w_c$  上进行前向传播,得到第  $k$  层(分割层)模型的计算结果  $x_k = \text{Forward}(x_0, x_c)$ , 计算结果也称为激活信息(activation)或中间特征(intermediate feature),也可将模型计算过程看作  $x_k = f_1(x_0)$ , 其中  $f_1$  代表底层模型所对应的函数,最后将计算结果上传给服务器。

(2) 服务器接收到激活信息  $x_k$  后,以其为输入,继续前向传播流程计算  $y = f_2(f_1(x_k))$ , 之后计算损失函数  $\text{Loss} = \text{CE}(y, y_{\text{true}})$ , 并反向传播更新顶层模型  $w_s$ , 得到分割层反向传播的梯度  $g_k = \text{BackProp}(x_k, x_s)$ , 并将其发送给客户端, 其中  $f_2$  代表顶层模型对应函数,  $y$  代表顶层模型输出,

$y_{\text{true}}$  代表真实标签。

(3) 客户端接收数据后, 使用服务器的梯度  $g_k$  继续执行反向传播过程并更新底层模型  $w_c$ 。

由于不同类型分割学习在训练过程中, 对激活信息和梯度的计算方式相似, 因此在后续对其他分割学习框架的流程描述中, 将着重介绍模型结构和传输流程的区别, 忽略其中的计算细节。

在传统分割学习框架中, 服务器与客户端每一轮训练需要进行一轮数据传输, 开销较小。但当标签是由客户端共享给服务器时, 会暴露标签的分布信息, 隐私性较差。因此, 此类框架适用于数据分布均匀、对实时性要求高的任务。

### 1.2.2 标签不共享的分割学习

标签不共享的分割学习在传统分割学习保护原始数据隐私的基础上, 进一步保护了标签信息的隐私性。标签不共享的分割学习把模型分为3份, 将直接参与原始数据和目标函数计算的部分模型交给客户端, 客户端不再需要向服务器共享标签信息, 双方在不共享原始数据和标签信息的情况下, 完成模型训练。

在标签不共享的分割学习中, 客户端拥有数据、标签、底层模型和顶层模型, 服务器拥有中间模型。标签不共享分割学习框架如图2所示。

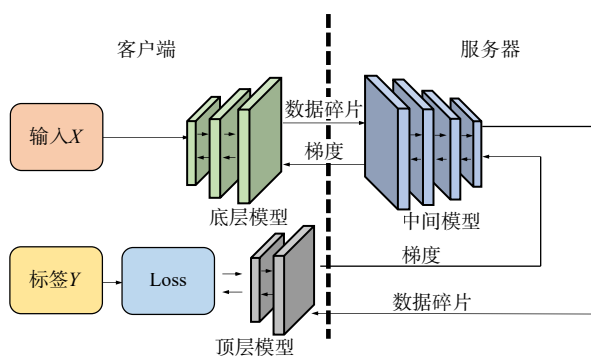


图2 标签不共享的分割学习框架  
Figure 2 Split learning framework without label sharing

标签不共享的分割学习训练流程分为以下4步: ①客户端向服务器传输底层模型输出的激活信息; ②服务器向客户端传输中间模型输出的激活信息; ③客户端向服务器中间模型传输顶层模型梯度信息; ④服务器向客户端底层模型传输中间模型输出的梯度信息。

在标签不共享的分割学习框架下, 服务器无

法直接接触到数据和标签, 因此具有更高的隐私性, 但是每一轮训练都需要进行两轮数据传输, 开销较大。因此, 此类分割学习框架适用于数据分布会泄露隐私、对实时性要求较低的任务。

### 1.2.3 纵向数据切分的分割学习

纵向数据切分的分割学习是多客户端的分割学习。在此框架下, 多个客户端可以在不共享原始数据的情况下训练分布式模型。例如, 在多家专科医疗机构中, 包括眼科、脑科和神经科等, 希望联合分析同一个患者在不同机构的检测报告, 做出统一的诊断。但不同机构的检测报告属于商业机密和个人隐私, 无法实现相互共享, 而应用纵向数据切分的分割学习就可以解决此问题。

在该框架下, 每个客户端拥有底层模型和同一个数据对象的不同特征, 服务器负责聚合特征并通过统一的顶层模型帮助所有客户端进行训练。纵向数据切分的分割学习框架如图3所示。

纵向数据切分的分割学习训练流程分为2步: ①每个客户端分别向服务器传输底层模型输出的激活信息; ②服务器接收到所有的激活信息后进行特征聚合作为顶层模型的输入, 向每个客户端传输顶层模型输出的梯度信息。

纵向数据切分的分割学习适用于数据的不同特征分散在多个客户端的场景, 解决多数据聚合问题, 在具体应用中, 经常与纵向联邦学习结合使用, 解决同数据对象的多角度信息多源训练问题。

## 2 分割学习存在的隐私安全问题

在分割学习的训练过程中客户端和服务器需要传输计算的中间结果, 这些中间结果中包含着隐私信息, 如原始数据和标签, 因此可能会导致机密数据泄露。本节总结了目前已有的针对分割学习训练过程的两大类攻击方法: 原始数据重建攻击和标签推断攻击。部分工作同时包含了两类攻击方法, 对同一工作的两类攻击方法分别在2.1节和2.2节中介绍; 部分工作分别介绍了攻击方法和对应的防御方法, 其中防御方法将在第3节中介绍。分割学习训练阶段攻击方法总结如表1所示。

### 2.1 原始数据重建攻击

当分割学习训练时, 客户端的模型可以被视

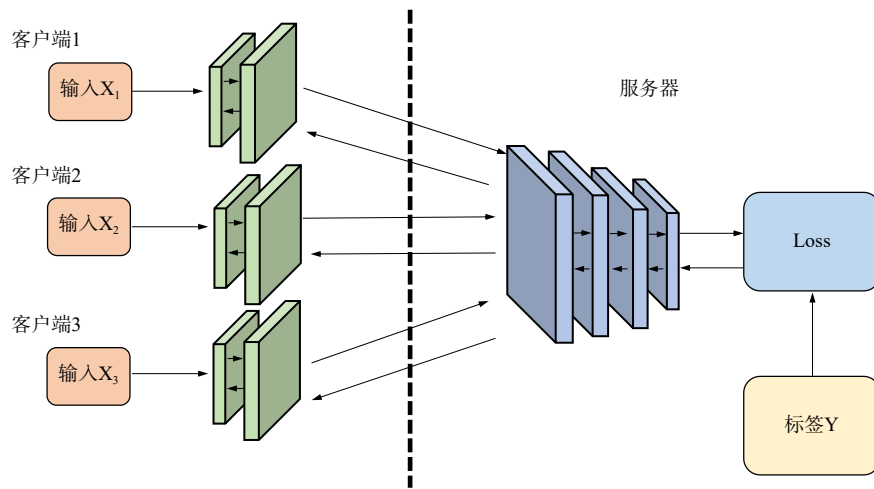


图3 纵向数据切分的分割学习框架  
Figure 3 Split learning framework for vertically partitioned data

表1 分割学习训练阶段攻击方法总结  
Table 1 Summary of attack methods in split learning training phase

| 攻击类型        | 参考文献   | 攻击手段              | 攻击形式 | 攻击条件                | 神经网络类型    |
|-------------|--------|-------------------|------|---------------------|-----------|
| 原始数据重建攻击    | 文献[21] | 优化影子模型并逆推中间结果     | 被动   | 了解原始模型结构            | 图像分类网络    |
|             | 文献[22] | 优化影子模型并逆推中间结果     | 被动   | 拥有少量同任务带标签数据        | 图像分类网络    |
|             | 文献[23] | 遍历搜索特征空间并线性求解     | 被动   | 拥有数据集知识且特征维度低       | 二分类纵向联邦网络 |
|             | 文献[24] | 篡改更新使客户端模型更新为恶意模型 | 主动   | 拥有同任务类型的公开数据集       | 图像分类网络    |
| 标签推断攻击      | 文献[25] | 微调底层模型使其可以用于直接分类  | 被动   | 底层模型训练足够好且拥有少量带标签数据 | 图像分类网络    |
|             | 文献[26] | 正则化与统计分类          | 被动   | 样本数量不均衡             | 二分类网络     |
|             | 文献[27] | 无监督聚类             | 被动   | 拥有少量带标签数据           | 图像分类网络    |
|             | 文献[28] | 添加分类层并利用半监督学习微调   | 被动   | 模型训练完成且拥有少量带标签数据    | 图像分类网络    |
|             | 文献[29] | 根据链式规则求解联立方程组     | 被动   | —                   | 二分类网络     |
|             | 文献[21] | 优化影子模型并逆推中间结果     | 被动   | 拥有模型结构              | 图像分类网络    |
| 基于优化的标签推断攻击 | 文献[30] | 监督学习              | 被动   | 了解模型和数据集关键属性        | 分类网络      |
|             | 文献[22] | 优化影子模型并逆推中间结果     | 被动   | 拥有少量同任务带标签数据        | 图像分类网络    |

为将原始数据映射到高维度特征空间的映射函数  $f$ 。客户端将原始数据输入本地模型进行计算，得到模型输出，即激活信息。之后，客户端将激活信息发送给服务器继续执行训练流程。原始数据重建攻击的攻击者为服务器，当其接收到客户端传输的激活信息后，可以利用激活信息中隐含的信息来学习客户端的近似映射函数  $f'$  进而逆向出逆映射函数  $f'^{-1}$ ，或者通过控制训练过程进而修改映射函数  $f$  为  $f'$ ，最终通过学习到的逆映射函数  $f'^{-1}$  将激活信息重建为原始数据。

### 2.1.1 模型反转攻击

模型反转攻击 (model inversion attack, MIA) 是黑盒被动攻击方法，其威胁模型针对的是客户端和半诚实的服务器利用分割学习共同训练的模型，其中服务器严格执行协议，且无法对客户端进行任意查询，但会使用收集到的数据重建客户端的输入。由于不同的攻击方法威胁模型对于模型结构和服务器是否拥有训练集的知识假设不同，在之后对攻击方法的介绍中会详细介绍其威胁模型的特点。

Erdoğan 等<sup>[21]</sup>提出两步的基于优化的模型反转攻击 Unsplit, 假设服务器清楚客户端的模型结构, 但不清楚具体参数, 也没有任何客户端的训练数据。

第一步: 在训练前, 攻击者随机初始化一个和客户端模型结构相同的影子模型  $\tilde{\theta}_1$ , 之后在每轮分割学习训练过程中, 随机初始化一批伪训练数据  $\tilde{x}$ , 以式 (1) 和式 (2) 为目标函数分别优化伪训练数据和影子模型直至收敛, 在一段训练流程后即可得到和客户端模型  $\theta_1$  结构基本相同的模型。

$$\tilde{x}^* = \operatorname{argmin}_{\tilde{x}} \operatorname{MSE}(\tilde{f}(\theta_1, \tilde{x}), f(\theta_1, x)) + \lambda \operatorname{TV}(\tilde{x}) \quad (1)$$

$$\tilde{\theta}_1^* = \operatorname{argmin}_{\tilde{\theta}} \operatorname{MSE}(\tilde{f}(\tilde{\theta}_1, \tilde{x}), f(\theta_1, x)) \quad (2)$$

其中,  $f$  和  $\tilde{f}$  代表模型运算,  $x$  代表原始数据,  $\operatorname{MSE}$  (mean-square error) 为均方误差损失函数,  $\operatorname{TV}$  (total variation)<sup>[31]</sup> 用来约束重建的数据, 使其更自然。

第二步: 服务器在得到影子模型之后, 就可以学习到影子模型对应的逆映射模型, 利用逆映射模型和激活信息即可反推出客户端输入的原始数据。Gao 等<sup>[22]</sup>提出了一个更有效的功能与数据窃取伪客户端攻击 PCAT (pesudo client authentication throttling) 方案, 其目标是窃取一个和客户端模型函数功能相似的影子模型, 假设服务器不了解客户端模型结构和参数, 但拥有少量带标签的同学习任务的数据 (私人训练集的 0.1%~0.5%)。攻击者首先随机初始化一个影子模型, 之后在每一轮分割学习过程中, 使用和客户端同标签的数据对影子模型训练就可以借助服务器顶层模型拥有的知识帮助提高影子模型在函数级别上模拟客户端底层模型的映射能力。当影子模型训练完毕之后就可以获取对应的逆映射模型来恢复原始数据。此方法主要是利用服务器底层模型的能力, 帮助优化影子模型, 最终影子模型和客户端底层模型具有相似的函数化能力。

以上两种基于优化的方法需要服务器先获取或者模拟客户端的底层模型, 因此原始数据恢复的效果取决于服务器影子模型和客户端底层模型的相似程度, 而在未知输入和未知模型参数的情

况下, 训练出和客户端底层模型相似的模型是困难的。实验结果表明, 在底层模型较浅、学习任务较简单的情况下此类方法攻击效果较好, 但是当底层模型较深或学习任务复杂时, 攻击效果将会变差。如图 4 所示<sup>[22]</sup>, 在 CIFAR-10 数据集上, Unsplit 攻击方法在客户端底层模型层数为 2 时, 重建出的图像已经无法分辨; PCAT 表现相对较好, 但重建出的图像十分模糊, 且随着任务变难和底层模型层数增加, 攻击效果也随之变差。

### 2.1.2 特征重建攻击

特征重建攻击 (feature reconstruction attack, FRA) 是黑盒被动攻击, Ye 等<sup>[23]</sup>提出对分割学习与纵向联邦学习相结合场景中的二分类任务的攻击, 其威胁模型假设在两方参与的纵向联邦学习中, 持有部分底层模型和特征的参与者为被动方, 持有部分底层模型和完整顶层模型、特征和标签的参与者为主动方, 被动方和主动方利用双方的特征共同执行模型训练, 被动方将经本地模型计算得到的激活信息交给主动方, 主动方计算本地特征并聚合双方的激活信息用于之后的训练流程。在此场景下主动方作为攻击者, 希望重建出被动方的原始输入特征。本文证明了特征重建攻击是 NP 难 (NP-hard) 的问题, 但是在特征维度较小且攻击者了解数据集知识的前提下, 仍然可以采用基于搜索的方法根据激活信息逆向搜索原始输入, 通过求解线性等式方程或线性回归方程, 遍历所有可能的解空间并利用已有的知识帮助缩小解集, 最终可以重建出被动方的原始输入特征。但是此方法仅对二分类任务有效, 只有主动方拥有被动方数据集知识的情况下, 才可以发动攻击; 且当特征维度高到一定程度时, 求解会变得困难, 在最坏的情况下可能无法求解。

### 2.1.3 特征空间劫持攻击

Pasquini 等<sup>[24]</sup>提出的特征空间劫持攻击 (feature space hijacking attack, FSHA) 是黑盒主动攻击, 其威胁模型是客户端和恶意服务器执行分割学习协议, 服务器不严格执行协议, 可以篡改传输数据内容。

FSHA 假设服务器不了解客户端的模型结构和参数, 不需要拥有客户端的数据集, 但是拥有



图4 Unsplit和PCAT在CIFAR-10数据集上的重建效果  
Figure 4 Data reconstruction on CIFAR-10 of Unsplit and PCAT

一个和客户端数据集同类型的公开数据集  $X_{\text{pub}}$ ，例如，若目标任务为人脸识别，目标任务使用 UTK-Face 数据集，辅助数据集可以使用 CelebA-HQ 数据集。服务器利用辅助数据集在本地以式 (3) 为目标函数训练可以将数据映射为高维特征的编码器  $\tilde{f}$  和可以将高维特征低损耗映射回数据本身的解码器的自动编码器  $\tilde{f}^{-1}$ 。

$$L_{\tilde{f}\tilde{f}^{-1}} = d(\tilde{f}^{-1}(\tilde{f}(X_{\text{pub}})), X_{\text{pub}}) \quad (3)$$

其中， $d$  代表任意的距离损失函数，如 MSE 函数。

同时，服务器会以式 (4) 为目标函数训练判别器  $D$  来区分激活信息  $f(X_{\text{priv}})$  和编码器生成的高维特征  $\tilde{f}(X_{\text{pub}})$ 。

$$L_D = \log(1 - D(\tilde{f}(X_{\text{pub}}))) + \log(D(f(X_{\text{priv}}))) \quad (4)$$

之后服务器会篡改客户端的更新流程，向客户端发送使用式 (5) 计算出的梯度，使客户端模型的训练朝服务器预定的方向进行。

$$L_f = \log(1 - D(f(X_{\text{priv}}))) \quad (5)$$

最终，恶意服务器可以以高概率令客户端模型的输出空间和编码器的输出空间重叠，然后使用解码器便可以将激活信息恢复为客户端的原始数据。

从文献[24]数据重建结果来看，FSHA 是一种非常强力的攻击手段，本文分别对不同深度的底层模型进行了实验，结果表明即使在底层模型较深时，重建出来的图像仍然具有明显的辨识度。相较 Unsplit 和 PCAT 等被动攻击，FSHA 不仅重建效果好，还会损害客户端的底层模型，使其丧失原有功能，因此危害性更强。

## 2.2 标签推断攻击

标签推断攻击的攻击目标是无标签的参与方想要去获取数据对应的标签。本节根据攻击原理将其分为基于数据分布的标签推断攻击和基于优化的标签推断攻击。

### 2.2.1 基于数据分布的标签推断攻击

基于数据分布的标签推断攻击利用数据本身的分布特点或模型对数据分布的分类能力，根据同类型数据在分布上的聚类特性实施攻击。

Zheng 等<sup>[25]</sup>提出，当客户端的底层模型训练得足够好时，能够获得分离不同类别的样本和聚集同一类别的样本的能力，其输出结果已经变得和标签高度相关，所以仅需要使用部分带标签的数据就可以将底层模型微调成高性能的分类器。Li 等<sup>[26]</sup>发现，在正负样本数量不均衡的二分类任务下，正负样本产生的梯度的分布同样是不均衡的，因此可以通过对正则化 (norm-based) 梯度的分布进行统计，判断出样本的正负标签。虽然该方法以 AUC (area under curve) 接近 1 的成功概率在二分类数据集下实施攻击，但是此方法的条件限制过于严格，当样本数量均衡时，便难以实现攻击。Liu 等<sup>[27]</sup>观察到在分割学习中，不论是激活信息还是梯度，经过 K-means 等无监督学习方法处理后，在空间上都呈现出聚类的特性，如图 5 所示，对 CIFAR-10 数据集的中间结果进行主成分分析 (principal components analysis, PCA) 技术降维，并在三维空间坐标系中绘制出来，数据呈现出高度聚合的特点。因此，仅仅需要少量有标签的数据 (每类最少一个数据) 就可

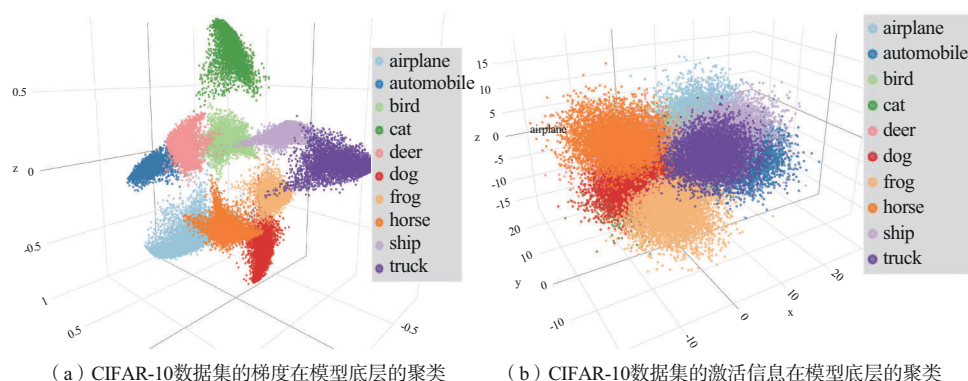


图5 CIFAR-10数据集的梯度和激活信息在模型底层的聚类表现  
Figure 5 Clustering results of gradients and activations on CIFAR-10 at the bottom of model

以为聚类打上标签, 对于每一个新数据, 即使不了解客户端模型的信息也可以推断出其对应的标签。

从图5中可以看出, 虽然中间结果在分布上是聚类的, 但不同类数据在区间边界处有部分重叠, 任务越复杂, 重叠区间越大。因此对决策边界附近的数据来说, 很容易被分类到错误的标签上, 想要更精确地标注聚类就需要更多数据。

Fu等<sup>[28]</sup>发现, 在使用以分割学习训练框架为基础的纵向联邦学习进行训练的过程中, 当模型训练完成后, 半诚实攻击者只需在客户端底层模型后增加一层模型作为分类器, 少量带标签的辅助数据通过半监督学习方法对模型进行微调, 就可以获取完整的带有分类功能的模型用来推断标签, 因此, 此攻击方法也称为模型完成攻击(model completion attack)。在CIFAR-10数据集上, 攻击者仅需要使用不到1%的辅助数据, 即可训练出具有64%测试集准确率的标签推断分类器。由于此攻击方法在模型基本训练完毕的情况下才会生效, 因此本文提出了主动攻击方案, 恶意客户端通过设计自适应的本地优化器来扩大服务器传输给底层模型的梯度, 增加梯度下降速度, 加速模型的收敛, 更快地完成训练, 实现攻击目的。

仅针对二分类任务, Yuan等<sup>[29]</sup>提出了基于链式规则的确定性攻击算法, 通过建立并求解关于标签、模型参数和梯度之间的联立方程组, 可以根据聚合梯度以100%精度推断标签。

### 2.2.2 基于优化的标签推断攻击

基于优化的标签推断攻击将标签推断攻击看

作一个可学习的任务, 因此可以利用优化算法训练出标签推断模型来实施攻击。Erdoğan等<sup>[21]</sup>在Unsplit基础上提出的标签推断攻击的方法和其重建原始数据的方法类似, 通过随机初始化可学习的影子模型 $\tilde{\theta}_2$ , 在每一轮训练中随机初始化标签 $\tilde{y}$ , 使用式(6)所示的损失函数, 利用梯度信息训练推断客户端的顶层模型 $\theta_2$ , 进而推断标签 $\tilde{y}^*$ 。

$$\tilde{y}^* = \operatorname{argmin}_{\tilde{y}} \cdot$$

$$\left( \frac{\partial L(f_2(f_1(x)), y)}{\partial \theta_2}, \frac{\partial L(\tilde{f}_2(f_1(x)), \tilde{y})}{\partial \tilde{\theta}_2} \right) \quad (6)$$

Kariyappa等<sup>[30]</sup>提出, 半诚实攻击者在了解模型和数据集的关键属性(梯度匹配、标签分布、伪标签熵和伪标签准确率)时, 可以将标签推断攻击设计为监督学习的任务, 根据上述的关键属性, 设置初始化模型、标签向量、激活信息及对应的梯度信息, 利用精心设计的目标函数就可以训练出可以推断标签的替代模型, 其在顶层模型层数为2和3时, 在图像分类数据集上的表现均优于Unsplit和基于K-Means聚类的方法, 且在二分类数据集上和Norm-Based<sup>[26]</sup>的攻击成功率相近。但是该方法对于数据集和模型的关键属性的假设较强, 现实中很难直接实现。在Gao等<sup>[22]</sup>提出的PCAT中, 由于攻击者需要用标签来确定每轮训练的数据, 而在标签不共享的分割学习中, 服务器无法获取标签, 因此他们提出攻击者可以构建伪顶层模型来替代真实的顶层模型, 服务器将客户端的激活信息送入伪顶层模型来推断私人标签, 最终用于数据重建攻击。在MNIST(modified national institute of standards and technology)和CIFAR-10数据集上, 当Unsplit<sup>[21]</sup>在顶

层模型层数较少时准确率较高, 而随着层数增加, 准确率下降明显, PCAT<sup>[22]</sup>的攻击效果随着顶层模型层数增加也有所下降, 但下降幅度较小。根据上述实验结果可以总结出, 攻击者对数据集和模型的了解程度越高, 标签推断的成功率就越高。

### 3 防御方法分类及原理

针对分割学习中存在的隐私风险, 研究者从不同角度出发、利用不同技术提出了多种解决方案, 本节首先分析分割学习数据隐私受到攻击的原因, 然后介绍常用的隐私保护技术的原理, 最后针对不同攻击方法, 归纳总结已有或可行的防御方法。

#### 3.1 分割学习隐私风险分析

分割学习面临的隐私风险主要来自参与方之间传输的计算中间结果。恶意攻击者可以利用中间结果恢复出原始输入或标签信息。中间结果包括激活信息和梯度, 它们都是由原始数据和标签分别经模型计算得到的, 因此隐含着原始数据和标签信息。若将模型看作映射函数 $f$ , 那么攻击者只需要恢复出相似的映射 $f' \approx f$ , 就可以逆向构造出逆映射函数 $f'^{-1} \approx f^{-1}$ , 或者通过劫持训练过程直接构造出逆映射函数 $f^{-1}$ 来恢复出原始数据或标签信息。相较利用逆映射函数将中间结果逆推得到原始数据和标签信息, 中间结果本身的性质让标签推断变得更加容易。训练好的神经网络分类模型具有强大的聚类能力, 中间层的计算结果和梯度信息在高维特征空间内都具有一定的聚类特性, 攻击者只需利用少量的带标签数据就可以构建一个完整的分类器去推断标签。

因此, 为了解决分割学习存在隐私风险的问题, 需要增加攻击者从中间结果提取信息的难度, 并检查是否存在篡改梯度信息的恶意服务器。

#### 3.2 常用隐私保护技术

为了解决分割学习中数据隐私泄露的问题, 现有的方案会用到差分隐私 (differential privacy, DP)、多方安全计算 (secure multi-party computation, SMPC)、同态加密 (homomorphic encryption, HE) 和函数加密 (functional encryption, FE) 等密码学方法保护中间结果, 此类方法在其他领域的研究

已广泛应用于隐私保护机器学习<sup>[45]</sup>中。为了在下文中更好地描述分割学习的防御方法, 在本小节内对上述的密码学方法进行了简单的介绍。

**定义 1** 差分隐私<sup>[46]</sup>。设随机化算法 $A$ , 分别作用于两个相邻数据集 $D_1, D_2$ , 得到的输出 $S$ 难以区分, 即 $A$ 满足

$$\Pr[A(D_1) \in S] \leq e^\epsilon \Pr[A(D_2) \in S] + \delta \quad (7)$$

其中,  $\epsilon$ 代表隐私预算,  $\delta$ 代表允许的误差。

**定义 2** 多方安全计算<sup>[47-48]</sup>。设 $n$ 个参与方 $p_1, p_2, \dots, p_n$ , 拥有明文数据 $m_1, m_2, \dots, m_n$ , 在不暴露各自输入的前提下, 计算共同协商的函数 $y = f(m_1, m_2, \dots, m_n)$ 。

**定义 3** 同态加密<sup>[49]</sup>。假设 $n$ 个明文数据 $m_1, m_2, \dots, m_n$ , 在同态加密条件下, 满足

$$\begin{aligned} \text{Dec}(\text{Enc}(m_1) \odot \text{Enc}(m_2) \odot \dots \odot \text{Enc}(m_n)) \\ = m_1 \odot m_2 \odot 1 \dots \odot m_n \end{aligned} \quad (8)$$

其中,  $\text{Enc}$ 为加密算法,  $\text{Dec}$ 为解密算法,  $\odot$ 代表同态计算。

**定义 4** 函数加密<sup>[50]</sup>。拥有解密密钥的用户, 可以获取函数 $y = f(x)$ 计算结果的值而不会获取任何关于明文输入 $x$ 的信息。

#### 3.3 防御方法

在本节中, 分别介绍分割学习针对模型反转攻击、特征重建攻击、特征空间劫持攻击和标签推断攻击的防御方法, 如表 2 所示。

##### 3.3.1 模型反转攻击的防御方法

模型反转攻击的方法按照原理分类大致可以归为 7 种: 基于噪声添加、基于模型深度和训练方法、基于迁移学习、基于数据混合、基于数据相关性、基于二值化和基于密码学的方法。

基于噪声添加的防御方法通过客户端在激活信息处添加随机噪声掩码来扰乱服务器的数据重建。差分隐私<sup>[46]</sup>是可证明安全性的噪声添加技术, 通过添加高斯噪声、拉普拉斯噪声等为数据添加随机性, 基于差分隐私噪声的方法已经广泛应用到各类防御方法中。

基于模型深度和训练方法的防御方法通过分配给客户端更深的底层模型或利用随机化训练方法来提高激活信息的复杂性。Abuadba 等<sup>[32]</sup>在研究中发现, 使用分割学习训练用于信号处理的一维卷积神经网络时, 分割学习无法提供足够的隐

表2 分割学习训练阶段防御方法总结  
Table 2 Summary of defense methods in split learning training phase

| 攻击类型     | 参考文献      | 防御方法                        | 神经网络类型         |
|----------|-----------|-----------------------------|----------------|
| 原始数据重建攻击 | 文献[32]    | 差分隐私、增加模型深度                 | 一维信号处理 CNN     |
|          | 文献[5]     | 差分隐私、增加模型深度、Dropout         | 医疗图像分类网络       |
|          | 文献[33]    | 使用迁移学习专家模型初始化               | SplitFed 训练框架  |
|          | 文献[7]     | 数据与标签混合                     | 分类 Transformer |
|          | 文献[34]    | 差分隐私、数据与标签混合                | 图像分类网络         |
|          | 文献[35-37] | 减少原始数据与中间结果的相关性             | 分类网络           |
|          | 文献[38]    | 部分网络与数据二值化                  | 分类网络           |
|          | 文献[39]    | 模型末端增加二值化哈希层                | 分类网络           |
|          | 文献[40]    | 利用 HE、MPC 保护中间结果聚合          | —              |
|          | 文献[41]    | 利用函数加密顶层模型运算                | —              |
| 特征重建攻击   | 文献[23]    | 添加可以误导攻击者的特征                | 二分类网络          |
| 特征空间劫持攻击 | 文献[38]    | 部分网络与数据二值化                  | 分类网络           |
|          | 文献[42]    | 间歇发送伪造数据以检测是否存在攻击           | 分类网络           |
| 标签推断攻击   | 文献[29]    | 差分隐私                        | 分类网络           |
|          | 文献[26]    | 添加噪声平衡正负样本数量                | 二分类网络          |
|          | 文献[25]    | 拉远同标签样本在特征空间上的距离            | 分类网络           |
|          | 文献[28]    | 多种防御策略                      | 纵向联邦分类网络       |
|          | 文献[43]    | 训练影子模型并利用秘密共享技术在两个模型间分享参数   | 纵向联邦网络         |
|          | 文献[44]    | 添加噪声阻碍攻击者模型更新               | 分类网络           |
|          | 文献[34]    | 生成与原始数据相关的混淆标签以防止攻击者推断出真实标签 | 分类网络           |

私安全性。他们分别使用逆视图、距离相关（distance correlation）系数<sup>[51]</sup>和动态时间规整（dynamic time warping, DTW）<sup>[52]</sup>作为衡量标准，计算原始数据和激活信息之间的距离，其中逆视图方法所还原出的图像如图 6 所示<sup>[32]</sup>。

研究发现经过一维卷积层计算得到的激活信息和原始数据之间的距离较小，隐私保护的性能弱，容易受到数据重建攻击，因此提出了两种方法来增强分割学习的隐私性。①增加底层模型的深度：底层模型越深，其所对应的映射函数就越复杂，原始数据和激活信息之间的距离就越大，攻击者就越难攻击。②差分隐私：通过在底层模型的分割层的激活信息处添加差分隐私噪声，在尽可能少地影响训练过程的同时，扰乱攻击过程，降低攻击成功率。作为早期研究分割学习数据隐私性的文章，其提出的方法，本质是让激活信息更加复杂，攻击者难以从中抽取有效信息，

但是增加隐藏层数量会加剧客户端的运算压力，而添加差分隐私噪声会使准确率降低。Roth 等<sup>[5]</sup>提出在医疗图像分割任务的分割学习中，增加模型深度、应用差分隐私和使用 Dropout 都可以帮助客户端抵抗数据重建攻击。其中，Dropout 为模型提供了更多的随机性，提高了服务器重建原始数据的难度。

目前基于迁移学习的防御方法的相关研究较少。Li 等<sup>[33]</sup>借助迁移学习防御 SplitFed 中可能受到的模型反转攻击。在预训练阶段，高算力的专家服务器会使用和目标数据集同分布的辅助数据集训练高性能的特征提取器，如目标数据集为 CIFAR-10，则使用 CIFAR-100 作为辅助数据集。使用迁移学习技术，利用特征提取器初始化客户端的模型，模型训练过程就相当于在微调特征提取器。利用此方法初始化的模型拥有辅助数据集的知识，在训练中，即使模型受到原始数据重建

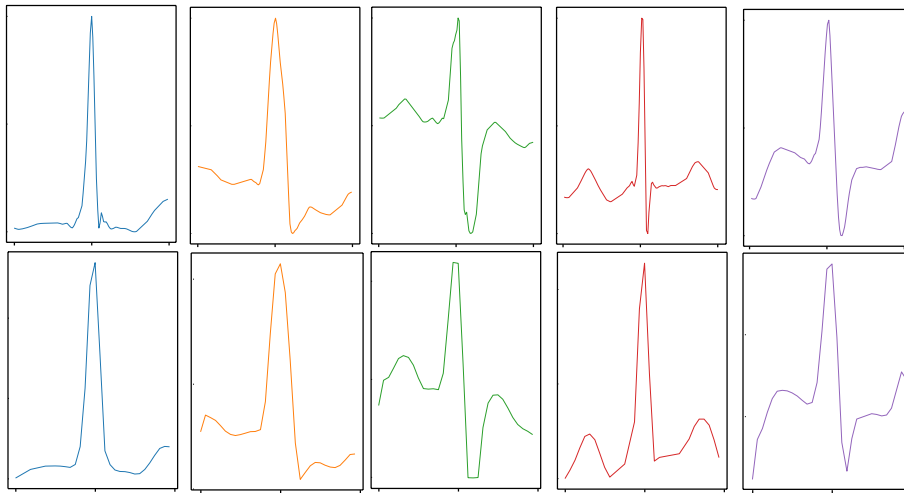


图6 原始输入(上)和第二层卷积的单通道逆视图(下)

Figure 6 Raw input and visual invertibility from one of the channels in the output from the second convolutional layer

攻击, 攻击者从激活信息中提取的信息包含了大量的辅助数据集信息, 使重建的数据接近于辅助数据集中的数据, 隐藏了真实数据。此方法可以加速模型收敛过程, 并且防止了训练早期可能存在的攻击。但是此方法的假定过强, 因为现实中获取完整的同分布辅助数据集是困难的。

基于数据混合的防御方法将多个数据样本和对应标签分别混合, 服务器使用混合数据进行训练。Xiao 等<sup>[34]</sup>提出多重激活和标签混合 (multiple activations and labels mix, MALM) 技术, 利用样本的多样性, 使用如下的混合算法:

$$\text{Mix}(A, A', \lambda) = \lambda \cdot A + (1 - \lambda) \cdot A' \quad (9)$$

其中,  $A$ 、 $A'$  代表激活值,  $\lambda$  代表混合系数。

通过生成混合激活值 (激活信息) 来降低中间特征与原始数据间的距离相关性, 从而降低重建攻击的风险。经过多次混合后的数据在视觉上是难以分辨其内容的, 并且混合数据对应的混合标签也难以为攻击者提供足够的信息来恢复出原始数据, 即使攻击者能力足够强, 恢复出来的数据仍是多张图像的叠加, 难以从中拆分出特定的图像。从实验结果来看, 在混合次数较低时, 训练出来的模型准确率与正常训练的模型准确率相近; 而在标准 CIFAR-10 和 CIFAR-100 数据集上使用混合次数更高的数据训练出的模型准确率往往会降低。Oh 等<sup>[7]</sup>提出视觉 Transformer 模型<sup>[53]</sup> (vision Transformer, ViT), 由于 Transformer 使用注意力层作为计算单元对数据进行处理, 而注意力

机制与传统卷积神经网络的卷积和池化的计算方式不同, 导致激活信息与原始数据的相似度较高, 因此利用分割学习训练 Transformer 的过程中, 数据隐私泄露的风险相较传统卷积神经网络更高。

本文改进了传统的数据混合方法 CutMix<sup>[54]</sup>, 进而提出了应用于分割学习的 CutMixSL 和带差分隐私的 DP-CutMixSL。客户端随机切分激活信息生成数据补丁 (PATCH), 并添加高斯噪声, 之后将其发送给服务器, 服务器将多个数据补丁和对应的标签信息拼接, 使用这些拼接的数据继续训练; 同时, 本文证明了多个数据补丁拼接成的新数据中蕴含的噪声符合差分隐私原理, 保证了数据隐私性。从实验结果来看, 使用 CutMixSL 训练出来的模型比使用传统分割学习训练出的模型准确率更高。数据混合是值得进一步探索的方法, 神经网络具有强大的学习能力, 因而可以学习到混合数据和混合标签之间的关联, 但相关工作仅在如 MNIST 和 CIFAR-10 等简单数据集上进行实验, 该方法在复杂数据集上的表现仍需要进一步检验。

基于数据相关性的防御方法也是目前的研究重点, 如 MIT 的 Vepakomma 等<sup>[35-37]</sup>提出的方法证明了减少原始数据和其中间特征之间的相关性可以有效防止数据重建攻击, 因此他们尝试在损失函数中添加对原始数据和中间特征之间距离相关性的约束条件, 在训练中减少原始数据和激活信

息的相关性，损失函数如下：

$$\text{Loss} = \alpha \text{DCOR}(X, Z) + (1 - \alpha) \text{CE}(Y_{\text{true}}, Y) \quad (10)$$

其中， $X$ 代表输入数据， $Z$ 代表激活信息， $Y_{\text{true}}$ 代表真实标签， $Y$ 代表预测标签， $\text{CE}$ 代表交叉熵， $\text{DCOR}$ 代表距离相关性， $\alpha$ 代表相关性系数。但是，此方法需要修改服务器计算的损失函数，因此服务器可能会使用自适应的攻击手段，并且未对模型训练中真实的攻击方法做实验验证其有效性，无法有效地说明真实攻击中防御效果的好坏。

基于二值化的防御方法利用二值量化的方法减少信息泄露。在Pham等<sup>[38]</sup>提出的二值化分割学习（binarizing split learning, B-SL）方法中，客户端对部分模型执行二值化运算，借助二值化运算消除激活信息中的信息荷载，减少隐私泄露量。在训练时，使用如下的损失函数：

$$\begin{aligned} \text{Loss} &= \lambda E_1 + (1 - \lambda) E_2 \\ &= \lambda L_1(a^{(l)}, y) + (1 - \lambda) L_2(a^{(l)}, x) \quad (11) \end{aligned}$$

通过降低激活信息和原始数据之间的关联性进一步限制隐私泄露，其中， $L_1$ 代表训练函数如交叉熵， $a^{(l)}$ 代表预测的标签， $L_2$ 代表任意距离衡量函数， $a^{(l)}$ 代表激活信息， $x$ 代表原始数据。本文还利用了差分隐私技术，进一步保护隐私。从实验结果上来看，使用B-SL训练出的模型和正常模型的准确率相近，在CIFAR-10和Fashion MNIST上分别有0.2%和0.7%的差距，在MNIST上准确率相同。Qiu等<sup>[39]</sup>提出客户端在底层模型后增加一层哈希层，通过Sign()函数将模型输出变为二值化的哈希码，哈希化的数据信息荷载低，因此服务器难以通过哈希码重建原始数据。但是此方法会导致信息损失，进而影响模型整体准确率，服务器使用如下损失函数以提高准确率：

$$\text{Loss} = \text{CE}(f_2(H), y) + \text{dist}(H, \text{Targets}) \quad (12)$$

其中， $f_2$ 代表底层模型， $\text{dist}()$ 代表任意距离函数， $H$ 代表经过二值化的激活信息， $y$ 代表真实

标签，Targets代表预定义的二值码。通过减少哈希码与预定义的二值码之间的距离来提升不同数据的哈希码之间的关联性，进而提高准确率。从图7可以明显看出<sup>[38]</sup>，使用4-bit哈希码重建出的数据已经无法用肉眼分辨其真实内容。

但是，在MNIST和CIFAR10数据集上，使用该方法与不使用该方法训练模型的准确率相比分别有1.24%和5.39%的差距。以上两种方法都是使用二值化的方法减少激活信息中的信息量，阻碍服务器提取关键特征，但是二值化的方法在去除信息荷载的同时，也会减少模型学习到的特征的数量和精度，对模型准确率产生影响，尤其是在复杂数据集上，已有的方法仅在较简单数据集上做实验，面对复杂任务如ImageNet等数据集时的表现未知。

上述的防御方法大多从神经网络的结构和其训练过程出发，而基于密码学的防御方法可以在不修改模型和训练策略的前提下，保证数据隐私。Zhou等<sup>[40]</sup>提出使用同态加密或多方安全计算保护多方分割学习客户端底层模型的运算和激活信息的聚合过程，防止服务器通过激活信息区分客户端，并在聚合的激活信息上应用差分隐私噪声进一步防御数据重建攻击。上述方法在客户端上进行加密运算，同样，在服务器上进行加密运算也可以保护客户端的数据隐私，即客户端对激活信息进行加密，服务端在加密的情况下完成整个分割学习的训练流程。目前已有许多基于同态加密和多方安全计算的隐私保护机器学习训练<sup>[41,53]</sup>的研究工作，可以应用到分割学习中保护数据隐私。Xu等<sup>[56]</sup>发现传统的基于同态加密或多方安全计算的隐私保护机器学习训练方法开销大，不适合实际应用，因此提出了基于函数加密的多数据源隐私保护机器学习训练方法，将模型训练的计算过程看作一个特定函数，客户端利用函数加密将加密的数据交给服务器，服务器就可

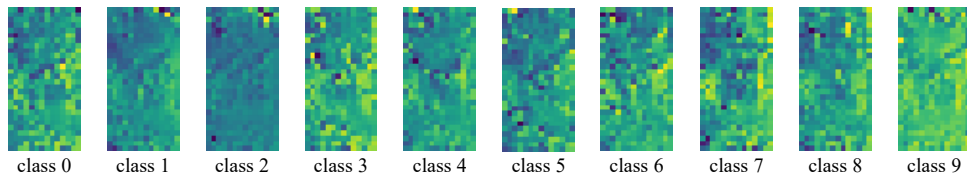


图7 MNIST上使用4-bit哈希码重建的受B-SL保护的数据  
Figure 7 Reconstruction results on MNIST with 4-bit hash code protected by B-SL

以在不解密的情况下完成整个模型的训练, 本文提出, 此方法也可以直接应用到分割学习中, 服务器仅接收到加密的激活信息, 保护了数据隐私。使用基于密码学的防御方法基本不会影响模型准确率, 并且具有可证明的隐私安全性, 但是密码学运算效率低, 密文计算时间远长于明文计算时间, 并且需要耗费庞大的计算资源和占用巨额通信开销, 不适用于现实应用, 因此基于密码学的防御方法需要被进一步探索。

### 3.3.2 特征重建攻击的防御方法

Ye 等<sup>[23]</sup>认为, 向激活信息添加高斯噪声的方法实质上是通过误导攻击者找到一个随机的错误解, 此方法虽然有效, 但是激活信息和采样噪声之间复杂的相互作用会降低模型的准确率。因此他们提出了特征保护算法, 等级降低技术 (rank reduction technique) 来扰动被动方的激活信息, 通过误导主动方, 使其获得一个预先编造的二进制特征。由于此方法不涉及随机化的噪声, 因此对模型准确率的影响较小。实验结果验证了该观点, 此防御方法可以高效地抵御攻击, 并且基本不会影响模型的准确率。

### 3.3.3 特征空间劫持攻击的防御方法

与模型反转攻击不同, 特征空间劫持攻击是主动攻击, 服务器利用辅助数据集篡改梯度信息, 实现影响客户端模型更新方向的目的。

由于 FSHA 不通过提取中间结果的信息来达

到攻击目的, 因此传统的利用差分隐私的方法难以有效地防止此类主动攻击, Pham 等<sup>[57]</sup>通过实验证明, 差分隐私不能完全防御特征空间劫持攻击, 最多可以延迟其收敛。Pham 等<sup>[38]</sup>通过实验证明了 B-SL 对 FSHA 攻击具有一定的防御效果。为了防御此类攻击, 研究者从伪造的梯度入手, 来识别是否存在攻击。Erdoğan 等<sup>[42]</sup>提出了客户端主动检测攻击的方法 SplitGuard, 客户端间歇性地向服务器发送带有假标签的数据, 由于此类数据是伪造的假数据, 其在训练过程中产生的梯度信息与常规数据不同, 当客户端模型在学习预期任务时, 假梯度和常规梯度之间的角度将高于常规梯度的两个随机子集之间的角度, 并且假梯度的幅度将高于常规梯度。据此可以判断出是否存在 FSHA 攻击。实验结果表明, SplitGuard 可以以极高的概率检测出 FSHA 攻击, 但是, 此方法难以防御自适应攻击, 由于攻击者同样可以获得防御需要用到的梯度信息, 也就可以利用假梯度和常规梯度的区别来判断客户端是否在探测攻击。Fu 等<sup>[58]</sup>提出的 SplitSpy 可以令攻击者检测出客户端发送的带假标签的数据, 攻击者通过维护执行预定任务的影子模型, 利用影子模型在训练中产生的梯度信息作为标准, 检测客户端发送的数据是否为带假标签的数据, 即可逃避 SplitGuard 的检测机制, 从图 8 所示的实验结果中可以看出<sup>[58]</sup>, SplitSpy 可以有效绕过 SplitGuard 的防

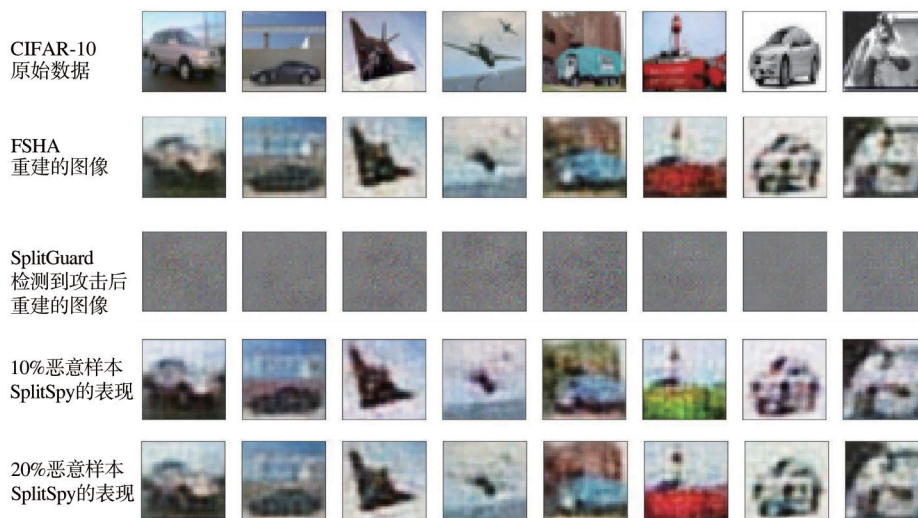


图 8 CIFAR-10 原始数据、FSHA 重建的图像、SplitGuard 检测到攻击后重建的图像, 10% 和 20% 恶意样本下 SplitSpy 的表现  
Figure 8 CIFAR-10 original data, reconstructed image by FSHA, reconstructed image when attack is detected by SplitGuard and reconstructed image with SplitSpy with 10% and 20% detecting samples

御。同时,他们还提出了一种被动的攻击检测方法 Gradients Scrutinizer,它依赖于来自预期模型的梯度和来自恶意模型的梯度之间的内在差异。对于预期模型,同标签样本的梯度之间的预期相似性与跨标签样本的梯度之间的预期相似性不同,而对于恶意模型,它们是相同的。这种内在的可区分性使 Gradients Scrutinizer 能够有效地检测出特征空间劫持攻击而不影响良性模型的正常训练,因此可以很好地防御 FSHA 攻击和针对其的自适应攻击。

针对被动攻击的防御方法如差分隐私无法有效防御 FSHA, B-SL 虽然在底层模型层数为1时具有一定的防御效果,但是当底层模型变深后,防御效果极有可能变差。因此,针对 FSHA 的防御方法大多以检测为主,客户端需要保存真实数据的梯度信息来判断恶意攻击是否存在,已有方法虽然已经可以很好地防御此类攻击,但是每一轮训练都需要完整的检测数据,会带来额外的时延,这在训练过程中是不可忽视的。

### 3.3.4 标签推断攻击的防御方法

为了更好地总结和归纳防御方法的性质,在本节中,分别总结了针对基于数据分布的标签推断攻击和基于优化的标签推断攻击的防御方法。

基于数据分布的标签推断攻击利用模型的分类能力和同类数据的中间结果(激活信息或梯度)分布的相似性,就可以通过统计分析的方法获取标签<sup>[26]</sup>或使用带标签数据执行半监督学习训练推断标签的影子模型<sup>[27-28]</sup>。针对二分类任务, Yuan 等<sup>[29]</sup>采用差分隐私技术保护标签,提出了基于差分隐私随机梯度下降(differentially private stochastic gradient descent, DP-SGD)的差分隐私下界估计算法用来帮助决定噪声添加量。Li 等<sup>[26]</sup>针对基于正则化梯度正负分布的攻击方法提出了 Marvell 防御方法,标签拥有方通过在梯度上添加噪声,令正负梯度分布均衡,使半诚实客户端无法通过梯度分布的正负样本推断样本标签。针对多分类任务, Zheng 等<sup>[25]</sup>从物理学电荷同性相斥的原理中获取了灵感,将模型对同标签数据的输出在特征空间分布上拉远,使模型输出结果在空间上接近于决策边界,进而使不同标签的数据在空间上交叠在一起。这种方法可以使攻

击者在利用少量带标签数据微调模型的时候产生较大误差,但从实验结果来看,模型准确率和攻击者攻击成功率大致呈现正比关系,具体使用时需要对隐私性和模型准确率做一定的权衡。

为了防御纵向联邦学习中的标签推断攻击, Fu 等<sup>[28]</sup>针对自己提出的模型完成攻击也提出了相应的防御方法,该文献提出可以使用以下的防御方法。①服务器向梯度中添加噪声。②梯度压缩:服务器仅向客户端传输绝对值最大的部分梯度(约为梯度总量的10%)。③隐私保护机器学习:应用差分隐私和梯度随机选择技术保护数据隐私。④离散随机梯度下降算法(DiscreteSGD):他们在 SignSGD<sup>[59]</sup>的基础上提出了 DiscreteSGD,首先根据上一轮的梯度分布将服务器切分为多个子区间,再将本轮梯度按照其所在的子区间映射到区间端点处,最后将离散的梯度信息发送给客户端。Wang 等<sup>[43]</sup>提出了新型的分布式训练的框架,其将潜在攻击者的特征数据以秘密共享的方式分享给影子模型,影子模型和潜在攻击者的底层模型的输出结果共同组成了原始数据的激活信息。因此潜在攻击者仅能获取部分梯度,从而打破了梯度和训练数据之间的关联性,使攻击者即使接收到训练的梯度,也无法使用底层模型推断标签,但是训练影子模型会增加至少一倍的训练开销。

基于优化的标签推断攻击方法将标签推断攻击当作监督学习任务,通过训练推断标签的影子模型执行攻击<sup>[22,30-31]</sup>。Yang 等<sup>[44]</sup>提出了基于梯度扰动的分割学习框架,标签拥有方通过在梯度上添加方向性的噪声扰乱梯度和模型更新方向来阻碍攻击者实施标签推断攻击。Xiao 等<sup>[34]</sup>利用 MALM 技术,通过生成与原始数据相关的混淆标签以防止攻击者推断出真实标签。

相比原始数据重建攻击,标签推断攻击难度较低,因此更难以防御。已有的防御方法大多会导致模型准确率的下降且无法以高概率进行防御攻击,因此需要研究者探索更先进的防御方法,为标签隐私提供更有力的保护。

## 4 讨论

分割学习是隐私保护机器学习方法,具有广泛的应用潜力,尤其是在智慧医疗和边缘计算等

对隐私性有要求的场景中有着出色的表现,成为热门的研究领域。但是分割学习本身仍存在安全性不可证明、传输内容可能会导致隐私泄露等问题。在实际应用中,分割学习会与纵向联邦学习等技术相结合使用,因此也会遭受不同种类的攻击。研究攻击手段才能更好地采用相应的防御方法,以应对可能出现的挑战,该领域未来发展的趋势主要如下。

#### 4.1 攻击方法未来发展趋势

1) 探索对复杂任务、复杂模型的被动攻击方法。模型反转攻击属于黑盒被动攻击,半诚实服务器在不了解客户端模型信息的情况下逆推客户端模型并重建原始数据。目前的模型反转攻击方法 Unsplit 和 PCAT 仅能在底层模型浅和学习任务简单的情况下以较高精度重建原始数据,而现实情况中的目标任务可能更复杂、模型可能更庞大,因此未来的攻击方法应着重于复杂模型和复杂任务的研究。

2) 研究新型主动攻击方法。已有的主动攻击方法如 FSHA 不受模型深度和任务复杂度的限制,甚至直接对模型本身造成危害。针对此类威胁,最有效的防御方法是通过检测是否存在攻击,及早发现问题。已有的方法虽然已经可以有效地检测 FSHA,但随着研究的深入,双方在攻防博弈的过程中会产生更精妙的攻击方法和理论。

#### 4.2 防御方法未来发展趋势

1) 结合分割学习的特性制定防御方法。在分割学习的不同训练阶段,模型的能力不同,容易受到的攻击也不同,如在基于数据分布的标签推断攻击中,梯度在模型训练早期就拥有聚类特性,而激活信息在模型训练晚期才拥有聚类特性,因此需要进一步发掘分割学习的特性帮助制定相应的防御方法,提高防御效果。

2) 在训练的不同节点采用不同的防御方法。由于现有的防御方法无法单独抵抗所有种类的攻击,因此需要结合不同的防御方法构建整体的防御策略。

3) 提升基于噪声添加的防御方法性能。基于噪声的防御方法能够有效应对多种被动攻击。添加的随机噪声通常会降低模型准确率,但是 Ye 等<sup>[23]</sup>提出的确定性的噪声添加方法基本不影响

模型表现,因此,如何在保持模型准确率的前提下设计出有效的噪声添加策略是值得研究的问题。

4) 探索基于密码学加密算法的防御方法。同态加密和多方安全计算等加密算法已经广泛应用于隐私保护机器学习中,其可证明的安全性与分割学习不可证明的安全性相比有天然的优势,但是计算效率低仍难以在实际中应用,结合分割学习和密码学算法可以在保证安全性的基础上减少加密运算的计算量和计算时延。目前仅有少量工作利用密码学加密算法保护分割学习,如何高效地结合二者仍需要进一步研究。

5) 加强防御方法验证。已有的防御方法,如基于二值化和数据混合的防御方法,仅在 MNIST、CIFAR-10 和 CIFAR-100 等简单数据集上验证了其防御效果,因此还需要在更复杂的任务上验证其准确率和真实攻击下的防御效果。

6) 结合实际应用场景。随着边缘计算和纵向联邦学习在工业领域的发展,分割学习慢慢成为工业领域关注的热点,但分割学习仍处于研究早期。工业领域庞大而复杂,有各种不同的应用场景,需要在不同的威胁模型下研究分割学习的防御方法。

## 5 结束语

分割学习在机器学习中保护数据隐私安全和解决算力限制等方面的优秀性能使其成为工业领域关注的热点。本文对分割学习数据隐私问题的研究做了归纳和总结,归纳总结了分割学习的定义、原理、分类、存在的攻击方法和相应的防御方法,并展望了分割学习数据隐私研究的未来发展趋势。

#### 参考文献:

- [1] KONEČNÝ J, MCMAHAN H B, YU F X, et al. Federated learning: strategies for improving communication efficiency[EB]. arXiv Preprint arXiv: 1610.05492, 2016.
- [2] VEPAKOMMA P, GUPTA O, SWEDISH T, et al. Split learning for health: distributed deep learning without sharing raw patient data[EB]. arXiv preprint arXiv: 1812.00564, 2018.
- [3] POIROT M G, VEPAKOMMA P, CHANG K, et al. Split Learning for collaborative deep learning in healthcare[EB]. arXiv Preprint arXiv: 1912.12115, 2019.

- [4] HA Y J, LEE G S, YOO M, et al. Feasibility study of multi-site split learning for privacy-preserving medical systems under data imbalance constraints in COVID-19, X-ray, and cholesterol dataset[J]. *Scientific Reports*, 2022, 12: 1534.
- [5] ROTH H R, HATAMIZADEH A, XU Z Y, et al. Split-U-Net: preventing data leakage in split learning for collaborative multi-modal brain tumor segmentation[C]//*Proceedings of International Workshop on Distributed, Collaborative, and Federated Learning, Workshop on Affordable Healthcare and AI for Resource Diverse Global Health*. Cham: Springer, 2022: 47-57.
- [6] PARK S, KIM G, KIM J, et al. Federated split vision transformer for COVID-19 CXR diagnosis using task-agnostic training[EB]. *arXiv Preprint arXiv*: 2111.01338, 2021.
- [7] OH S, PARK J, BAEK S, et al. Differentially private CutMix for split learning with vision transformer[EB]. *arXiv Preprint arXiv*: 2210.15986, 2022.
- [8] SHAN C Q, JIAO H Y, FU J. Towards representation identical privacy-preserving graph neural network via split learning[EB]. *arXiv Preprint arXiv*: 2107.05917, 2021.
- [9] JIANG L L, WANG Y X, ZHENG W Y, et al. LSTMSPLIT: effective split learning based LSTM on sequential time-series data[EB]. *arXiv Preprint arXiv*: 2203.04305, 2022.
- [10] 陈晋音, 李荣昌, 黄国瀚, 等. 纵向联邦学习方法及其隐私和安全综述[J]. *网络与信息安全学报*, 2023, 9(2): 1-20.  
CHEN J Y, LI R C, HUANG G H, et al. Survey on vertical federated learning: algorithm, privacy and security[J]. *Chinese Journal of Network and Information Security*, 2023, 9(2): 1-20.
- [11] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning[J]. *ACM Transactions on Intelligent Systems and Technology*, 2019, 10(2): 1-19.
- [12] VEPAKOMMA P, SWEDISH T, RASKAR R, et al. No Peek: a survey of private distributed deep learning[EB]. *arXiv Preprint arXiv*: 1812.03288, 2018.
- [13] THAPA C, CHAMIKARA M A P, CAMTEPE S A. Advancements of federated learning towards privacy preservation: from federated learning to split learning[M]. REHMAN MHU, GABER MM. *Federated Learning Systems*. Cham: Springer, 2021.
- [14] THAPA C, CAMTEPE S. Precision health data: requirements, challenges and existing techniques for data security and privacy[J]. *Computers in Biology and Medicine*, 2021, 129: 104130.
- [15] USYNIN D, ZILLER A, MAKOWSKI M, et al. Adversarial interference and its mitigations in privacy-preserving collaborative machine learning[J]. *Nature Machine Intelligence*, 2021, 3: 749-758.
- [16] MA C, LI J, WEI K, et al. Trusted AI in multi-agent systems: an overview of privacy and security for distributed learning[J]. *arXiv Preprint arXiv*: 2202.09027, 2022.
- [17] SOYKAN EU, KARAÇAY L, KARAKOÇ F, et al. A survey and guideline on privacy enhancing technologies for collaborative machine learning[J]. *IEEE Access*, 2022, 10: 97495-97519.
- [18] VEPAKOMMA P, RASKAR R. Split learning: a resource efficient model and data parallel approach for distributed deep learning[M]. LUDWIG H, BARACALDO N. *Federated Learning*. Cham: Springer, 2022.
- [19] THAPA C, MAHAWAGA ARACHCHIGE P C, CAMTEPE S, et al. SplitFed: when federated learning meets split learning[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(8): 8485-8493.
- [20] CHOPRA A, SAHU S K, SINGH A, et al. AdaSplit: adaptive trade-offs for resource-constrained distributed deep learning[EB]. *arXiv Preprint arXiv*: 2112.01637, 2021.
- [21] ERDOĞAN E, KÜPÇÜ A, ERCÜMENT ÇİÇEK A. UnSplit: data-oblivious model inversion, model stealing, and label inference attacks against split learning[C]//*Proceedings of the 21st Workshop on Privacy in the Electronic Society*. New York: ACM, 2022: 115-124.
- [22] GAO X, ZHANG L. PCAT: functionality and data stealing from split learning by pseudo-client attack[C]//*Proceedings of 32nd USENIX Security Symposium(USENIX Security)*. [S.l.:s.n.], 2023.
- [23] YE P, JIANG Z F, WANG W, et al. Feature reconstruction attacks and countermeasures of DNN training in vertical federated learning[EB]. *arXiv Preprint arXiv*: 2210.06771, 2023.
- [24] PASQUINI D, ATENIESE G, BERNASCHI M. Unleashing the tiger: inference attacks on split learning[C]//*Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. New York: ACM, 2021.
- [25] ZHENG F, CHEN C C, YAO B H, et al. Making split learning resilient to label leakage by potential energy loss[EB]. *arXiv preprint arXiv*: 2210.09617, 2022.
- [26] LI O, SUN J K, YANG X, et al. Label leakage and protection in two-party split learning[EB]. *arXiv Preprint arXiv*: 2102.08504, 2021.
- [27] LIU J L, LYU X C, CUI Q M, et al. Similarity-based label inference attack against training and inference of split learning[EB]. *arXiv Preprint arXiv*: 2203.05222, 2022.
- [28] FU C, ZHANG X, JI S, et al. Label inference attacks against vertical federated learning[C]//*Proceedings of 31st USENIX Security Symposium(USENIX Security)*. Boston, [S.n.], 2022.
- [29] YUAN S, XUE M, WANG K, et al. Deep gradient attacks with strong DP-SGD lower bounds for label privacy[C]//*Proceedings of International Conference on Learning Representations Workshop(ICLRW)*. [S.l.:s.n.], 2021.
- [30] KARIYAPPA S, QURESHI M K. ExPloit: extracting private labels in split learning[C]//*Proceedings of the 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. Piscataway: IEEE Press, 2023: 165-175.
- [31] RUDIN L I, OSHER S, FATEMI E. Nonlinear total variation based noise removal algorithms[J]. *Physica D*, 1992, 60: 259-268.
- [32] ABUADDBA S, KIM K, KIM M, et al. Can we use split learning on 1D CNN models for privacy preserving training[C]//*Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*. New York: ACM, 2020.
- [33] LI J T, RAKIN A S, CHEN X, et al. ResSFL: a resistance transfer

- framework for defending model inversion attack in split federated learning[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 10184-10192.
- [34] XIAO D Y, YANG C G, WU W G. Mixing activations and labels in distributed training for split learning[J]. IEEE Transactions on Parallel and Distributed Systems, 2022, 33(11): 3165-3177.
- [35] VEPAKOMMA P, SINGH A, GUPTA O, et al. NoPeek: information leakage reduction to share activations in distributed deep learning[C]//Proceedings of the 2020 International Conference on Data Mining Workshops (ICDMW). Piscataway: IEEE Press, 2020: 933-942.
- [36] VEPAKOMMA P, GUPTA O, DUBEY A, et al. Reducing leakage in distributed deep learning for sensitive health data[EB]. arXiv Preprint arXiv: 1812.00564, 2019.
- [37] VEPAKOMMA P, SINGH A, ZHANG E, et al. NoPeek-Infer: preventing face reconstruction attacks in distributed inference after on-premise training[C]//Proceedings of the 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). Piscataway: IEEE Press, 2021: 1-8.
- [38] PHAM N D, ABUADBBA A, GAO Y, et al. Binarizing split learning for data privacy enhancement and computation reduction[EB]. arXiv Preprint arXiv: 2206.04864, 2022.
- [39] QIU P Y, ZHANG X H, JI S L, et al. HashVFL: defending against data reconstruction attacks in vertical federated learning[EB]. arXiv Preprint arXiv: 2212.00325, 2022.
- [40] ZHOU J, ZHENG L F, CHEN C C, et al. Toward scalable and privacy-preserving deep neural network via algorithmic-cryptographic co-design[J]. ACM Transactions on Intelligent Systems and Technology, 2022, 13(4): 1-21.
- [41] DEMMLER D, SCHNEIDER T, ZOHNER M. ABY-A framework for efficient mixed-protocol secure two-party computation[C]//Proceedings of 2015 Network and Distributed System Security Symposium. Reston, VA: Internet Society, 2015.
- [42] ERDOGAN E, KÜPÇÜ A, CICEK A E. SplitGuard: detecting and mitigating training-hijacking attacks in split learning[C]//Proceedings of the 21st Workshop on Privacy in the Electronic Society. New York: ACM, 2022.
- [43] WANG Y L, LV Q Z, ZHANG H, et al. Beyond model splitting: preventing label inference attacks in vertical federated learning with dispersed training[J]. World Wide Web, 2023, 26(5): 2691-2707.
- [44] YANG X, SUN J K, YAO Y S, et al. Differentially private label protection in split learning[EB]. arXiv Preprint arXiv: 2203.02073, 2022.
- [45] 谭作文, 张连福. 机器学习隐私保护研究综述[J]. 软件学报, 2020, 31(7): 2127-2156.
- TAN Z W, ZHANG L F. Survey on privacy preserving techniques for machine learning[J]. Journal of Software, 2020, 31(7): 2127-2156.
- [46] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[M]. Theory of Cryptography. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.
- [47] DOLEV D, YAO A. On the security of public key protocols[J]. IEEE Transactions on Information Theory, 1983, 29(2): 198-208.
- [48] 苏冠通, 徐茂桐. 安全多方计算技术与应用综述[J]. 信息通信技术与政策, 2019, (5): 19-22.
- SU G T, XU M T. Survey of secure multiparty computing technology and application[J]. Information and Communication Technology and Policy, 2019, (5): 19-22.
- [49] GENTRY C. Fully homomorphic encryption using ideal lattices [C]//Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing(STOC). Bethesda, [s.n.], 2009.
- [50] BONEH D, SAHAI A, WATERS B. Functional encryption: definitions and challenges[M]. ISHAI Y, ed. Theory of Cryptography. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.
- [51] WANG R, KARIMI A H, GHODSI A. Distance correlation autoencoder[C]//Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2018: 1-8.
- [52] SENIN P. Dynamic time warping algorithm review[J]. Information and Computer Science Department University of Hawaii at Manoa Honolulu, 2008, 855(1-23): 40.
- [53] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB]. arXiv Preprint arXiv: 2010.11929, 2020.
- [54] YUN S, HAN D, CHUN S, et al. CutMix: regularization strategy to train strong classifiers with localizable features[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 6022-6031.
- [55] MOHASSEL P, ZHANG Y P. SecureML: a system for scalable privacy-preserving machine learning[C]//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 19-38.
- [56] XU R H, JOSHI J, LI C. NN-EMD: efficiently training neural networks using encrypted multi-sourced datasets[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(4): 2807-2820.
- [57] GAWRON G, STUBBINGS P. Feature space hijacking attacks against differentially private split learning[EB]. arXiv Preprint arXiv: 2201.04018, 2022.
- [58] FU J Y, MA X J, ZHU B B, et al. Focusing on pinocchio's nose: a gradients scrutinizer to thwart split-learning hijacking attacks using intrinsic attributes[C]//Proceedings of 2023 Network and Distributed System Security Symposium. Reston, VA: Internet Society, 2023.
- [59] BERNSTEIN J, WANG Y X, AZIZZADENESHELI K, et al. SignSGD: compressed optimisation for non-convex problems [EB]. arXiv Preprint arXiv: 1802.04434, 2018.

## [作者简介]



秦轶群（1997- ），男，辽宁丹东人，华中科技大学硕士生，主要研究方向为人工智能安全、隐私保护。



胡平一（1997- ），男，湖北武汉人，华中科技大学博士生，主要研究方向为分布式机器学习安全、机器学习隐私保护。



马晓静（1983- ），女，安徽金寨人，博士，华中科技大学副教授、博士生导师，主要研究方向为多媒体安全、人工智能安全和隐私保护。



徐鹏（1981- ），男，湖北武汉人，华中科技大学教授、博士生导师，主要研究方向为格密码学、可搜索加密、人工智能安全。



付佳韵（1997- ），女，湖北武汉人，华中科技大学博士生，主要研究方向为隐私保护的智能数据处理技术、人工智能安全。



金海（1966- ），男，上海人，博士，华中科技大学教授、博士生导师，主要研究方向为计算机系统结构、虚拟化技术、集群计算、网格计算、并行与分布式计算、对等计算普适计算、语义网、存储与安全。