https://crad.ict.ac.cn
63(3): 710−734, 2026

拆分学习系统的隐私攻击和防御技术综述

刘雅欣[1,2] 王 鹃[1,2] 杨梦达[1,2] 李子昂[1,2] 易文哲[1,2] 徐枭洋[1,2] 陈美会[3] 韦 韬[3]
[1](武汉大学国家网络安全学院 武汉 430040)
[2](空天信息安全与可信计算教育部重点实验室(武汉大学) 武汉 430040)
[3](蚂蚁集团平台技术事业群 杭州 310023)
(yaxin.liu@whu.edu.cn)

Survey of Privacy Attack and Defense Techniques for Split Learning Systems

Liu Yaxin[1,2], Wang Juan[1,2], Yang Mengda[1,2], Li Zi'ang[1,2], Yi Wenzhe[1,2], Xu Xiaoyang[1,2], Chen Meihui[3], and Wei Tao[3]
[1](*School of Cyber Science and Engineer*, *Wuhan University*, *Wuhan* 430040)
[2](*Key Laboratory of Aerospace Information Security and Trusted Computing* (*Wuhan University*), *Ministry of Education*, *Wuhan* 430040)
[3](*Platform Technology Business Group*, *Ant Groupy*, *Hangzhou* 310023)

Abstract Split learning is an emerging distributed learning technique, whose main idea is to split the complete machine learning model and deploy it on the client and server respectively. During the training and inference process of the system, the client's data are kept locally and only the encoded intermediate features are passed to the server, thus protecting the client's data privacy to a certain extent, while alleviating the computational load of the model's end-side operation. As split learning technology widens its application across various domains, various privacy attacks targeting split learning systems have emerged incessantly. Attackers can leverage intermediate information such as intermediate features and gradients from split layers to reconstruct users' private data or infer their private information, posing a severe threat to data privacy. Currently, academia lacks a systematic and comprehensive overview of research achievements in split learning, with some studies confusing it with federated learning technology or offering insufficiently detailed summaries. Therefore, this paper aims to fill this gap by comprehensively summarizing relevant attack and defense techniques in split learning, providing guidance for subsequent research and development. Firstly, we introduce the definition of split learning technology, its training and inference processes, and outline its various extended architectures. Subsequently, we analyze the threat model of split learning systems and summarize the fundamental concepts, implementation stages, and existing schemes of reconstruction attacks, as well as inference attacks such as attribute inference, membership inference, and label inference targeting split learning systems. Furthermore, we summarize corresponding defense techniques, encompassing methods like anomaly detection, regularization defense, noise addition, adversarial representation training, and feature pruning. Finally, we discuss research challenges and future directions in addressing privacy and security issues in split learning.

Key words split learning; data privacy; reconstruction attack; inference attack; defense method

摘 要 拆分学习是一种新兴的分布式学习技术,其主要思想是将完整的机器学习模型进行拆分,并分别部署于客户端和服务器。在系统的训练和推理过程中,客户端的数据保留在本地,只向服务器传递编码后的中间特征,因此在一定程度上保护了客户端的数据隐私,同时缓解了模型端侧运行的计算负荷。随着拆分学习技术在多个领域的广泛应用,针对拆分学习系统的各种隐私攻击也层出不穷,攻击者能利用中间特征和分割层的梯度等中间信息重构出用户隐私数据或者推断出其隐私信息,严重危及数据的隐私

收稿日期: 2024-08-29; **修回日期**: 2025-10-21
通信作者: 王鹃(jwang@whu.edu.cn)

性。目前，学术界尚缺乏针对拆分学习研究成果的系统性、全面性综述，部分研究将其与联邦学习技术混淆，或总结不够详尽具体。因此，为填补这一空白，需全面总结拆分学习的相关攻击与防御技术，为后续研究发展提供指导。首先介绍拆分学习技术的定义以及其训练和推理过程，并对其多种扩展架构进行概述。随后分析拆分学习系统的威胁模型，并针对拆分学习系统的重构攻击和属性推理、成员推理、标签推理等推理攻击的基本概念、实施阶段和现有方案进行总结归纳。接着总结相应的防御技术，包括异常检测、正则化防御、噪声混淆、对抗性表征学习、特征裁剪等方法。最后探讨拆分学习中隐私安全问题的研究挑战和未来研究方向。

关键词 拆分学习；数据隐私；重构攻击；推理攻击；防御方法

中图法分类号 TP181；TP309

DOI: 10.7544/issn1000-1239.202440703 **CSTR:** 32373.14.issn1000-1239.202440703

机器学习（machine learning, ML）技术被广泛应用于社会生活的各个领域，其中深度学习（deep learning, DL）技术，尤其是深度神经网络（deep neural network, DNN），因其卓越表现深受欢迎，目前在机器翻译[1]、智慧医疗[2]、智能推荐系统[3-5]和自动驾驶[6-8]等方面均有成熟应用。机器学习模型往往需要大量的训练数据才能达到令人满意的性能，因此传统的机器学习应用通常由服务提供商预先收集用户数据，并上传到云服务器进行训练，随后发布API供用户使用，这种方式被称为机器学习即服务（machine learning as a service, MLaaS）。然而随之而来的2个问题限制了机器学习即服务的进一步推广：由于数据需要频繁地上传至云服务器进行处理，导致了通信开销大，尤其在实时性要求高的应用如自动驾驶中，网络瓶颈可能会严重影响系统性能；同时，随着个人对隐私保护意识的增强，加之多起数据泄露事件和严格的数据保护法规的实施，用户更倾向于在本地保留数据，这限制了数据共享和集中式机器学习的应用。

为了应对以上2个问题，拆分学习（split learning, SL）应运而生。拆分学习技术将一个复杂的机器学习模型拆分并分别部署于客户端和服务器端。客户端通常只执行前半部分浅层网络的简单计算，后半部分网络的计算被云端负载，兼顾客户端计算资源受限的同时，最大限度地利用了云服务器的计算资源。同时，拆分学习系统将客户端的数据保留在本地，通过客户端网络对原始查询数据进行编码，将编码后的中间特征发送给服务器，提供了一定程度的数据隐私保护。

拆分学习自从2018年由Gupta等人[9]提出以来，已经在医疗[10-11]和物联网[12]等领域中得到应用。在近期的工作中，拆分学习也应用于不同种类的神经网络模型架构，如GNN[13]，RNN[14]，Transformer[15]。与此同时，学术界对其鲁棒性和所面临的隐私威胁也进行了相应的研究，编码后的中间特征能在一定程度上保护隐私，然而其也为攻击者提供了新的攻击面，同理从拆分层下发的梯度也蕴含隐私信息。基于此，许多学者研究了针对拆分学习系统的隐私攻击，如数据重构攻击（data reconstruction attack）[16-25]、属性推理攻击（attribute/property inference attack）[26-28]、成员推理攻击（membership inference attack）[29]和标签推理攻击（label inference attack）[30-34]等。具体而言，数据重构攻击旨在近似还原或重构出模型的原始输入，直接暴露了用户的隐私数据。属性推理攻击企图推断出用户数据中与任务无关的某些敏感属性。例如，用户上传照片至用于推断年龄的深度学习模型中，而攻击者可能将用户上传的中间特征用于预测人种等敏感信息。成员推理攻击意图推断某个特定的数据样本是否曾被用作该模型的训练数据。例如，医疗机构使用包含患者医疗记录的数据集来训练疾病诊断模型。如果攻击者成功推断出某个人的数据被用于该模型的训练，那么他们就可能进一步推断出这个人患有某种疾病或其他敏感信息。标签推理攻击的意图是揭露用户隐私数据的敏感标签。例如，在一个在线购物平台中，用户的购买行为被记录并用于训练推荐系统。这些购买行为被标记为不同的标签，如“服装”“电子产品”等。如果攻击者能够获取到这些标签信息，他们可能会利用这些标签来推断用户的购买习惯、兴趣爱好甚至身份信息。这些攻击对用户隐私造成了严重威胁，影响了拆分学习系统的实际部署。相应的，学术界也提出了一系列防御方案，如异常检测[35-36]、正则化防御[37-41]、噪声混淆[30-31,42-45]、对抗性表征学习[46-50]和特征裁剪[23,51]等。

本文对拆分学习系统的隐私威胁进行了详细的

分析,并对目前学术界提出的隐私攻击和防御技术进行了系统的归纳和总结,为以后的研究提供了参考和方向。图 1 展示了拆分学习系统的隐私攻击和防御技术的基本分类。

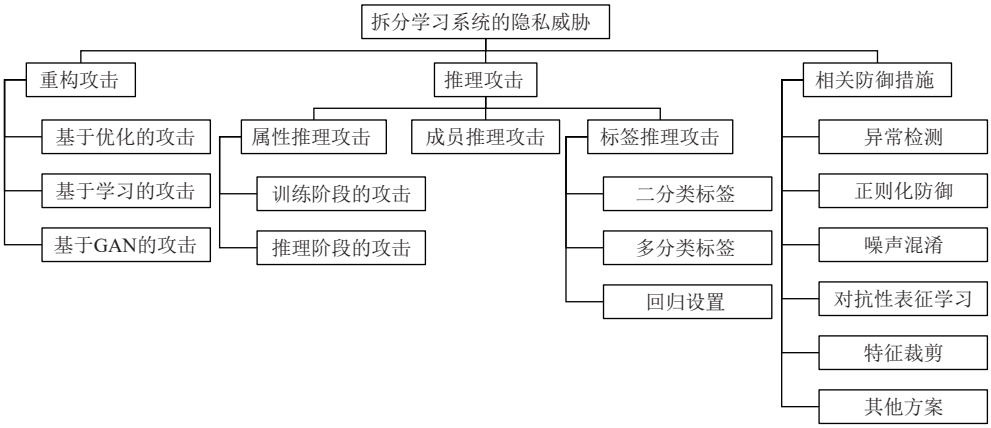


Fig. 1 Classification of privacy attacks and defense techniques for split learning systems

图 1 拆分学习系统的隐私攻击与防御技术分类

1 拆分学习概述

本节首先概述了拆分学习的概念和拆分学习系统的训练及推理过程,然后介绍了拆分学习常用的扩展架构及其特点。此外本节以智慧医疗为例,探讨了拆分学习在实际场景中的应用及其面临的现实威胁和防御技术的实际使用。最后讨论了拆分学习与联邦学习等相似技术的区别和各自的优缺点。

1.1 定 义

1.1.1 拆分学习系统的训练阶段

在单个客户端和单个服务器参与的两方拆分学习系统中,完整的深度神经网络模型M被划分为客户端模型M_{C}和服务器端模型M_{S},如图 2 所示。设有一个含有N个样本的训练数据集$D^{\mathrm{train}}=\{x_i^{\mathrm{train}}, y_i^{\mathrm{train}}\}_{i=1}^{N}$,其中客户端拥有训练样本$x_i^{\mathrm{train}}$,而服务器拥有标签$y_i^{\mathrm{train}}$。在每次训练迭代中,客户端将私有训练数据$x_i^{\mathrm{train}}$输入到$M_{\mathrm{C}}$并计算输出特征值$z=M_{\mathrm{C}}\left(x_i^{\mathrm{train}}\right)$,然后将$z$发送到服务器。服务器$M_{\mathrm{S}}$继续前向传播$z$并计算其损失函数$\mathcal{L}$,随即进行反向传播计算得到梯度。然后,服务器应用基于梯度下降的优化算法更新M_{S},并继续将反向传播到切割层的梯度值g发送到客户端,客户端在本地继续对该梯度反向传播并更新M_{C}。以上述方式持续迭代,直到模型收敛至θ^*。该过程用公式表示为

$$\theta^*=\arg\min_{\theta}\sum_{i=1}^{N}\mathcal{L}\left(y_i^{\mathrm{train}}, M_{\mathrm{S}}\left(M_{\mathrm{C}}\left(x_i^{\mathrm{train}}\right)\right)\right)。\tag{1}$$

在整个训练阶段,客户端保留其私有训练样本和模型在本地,仅共享客户端模型最后 1 层输出的中间特征z。

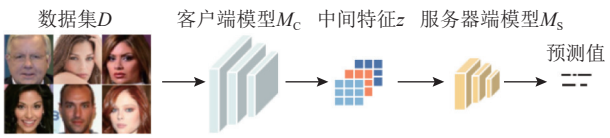


Fig. 2 Split learning system for two-party participants

图 2 两方参与者的拆分学习系统

1.1.2 拆分学习系统的推理阶段

拆分学习模型在完成训练后被服务提供商部署到端云协同推理系统中开放给用户使用,并进入到拆分学习系统的推理阶段。在本文中,将拆分学习系统的训练阶段称为拆分学习,而将拆分学习系统的推理阶段称为拆分推理(split inference, SI),拆分推理直接继承拆分学习所带来的各项优势。

拆分推理仅计算模型的前向传播,在单个客户端的拆分学习系统中,用户使用自己的数据集$D^{\mathrm{test}}=\{x_i^{\mathrm{test}}\}_{i=1}^{N}$进行查询,先由客户端生成中间特征$z^*=M_{\mathrm{C}}\left(x_i^{\mathrm{test}}\right)$,并将中间特征$z^*$传输到服务器。然后,服务器计算推理输出$y_i^{\mathrm{test}}=M_{\mathrm{S}}(z^*)$,并将预期结果$y_i^{\mathrm{test}}$发送回客户端。该过程可以形式化表示为

$$y_i^{\mathrm{test}}=M_{\mathrm{S}}\left(M_{\mathrm{C}}\left(x_i^{\mathrm{test}}\right)\right)。\tag{2}$$

1.2 扩展架构

除图 2 所示的两方参与者的基础架构外,Gupta 等人[9]和 Vepakomma 等人[10]还提出了具有不同特点以及隐私属性的扩展架构。

1)U 型架构

如图 3 所示,机器学习模型被划分为 3 个部分:客户端持有模型的最底层和最顶层部分以及服务器

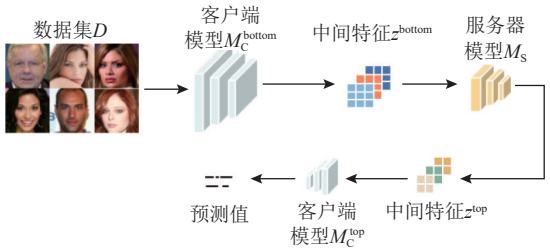


Fig. 3 U-shape architecture

图 3 U 型架构

持有模型的中间部分。在训练和推理过程中，客户端保留其私有训练样本在本地，并且不需要与服务器共享标签数据。U 型架构适用于标签包含了敏感信息的情况，比如说医疗系统数据库中将患者所患的疾病作为标签的情况[11]。

2）水平多客户端架构

在$n(n \geqslant 2)$个客户端参与的协作学习设置下，每个客户端拥有其各自的私人数据集$\{D_1, D_2, \cdots, D_n\}$，通常为同分布数据。假设特征空间为$X$，标签空间为$Y$，样本 ID 空间为$I$，即$D_k = (X_k, Y_k, I_k)$，各参与者所拥有的私有数据集的关系为：

$$X_i = X_j,\ Y_i = Y_j,\ I_i \neq I_j,\ \forall D_i, D_j,\ i \neq j。\tag{3}$$

在多客户端的拆分学习系统中需要实施基于循环顺序的训练协议，如图 4 所示。客户端被放置在 1 个循环列表中，以循环方式与服务器交互。在每一次循环中，仅 1 个客户端使用其本地的私有数据完成 1 轮模型训练任务。该客户端将最新的权重进行共享，而重复该过程的下一个客户端在训练开始之前会先同步最新权重。该过程持续迭代至模型收敛。

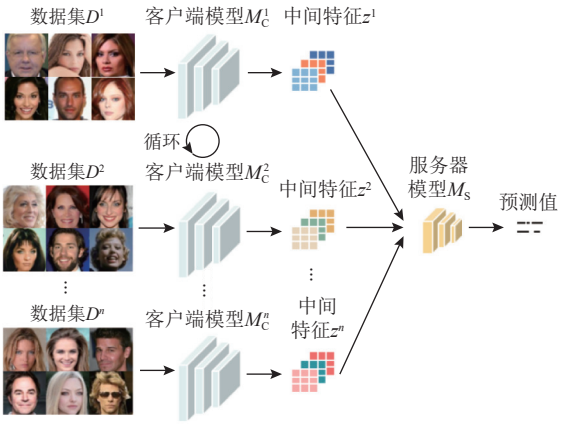


Fig. 4 Horizontol multiple clients architecture

图 4 水平多客户端架构

3）垂直多客户端架构

垂直拆分学习（vertical split learning, VSL）的概念来自垂直联邦学习（vertical federated learning, VFL）和拆分学习，适用于多个参与方的数据集拥有不同的数据特征但样本 ID 相同的场景。假设有$n(n \geqslant 2)$个参与方拥有其各自的私人数据$\{D_1, D_2, \cdots, D_n\}$，特征空间为$X$，标签空间为$Y$，样本 ID 空间为$I$，即$D_k = (X_k, Y_k, I_k)$，各参与者所拥有的私有数据集的关系为：

$$X_i \neq X_j,\ Y_i \neq Y_j,\ I_i = I_j,\ \forall D_i, D_j,\ i \neq j。\tag{4}$$

垂直拆分学习的系统架构如图 5 所示。需要注意的是，服务器拥有可训练的服务端模型。服务器聚合来自不同客户端的中间特征，并得到预测输出。

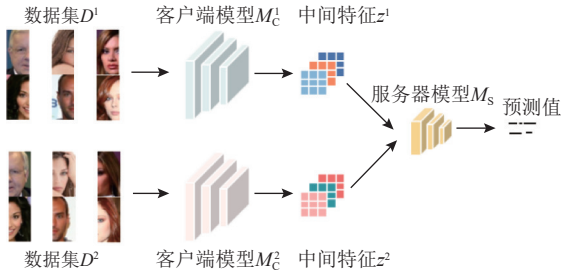


Fig. 5 Vertical multiple clients architecture

图 5 垂直多客户端架构

4）拆分联邦学习

如图 6 所示，与拆分学习相同，拆分联邦学习（split federated learning, SFL）[52]将深度学习模型拆分成客户端模型和服务器端模型，该服务器称为主服务器（main sever）。除主服务器外，拆分联邦学习还引入了一个联邦服务器（fed sever）。所有客户端并行执行其客户端模型的前向传播，并将它们输出的中间特征传递给主服务器。然后主服务器在其服务器端模型上分别处理前向传播和反向传播，并将中间特征的梯度发送给各自的客户端进行反向传播。随后主服务器通过 FedAvg 算法[53]更新其模型。而每个客户端在接收到中间特征的梯度后，在其客户端本地模型上执行反向传播并计算相应梯度，并将梯度发送到联邦服务器。联邦服务器执行 FedAvg 算法，并将聚合后的梯度发送回所有参与的客户端。这种设计避免了每个客户端顺序更新带来的开销。

1.3 实际应用场景

本文以智慧医疗为例，探讨了拆分学习在实际场景中的典型应用，并展示了其攻防技术的具体实践。

在智慧医疗场景中，由于数据资源的限制，某医院需要借助医疗中心提供的专业模型对医疗影像数据进行诊断。然而，患者通常不希望将自己的医疗图像和标签信息公开给不信任的第三方。为此，医院可以与医疗中心达成协议，采用 U 型架构的拆分学习系统以保护原始图像和标签信息。如图 7 所示，

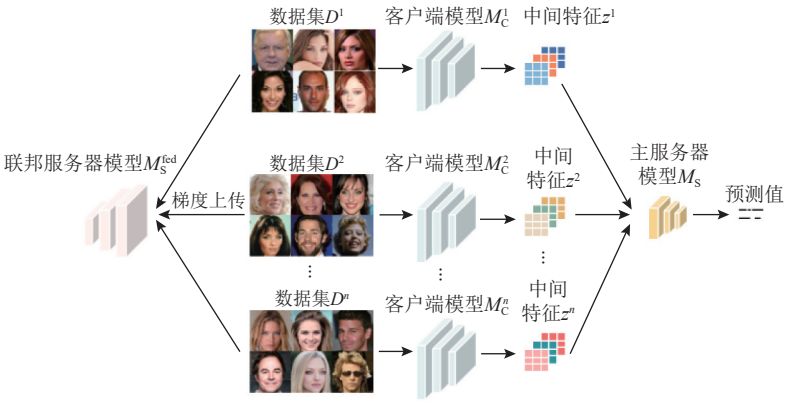


Fig. 6 Split federated learning architecture

图 6 拆分联邦学习架构

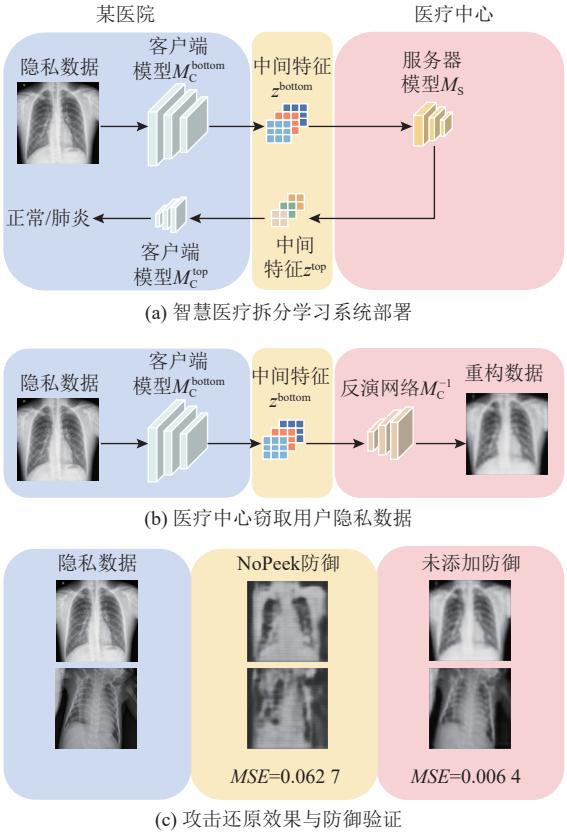


Fig. 7 The application of split learning in smart healthcare

图 7 拆分学习在智慧医疗中的应用

U 型架构将拆分学习模型分为 3 个部分：模型的最底层和最顶层由医院保留在本地，分别记为 M_C^{bottom} 和 M_C^{top}，而模型的中间层由医疗中心持有，记为 M_S。这种配置不需要共享原始数据和标签，与智慧医疗的隐私要求高度契合。

从前向传播的角度来看，M_C^{bottom} 接收原始输入数据，并将其初步编码为中间特征 z^{bottom}。随后，M_S 接收 z^{bottom}，并进行大部分计算，最终将 2 次编码后的中间表示 z^{top} 返回给医院的顶层模型 M_C^{top}。M_C^{top} 则在其接收到的 z^{top} 上执行最终的解码计算。

我们在 Chest X-ray 数据集[54]上执行了 2 分类任务，以模拟真实的智慧医疗场景。Chest X-ray 数据集包含大量的胸部 X 光图像，主要用于肺部疾病的诊断。我们采用了 ResNet-18 模型[55]，并将其第 4 个残差块及其之前的层定义为 M_C^{bottom}，将第 5~8 个残差块定义为 M_S，最后的全连接层则作为 M_C^{top}。

在实际应用中，我们关注拆分学习所面临的攻击与防御问题。假设医疗中心是诚实但好奇的，它试图通过医院上传的 z^{bottom} 来重构患者的医疗图像 x，而重构攻击是拆分学习系统中最大的隐私威胁。医疗机构拥有与医院不重叠的 30% 的同分布辅助数据，通过构建由一系列反卷积层组成的反演网络 M_C^{-1}[22]，将 z^{bottom} 反向映射回隐私数据 $\hat{x}=M_C^{-1}\left(z^{bottom}\right)$。而为了防止患者医疗数据泄露，医院在其模型上叠加了 NoPeek 防御[37]，通过添加额外的距离相关性损失项[56]，在维持模型效用性的同时使中间特征 z^{bottom} 与原始输入数据 x 的距离相关性减小，从而防止攻击者利用 z^{bottom} 重构出隐私数据，拆分学习模型训练的损失函数为：

$$\mathcal{L}=\alpha DCOR\left(x,z^{bottom}\right)+\mathcal{L}_{task}, \tag{5}$$

其中 $DCOR\left(x,z^{bottom}\right)$ 是 NoPeek 防御的距离相关性损失，$\mathcal{L}_{task}$ 则是模型预测任务损失，α 是控制 NoPeek 防御效果的超参数。

如图 7 所示，在没有任何防御措施的情况下，医疗中心能轻易地揭露患者的胸片图像，重构图像与原始图像的均方误差(mean squared error, MSE)。而当客户端 M_C^{bottom} 施加了 NoPeek 防御后，医疗机构只能重构出模糊不清的图像，其 MSE 比未添加防御时的 MSE 明显增加，这种程度的数据不足以让医疗机构揭露出关于患者的其他信息。

1.4 与联邦学习的区别及优势

联邦学习[53,57]与拆分学习[9-10]的应用场景多有交叉，为了避免混淆，本节首先说明了联邦学习的定义，然后将联邦学习和拆分学习进行比较，分析它们的异同之处及各自的优势，并在表1中进行总结。

在联邦学习系统中，有多个客户端和1个服务器。每个客户端在自己的数据上运行整个深度学习模型的副本，计算训练梯度并将梯度发送给服务器。服务器接收来自每个客户端的梯度更新，并在不了解任何参与者的信息的情况下执行安全聚合，随后将梯度的聚合结果发送给各个客户端。客户端使用聚合后的梯度更新本地客户端模型，这个过程一直持续到收敛。

联邦学习和拆分学习都是主流的分布式学习技术，二者的相似之处在于：1)都保留原始隐私数据在本地；2)都适用于多客户端的分布式场景；3) 都具有水平和垂直架构，即横向联邦和水平拆分、纵向联邦和垂直拆分。

联邦学习和拆分学习也存在明显的不同：1)在联邦学习中，客户端和服务器都需要存储完整的模型副本。而在拆分学习中，以切割层为界，客户端存储模型的前半部分，而服务器存储模型的后半部分。2)联邦学习系统通常拥有2个或以上的客户端，而拆分学习存在如图2所示的单客户端情况。3)联邦学习系统的多个客户端可以并行更新，而拆分学习系统的多个客户端需要执行顺序更新，等待开销较大。4)在客户端与服务器通信方面，联邦学习的客户端需要将训练梯度上传给服务器，并接收来自服务器的聚合后的梯度；而拆分学习的客户端需要将样本被编码后的中间特征上传给服务器，并接收服务器下发的切割层梯度。

联邦学习与拆分学习具有不同的优势，本文主要在模型效率和通信开销，以及模型性能方面对二者进行比较：1)在模型效率和通信开销方面，目前已经有研究关于如何拆分模型以在模型效率、通信开销和隐私性之间进行权衡[58-60]。与联邦学习相比，拆分学习显著减少了所需的带宽[61-62]，因为对于大规模深度神经网络，模型中间层的特征始终比整个网络的梯度或权重更加紧凑。此外，拆分学习的客户端的计算负担比联邦学习造成的计算负担更小，因为拆分学习的客户端仅需对网络的一部分而不是对整个网络执行前向或反向传播计算。2)在模型性能方面，Gao等人[63]在物联网设置中提供联邦学习和拆分学习的经验比较，通过实验证明在数据分布不平衡的情况下，拆分学习的学习性能优于联邦学习，但在极端非独立同分布的情况下比联邦学习差。此外，当拆分学习模型可以收敛时，它总是比联邦学习模型收敛得快。

个性化联邦学习(personalized federated learning)是在联邦学习框架基础上发展的一种新技术，旨在解决传统联邦学习方法面临的在高度异构的数据上收敛性差和缺乏个性化解决方案的问题[64]。联邦元学习(federated meta-learning)将联邦学习与元学习算法相结合，为每个客户端训练个性化模型，从而降低模型在不同客户端上的表现差异，提高模型的公平性[65]。实际上，联邦元学习可以视为个性化联邦学习中的一种技术。在这三者之间，可以看作是递进的包含关系，即联邦元学习⊂个性化联邦学习⊂联邦学习。个性化联邦学习和联邦元学习都是学习方法而不是安全框架，本文的重点讨论是拆分学习这种安全框架。

2 拆分学习系统中的隐私攻击技术

本节首先介绍了针对拆分学习系统的隐私威胁模型，然后详细介绍了针对拆分学习系统的各种隐私攻击技术如数据重构攻击、属性推理攻击、成员推理攻击和标签推理攻击等，并在最后进行总结。图8展示了不同攻击的目标以及攻击方法。

2.1 威胁模型分析

现有研究主要集中于服务器接收客户端上传的中间特征，并充当攻击者对客户端隐私数据进行重

Table 1 Comparison of Federated Learning and Split Learning

表1 联邦学习和拆分学习的比较

分布式技术	通信内容		客户端数量	模型架构		是否支持垂直架构	优点
	客户端→服务器	服务器→客户端		客户端	服务器		
联邦学习	模型训练梯度	聚合后的梯度	≥2	完整模型		是	在极端非独立同分布下性能更优
拆分学习	样本编码后的中间特征	切割层梯度	≥1	切割层前半部分模型	切割层后半部分模型	是	通信带宽小、客户端计算负担小、收敛快、在数据分布不平衡情况下性能更优

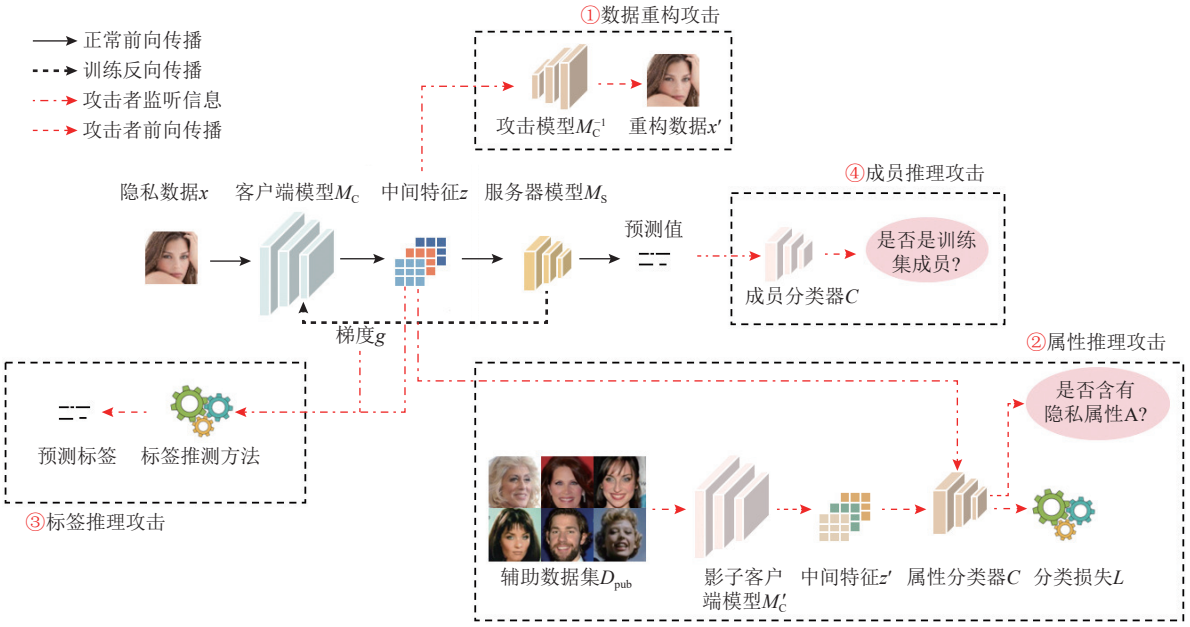


Fig. 8 Privacy attack technologies of split learning systems

图 8 拆分学习系统的隐私攻击技术

构和推理等隐私攻击。为此,本文从 3 个方面对拆分学习系统所面临的威胁模型进行了概括。

1)攻击者的目标

攻击者的目标可以具体分为数据重构、属性推理、成员推理和标签推理 4 类。数据重构旨在重构出客户端所持有的隐私数据样本。属性推理旨在推断出与任务无关的模型训练数据隐私属性,如通过性别分类模型来推断单个样本的人种、年龄等。成员推理旨在推知样本是否在目标模型的训练集中。标签推理旨在推断出客户端特定隐私数据样本所对应的标签。

2)攻击者的可信性

攻击者的可信性是指攻击者是恶意的(malicious)还是半诚实的(honest but curious)。半诚实的攻击者忠实地遵循拆分学习系统的训练或推理协议,不会破坏正常的拆分学习或拆分推理过程,但对系统中的某些隐私信息具有好奇心,被动地通过已知信息来推断隐私。恶意的攻击者会主动干扰系统的正常运行以达成攻击目的。此外,与半诚实的服务器不同,服务器的恶意行为通常很容易被客户端察觉。

3)攻击者的能力

攻击者的能力是指攻击者能获取的目标模型的信息或能对目标模型执行的操作,我们根据攻击者的能力在实际拆分学习系统中的可达性对其进行分类和排序:

L0. 攻击者只能获得基本信息。基本信息是指在拆分学习系统执行流程中服务器能获取的信息,也是进行攻击的必要条件,包括客户端上传的中间特征、反向传播中服务器计算的切割层的梯度和服务器输出的标签种类。

L1. 攻击者能够获取数据分布信息。数据分布信息是指客户端隐私数据集的统计特征。在拆分学习中,客户端不被强制要求向服务器披露其数据分布。然而,考虑到服务器对任务有一定了解,它可能会利用开源数据集来估计这些分布。例如,如果服务器知道客户端拥有人脸数据,它可以利用类似 CelebA[66] 和 FFHQ[67] 的数据集来进行攻击模型的预训练。

L2. 攻击者可以使用随机样本对目标模型进行查询。拆分学习系统部署后,进入了拆分推理阶段。此时,服务器可以伪装成使用 API 的普通用户,向客户端发送查询请求,以获取随机样本对应的中间特征。通过多次随机查询,攻击者能够推知客户端模型的表示偏好等信息,并根据这些信息构建更接近目标客户端的影子模型[22]。

L3. 攻击者能够获取模型架构信息。模型架构信息是指客户端模型的结构和组成部分,包括层的类型、层的顺序和连接方式,以及模型的深度和宽度。一些拆分学习系统服务往往基于经典模型架构,或者明确说明模型架构。服务器可以据此推测切割层之前的客户端模型架构,从而提升对模型敏感的攻击方法的性能[16]。

L4. 攻击者能够获取标签分布信息。标签分布信息是指隐私数据中各个类别标签出现的频率或比例。作为模型的输出结果，标签分布通常不会被直接披露，攻击者只能通过推测，例如猜测类别的平衡性，来大致获取标签分布的信息。

L5. 攻击者能够获取模型参数信息。模型参数信息是指模型经过训练后得到的各种参数值，包括权重和偏置等。这些参数决定了模型的行为和性能，通常不会被公开，并且难以被推测得到。

L6. 攻击者能够实施恶意行为。恶意行为涉及对拆分学习系统的协议、模型和其他组件的主动操作，如修改训练梯度以提高重构攻击效果。然而，这些恶意行为通常会导致拆分学习系统出现异常表现，因而容易被检测到，因此缺乏隐蔽性和实际应用性[17]。

综上所述，在实际拆分学习系统中，L0 和 L1 是合理的假设。L2 仅在拆分推理场景下被认为是合理的假设，而在拆分学习系统的训练阶段，攻击者无法查询目标模型。L3, L4, L5 被认为是强假设，因为在实际中客户端无需向服务器提供这些信息，尽管攻击者可能在某些情况下近似推测或获得这些信息。此外，L6 由于隐蔽性不足而被认为是过度假设。

2.2 重构攻击

本节介绍了重构攻击（reconstruction attack）的定义，并按照攻击手段的差异将攻击方法进行分类，随后对不同类型的攻击进行介绍和分析，并在表 2 中对攻击的实施阶段、攻击手段、方法类别和前提假设进行总结。

2.2.1 重构攻击的定义和分类

拆分学习场景下的重构攻击通常是指在服务器端的攻击者利用中间特征等信息重构出客户端模型的输入数据。重构得到的数据是与输入数据相似的数据样本，也被称为数据重构攻击（data reconstruction attack），在后文中数据重构攻击与重构攻击的概念可以混用。

学术界还有另一类模型反演攻击（model inversion attack）容易与数据重构攻击混淆，二者的主要区别是：数据重构的目标是还原出与目标模型的输入数据完全一致的数据样本；而模型反演的目的是还原出与目标模型的训练数据相同的数据样本。

通常通过计算重构样本和原始输入样本之间的均方误差、峰值信噪比（peak signal-to-noise ratio, PSNR）[68]、结构相似性（structural similarity index, SSIM）[69] 和学习感知图像块相似度（learned perceptual image patch similarity, LPIPS）[70] 等指标来评估重构攻击的效果。

重构攻击是深度学习领域中最严峻的隐私威胁之一。学术界已广泛研究了在传统深度学习领域中多种基于不同原理的重构攻击方案。拆分学习将模型分割部署于客户端与服务器，并在通信中交换中间特征和梯度，引入了新的攻击面，激发了学术界对拆分学习系统重构攻击的探索。本文根据现有攻击方法在攻击手段上的差异性，将针对拆分学习系统的重构攻击分为 3 类：

1）基于优化的攻击。优化导向的攻击方法起始

Table 2 Summary of Reconstruction Attacks

表 2 重构攻击总结

攻击阶段	方法	提出年份	攻击手段	方法分类			攻击假设
				优化	学习	GAN	
训练	UnSplit[16]	2022	交替优化随机样本和模型参数	√			L0，L3
	FSHA[17]	2021	通过虚假梯度劫持目标模型训练过程		√		L0，L1，L6
	PCAT[18]	2023	利用服务器指导伪客户端训练		√		L0，L1
	SDAR[19]	2023	使用判别器指导模拟器和解码器训练		√		L0，L1
	FORA[21]	2024	学习目标模型对中间特征的表示偏好		√		L0，L1
推理	rMLE[22]	2019	优化随机样本	√			L0，L3，L5
	Shadow Model + rMLE[22]	2019	训练影子客户端	√			L0，L1，L3
	LM[23]	2021	优化模型参数	√			L0，L3，L5
	IN[22]	2019	训练反演网络		√		L0，L1，L2
	Ginver[24]	2023	原始中间特征和经重建、编码后的中间特征互为对照		√		L0，L2
	GLASS[25]	2023	优化 StyleGAN 的 $\mathcal{Z}$ 和 $\mathcal{W}^+$ 空间潜编码			√	L0，L1，L3，L4
	GLASS++[25]	2023	学习将中间特征映射到 $\mathcal{W}^+$ 空间潜编码			√	L0，L1，L3，L4

于由随机噪声构成的初始样本,通过直接优化这些噪声或模型参数,逐步引导噪声样本逼近攻击目标,直至成功重构出所需结果。

2)基于学习的攻击。这类攻击通常依赖于辅助数据集来训练一个反演网络[22],该网络旨在模拟客户端模型的逆操作,直接将客户端模型生成的中间特征逆向映射回原始样本空间,从而实现数据重构。

3)基于生成式对抗网络(generative adversarial network, GAN)[71]的攻击。这类攻击借助了GAN模型的生成能力,通过优化GAN模型的潜编码来间接地优化重构样本,最终获得重构结果。

2.2.2 基于优化的重构攻击

根据优化对象的不同,基于优化的重构攻击可划分为样本优化攻击与模型参数优化攻击两大类。由于优化过程往往依赖于梯度下降等反向传播算法,攻击者必须事先了解客户端模型的架构和具体参数细节,这通常需要攻击者掌握L0, L3, L5等层级的信息,这类攻击方法的一个显著优势在于无需辅助数据集即可发动攻击。

1)样本优化攻击

He等人[22]首先基于优化的思想提出了正则化最大似然估计(regularized maximum likelihood estimation, rMLE)攻击,利用客户端模型M_C的参数和中间特征重构出输入数据x_0。He等人[22]的初始重构样本是随机噪声样本x,通过梯度下降逐步对x进行优化。其损失函数由重构数据中间特征$M_C(x)$与隐私数据中间特征$M_C(x_0)$的像素级欧几里得距离(Euclidean distance)以及总变分(total variation)$TV(x)$作为正则项来确保重构图像x满足自然图像的分布[70]:

$$x^* = \arg\min_x (ED(M_C(x_0), M_C(x))) + \lambda TV(x)。 \quad (6)$$

rMLE需要客户端模型M_C的参数来计算梯度对随机样本进行的优化,因此只适用于白盒场景。为了扩展rMLE的适用范围,He等人[22]进一步提出了在无查询场景下的攻击方法,即使用辅助数据集训练一个能模仿客户端行为的影子模型(shadow model)来替代M_C。然而,He等人[22]提出在无查询场景下的攻击要求攻击者知道目标模型的模型架构和输入数据的分布,否则重构图像的质量较差。

2)模型参数优化攻击

Singh等人[23]在rMLE[22]的基础上提出了基于自编码器的极大似然估计(likelihood maximization, LM)攻击。攻击者构建1个自编码器f_θ,其输入是固定的随机噪声ε,输出是重构图像。区别于rMLE通过优化随机样本得到重构结果,在LM中攻击者通过优化自编码器的参数得到重构结果

$$x^* = f_\theta^*(\varepsilon) = \arg\min_\theta ED(M_C(x_0), M_C(f_\theta(\varepsilon)))。 \quad (7)$$

3)混合优化攻击

Erdoğan等人[16]提出了能在拆分学习阶段进行攻击的UnSplit方法,通过交替地优化随机样本$\tilde{x}$和模型参数$\tilde{\theta}$达到重构隐私数据的目的,其优点是可以在十分有限的条件下实现数据重构,即攻击者只知道客户端模型的架构和输入数据的中间特征,除此以外攻击者不知道客户端模型的参数,也不知道任何关于客户端隐私数据集的知识,且不能向客户端发送查询请求。然而在UnSplit攻击中,由于模型参数$\tilde{\theta}$和输入$\tilde{x}$都是未知的,导致UnSplit可能因搜索空间太大而无法收敛。因此,一旦拆分学习系统的任务或客户端模型变得稍微复杂些,UnSplit的性能就会急剧下降或是无法成功攻击[18]。

2.2.3 基于学习的重构攻击

基于学习的重构攻击方法的核心思想是使用辅助数据集训练攻击模型,使其能将客户端输出的中间特征直接映射回输入数据。辅助数据集与隐私数据集同域,能帮助攻击者学习关于隐私数据集的先验信息,并更好地利用这些知识来重构隐私输入。以在重构攻击中常常用到的人脸数据集为例,人脸数据是高度对齐的,如CelebA[66]和FFHQ[67],虽然数据集的分布不同,但数据集中的人脸图像都有眼睛、鼻子、嘴巴等五官,这些信息帮助攻击模型限制了重构图像的范围。

基于学习的重构攻击对攻击者的前提条件设定颇为全面,这些多样化的假设成为了攻击者在不同情境下达成卓越攻击效果的关键因素。它们的共同之处在于,都预设了攻击者能够利用L0与L1层级的条件。依据攻击模型的差异以及攻击者能否实施恶意行为,本文将这些攻击进一步细分为反演网络攻击、恶意梯度攻击以及判别器指导攻击三大类。基于学习的重构攻击方法的一大优势在于其速度快且应用广泛,不仅在拆分学习与拆分推理的场景中得以应用,而且能够适应多种假设条件,因此成为目前应用最为普遍的一种攻击手段。

1)反演网络攻击

He等人[22]提出的反演网络(inverse-network, IN)攻击方法是基于学习的重构攻击方法的基础。他们使用辅助数据集训练第2个神经网络M_C^{-1}作为客户端模型M_C的逆,并通过M_C^{-1}将中间特征映射回输入数据$\hat{x} = M_C^{-1}(M_C(z))$。客户端模型$M_C$通常由卷积层构

成，其近似逆M_{C}^{-1}可考虑由反卷积网络组成。M_{C}^{-1}的训练过程如下所示：攻击者首先使用辅助数据集的样本x_{aux}查询M_{C}，并将得到的对应的中间特征z_{aux}作为输入馈送到M_{C}^{-1}；训练M_{C}^{-1}生成重构数据$\hat{x}_{\mathrm{aux}}=M_{\mathrm{C}}^{-1}(M_{\mathrm{C}}(z_{\mathrm{aux}}))$；通过优化$M_{\mathrm{C}}^{-1}$的参数，最小化原始输入数据$x_{\mathrm{aux}}$与重构数据$\hat{x}_{\mathrm{aux}}$间的欧几里得距离

$$(M_{\mathrm{C}}^{-1})^{*}=\arg\min_{\theta_{M_{\mathrm{C}}^{-1}}} ED\left(M_{\mathrm{C}}^{-1}(M_{\mathrm{C}}(z_{\mathrm{aux}})),\ x_{\mathrm{aux}}\right)。\tag{8}$$

鉴于IN攻击的效果很大程度上取决于辅助数据集的质量，为了降低对这种数据集的依赖，Yin等人[24]与Gao等人[18]在原有基础上进行了相应的改进。Yin等人[24]提出Ginver攻击，建议攻击者收集客户端上传的中间特征Z作为训练数据集，并利用自然进化策略（natural evolution strategies, NES）技术[72]在不知道损失函数的具体参数和结构的情况下来训练反演模型M_{C}^{-1}。使用得到的恢复图像$\hat{x}=M_{\mathrm{C}}^{-1}(z)$再对客户端模型$M_{\mathrm{C}}$进行查询，得到新的1组中间特征$\hat{z}=M_{\mathrm{C}}(\hat{x})$与原始中间特征$z$做对比来更新$M_{\mathrm{C}}^{-1}$，以减少对辅助数据集的依赖。Gao等人[18]采用另一个思路，他们提出的PCAT攻击利用训练过程中处于不同迭代轮次的服务器模型所学习到的知识来指导伪客户端模型$\tilde{M}_{\mathrm{C}}$训练。每当M_{C}上传中间特征与标签到服务器后，服务器从本地数据集中挑选出与客户端上传的中间特征标签相同的数据输入$\tilde{M}_{\mathrm{C}}$以完成前向传播，然后分别为M_{C}和$\tilde{M}_{\mathrm{C}}$计算更新梯度以更新模型。一旦$\tilde{M}_{\mathrm{C}}$学习到从输入到中间特征的特征空间的映射，服务器可以通过反演网络方法[22]得到重构结果。PCAT仅需要少量隐私训练数据就能实施重构攻击。

2）恶意梯度攻击

Pasquini等人[17]发现并利用了拆分学习协议的一个关键漏洞：服务器计算的损失函数与梯度对客户端而言是不透明的，这一特性使得恶意服务器能够主动干预分布式模型的学习过程，通过伪造合适的梯度，迫使客户端模型M_{C}收敛至攻击者指定的任意目标函数。针对此漏洞，他们提出了一种针对拆分学习阶段的主动攻击方案——特征空间劫持攻击（feature-space hijacking attack, FSHA）。

在FSHA攻击中，攻击者需要与客户端训练数据x_{priv}同域的辅助数据集x_{pub}，并训练3个网络：$\tilde{M}_{\mathrm{C}}$，$\tilde{M}_{\mathrm{C}}^{-1}$和$D$。其中，$\tilde{M}_{\mathrm{C}}$是为客户端网络$M_{\mathrm{C}}$动态定义目标特征空间$\tilde{Z}$的先导网络，$\tilde{Z}$空间更易于攻击者实现重构攻击。$\tilde{M}_{\mathrm{C}}^{-1}$是一个反演网络，用于从中间特征恢复输入数据。判别器D被训练来区分由M_{C}映射的特征空间Z和由$\tilde{M}_{\mathrm{C}}$诱导的中间特征空间$\tilde{Z}$。在攻击过程，判别器D替代服务器M_{S}计算并下发梯度给客户端，从而引导M_{C}学习隐私训练数据到由$\tilde{M}_{\mathrm{C}}$定义的$\tilde{Z}$空间的映射关系。攻击模型各组件的损失函数为：

$$\mathcal{L}_{\tilde{M}_{\mathrm{C}},\tilde{M}_{\mathrm{C}}^{-1}}=MSE\left(\tilde{M}_{\mathrm{C}}^{-1}\left(\tilde{M}_{\mathrm{C}}\left(x_{\mathrm{pub}}\right)\right),x_{\mathrm{pub}}\right),\tag{9}$$

$$\mathcal{L}_{D}=\mathrm{lb}\left(1-D\left(\tilde{M}_{\mathrm{C}}\left(x_{\mathrm{pub}}\right)\right)\right)+\mathrm{lb}\left(D\left(M_{\mathrm{C}}\left(x_{\mathrm{priv}}\right)\right)\right),\tag{10}$$

$$\mathcal{L}_{M_{\mathrm{C}}}=\mathrm{lb}\left(1-D\left(M_{\mathrm{C}}\left(x_{\mathrm{priv}}\right)\right)\right)。\tag{11}$$

然而由于FSHA服务器完全破坏了拆分学习的过程，使得客户端可以检测到这种攻击[35]。

3）判别器指导攻击

受FSHA的启发，后续研究也纷纷引入了判别器，在此框架下，判别器与生成器形成了一种对抗关系：判别器致力于将生成器产出的假数据与真实的隐私数据区分开来，而生成器则竭力迷惑判别器。通过持续的对抗训练，生成器能够学习到判别器所识别的特征，进而使其生成的数据愈发逼真。

Zhu等人[19]提出的SDAR攻击中引入了2个判别器d_1和d_2，判别器d_1用于区分客户端模型f和模拟器$\tilde{f}$生成的中间特征，判别器d_2用于区分由解码器$\tilde{f}^{-1}$生成的解码图片和辅助数据集中的真实图片。Xu等人[21]提出的FORA攻击在训练阶段使目标客户端和伪客户端同步进行训练，并引入判别器损失和多核最大平均差异（multi-kernel maximum mean discrepancy, MK-MMD）损失[73-74]以计算隐私中间特征和伪客户端产生的中间特征的差异，前者迫使伪客户端和目标客户端生成的中间特征不可区分，后者将2组特征对齐到一个共享空间并缩小其中的差异。其核心思想是使攻击者构建的伪客户端模型学习到真实客户端模型对中间特征的表示偏好以便于更容易地对中间特征进行重建。

2.2.4 基于GAN的重构攻击

基于GAN的重构攻击方法将GAN技术[71]应用于攻击架构之中。首先在合适的辅助数据集上训练GAN模型，提取辅助数据集上的先验知识来规范生成的重构图像。随后如式（12）所示，通过优化GAN的潜编码（latent code）z间接地优化重构图像$\hat{x}=GAN(z)$，最小化重构图像$\hat{x}$和隐私数据x间的MSE，得到最终的重构结果。

$$z^{*}=\arg\min_{z} MSE(M_{\mathrm{C}}(x),M_{\mathrm{C}}(GAN(z)))。\tag{12}$$

基于GAN的重构攻击是目前最先进的攻击手段，凭借GAN模型的强大的生成能力，可以得到高保真度的重构图像。

Li等人[25]将StyleGAN[75]用于重构攻击，提出了

一种基于 GAN 的潜编码空间搜索攻击(GAN-based latent space search attack, GLASS)以及增强的 GLASS++攻击来提高重构稳定性。Li 等人[25]利用了 StyleGAN 特有的$\mathcal{W}$和$\mathcal{W}^+$潜编码空间[76]来实现高质量的重构。对于 GLASS, 攻击者在第 1 步时通过优化$\mathcal{Z}$空间的潜编码z引入较大的抖动, 防止优化陷入局部最优。在第 2 步中, 攻击者将潜编码z通过 StyleGAN 的映射网络f得到$\mathcal{W}^+$空间上的潜编码w^+, 利用$\mathcal{W}^+$空间中高度解纠缠的特点所提供的良好的可编辑性[67,75,77]以及优化潜编码w^+提高重构质量。GLASS++则引入了一个编码器E学习从客户端模型M_C生成的中间特征到$\mathcal{W}^+$空间潜编码的映射。

2.3 推理攻击

在拆分系统中的推理攻击包括属性推理攻击(attribute/property inference attack)、成员推理攻击(membership inference attack)和标签推理攻击(label inference attack)。属性推理攻击旨在利用客户端上传的中间特征推断出与主任务无关的信息, 揭露客户端私有数据隐私。成员推理攻击根据目标模型输出的置信分等信息判断特定样本是否是隐私训练数据集的成员。标签推理攻击利用服务器下发的梯度等信息推断出样本所属的标签。在拆分场景中, 推理攻击常常作为重建攻击的下游任务出现。

2.3.1 属性推理攻击

在拆分学习系统中, 属性推理攻击的攻击目标通常是推断出单个样本的某些与任务无关的隐私属性, 如在一个人脸数据集上训练的模型原本用于预测人像的性别, 而在服务器端的攻击者试图通过客户端上传的中间特征揭露人像的种族、年龄、肤色等敏感信息。属性推断攻击可以发生在训练或是推理阶段。表 3 中总结了属性推理攻击的实施阶段、攻击手段和攻击假设等。

Table 3 Summary of Attribute Inference Attacks

表 3 属性推理攻击总结

攻击阶段	方法	攻击手段	攻击假设
训练	文献 [26]	通过 FedAvg 算法构建客户端影子模型	L0, L1, L3
	文献 [17]	通过虚假梯度劫持目标模型训练过程	L0, L1, L6
推理	文献 [27]	使用辅助数据中间特征训练敏感属性分类器	L0, L1, L2
	文献 [28]	将混淆的中间特征解码为影子模型内部纯净特征	L0, L1, L2

1)训练阶段的属性推理攻击

Dougherty 等人[26]提出在攻击者拥有辅助数据和目标客户端模型架构知识的情况下, 可以通过构建阴影模型来模仿目标客户端的行为, 并利用该阴影模型生成的辅助数据集的中间特征训练属性分类器。在目标模型的训练阶段, 攻击者构建K个影子模型$\widehat{\mathcal{F}}_C^i$, 每个影子模型在不同的辅助数据集上训练。在反向传播时冻结服务器模型参数, 只更新阴影模型, 最后使用 FedAvg 算法[53]对影子模型的参数进行加权聚合得到最终的聚合阴影模型$\bar{\mathcal{F}}_C$, 再继续训练属性分类器进行攻击。

Pasquini 等人[17]提出的 FSHA 攻击除了可用于前文介绍过的数据重构, 也可以用于推断单个样本的特定隐私属性。在属性推理攻击中, 攻击者可以通过将伪客户端模型$\tilde{M}_C$与敏感属性分类器C共同训练来调节特征空间$\tilde{Z}$, 强制中间特征泄露有关特定属性的信息。区别于重构攻击只要求无标签的辅助数据, 在属性推理中, 攻击者要求有标签的辅助数据对以便对敏感属性分类器C进行有监督训练。

2)推理阶段的属性推理攻击

Song 等人[27]利用了机器学习模型过度学习(overlearning)的特点实施属性推理攻击。攻击者使用辅助数据查询客户端模型M_C得到其中间特征, 再输入给攻击者自己构建的敏感属性分类器C学习中间特征到敏感属性的映射关系。考虑到拆分场景中, 为保护中间特征的隐私性, 经常会施加一些防御手段, Song 等人[27]还提出了一种防御破除方法, 攻击者构建未施加防御的影子模型以生成纯净的中间特征, 并训练一个防御去除模型T, 使经过T处理后的目标模型与影子模型生成的中间特征尽可能接近, 从而去除施加在隐私中间特征上的防御。

Wang 等人[28]认为不同分类器内的特征具有显著的分布差异, 训练防御去除模型T对不同分布下的混淆后的中间特征进行变换, 可能会导致T陷入局部最优, 导致最终攻击性能较差, 因此需要对 Song 等人[27]提出的防御破除方法进行改进。Wang 等人[28]通过观察发现基于对抗训练[78]和基于信息理论的防御[79-80]通常只作用于模型最后 1 层输出的特征上, 而对于模型前面层输出的内部特征的保护效果则一般。基于此他们提出的 AttrLeaks 攻击构建了 1 个有l层的卷积神经网络(convolutional neural networks, CNN)$\tilde{M}_C$作为客户端模型的替代, 其中$\tilde{M}_C$最后 1 层产生的特征$\hat{z}$具有与客户端上传的中间特征z相同的维度, 而在$\tilde{M}_C$的$l-n(n<l)$层输出的内部特征记为z_{aux}, 可近似看成未混淆的中间特征。

2.3.2 成员推理攻击

成员推理攻击只发生在拆分学习系统的推理阶

段，攻击者拥有部分测试样本，旨在判断这些样本是否属于目标模型的训练集，即成员还是非成员。在传统场景下的成员推理攻击，攻击者依靠置信分、标签等信息训练 2 分类模型，从而判断待测样本是否为训练集[81-86]，通常是通过训练一系列影子模型来模仿目标模型的输出，并构建 1 个 2 分类器以判别目标模型对成员样本和非成员样本的预测结果的差异。这种攻击会造成间接的用户隐私泄露风险，比如在医疗领域，如果攻击者知道受害者的数据被用于训练某种医疗诊断分类器，那么攻击者就可以直接推断受害者的健康状态。除了破坏受害者隐私之外，因为收集和标记训练数据集需要消耗大量资源，成员推理攻击还会严重侵害模型供应商的私有训练数据的知识产权。

然而目前鲜有工作讨论针对拆分学习框架的成员推理攻击，仅有 Chen 等人[29]对工业物联网（industrial Internet of things, IIoT）场景中的拆分推理系统进行了研究。他们认为现有的传统场景下的成员推理攻击要么假定攻击者可以查询系统的 API，要么假定可以获得目标模型的参数和目标训练数据集分布，这在 IIoT 场景下是不实际的。因此，Chen 等人[29]设计的攻击流程比传统场景下的属性推理攻击多了 1 个影子模型构建步骤。攻击者采用 transfer-inherit 算法，利用了迁移学习的思想使用辅助数据微调 1 个公共发布的预训练好的网络，以代替客户端模型。随后该伪客户端模型和服务器模型联合成为影子模型 M_{shadow}，以产生置信分来训练 2 分类器并判断特定样本是否为训练集成员。

2.3.3 标签推理攻击

标签推理攻击指非标签方意图提取标签所有者的私有标签。在拆分学习系统中，标签推理通常发生在客户端拥有数据而服务器拥有标签的垂直分区场景下。在这种场景下，客户端和服务器可以通过私有集交集算法[87-88]链接 2 个数据集中的相应条目，从而进行正常的推理任务。标签推理攻击通常由客户端发起，试图推断出特定训练集样本对应的隐私标签，通常假设客户端知道所有的标签种类以及预测任务，按照标签的种类数可以将其分为 2 分类、多分类和回归设置下的标签推理攻击。表 4 中对拆分学习系统中标签推理攻击的攻击阶段、标签种类、攻击方法和攻击假设等信息进行了总结。

1）2 分类标签推理攻击

Li 等人[30]提出了在 2 分类下的针对基于范数（norm-based）和基于方向（direction-based）的标签推理攻击。他们观察到在整个训练过程中，模型对判断正例为正的置信分低于判断负例为负的置信分，然而梯度向量的范数 $\|\nabla_z L_i(z)|z=M_{\text{C}}(x)\|_2$ 对于正例和负例都是相同的数量级。因此正实例的梯度范数 $\|g\|_2=|\tilde{p}_1-y|\times\|\nabla_z L_i(z)|z=M_{\text{C}}(x)\|_2$ 一般大于负实例的梯度范数，据此可以判断出标签。此外，梯度的方向也会泄露标签信息，即 1 对同类的示例梯度会导致正余弦相似度，而 1 对相反类的示例梯度会导致负余弦相似度。如果攻击者知道问题是类不平衡的，如正例比负例少，那么就可以确定每个示例的标签。

Table 4 Summary of Label Inference Attacks

表 4 标签推理攻击总结

方法	攻击阶段	标签种类	攻击方法	攻击假设
文献 [30]	训练	2 分类	通过梯度范数和方向推断标签	L0，L4
文献 [16]	训练	多分类	交替优化模型参数和虚拟标签	L0，L3
文献 [31]	训练	多分类	将交叉熵和标签先验作为正则	L0，L4
文献 [32]	训练/推理	多分类	将辅助数据的梯度或中间特征作为质心聚类	L0，L1
文献 [33]	推理	多分类	计算代理中间特征和隐私中间特征的 softmax 软标签	L0，L1
文献 [34]	训练	连续型	优化代理服务器和虚拟标签	L0，L1

2）多分类标签推理攻击

Erdoğan 等人[16]提出的 UnSplit 攻击需要处于客户端的攻击者随机初始化一个与服务器模型 M_{S} 具有相同体系结构的代理模型 $\tilde{M}_{\text{S}}$，通过优化代理模型参数 $\theta_{\tilde{M}_{\text{S}}}$ 和代理模型计算的标签 $\tilde{y}_i$ 以最小化代理模型产生的梯度和真实梯度之间的 *MSE*，从而学习私有标签。然而该方法只适用于 M_{S} 是单层网络的情况，且 Kariyappa 等人[31]测试了其对于 CIFAR10 的攻击，成功率仅为 10.99%，与随机猜测相当。

基于此，Kariyappa 等人[31]提出了 ExPLoit 攻击，对 UnSplit 中使用的 *MSE* 损失进行改进。他们假设攻击者知道真实标签的先验分布，要求代理标签的分布 $P_{\tilde{y}_i}$ 必须与真实标签先验 P_y 相匹配，通过最小化 $\mathcal{D}_{\text{KL}}(P_y\|P_{\tilde{y}_i})$ 来实现，且要求代理模型必须相对于代理标签具有较高的预测精度，可以通过最小化交叉熵 $E[H(\tilde{y}_i)+\mathcal{D}_{\text{KL}}(\tilde{y}_i\|\tilde{p}_i)]/H(P_y)$ 计算，其中用标签先验的熵 $H(P_y)$ 进行归一化，以确保其对标签类的数量和标签先验不敏感。将上面 2 个式子作为正则项添加到 UnSplit 的均方误差损失中，权衡了梯度损失项和交叉熵以及标签先验项的相对重要性。ExPLoit 攻击可以在大多数数据集上实现高准确率的标签推断。然而当服务器模型变复杂后，标签推断的准确率会大幅下降。

Liu 等人[32]提出在攻击者知道标签类别且对于每种标签类别拥有至少 1 个对应样本时，可以使用辅助数据作为质心，通过聚类的方法推断出隐私数据的标签。该方法在拆分系统的训练和推理阶段都可以进行攻击，在训练阶段对梯度进行聚类，在推理阶段对中间特征进行聚类。Zhou 等人[33]观察到在迭代训练过程中，同一类别中对象的中间特征变得更加紧密聚类，因此可以通过利用平均辅助数据中同类别的实例的中间特征来计算该类别的代理中间特征z^i_{agt}，进而通过将代理中间特征z^i_{agt}和目标样本的中间特征z_j一起进行 softmax 计算得出的软标签来表示该样本属于第i个类别的概率。

3）回归设置下的标签推理攻击

上述攻击是在标签为离散的场景下实施的，而 Xie 等人[34]则研究了在标签为连续分布的场景下的推理攻击。他们假设攻击者拥有一小部分已知标签的辅助数据$\mathcal{D}_{\text{known}}=\{x_j, y_j\}$，在黑盒场景下，攻击者需要构建代理服务器$\mathcal{M}_{\text{surr}}$和虚拟标签$y_i^{\text{dummy}}$，并通过减小损失函数$L_{\text{loss}}$的值对这二者进行迭代优化。$L_{\text{loss}}$由梯度距离损失$L_{\text{g}}$、训练准确率损失$L_{\text{t}}$和知识学习损失$L_{\text{k}}$构成。$L_{\text{g}}$是真实梯度和代理服务器产生的虚拟梯度间的距离损失，L_{t}计算虚拟标签和代理服务器预测的标签之间的绝对值平方，用以约束代理模型，尽可能地模仿服务器模型的行为，而L_{k}综合考虑L_{g}与L_{t}的和，使用辅助数据的真实标签代替虚拟标签进行计算。

2.4 隐私攻击技术总结

在本节中介绍了拆分学习系统中常见的隐私攻击，包括数据重构攻击、属性推理攻击、成员推理攻击和标签推理攻击。我们在此对攻击的优缺点以及与传统机器学习场景下的类似攻击进行对比。

数据重构攻击是拆分学习系统面临的最严重的一类攻击，与传统机器学习场景相比，攻击者能从客户端上传的中间得到更多的隐私信息。重构攻击方法繁多，但大致可以分为基于优化、基于学习和基于 GAN 的攻击方法三大类。基于优化的攻击方法一般要求在白盒条件下进行攻击，即要求攻击者掌握目标客户端模型架构(L3)和参数(L5)，较强的攻击假设限制了基于优化攻击的应用场景，但这类攻击的优势在于不需要辅助数据即可实施攻击(L2)。基于学习的攻击应用范围最广，He 等人[22]提出的反演网络攻击是几乎所有基于学习的数据重构攻击方法的基础。这类攻击通常都要求攻击者具有同分布辅助数据(L2)，且辅助数据质量直接影响重构效果，但在同等条件下其重构结果一般优于基于优化的攻击。基于 GAN 攻击的重构质量最高，但需要引入 GAN 模型，增加了攻击的复杂度。

推理攻击可以单独出现，也可以作为重构攻击的下游任务出现。拆分学习系统中的属性推理攻击和标签推理攻击的方法与普通机器学习场景中的攻击类似，区别在于攻击者一般通过分析客户端上传的中间特征而不是模型置信分来推知隐私属性。成员推理攻击目前在拆分学习系统中较为罕见，只有 Chen 等人[29]对此攻击进行了研究，与传统机器学习场景相比，拆分学习系统中的攻击者受到较大的限制。

3 拆分学习系统中的隐私防御技术

针对拆分学习系统中面临的隐私威胁，学术界已经提出了一系列隐私防御方案，如图 9 所示主要可以分为异常检测、噪声混淆、对抗性表征学习、特征裁剪和正则化防御 5 类，还有个别文献提出了其他方案，在本节中进行了系统的总结，并在图 9 中展示了各种防御方法的防御原理。异常检测的原理是使用统计和分析方法对模型的训练模式或相关事件进行甄别。其他方案的原理均是通过混淆或裁剪发送至服务器端的中间特征，使其重构得到的图片质量严重下降。

3.1 异常检测

异常检测是通过统计或分析的方法对模型的行为模式和数据等进行甄别，若检测为异常，则会及时预警并采取相应的防御措施。目前拆分学习系统中的异常检测都是客户端对服务器的检测，即依据检测的客户端是否受到 FSHA 攻击[17]将异常检测可以分为主动检测和被动检测。

1）主动检测

Erdoğan 等人[35]提出的 SplitGuard 方法是一种主动检测，其主要观察是，在正常的任务执行过程和攻击者执行 FSHA 攻击过程中，服务器下发的梯度具有不同的角度和幅度。因此客户端可以使用随机标签的数据作为假批次发送给服务器并获得相应梯度，对假批次和正常批次以及 2 个正常批次间的梯度的余弦值和向量平均幅度的绝对差进行运算得到分数SG。客户端通过比较SG是否超过特定阈值来判断是否正被攻击，并在攻击的早期停止训练。SplitGuard 可以有效地检测训练劫持攻击，同时最小化攻击者恢复的信息量。然而 SplitGuard 注入的假批次样本会导致模型精度下降，且容易受到名为 SplitSpy 的自

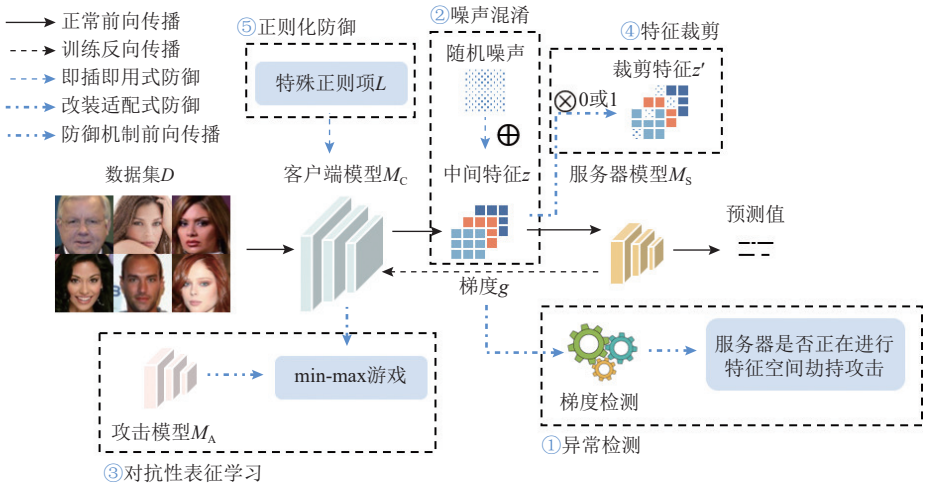


Fig. 9 Privacy defense technologies of split learning systems

图 9 拆分学习系统的隐私防御技术

适应劫持攻击的影响，在 SplitSpy 中，恶意服务器维护一个影子模型，该模型执行预期任务以检测错误标记的数据并逃避 SplitGuard。

2）被动检测

Fu 等人[36]观察到 FSHA 中的编码器和正常任务中的分类器之间存在根本区别：在前一种情况下使用标签，但在后一种情况下不使用标签，这种内在的区别导致了预期分类模型产生的分割层梯度与恶意自动编码器模型产生的分割层梯度之间有可观察到的行为差异。对于分类模型，具有相同标签的样本的梯度往往比具有不同标签的样本的梯度更为相似[89]，可以采用余弦相似度衡量。而来自攻击者的自动编码器的梯度在相同的标签样本和不同的标签样本之间没有这种区别。Fu 等人[36]提出的被动检测方法 Gradients Schrutinizer 利用这种内在差异来区分被劫持的训练和诚实训练。

3.2 正则化防御

正则化防御通过在原任务的损失函数上添加隐私正则项（式 (13)）使经过训练后的模型获得一定程度的隐私保护能力。正则化防御是非侵入性的，不需要改变系统架构，实现简单。

$$\mathcal{L}=\mathcal{L}_{\text{task}}+\alpha\mathcal{L}_{\text{norm}}。\tag{13}$$

1）NoPeek 防御

NoPeek 防御[37-38]的核心思想是通过添加额外的距离相关性（distance correlation）损失项[56]，在维持模型效用性的同时使中间特征与原始输入数据的距离相关性更小，同时确保其拥有足够模型准确分类的信息。NoPeek 的损失函数可以表示为

$$\mathcal{L}=\alpha_1 DCOR(x,z)+\alpha_2 CCE(y_{\text{true}},y),\tag{14}$$

等号右侧的第 1 项为距离相关性，第 2 项为正常的模型训练任务的交叉熵损失。

2）互信息正则化防御

Wang 等人[39]和 Samragh 等人[40]都采用了互信息正则化防御（mutual information regularization defense）。Wang 等人[39]将互信息 $I(x;z)$ 添加为模型训练目标的一个正则项以减少输入数据 x 和中间特征 z 之间的依赖性。有如下损失函数

$$L=\min_{\theta}-I(z;y)+\lambda I(x;z),\tag{15}$$

其中等号右侧的第 1 项使得模型将输入编码得到的中间层输出和真实标签的互信息增大，保证模型的预测能力；第 2 项通过最小化上界 $I(x;z)$，使得 x 和 z 之间的依赖性减少，提升对模型反演攻击的鲁棒性。而 Samragh 等人[40]则计算未施加防御的纯净的中间特征和混淆后的中间特征之间的互信息 $I(z',z)$，来最小化 z' 携带的关于 z 的信息，同时尽可能保持模型的效用。

3）MixCon 防御

MixCon 防御[41]聚焦于数据的分离性。其基本观察是数据点在隐编码空间越分离，越有助于模型分类，但是更大的分离会使重构模型输入变得容易，带来隐私威胁。基于以上观察，Li 等人[41]设计的 MixCon 防御使用一个新的损失项来平衡数据的分离性：

$$\mathcal{L}_{\text{MixCon}}=-1/\left(p\times|C|^2\right)(\mathcal{L}_1+\mathcal{L}_2),\tag{16}$$

$$\mathcal{L}_1=\sum_{i=1}^{p}\sum_{C_1\in C}\sum_{C_2\in C}dist(i,C_1,C_2),\tag{17}$$

$$\mathcal{L}_2=\beta/dist(i,C_1,C_2)。\tag{18}$$

$\mathcal{L}_1$ 惩罚数据点之间过大的距离，$\mathcal{L}_2$ 促使有效的数据分离性。

3.3 噪声混淆

噪声混淆是指在中间特征上施加精心构造的噪声，从而在保证目标模型原始任务准确率的基础上减少攻击者通过中间特征推理出的隐私信息，最常用的噪声是差分隐私噪声。差分隐私(differential privacy)[90]是建立在严格的数学理论基础之上的强隐私保护模型，能保证攻击者即便在具有最大背景知识的前提下，即已知数据库中除目标记录以外其他所有记录的信息也无法推测出目标记录的敏感信息。给定任意相邻数据集D和D'，对随机算法M及任意输出结果S，有不等式$Pr[M(D)\in S]\leqslant \exp(\varepsilon)\times Pr[M(D')\in S]+\delta$成立，则称算法$M$满足$(\varepsilon,\delta)$差分隐私。$\varepsilon$为差分隐私预算，$\delta$为扰动，$\varepsilon$和$\delta$越小，$M$在$D$和$D'$上输出的数据分布越接近。

1)无序噪声

Wang 等人[42]和 Titcombe 等人[43]都提出直接对中间特征施加拉普拉斯噪声(Laplacian noise)来防御。客户端首先进行 Nullification 操作，随机将输入数据的部分特征置零，然后对其中间特征进行归一化操作，并在归一化的特征上施加拉普拉斯噪声。由于客户端施加的噪声是随机的，模型必须添加噪声训练，以提升对噪声的鲁棒性。Kariyappa 等人[31]则提出向梯度叠加服从$\mathcal{N}(0,\sigma)$分布的高斯噪声来防御训练阶段的标签推理攻击。然而这种防御可能适用于更简单的数据集，但对于更复杂的数据集如 CIFAR10[91]，将标签推理攻击的准确度下降到 79% 会导致测试准确度降低 38 个百分点，这是不切实际的。而 Abuadbba 等人[45]则在拆分层输出的中间特征上应用差分隐私技术，其后果是模型准确率从 98.9% 显著降低到仅 50%。

2)启发式噪声

添加无序的噪声会显著降低推理的准确性，损害模型可用性。为了解决这个挑战，Mireshghallah 等人[44]提出一种端到端的噪声训练和推理框架 Shredder 来保护中间特征的隐私，其核心思想是通过对噪声进行训练，找到能满足隐私保护和模型效用权衡的噪声，并在推理阶段从训练好的噪声库中随机采样 1 组噪声叠加到中间特征中作为防御。为了衡量中间特征携带的隐私量并且减少计算成本，使用信噪比(signal to noise ratio, *SNR*)作为正则项，即计算中间特征和噪声规模的比值$SNR=E\left(a^2\right)/\sigma^2(n)$，该值越大，说明噪声越小，隐私信息蕴含越多。Li 等人[30]则提出了一种启发式随机扰动技术 Marvell，其核心思想是向梯度中添加非各向同性且协方差和样本相关的零均值高斯噪声，同时限制噪声的功率以防止模型效用大幅度下降。

3.4 对抗性表征学习

对抗性表征学习也即对抗训练是一种对对抗隐私攻击较为有效的防御方式，通常模拟了一个攻击者代理网络来与目标网络进行 min-max 游戏来进行训练[75]，使得目标网络能在一定程度上抵抗隐私攻击。对抗性表征学习通常可以很好地平衡模型效用和隐私之间的关系。

Xiao 等人[46]利用 GAN[75]来模拟攻击者进行数据重构，而防御者的目标是阻止攻击者重构出隐私数据，因此将攻击者的 *MSE* 损失和感知距离(perceptual distance)损失[92]当做正则项加入到原始模型来提升目标网络的鲁棒性。他们还对该防御方式的原理进行了探究，使用 t-SNE[93]对中间特征进行降维处理后发现，经过对抗训练后的特征变得更加分散，使得攻击者更难还原出原始图片。Liu 等人[47]提出了一种隐私对抗网络(PAN)，该训练框架有 4 个部分：编码器(E)、效用鉴别器(UD)、隐私重构器(PR)、隐私鉴别器(PD)。编码器相当于客户端模型，效用鉴别器类似于服务器端模型，隐私重构器是攻击者的数据重构网络，隐私鉴别器是攻击者的属性推断攻击的网络。类似地，Li 等人[48]提出的 DeepObfuscator 防御框架也包含 4 个部分，分别是位于模型前半部分的混淆器(obfuscator)、位于模型后半部分的分类器、攻击者的重构攻击和属性推理攻击网络。DeepObfuscator 的对抗训练算法迭代地训练混淆器防御重构攻击、属性推理攻击以及优化攻击模型。

Rezaeifar 等人[49]提出了一种利用对比学习(contrastive learning)[94]来防御重构攻击的方法，该框架主要存在编码器、混淆器和分类器 3 个部分。编码器是一个通过对比学习预训练好的模型，该部分的主要作用是将原始数据进行特征提取，以供后续的分类任务。混淆器基于信息论减少中间特征和原始数据之间的信息关联程度，以此来减少中间特征所蕴含的信息量。分类器的作用是用来做原始的分类任务，分类器和混淆器是联合训练的，这样可以保留原始分类任务精度的同时尽量减少中间特征蕴含的信息，进而降低隐私泄露风险。

Li 等人[50]提出了一种针对拆分学习训练阶段的 2 端式防御方法，动机源自 2 个方面：1)对抗训练需要一个很强的攻击者，但是使用比较复杂的模型需要较高的时间开销，对于拆分学习等资源受限场景不够实际；2)模型在对抗训练的前面几个迭代中对

于重构攻击依然很脆弱，这是由于鲁棒能力的增强是一个循序渐进的过程。基于此提出了2端式的防御手段，首先防御者会选取一个有计算资源的服务器，利用对抗训练算法预训练一个具有鲁棒性的模型。预训练过程中，防御者会初始化一个攻击能力很强的模型，用该模型进行对抗训练来增强客户端模型的鲁棒性。其次是将预训练好的模型进行迁移学习，并迁移到客户端模型的初始化中。同时，Li等人[50]发现了中间特征的尺寸越大，越被利用攻击，并基于这个观察提出的Bottleneck层通过编码器结构来压缩中间特征的尺寸。

3.5 特征裁剪

客户端浅层网络输出的中间特征通常含有较多的低级空间信息，并在深层网络中逐步转化为用于完成目标任务的高级语义信息。空间信息中常常含有大量冗余，对于目标任务的帮助较小，但是攻击者能利用这些冗余信息进行隐私攻击。因此，特征裁剪的思想是，把与目标任务无关的冗余的中间特征裁剪掉，使得攻击者难以利用裁剪后有限的中间特征进行隐私攻击。

He等人[51]提出利用Dropout的思想随机选择中间特征中的某些元素置零，可以在不显著降低模型有效性的同时提供对输入数据的混淆。区别于Dropout防御作用于单个元素，Singh等人[23]提出的DISCO防御直接裁剪掉整个通道的特征。DISCO的主体架构由预处理模块、动态通道剪裁网络以及对抗网络组成。预处理模块首先对原始图像进行切割划分，并将切割出的部分对齐为原图大小，依次输入客户端网络的第1个卷积层，随后按组进行平均聚合降维。其根本思路是使模型得到的图像是原始图像的切分，减少直接输入原图带来的空间信息冗余。动态通道剪裁网络接收客户端输入的中间特征，经由多个卷积层得到与特征通道数相同的得分，并通过超参数、剪裁阈值来控制隐私和效用的权衡。最后通过在对抗训练过程中模拟来自数据重构、敏感属性推理的攻击来优化动态通道剪裁网络和目标模型，在提升目标模型对隐私攻击的鲁棒性的同时，保持目标任务的精度。DISCO的特征剪裁网络为每个输入图像提供了一个独立的动态剪裁方案，有较强的隐私保护能力。

3.6 其他方案

除了异常检测、正则化防御、噪声混淆、对抗性表征学习和特征裁剪五大类隐私防御方案外，还有一些其他方案能提供隐私保护。

Osia等人[95]建议使用Siamese架构进行微调，其主要思想是使语义相似的点的表示彼此尽可能接近，而不同点的表示彼此远离。在微调后，具有相同主类的样本点会彼此接近，使攻击者更难获得敏感信息。例如，在性别判断问题中，在Siamese微调前，不同身份的“男性”类的图像在特征空间中较为分散，攻击者可以较容易地发现由中间特征到输入空间的反向映射函数，能够重建原始图像。在Siamese微调后，如果所有k个不同的男性图像都映射到特征空间中同一个点的邻域内，攻击者就被混淆。特征间的距离使用欧几里得距离进行衡量。在使用Siamese微调时，需要准备一个包含标记为相似或不相似的样本对的训练数据集，并在微调时对每一对样本应用对比损失函数，使得模型将被标记为相似的一对样本间的特征距离最小化，而被标记为不相似的一对样本间的特征距离最大化。

Wang等人[96]认为传统的基于对抗博弈的防御方法难以抵御自适应攻击者，为此他们提出了Crafter防御。Crafter防御与传统的基于混淆或基于裁剪的防御的不同之处在于，Crafter防御利用推土距离（earth mover's distance, EMD）作为其隐私保护损失，使得叠加了防御后的中间特征与原始中间特征很接近，但利用叠加防御后的中间特征重建出来的图像在身份感知上与原始中间特征截然不同。Crafter的目标损失函数涉及到隐私保护损失$\mathcal{L}_{\mathrm{p}}$、模型效用损失$\mathcal{L}_{\mathrm{u}}$和重构损失$\mathcal{L}_{\mathrm{inv}}$这3个损失函数的嵌套组合，并将其转换为最小最大博弈的规范的Wasserstein-GAN[97]公式形式。Crafter可以防御白盒和黑盒场景下的重构攻击，Wang等人[96]的评估发现，Crafter对增加攻击的优化轮次、使用不同的GAN网络作为攻击模型和使用多次查询得到的平均特征进行重构3种自适应攻击具有鲁棒性。

3.7 隐私防御技术评估

在本节中展示了在统一的实验设置下，不同防御方法对拆分学习系统所面临的最严重的隐私威胁——数据重构攻击的防御效果，并进行对比评估。

3.7.1 实验设置

1）目标系统。本文选用ResNet-18[55]作为两方参与者的拆分学习系统模型。将ResNet-18网络在第4个残差块后划分为2部分，分别部署在客户端和服务器上。

2）数据集和任务。实验采用2个经典数据集CIFAR10[91]和CelebA[66]。CIFAR10是包含10个类别的物品数据集，共有60 000张彩色图像，其中50 000

张用于训练, 10 000 张用于测试。CelebA 是人脸数据集, 每张人脸图像标注了 40 个不同的属性, 其中 162 770 张图像用于训练, 5 000 张图像用于测试。拆分学习系统任务是在 CIFAR10 数据集上进行 10 分类, 以及在 CelebA 数据集上进行 Attractive 属性的 2 分类。我们假设拆分学习系统拥有训练集作为隐私训练数据集, 而攻击者则拥有测试集作为同分布辅助数据集。

3)隐私训练设置。我们从每种防御类别中各选择了一种最主流的防御进行评估, 包括 NoPeek 防御、Shredder 防御、DeepObfuscator 防御、Disco 防御和 Siamese 防御。由于异常检测机制的不同, 故未将其纳入此次评估实验。无论使用何种防御机制, 拆分学习系统的训练轮次均设置为 80, 学习率设为 3E−4, 采用 Adam[98] 作为优化器。针对每种防御方法, 我们评估了 3 组不同超参数下的防御效果。

4)攻击设置。我们假设服务器端攻击者能够在拆分学习系统训练完成后对客户端进行查询并获取中间特征。攻击者随后构建反演网络[22] 以重构原始数据。该攻击网络的训练轮次为 100, 采用 Adam 优化器, 学习率设为 0.01。

5)量化指标。采用模型准确率(*ACC*)作为模型效用评估指标。采用结构相似性指数(*SSIM*)[69]、峰值信噪比(*PSNR*)[68]、感知相似性指标(*LPIPS*)[70] 和 *MSE* 等指标对重构图像的质量进行评估。其中 *SSIM*, *PSNR* 值越大或 *LPIPS*, *MSE* 值越小, 代表重构图像质量越高, 攻击效果越好; 反之, 防御效果越好。

3.7.2 防御效果评估

图 10 直观地展示了不同防御机制的效果, 而表 5 则列出了各防御方法的超参数设置及其量化指标。

从图 10 中可以看出, NoPeek 防御和 Siamese 防御在 CelebA 和 CIFAR10 数据集上均表现优异, 攻击者重构的图像完全未能泄露任何隐私信息。这 2 个防御的核心思想形成对比, NoPeek 防御旨在增加中间特征和输入数据间的距离相关性, 而 Siamese 防御旨在使同类数据的中间特征更加接近, 从而防止攻击者学到中间特征到输入数据的正确映射关系, 但实验证明这 2 种防御都可以达到较好的防御效果。Shredder 防御在多分类的 CIFAR10 数据集上提供了更好的隐私保护, 但在处理人脸数据集时, 则基本无法抵御自适应攻击者的攻击。同时, Shredder 防御所施加的噪声会显著干扰模型的效用。与 Shredder 相反, DeepObfuscator 防御在脸部数据集上的保护效果优于物品数据集。在 CelebA 数据集上, 重构的图像完全无法揭示隐私信息, 而在 CIFAR10 上, 重构图像仅能模糊地显示物品的轮廓。虽然 Disco 防御在抵御自适应攻击者的能力上稍逊一筹, 但仍能提供足够的保护。除了 Shredder 防御外, 其他防御方法对模型性能的影响都较小。

表 5 展示了不同超参数对各防御性能的影响。我们的研究发现, 根据模型的具体任务, 防御的超参数需要进行相应的调整以应对不同的情况。例如, NoPeek 防御在 2 个数据集上达到最佳防御效果时所



Fig. 10 Effectiveness of different defense methods

图 10 不同防御方法的作用效果

Table 5 Quantitative Indicators of Different Defense Methods

表 5 不同防御方法量化指标

防御方法	超参数	CelebA					CIFAR10				
		ACC↑	*MSE*↑	*LPIPS*↑	*SSIM*↓	*PSNR*↓	*ACC*↑	*MSE*↑	*LPIPS*↑	*SSIM*↓	*PSNR*↓
无防御	无	0.772	0.026	0.385	0.528	64.049	0.749	0.038	0.374	0.442	62.368
DeepObfuscator	$\alpha=0.1$	**0.77**	0.22	0.695	0.078	54.713	**0.75**	0.042	0.461	0.417	61.911
	$\alpha=0.5$	0.768	0.202	0.695	0.085	55.086	0.737	0.056	0.497	0.349	60.677
	$\alpha=0.9$	0.767	**0.246**	**0.702**	**0.064**	**54.232**	0.737	**0.07**	**0.544**	**0.265**	**59.703**
Disco	*pruning_ratio* = 0.1	0.768	0.026	0.386	0.532	64.016	**0.751**	0.034	0.367	0.502	62.861
	pruning_ratio = 0.5	**0.773**	0.034	0.437	0.461	62.778	0.749	0.047	0.399	0.418	61.42
	pruning_ratio = 0.99	0.77	**0.124**	**0.673**	**0.179**	**57.195**	0.717	**0.087**	**0.569**	**0.242**	**58.729**
NoPeek	$\alpha=0.2$	0.771	0.127	0.625	0.228	57.081	**0.749**	0.04	0.407	0.43	62.072
	$\alpha=0.5$	0.771	**0.392**	**0.692**	**0.038**	**52.204**	0.724	0.173	0.601	0.077	55.762
	$\alpha=0.8$	**0.773**	0.208	0.635	0.133	54.962	0.723	**0.361**	**0.614**	**0.028**	**52.57**
Siamese	$\alpha=0.005$	0.767	0.155	0.562	0.191	56.233	**0.738**	0.042	0.421	0.417	61.929
	$\alpha=0.5$	**0.778**	**0.315**	**0.636**	**0.058**	**53.156**	0.694	0.192	0.653	**0.036**	55.302
	$\alpha=0.95$	0.776	0.305	0.521	0.079	53.294	0.627	**0.23**	**0.685**	0.23	**54.515**
Shredder	$coeff=0.3$	**0.737**	**0.029**	**0.376**	0.515	**63.574**	0.506	**0.046**	**0.463**	0.393	**61.549**
	$coeff=0.5$	0.728	0.027	0.373	0.52	63.78	**0.511**	0.045	0.457	**0.391**	61.582
	$coeff=0.9$	0.733	0.028	0.372	**0.513**	63.724	0.506	0.043	0.455	0.399	61.764

注: “↑”表示数值越大，模型的准确度越高或防御效果越好，攻击效果越差; “↓”表示数值越小，模型的准确度越高或防御效果越好，攻击效果越差。黑体数值表示最优值。

需的超参数不同: 在 CelebA 数据集上, 当$\alpha=0.5$时效果最佳; 而在 CIFAR10 数据集上则是在$\alpha=0.8$时达到最优。大多数防御方法在增强防御的同时, 模型的效用也会随之下降, 但 Siamese 防御是个显著的例外。在人脸数据集中, 当 Siamese 防御达到最强防御效果时, 模型的效用也处于最高水平, 这得益于它在混淆同主类数据的中间特征时, 迫使不同主类数据的中间特征相距较远, 以便于模型进行分类。

3.8 隐私防御技术总结

表 6 中对不同防御方法的核心思想和特性进行了总结。综上所述, 应用异常检测、正则化防御、噪声混淆、对抗性表征学习、特征裁剪以及 Siamese 和 Crafter 防御等新兴技术可解决拆分学习中的隐私威胁, 从而增强拆分学习隐私保护。目前已有一些工作对这些隐私防御技术进行了较为系统的度量, 如 Li 等人[25] 和 Yang 等人[99] 基于数据重构攻击并通过

Table 6 Summary of Privacy Defense Technologies

表 6 隐私防御技术总结

文献	防御方式	核心思想	是否支持训练阶段	是否提供可证明安全性	隐私保护能力和模型效用性权衡
文献 [35−36]	异常检测	检测不同样本的梯度间的角度和幅度是否超过特定阈值	√		有较强的保护能力
文献 [37−38]	NoPeek	减小特征与输入的距离相关性	√	√	保护能力较强
文献 [39−40]	互信息正则化	减少特征和输入的互信息		√	有一定保护能力
文献 [41]	MixCon	减少特征空间分离性	√	√	保护能力受限
文献 [30−31, 42−45]	噪声混淆	引入随机性噪声	√	√	保护能力较弱，对模型效用性影响较大
文献 [46−50]	对抗性表征学习	基于 min-max 游戏的对抗训练	√		有较强保护能力
文献 [23, 51]	特征裁剪	减裁特征信息量			有一定保护能力
文献 [95]	Siamese	减少特征空间的分离性	√	√	有较强的保护能力
文献 [96]	Crafter	使叠加了防御的中间特征与原始中间特征很接近，但重构图像在身份感知上截然不同		√	有较强的隐私保护能力

实验展示了不同防御机制的保护效果。

这些隐私防御技术在为拆分学习隐私需求提供有效保障的同时，往往也会在一定程度上付出相应的代价。例如，在保证隐私性的同时牺牲了系统的效率，或是降低了模型的准确性。因此，如何在隐私性和模型效用之间做权衡，是拆分学习系统隐私防御技术发展的未来研究重点。

4 未来展望

目前，拆分学习应用广泛，与联邦学习共同成为主流的联合推理方法。与联邦学习相比，拆分学习只需要在客户端运行局部模型，而不需要整个模型，大大降低了客户端的性能负担，使得模型在物联网等小设备上运行成为可能。在工业界，目前支持拆分学习的框架有蚂蚁隐语[100]、华为 MindSpore[101]、FATE[102]、PySyft[103] 等，而支持简单纵向联邦或水平联邦学习的框架 FederatedScope[104] 等，在未来会有更多的框架支持拆分学习。虽然对拆分学习系统的隐私攻击和防御技术的研究已经取得了一定的成果，但是目前这些研究还处于初期阶段，还有许多问题亟待解决，其中以下 5 类问题值得进一步研究。

1)隐私和效用并重的创新拆分学习系统

现有的针对拆分学习系统的隐私防御方案几乎都会在一定程度上降低模型的效用，然而在实际生产过程中，对模型准确率的要求十分严格，哪怕是 1% 的准确率下降也是不可接受的。因此，处理好拆分学习系统的隐私性和有效性之间的矛盾是拆分学习在大规模应用之前面临的重要问题。近年来，隐私计算技术日趋成熟，已经有工作尝试将可信执行环境(trusted execution environment, TEE)[105]、同态加密(homomorphic encryption, HE)[106-107] 和多方安全计算(secure multi-party computation, MPC)[108] 等技术应用于人工智能模型的部署之中。然而这些隐私计算框架虽然在一定程度上对模型的隐私和安全起到保护作用，但是它们在模型计算效率、系统能耗方面仍然存在很大的上升空间。后续的工作可以考虑将可信执行环境、同态加密和多方安全计算等隐私计算技术和本文所介绍的针对拆分学习系统的隐私防御方案相结合，建立更加稳定、低能耗并且具备更强隐私保护能力的计算环境。

2)低成本和高隐蔽的隐私攻击新挑战

目前针对拆分学习系统的隐私攻击普遍依赖于强假设，除了少数攻击方案外，基本都需要使用与隐私数据集同分布的辅助数据或是对目标模型进行多次查询来训练攻击模型。在实际应用中，尤其是在医疗、金融等隐私性要求高的场景下，同分布数据集是很难获取的，协议甚至要求系统中各参与方不能保存中间数据。而较多的查询次数导致攻击的隐蔽性差，容易被检测出来。因此探索成本更低和隐蔽性更强的攻击方案和在现实场景下实施具有可行性的攻击是目前对拆分学习系统隐私攻防研究所面临的重要问题。针对上述问题，使用先进的生成模型如 GAN[67,75,77] 或是扩散模型(diffusion model)[109] 是一个可能的解决方案，生成模型的优势在于可以生成未在训练集中见过的数据样本，有机会在一定程度上弥合数据分布的差异。本文介绍的工作也已经将 GAN 应用到隐私攻击方案之中[25]，然而如何在查询次数受限以及无法得到同分布的辅助数据的情况下提高攻击的准确率和鲁棒性仍然是未来需要多加探索的问题。

3)拆分学习的前瞻性隐私保护策略

目前，只有 Erdoğan 等人[35] 和 Fu 等人[36] 对于拆分学习系统的异常检测做出了初步的探索，且只能对损害模型性能的主动攻击进行检测。而本文中所提到的隐私攻击除了 FSHA[17] 外都是被动攻击，攻击过程更加隐蔽且难以被察觉。而在安全领域，预防大于阻止，因此除了在攻击阶段进行异常检测，也可以在攻击之前采取预防措施。例如一些白盒场景下的隐私攻击，通常假设攻击者拥有目标模型的完整使用权或是完全了解目标模型的架构和参数，而模型提供商则可以在模型部署前对模型参数进行加密或是采取模型水印[110-111] 等方式来保护模型的知识产权，防止模型被用于不正当的目的。此外，Song 等人[27] 的工作利用了机器学习模型过度学习的特点实施属性推理攻击，Haim 等人[112] 和 Buzaglo 等人[113] 的工作也证明了机器学习模型的参数会记录一些敏感样本，而攻击者有机会从模型参数中还原敏感数据。未来在考虑模型效用的同时，还需要考虑如何将记录了隐私数据的模型参数进行剪枝，或是使模型忘却掉(machine unlearning)[114] 这些记住的样本。

4)拆分学习防御技术的综合评估框架

除了能够提供具有可证明安全性的防御方案，如基于差分隐私[30-31,42-45] 和基于信息论[41,95,99,105-107] 的防御方案外，大多数其他防御方案通常依赖于攻击效果来间接评估其效用。这种评估方式虽然在某种程度上能够反映防御方案的实际表现，但存在明显的局限性：不同的攻击者即使使用相同的攻击方案，

训练出的攻击模型能力仍然存在显著差异，这导致了通过攻击效果来评估防御效用的方法一直备受争议，评估结果的准确性在很大程度上取决于攻击者的能力和策略，而并非完全反映防御方案本身的优劣。因此，未来需要寻求一种更加客观、统一的评估方法，例如利用信息学理论或数学方法制定一个统一的评估指标，对隐私信息的泄露量进行准确评定。通过这种方法可以更加客观地评估不同防御方案在保护隐私信息方面的效果，从而为隐私保护技术的发展提供更加坚实的理论基础和实践指导。

5）围绕大模型的拆分学习革新

目前，具有大规模参数和复杂计算结构的大模型因其强大功能在数据挖掘和自然语言处理等多个领域得到了广泛的应用。以 GPT 为代表的大语言模型（large language model, LLM）由于其强大的文本处理能力而受到关注[115]。然而随着大模型便利性的提升，对数字化和文字作品版权泄露的担忧也日益加剧，将拆分学习技术应用于大模型上可能是一个潜在的解决方案。目前，已有研究尝试对如 BERT 等语言模型生成的词嵌入进行重构攻击，以还原出原始输入文本[116-118]。然而这一领域的工作尚处于初期探索阶段，攻击效果尚未达到理想水平，且尚未扩展至大语言模型领域。因此，未来的研究可以进一步探索如何对大模型进行高效拆分，并同步研究相关的隐私攻击与防御技术，以在保护知识产权的同时，为大模型提供更广泛、更安全的应用前景。

5 总　　结

随着拆分学习的研究发展和广泛应用，拆分学习模型的隐私安全问题吸引了许多学者的兴趣和关注，产生了不少瞩目的研究成果，但目前相关的研究还处于初级阶段，尚有许多关键问题亟待解决。本文在充分调研和深入分析的基础上，对拆分学习系统在隐私安全领域最新的研究成果进行综述，系统总结了拆分学习存在的隐私安全问题和隐私攻击技术，并对现有的防御方法进行科学的分类和分析。同时，本文也指出了当前拆分学习系统在隐私安全领域尚未解决的问题，并探讨了未来的研究方向。

作者贡献声明：刘雅欣确定综述主题，奠定综述框架，负责论文主体撰写；王鹃负责终审论文，提供整体修改意见，并编辑润色内容；杨梦达和李子昂负责梳理针对拆分学习系统隐私攻击的相关文献，并提供全面的分析描述；易文哲、徐枭洋负责梳理防御部分的文献，并提供现状描述与理论探讨分析；陈美会负责整合并制作论文的图表；韦韬负责未来展望部分撰写，并提供未来研究建议。

致谢：这项工作受到蚂蚁集团“拆分学习场景下的攻防研究”产学研项目的支持。

参　考　文　献

[1] Briakou E, Cherry C, Foster G. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM’s translation capability [J]. arXiv preprint, arXiv: 2305.10266, 2023

[2] Yan Siyuan, Yu Zhen, Zhang Xuelin, et al. Towards trustable skin cancer diagnosis via rewriting model’s decision [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11568−11577

[3] Bae H K, Ahn J, Lee D, et al. LANCER: A lifetime-aware news recommender system [C] //Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 4141−4148

[4] Cho J, Hyun D, won Lim D, et al. Dynamic multi-behavior sequence modeling for next item recommendation [C] //Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 4199−4207

[5] Lin Dongding, Wang Jian, Li Wenjie. Cola: Improving conversational recommender systems by collaborative augmentation [C] //Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 4462−4470

[6] Li Xiang, Yin Junbo, Shi Botian, et al. Lwsis: Lidar-guided weakly supervised instance segmentation for autonomous driving [C] //Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 1433−1441

[7] Li Zhuopeng, Li Lu, Zhu Jianke. Read: Large-scale neural scene rendering for autonomous driving [C] //Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 1522−1529

[8] Wu Hai, Wen Chenglu, Li Wei, et al. Transformation-equivariant 3D object detection for autonomous driving [C] //Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 2795−2802

[9] Gupta O, Raskar R. Distributed learning of deep neural network over multiple agents [J]. Journal of Network and Computer Applications, 2018, 116: 1−8

[10] 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

[11] Poirot M G, Vepakomma P, Chang Ken, et al. Split learning for collaborative deep learning in healthcare [J]. arXiv preprint, arXiv: 1912.12115, 2019

[12] Ayad A, Renner M, Schmeink A. Improving the communication and computation efficiency of split learning for IoT applications [C] // Proc of the 2021 IEEE Global Communications Conf (GLOBECOM). Piscataway, NJ: IEEE, 2021: 1−6

[13] He Chaoyang, Balasubramanian K, Ceyani E, et al. Fedgraphnn: A federated learning system and benchmark for graph neural networks [J]. arXiv preprint, arXiv: 2104.07145, 2021

[14] Abedi A, Khan S S. FedSL: Federated split learning on distributed sequential data in recurrent neural networks[J]. Multimedia Tools and Applications, 2024, 83(10): 28891−28911

[15] Park S, Kim G, Kim J, et al. Federated split task-agnostic vision transformer for COVID-19 CXR diagnosis[J]. Advances in Neural Information Processing Systems, 2021, 34: 24617−24630

[16] Erdoğan E, Küpçü A, Çiçek A E. UnSplit: Data-oblivious model inversion, model stealing, and label inference attacks against split learning [C] //Proc of the 21st Workshop on Privacy in the Electronic Society. New York: ACM, 2022: 115−124

[17] Pasquini D, Ateniese G, Bernaschi M. Unleashing the tiger: Inference attacks on split learning [C] //Proc of the 16th ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2021: 2113−2129

[18] Gao Xinben, Zhang Lan. PCAT: Functionality and data stealing from split learning by pseudo-client attack [C] // Proc of the 32nd USENIX Security Symp (USENIX Security 23). Berkeley, CA: USENIX Association, 2023: 5271−5288

[19] Zhu Xiaochen, Luo Xinjian, Wu Yuncheng, et al. Passive inference attacks on split learning via adversarial regularization [J]. arXiv preprint, arXiv: 2310.10483, 2023

[20] Luo Sida, Yu Fangchao, Wang Lina, et al. Feature sniffer: A stealthy inference attacks framework on split learning [C] //Proc of the 32nd Int Conf on Artificial Neural Networks. Berlin: Springer, 2023: 66−77

[21] Xu Xiaoyang, Yang Mengda, Yi Wenzhe, et al. A stealthy wrongdoer: Feature-oriented reconstruction attack against split learning [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 12130−12139

[22] He Zecheng, Zhang Tianwei, Lee R B. Model inversion attacks against collaborative inference [C] //Proc of the 35th Annual Computer Security Applications Conf. New York: ACM, 2019: 148−162

[23] Singh A, Chopra A, Garza E, et al. DISCO: Dynamic and invariant sensitive channel obfuscation for deep neural networks [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 12125−12135

[24] Yin Yupeng, Zhang Xianglong, Zhang Huanle, et al. Ginver: Generative model inversion attacks against collaborative inference [C] //Proc of the ACM Web Conf 2023. New York: ACM, 2023: 2122−2131

[25] Li Zi'ang, Yang Mengda, Liu Yaxin, et al. GAN you see me? Enhanced data reconstruction attacks against split inference[J]. Advances in Neural Information Processing Systems, 2023, 36: 54554−54566

[26] Dougherty S, Kumar A, Hou Jie, et al. A stealthy inference attack on split learning with a split-fuse defensive measure [C] //Proc of the 2023 IEEE Conf on Communications and Network Security (CNS). Piscataway, NJ: IEEE, 2023: 1−9

[27] Song Congzheng, Shmatikov V. Overlearning reveals sensitive attributes [J]. arXiv preprint, arXiv: 1905.11742, 2019

[28] Wang Zhibo, Liu Kaixin, Hu Jiahui, et al. Attrleaks on the edge: Exploiting information leakage from privacy-preserving co-inference[J]. Chinese Journal of Electronics, 2023, 32(1): 1−12

[29] Chen Hanxiao, Li Hongwei, Dong Guishan, et al. Practical membership inference attack against collaborative inference in industrial IoT[J]. IEEE Transactions on Industrial Informatics, 2020, 18(1): 477−487

[30] Li Runliang, Sun Jiankai, Yang Xin, et al. Label leakage and protection in two-party split learning [J]. arXiv preprint, arXiv: 2102.08504, 2021

[31] Kariyappa S, Qureshi M K. ExPLoit: Extracting private labels in split learning [C] //Proc of the 2023 IEEE Conf on Secure and Trustworthy Machine Learning (SaTML). Piscataway, NJ: IEEE, 2023: 165−175

[32] Liu Junlin, Lyu Xinchen, Cui Qimei, et al. Similarity-based label inference attack against training and inference of split learning[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 2881−2895

[33] Zhou Chenchen, Cao Hongbo, Zhao Yingying, et al. Label inference attack based on soft label towards two-party split learning [C] //Proc of the 14th IEEE Int Symp on Parallel Architectures, Algorithms and Programming (PAAP). Piscataway, NJ: IEEE, 2023: 1−6

[34] Xie Shangyu, Yang Xin, Yao Yuanshun, et al. Label inference attack against split learning under regression setting [J]. arXiv preprint, arXiv: 2301.07284, 2023

[35] Erdoğan E, Küpçü A, Cicek A E. SplitGuard: Detecting and mitigating training-hijacking attacks in split learning [C] //Proc of the 21st Workshop on Privacy in the Electronic Society. New York: ACM, 2022: 125−137

[36] Fu Jiayun, Ma Xiaojing, Zhu Bin, et al. Focusing on Pinocchio's nose: A gradients scrutinizer to thwart split-learning hijacking attacks using intrinsic attributes [EB/OL]. Fredericksburg, VA: Internet Society, 2023[2025-08-30]. https://www.researchgate.net/profile/Bin-Zhu/publication/369192403_Focusing_on_Pinocchio's_Nose_A_Gradients_Scrutinizer_to_Thwart_Split-Learning_Hijacking_Attacks_Using_Intrinsic_Attributes/links/648b04a1712bd82962254579/Focusing-on-Pinocchios-Nose-A-Gradients-Scrutinizer-to-Thwart-Split-Learning-Hijacking-Attacks-Using-Intrinsic-Attributes.pdf

[37] Vepakomma P, Singh A, Gupta O, et al. NoPeek: Information leakage reduction to share activations in distributed deep learning [C] //Proc of the 2020 Int Conf on Data Mining Workshops (ICDMW). Piscataway, NJ: IEEE, 2020: 933−942

[38] Vepakomma P, Singh A, Zhang E, et al. Nopeek-infer: Preventing face reconstruction attacks in distributed inference after on-premise

training［C］//Proc of the 16th IEEE Int Conf on Automatic Face and Gesture Recognition (FG'2021). Piscataway, NJ: IEEE, 2021: 1−8

［39］ Wang Tianhao, Zhang Yuheng, Jia Ruoxi. Improving robustness to model inversion attacks via mutual information regularization［C］//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 11666−11673

［40］ Samragh M, Hosseini H, Triastcyn A, et al. Unsupervised information obfuscation for split inference of neural networks［J］. arXiv preprint, arXiv: 2104.11413, 2021

［41］ Li Xiaoxiao, Huang Yangsibo, Peng Binghui, et al. MixCon: Adjusting the separability of data representations for harder data recovery［J］. arXiv preprint, arXiv: 2010.11463, 2020

［42］ Wang Ji, Zhang Jianguo, Bao Weidong, et al. Not just privacy: Improving performance of private deep learning in mobile cloud［C］//Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 2407−2416

［43］ Titcombe T, Hall A J, Papadopoulos P, et al. Practical defences against model inversion attacks for split neural networks［J］. arXiv preprint, arXiv: 2104.05743, 2021

［44］ Mireshghallah F, Taram M, Ramrakhyani P, et al. Shredder: Learning noise distributions to protect inference privacy［C］//Proc of the 25th Int Conf on Architectural Support for Programming Languages and Operating Systems. New York: ACM, 2020: 3−18

［45］ Abuadbba S, Kim K, Kim M, et al. Can we use split learning on 1D CNN models for privacy preserving training［C］//Proc of the 15th ACM Asia Conf on Computer and Communications Security. New York: ACM, 2020: 305−318

［46］ Xiao Taihong, Tsai Y H, Sohn K, et al. Adversarial learning of privacy-preserving and task-oriented representations［C］//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 12434−12441

［47］ Liu Sicong, Du Juanzhao, Shrivastava A, et al. Privacy adversarial network: Representation learning for mobile data privacy［J］. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2019, 3(4): 1−18

［48］ Li Ang, Guo Jiayi, Yang Huanrui, et al. DeepObfuscator: Obfuscating intermediate representations with privacy-preserving adversarial learning on smartphones［C］//Proc of the Int Conf on Internet-of-Things Design and Implementation. New York: ACM, 2021: 28−39

［49］ Rezaeifar S, Voloshynovskiy S, Asgari Jirhandeh M, et al. Privacy-preserving image template sharing using contrastive learning［J］. Entropy, 2022, 24(5): 643

［50］ Li Jingtao, Rakin A S, Chen Xing, et al. Ressfl: A resistance transfer framework for defending model inversion attack in split federated learning［C］//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10194−10202

［51］ He Zecheng, Zhang Tianwei, Lee R B. Attacking and protecting data privacy in edge–cloud collaborative inference systems［J］. IEEE Internet of Things Journal, 2020, 8(12): 9706−9716

［52］ Thapa C, Arachchige P C M, Camtepe S, et al. Splitfed: When federated learning meets split learning［C］//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 8485−8493

［53］ McMahan B, 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. New York: PMLR, 2017: 1273−1282

［54］ Wang Xiaosong, Peng Yifan, Lu Le, et al. ChestxRay8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases［C］//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2097−2106

［55］ He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition［C］//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

［56］ Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances［J］. Annals of Statistics, 2007, 35(6): 2769−2794

［57］ 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

［58］ Kang Yiping, Hauswald J, Gao Cao, et al. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge［J］. ACM SIGARCH Computer Architecture News, 2017, 45(1): 615−629

［59］ Eshratifar A E, Abrishami M S, Pedram M. JointDNN: An efficient training and inference engine for intelligent mobile cloud computing services［J］. IEEE Transactions on Mobile Computing, 2019, 20(2): 565−576

［60］ Banitalebi-Dehkordi A, Vedula N, Pei Jian, et al. Auto-split: A general framework of collaborative edge-cloud AI［C］//Proc of the 27th ACM SIGKDD Conf on Knowledge Discovery & Data Mining. New York: ACM, 2021: 2543−2553

［61］ Singh A, Vepakomma P, Gupta O, et al. Detailed comparison of communication efficiency of split learning and federated learning［J］. arXiv preprint, arXiv: 1909.09145, 2019

［62］ Thapa C, Chamikara M A P, Camtepe S A. Advancements of Federated Learning Towards Privacy Preservation: From Federated Learning to Split Learning［M］//Federated Learning Systems: Towards Next-Generation AI. Berlin: Springer, 2021: 79−109

［63］ Gao Yansong, Kim M, Abuadbba S, et al. End-to-end evaluation of federated learning and split learning for Internet of things［J］. arXiv preprint, arXiv: 2003.13376, 2020

［64］ Tan A Z, Yu Han, Cui Lizhen, et al. Towards personalized federated learning［J］. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(12): 9587−9603

［65］ Zhang Chuanyao, Si Shijing, Wang Jianzong, et al. A review of federated meta-learning［J］. Big Data Research, 2023, 9(2): 2096−0271 (in Chinese)
(张传尧, 司世景, 王健宗, 等. 联邦元学习综述［J］. 大数据, 2023, 9(2): 2096−0271)

［66］ Liu Jian, Juuti M, Lu Yao, et al. Oblivious neural network predictions via minionn transformations［C］//Proc of the 24th ACM

SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 619−631

[67] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4401−4410

[68] Hore A, Ziou D. Image quality metrics: PSNR vs SSIM [C] //Proc of the 20th Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2010: 2366−2369

[69] Wang Zhou, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600−612

[70] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 586−595

[71] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [EB/OL]. 2014[2025-08-25]. https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

[72] Wierstra D, Schaul T, Glasmachers T, et al. Natural evolution strategies[J]. The Journal of Machine Learning Research, 2014, 15(1): 949−980

[73] Gretton A, Sejdinovic D, Strathmann H, et al. Optimal kernel choice for large-scale two-sample tests [EB/OL]. 2012[2025-08-30]. https://proceedings.neurips.cc/paper/2012/hash/dbe272bab69f8e13f14b405e038deb64-Abstract.html.

[74] Long Mingsheng, Cao Yue, Wang Jianming, et al. Learning transferable features with deep adaptation networks [C] //Proc of the 32nd Int Conf on machine learning. New York: PMLR, 2015: 97−105

[75] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of StyleGAN [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8110−8119

[76] Zhu Peihao, Abdal R, Qin Yipeng, et al. Improved StyleGAN embedding: Where are the good latents [J]. arXiv preprint, arXiv: 2012.09036, 2020

[77] Karras T, Aittala M, Laine S, et al. Alias-free generative adversarial networks[J]. Advances in Neural Information Processing Systems, 2021, 34: 852−863

[78] Xie Qizhe, Dai Zihang, Du Yulun, et al. Controllable invariance through adversarial feature learning [EB/OL]. 2017[2025-08-30]. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8cb22bdd0b7ba1ab13d742e22eed8da2-Abstract.html.

[79] Moyer D, Gao Shuyang, Brekelmans R, et al. Invariant representations without adversarial training [EB/OL]. 2018[2025-08-30]. https://proceedings.neurips.cc/paper/2018/hash/415185ea244ea2b2bedeb0449b926802-Abstract.html.

[80] Osia S A, Taheri A, Shamsabadi A S, et al. Deep private-feature extraction[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 32(1): 54−66

[81] Shokri R, Stronati M, Song Congzheng, et al. Membership inference attacks against machine learning models [C] //Proc of the 38th IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2017: 3−18

[82] Leino K, Fredrikson M. Stolen memories: Leveraging model memorization for calibrated White-Box membership inference [C] //Proc of the 29th USENIX Security Symp (USENIX Security'20). Berkeley, CA: USENIX Association, 2020: 1605−1622

[83] Song Liwei, Shokri R, Mittal P. Privacy risks of securing machine learning models against adversarial examples [C] //Proc of the 26th ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 241−257

[84] Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning [C] //Proc of the 40th IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2019: 739−753

[85] Liu Yiyong, Zhao Zhengyu, Backes M, et al. Membership inference attacks by exploiting loss trajectory [C] //Proc of the 29th ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2022: 2085−2098

[86] Carlini N, Chien S, Nasr M, et al. Membership inference attacks from first principles [C] //Proc of the 29th IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2022: 1897−1914

[87] Kolesnikov V, Kumaresan R, Rosulek M, et al. Efficient batched oblivious PRF with applications to private set intersection [C] //Proc of the 23rd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 818−829

[88] Pinkas B, Schneider T, Zohner M. Scalable private set intersection based on OT extension[J]. ACM Transactions on Privacy and Security, 2018, 21(2): 1−35

[89] Liu Junlin, Lyu Xinchen. Clustering label inference attack against practical split learning [J]. arXiv preprint, arXiv: 2203.05222, 2022

[90] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy [C] //Proc of the 23rd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 308−318

[91] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images [EB/OL]. 2009[2025-07-20]. http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

[92] Dosovitskiy A, Brox T. Generating images with perceptual similarity metrics based on deep networks [EB/OL]. 2016[2025-08-30]. https://proceedings.neurips.cc/paper/2016/hash/371bce7dc83817b7893bcdeed13799b5-Abstract.html.

[93] Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579−2605

[94] Chen Ting, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations [C] //Proc of the 37th Int Conf on Machine Learning. New York: PMLR, 2020: 1597−1607

[95] Osia S A, Shamsabadi A S, Sajadmanesh S, et al. A hybrid deep learning architecture for privacy-preserving mobile analytics[J]. IEEE Internet of Things Journal, 2020, 7(5): 4505−4518

[96] Wang Shiming, Ji Zhe, Xiang Liyao, et al. Crafter: Facial feature crafting against inversion-based identity theft on deep models [J]. arXiv preprint, arXiv: 2401.07205, 2024

[97] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks [C] //Proc of the 34th Int Conf on Machine Learning. New York: PMLR, 2017: 214−223

[98] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint, arXiv: 1412.6980, 2014

[99] Yang Mengda, Li Ziang, Wang Juan, et al. Measuring data reconstruction defenses in collaborative inference systems [J]. Advances in Neural Information Processing Systems, 2022, 35: 12855−12867

[100] Ant Group. SecretFlow [EB/OL]. [2025-07-20]. https://www.secretflow.org.cn/zh-CN/

[101] Huawei. MindSpore [EB/OL]. [2025-07-20]. https://www.mindspore.cn/

[102] WeBank. FATE [EB/OL]. [2025-07-20]. https://fate.readthedocs.io/en/latest/zh/

[103] OpenMined. PySyft [EB/OL]. [2025-07-20]. https://github.com/OpenMined/PySyft

[104] Alibaba. FederatedScope [EB/OL]. [2025-07-20]. https://federatedscope.io/

[105] Ng L K L, Chow S S M, Woo A P Y, et al. Goten: GPU-outsourcing trusted execution of neural network training [C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 14876−14883

[106] Han K, Hong S, Cheon J H, et al. Logistic regression on homomorphic encrypted data at scale [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 9466−9471

[107] Choi H, Woo S S, Kim H. Blind-touch: Homomorphic encryption-based distributed neural network inference for privacy-preserving fingerprint authentication [C] //Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 21976−21985

[108] Chen Chaochao, Zhou Jun, Wang Li, et al. When homomorphic encryption marries secret sharing: Secure large-scale sparse logistic regression and applications in risk control [C] //Proc of the 27th ACM SIGKDD Conf on Knowledge Discovery & Data Mining. New York: ACM, 2021: 2652−2662

[109] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [J]. Advances in Neural Information Processing Systems, 2020, 33: 6840−6851

[110] Zhang Jie, Chen Dongdong, Liao Jing, et al. Model watermarking for image processing networks [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 12805−12812

[111] Charette L, Chu Lingyang, Chen Yizhou, et al. Cosine model watermarking against ensemble distillation [C] //Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 9512−9520

[112] Haim N, Vardi G, Yehudai G, et al. Reconstructing training data from trained neural networks [J]. Advances in Neural Information Processing Systems, 2022, 35: 22911−22924

[113] Buzaglo G, Haim N, Yehudai G, et al. Deconstructing data reconstruction: Multiclass, weight decay and general losses [J]. Advances in Neural Information Processing Systems, 2023, 36: 51515−51535

[114] Nguyen T T, Huynh T T, Ren Zhao, et al. A survey of machine unlearning [J]. arXiv preprint, arXiv: 2209.02299, 2022

[115] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners [J]. Advances in Neural Information Processing Systems, 2020, 33: 1877−1901

[116] Kugler K, Münker S, Höhmann J, et al. Invbert: Reconstructing text from contextualized word embeddings by inverting the BERT pipeline [J]. arXiv preprint, arXiv: 2109.10104, 2021

[117] Morris J X, Kuleshov V, Shmatikov V, et al. Text embeddings reveal (almost) as much as text [J]. arXiv preprint, arXiv: 2310.06816, 2023

[118] Morris J X, Zhao Wenting, Chiu J T, et al. Language model inversion [J]. arXiv preprint, arXiv: 2311.13647, 2023



Liu Yaxin, born in 2001. Master candidate. Her main research interests include deep learning, AI privacy security.

刘雅欣, 2001 年生。硕士研究生。主要研究方向为深度学习、人工智能隐私安全。



Wang Juan, born in 1976. PhD, professor, PhD supervisor. Senior member of CCF. Her main research interests include system and trusted computing, AI security and vulnerability mining, cloud computing, and IoT and SDN security.

王　鹃, 1976 年生。博士，教授，博士生导师。CCF 高级会员。主要研究方向为系统与可信计算、人工智能安全与漏洞挖掘、云计算、物联网和 SDN 安全。



Yang Mengda, born in 1997. PhD. His main research interests include trustworthy deep learning and confidential computing.

杨梦达, 1997 年生。博士。主要研究方向可信深度学习、可信计算。



Li Zi'ang, born in 2000. PhD candidate. His main research interest includes trustworthy deep learning.

李子昂, 2000 年生。博士研究生。主要研究方向为可信深度学习。



Yi Wenzhe, born in 2001. PhD candidate. His main research interests include trustworthy deep learning and machine learning privacy security.

易文哲, 2001 年生。博士研究生。主要研究方向为可信深度学习、机器学习隐私安全。



Xu Xiaoyang, born in 1999. PhD candidate. His main research interests include machine learning privacy security and federated learning.

徐枭洋, 1999 年生。博士研究生。主要研究方向为机器学习隐私安全、联邦学习。



Chen Meihui, born in 1990. PhD. Her main research interests include data security, AI security, and privacy computing.

陈美会, 1990 年生。博士。主要研究方向为数据安全、AI 安全、隐私计算。



Wei Tao, born in 1975. PhD, visiting professor of Peking University. Member of CCF. His main research interests include data security, cyber security, AI security, and privacy computing.

韦　韬, 1975 年生。博士，北京大学客座教授。CCF 会员。主要研究方向为数据安全、网络安全、AI 安全、隐私计算。