

## 联邦学习深度梯度反演攻防研究进展

孙钰 严宇 崔剑\* 熊高剑 刘建华

<sup>①</sup>(北京航空航天大学网络空间安全学院 北京 100191)

<sup>②</sup>(空天网络安全工业和信息化部重点实验室(北京航空航天大学) 北京 100191)

**摘要:** 联邦学习作为一种“保留数据所有权, 释放数据使用权”的分布式机器学习方法, 打破了阻碍大数据建模的数据孤岛。然而, 联邦学习在训练过程中只交换梯度而不交换训练数据的特点并不能保证用户训练数据的机密性。近年来新型的深度梯度反演攻击表明, 敌手可从共享梯度中重建用户的私有训练数据, 从而对联邦学习的私密性产生了严重威胁。随着梯度反演技术的演进, 敌手从深层网络恢复大批量原始数据的能力不断增强, 甚至对加密梯度的隐私保护联邦学习(PPFL)发起了挑战。而有效的针对性防御方法主要基于扰动变换, 旨在混淆梯度、输入或特征以隐藏敏感信息。该文首先指出了隐私保护联邦学习的梯度反演漏洞, 并给出了梯度反演威胁模型。之后从攻击范式、攻击能力、攻击对象3个角度对深度梯度反演攻击进行详细梳理。随后将基于扰动变换的防御方法依据扰动对象的不同分为梯度扰动、输入扰动、特征扰动3类, 并对各类方法中的代表性工作进行分析介绍。最后, 对未来研究工作进行展望。

**关键词:** 联邦学习; 梯度反演; 数据重建; 标签恢复; 扰动变换

中图分类号: TN918; TP301

文献标识码: A

文章编号: 1009-5896(2024)02-0428-15

DOI: [10.11999/JEIT230541](https://doi.org/10.11999/JEIT230541)

## Review of Deep Gradient Inversion Attacks and Defenses in Federated Learning

SUN Yu YAN Yu CUI Jian XIONG Gaojian LIU Jianhua

<sup>①</sup>(School of Cyber Science and Technology, Beihang University, Beijing 100191, China)

<sup>②</sup>(Key Laboratory of Ministry of Industry and Information Technology for Cyberspace Security  
(Beihang University), Beijing 100191, China)

**Abstract:** As a distributed machine learning approach that preserves data ownership while releasing data usage rights, federated learning overcomes the challenge of data silos that hinder large-scale modeling with big data. However, the characteristic of only sharing gradients without training data during the federated training process does not guarantee the confidentiality of users' training data. In recent years, novel deep gradient inversion attacks have demonstrated the ability of adversaries to reconstruct private training data from shared gradients, which poses a serious threat to the privacy of federated learning. With the evolution of gradient inversion techniques, adversaries are increasingly capable of reconstructing large volumes of data from deep neural networks, which challenges the Privacy-Preserving Federated Learning (PPFL) with encrypted gradients. Effective defenses mainly rely on perturbation transformations to obscure original gradients, inputs, or features to conceal sensitive information. Firstly, the gradient inversion vulnerability in PPFL is highlighted and the threat model in gradient inversion is presented. Then a detailed review of deep gradient inversion attacks is conducted from the perspectives of paradigms, capabilities, and targets. The perturbation-based defenses are divided into three categories according to the perturbed objects: gradient perturbation, input perturbation, and feature perturbation. The representative works in each category are analyzed in detail. Finally, an outlook on future research directions is provided.

**Key words:** Federated learning; Gradient inversion; Data reconstruction; Label restoration; Perturbation transformation

收稿日期: 2023-06-01; 改回日期: 2023-12-01; 网络出版: 2023-12-23

\*通信作者: 崔剑 [cuijianw@buaa.edu.cn](mailto:cuijianw@buaa.edu.cn)

基金项目: 国家自然科学基金(32071775)

Foundation Item: The National Natural Science Foundation of China (32071775)

## 1 引言

机器学习在自然语言处理、计算机视觉、推荐系统等领域取得了巨大的成功，并不断推动传统产业和社会生活的变革<sup>[1,2]</sup>。机器学习是数据驱动的，训练过程需要融合不同用户的本地数据。然而，随着欧盟《数据隐私保护条例》、美国《美国数据隐私和保护法》、我国《网络安全法》《数据安全法》和《个人信息保护法》等保护数据隐私法规的颁布实施，“数据孤岛”成为制约大数据建模的关键影响因素。为破解隐私保护与数据要素流动相悖之局，方滨兴院士<sup>[3]</sup>提出了“数据不动程序动、数据可用不可见”的数据要素安全流动机理。作为一种“保留数据所有权，释放数据使用权”的机器学习方法，联邦学习<sup>[4]</sup>允许多个数据持有方在满足隐私保护、数据安全和相关法规，并可协同进行数据开发利用和机器学习建模<sup>[5]</sup>，在有效破解“数据孤岛”问题的同时，保证了数据“可用不可见”<sup>[6]</sup>。近年来，联邦学习技术在数据要素市场化的政策下得到了快速发展，形成了FATE<sup>[7]</sup>，PaddleFL<sup>[8]</sup>，TensorFlow Federated<sup>[9]</sup>，PySyft<sup>[10]</sup>等开源框架，并在工业<sup>[11]</sup>、医疗<sup>[12,13]</sup>、物联网<sup>[14]</sup>、金融<sup>[15,16]</sup>等数据要素流通领域涌现出众多实际应用。

作为一种分布式机器学习技术，联邦学习将多个客户端在中心服务器的协调下共同训练一个全局的神经网络模型。联邦学习中的每个客户端不直接交换他们的私有训练数据，而是上传其在本地的训练集上产生的局部梯度并由中心服务器对其进行聚合。联邦学习的训练过程如下：中心服务器首先向每个参与训练的客户端发送一个全局的初始化神经网络模型。客户端使用本地数据完成训练后，将其产生的局部梯度上传至服务器。然后，服务器聚合每个客户端上传的局部梯度，并将聚合得到的全局梯度发送给每个客户端。最后，各客户端使用全局梯度更新自己的本地模型。通过不断迭代上述过程直至模型收敛，完成训练。

联邦学习仅上传梯度不上传原始数据的特性在一定程度上保护了数据隐私，但在用户与服务器进行梯度交互时暴露出新的攻击面。从输入与标签中推导而来的梯度隐藏着可用于反演原始数据的重要信息，因此梯度反演敌手可利用交互的梯度信息重建原始训练数据。早期的梯度反演仅能对训练集中是否包含特定数据样本(成员推断)<sup>[17]</sup>、部分训练数据的属性(属性推断)<sup>[18]</sup>、训练数据类别代表<sup>[19]</sup>等信息进行推理，而自2019年起出现的深度梯度反演攻击<sup>[20]</sup>可从梯度信息中恢复出原始数据(数据重建)及其标签(标签恢复)，完全暴露了参与方的敏感数据，极具威胁性。

为了保护用户的数据隐私，隐私保护联邦学习(Privacy-Preserving Federated Learning, PPFL)通常采用同态加密、安全多方计算等密码学技术对客户端本地训练产生的局部梯度进行处理，使得明文梯度信息在传输及聚合的过程中不会被服务器或未参与训练的敌手获取。该方案不仅带来了巨大的计算开销，并且随着梯度反演技术的演进，敌手从深层网络恢复大批量原始数据的能力不断增强，半诚实用户可对聚合后梯度进行反演，PPFL的安全性受到严重威胁。因此，为应对能力愈发强大的梯度反演攻击，各类基于扰动变换的针对性防御技术纷纷涌现，根据扰动对象的不同可分为梯度扰动、输入扰动、特征扰动，分别用于混淆梯度、输入或特征以隐藏敏感信息。

本文的结构如下：第2节简要介绍采用密码学技术的PPFL中存在的梯度反演漏洞及梯度反演威胁模型。第3节从攻击范式、攻击能力、攻击对象3个角度对梯度反演进行分类，并对深度梯度反演攻击进行介绍。第4节介绍基于扰动变换的梯度反演防御，并根据扰动对象的不同对其进行分类。第5节对梯度反演攻防的未来研究方向进行展望。

## 2 梯度反演威胁模型

当前PPFL通过使用基于密码技术的安全聚合方案，使得原始梯度对服务器不可见，聚合过程不泄露用户的局部梯度信息，具体包括同态加密<sup>[21-25]</sup>、秘密分享<sup>[26-30]</sup>等技术。同态加密允许直接对密文进行计算以生成密文结果，且解密结果与在对应的明文上进行计算的结果相同，已在许多联邦学习方法中得到了广泛的应用。Phong等人<sup>[21]</sup>利用同态加密技术对局部梯度进行保护，防止服务器获得明文梯度。Zhang等人<sup>[23]</sup>提出了基于批量同态加密的Batch-Crypt，将一批梯度统一编码后再加密以降低计算开销。Zhang等人<sup>[25]</sup>提出了基于同态加密的PEFL，利用分布式选择性随机梯度下降方法来降低计算开销。秘密分享技术将秘密以适当的方式拆分，拆分后的每一个份额由不同的参与者管理，单个参与者无法恢复秘密信息，只有若干个参与者一同协作才能重构秘密。Bonawitz等人<sup>[26]</sup>利用秘密分享机制为梯度添加掩码，用户数量满足门限后掩码可互相抵消，服务器只能获得聚合后的全局梯度，但不能显式地访问任何本地更新。Xu等人<sup>[27]</sup>提出VerifyNet，为服务器添加了可验证功能，并设计秘密分享协议用于保护客户端的局部梯度。

当前PPFL方案的威胁模型中只考虑了半诚实(honest but curious)服务器和外部敌手，忽视了来自半诚实用户及恶意服务器的深度梯度反演威胁。

实际上,随着梯度反演敌手能力的不断增强,通过构造恶意模型权重破坏安全聚合、基于强先验知识对大批量数据的平均梯度进行反演已成为可能,仅采用密码学技术对客户端产生的局部梯度进行保护已远远不够。明确联邦学习中的梯度反演威胁模型,刻画梯度反演敌手的攻击目标以及所具备的知识和能力,是所有梯度反演防御方案设计的基础。因此,本文将梯度反演威胁模型扩展至半诚实/恶意服务器和半诚实用户,如图1所示。根据攻击能力的强弱,即是否可以恶意篡改模型结构与权重,将梯度反演敌手分为(1)被动攻击者:半诚实服务器和半诚实用户,(2)主动攻击者:恶意服务器。攻击者所具备的知识和能力具体如下:

(1)已知模型的结构、权重以及批次大小(batch size)等训练超参数。部分攻击方法还假设敌手拥有批归一化(Batch Normalization, BN)统计信息、辅助数据集、预训练的生成对抗网络(Generative Adversarial Network, GAN)等辅助信息。

(2)服务器可获得各个用户本地训练批次上的平均梯度,而半诚实用户及安全聚合场景下的服务器可获得聚合后的全局梯度,并尝试从梯度中反演出原始训练数据。

(3)半诚实服务器、半诚实用户会忠实地执行联邦学习训练流程;但恶意服务器可通过恶意构造模型的结构及权重进一步诱导用户的隐私数据泄露。

### 3 深度梯度反演攻击研究进展

本文从以下3个角度对梯度反演攻击进行了分类,如图2所示。本节主要从攻击范式角度分别在3.1节与3.2节对迭代优化法和解析法两类攻击范式展开介绍。

(1)攻击范式:迭代优化法将待解数据作为可训练参数,以最小化待解数据产生的梯度与梯度真值的距离(梯度匹配目标)和对待解输入的正则化约

束为目标,对待解数据进行迭代优化直至重建出原始训练数据。而解析法则通过建立梯度与训练数据间的线性关系、分类层梯度与标签间的关系等方法求解训练数据及标签。本节主要从攻击范式的角度分别对两类梯度反演攻击进行介绍。

(2)攻击能力:被动攻击者(半诚实服务器、半诚实用户)忠实地执行训练流程,尝试从其得到的平均梯度中重建训练数据。主动攻击者(恶意服务器)通过构造恶意的模型结构、权重进一步诱导用户的隐私数据泄露。

(3)攻击对象:梯度反演的攻击对象包括数据重建和标签恢复。当前的梯度反演攻击研究集中在数据重建,包括计算机视觉场景中的图像数据重建和自然语言处理场景中的文本、语音数据重建等。

#### 3.1 基于迭代优化的梯度反演攻击

在图像数据重建方面,Zhu等人<sup>[20]</sup>最早提出梯度反演攻击DLG(Deep Leakage from Gradients),其以模型参数在待解数据和标签上的梯度 $\nabla \mathbf{W}'$ 与梯度真值 $\nabla \mathbf{W}$ 间的欧式距离作为梯度匹配目标,在上述梯度匹配目标下通过L-BFGS优化器同时对待解数据 $\mathbf{x}_{\text{rec}}$ 和待解标签 $\mathbf{y}_{\text{rec}}$ 进行迭代优化,优化目标如式(1)所示。DLG将待解数据和待解标签同时作为可训练参数进行优化,二者的重建过程会相互影响,因此仅在批次大小不大于8且图片分辨率在 $64 \times 64$ 以内时能取得优异的重建效果。另外,DLG仅能在采用均匀分布初始化模型权重的情况下对模型训练早期阶段产生的梯度进行攻击。

$$\arg \min_{\mathbf{x}_{\text{rec}}, \mathbf{y}_{\text{rec}}} \|\nabla \mathbf{W}' - \nabla \mathbf{W}\|^2 \quad (1)$$

Wang等人<sup>[31]</sup>观察到在相同训练数据下,与均匀分布初始化的模型权重的训练梯度相比,正态分布初始化的模型权重的训练梯度会更趋于0,并且随机初始化的待解数据产生的梯度的幅值远小于梯度真值。因此在正态分布初始化模型权重的情况

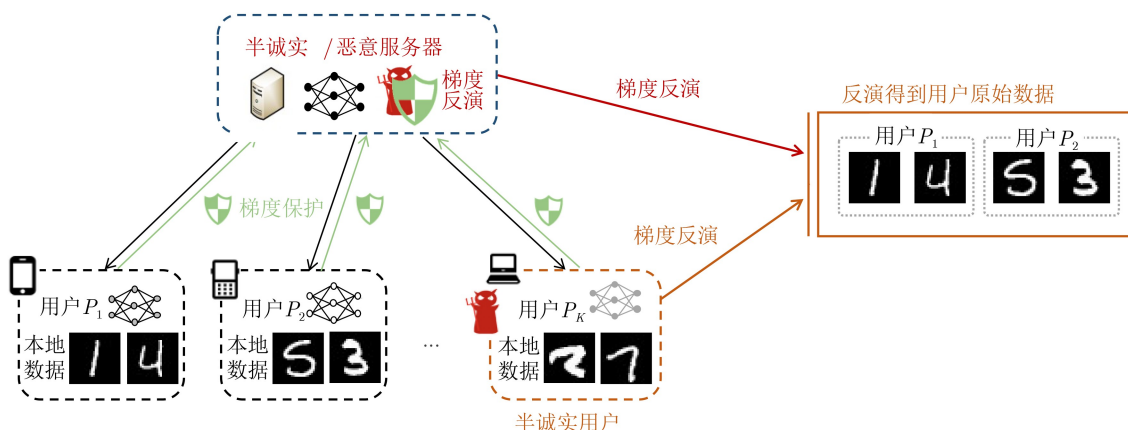


图1 梯度反演威胁模型

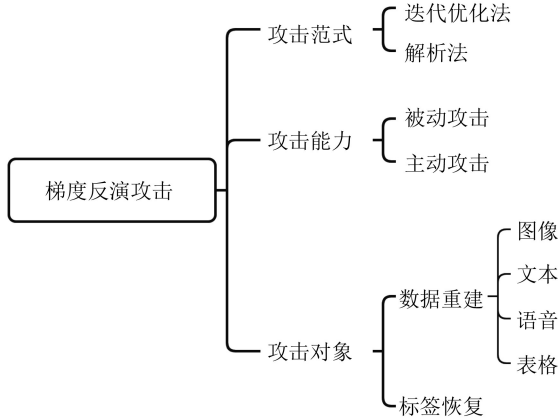


图2 梯度反演攻击分类

下，DLG采用欧式距离作为梯度匹配目标会造成数据重建过程主要由小部分幅值较大的梯度驱动，从而攻击失效。针对上述问题，Wang等人<sup>[31]</sup>提出了一种自适应的攻击方法SAPAG，其使用了基于加权高斯核的梯度匹配目标如式(2)所示，使得数据重建过程由大部分梯度驱动，而不是由一小部分幅值较大的梯度驱动。其中 $Q$ 为梯度匹配中各层的权重(随着层数的增加而减小)， $\sigma^2$ 为每层梯度真值的方差。与DLG<sup>[20]</sup>相比，SAPAG在不同网络、不同模型权重初始化方法、不同训练阶段上具有更好的泛化能力。

$$D(\nabla \mathbf{W}', \nabla \mathbf{W}) = Q \cdot \left( 1 - \exp \left( \frac{-\|\nabla \mathbf{W}' - \nabla \mathbf{W}\|^2}{\sigma^2} \right) \right) \quad (2)$$

Wei等人<sup>[32]</sup>评估了不同的攻击参数以及联邦学习中的超参数对梯度反演攻击的影响，包括待解数据的初始化方法、攻击终止条件、优化方法、批次大小、输入图片的分辨率、激活函数、梯度稀疏率，并且提出了CPL(Client Privacy Leakage)攻击，在恢复出标签后，采用与原始数据相同类别的图片初始化待解数据，并在目标函数中加入了标签正则化以提高重建数据在迭代优化过程中的稳定性，如式(3)所示。

$$\arg \min_{\mathbf{x}_{\text{rec}}} \|\nabla \mathbf{W}' - \nabla \mathbf{W}\|^2 + \alpha \|\mathbf{F}(\mathbf{x}_{\text{rec}}) - \mathbf{y}_{\text{rec}}\|^2 \quad (3)$$

Geiping等人<sup>[33]</sup>认为在梯度匹配中梯度的方向比幅值更重要，提出IG(Inverting Gradients)算法将梯度匹配目标替换为余弦相似度，并引入了自然图像先验全变分(Total Variance, TV)正则项，以保障重建图像的局部平滑性，总的优化目标如式(4)所示。该方法显著改进了在分辨率为 $224 \times 224$ 的ImageNet数据集上对单张图像的重建效果，但对批次大小大于8的批量恢复仍然十分困难。

$$\arg \min_{\mathbf{x}_{\text{rec}}} \left( 1 - \frac{\nabla \mathbf{W}' \cdot \nabla \mathbf{W}}{\|\nabla \mathbf{W}'\| \cdot \|\nabla \mathbf{W}\|} + \alpha \cdot \text{TV}(\mathbf{x}_{\text{rec}}) \right) \quad (4)$$

全变分正则对输入数据的约束较弱，为在大批量高分辨率图片上取得更好的重建效果则需引入更强的数据先验，包括对批归一化层统计量的正则约束和利用生成对抗技术预训练编码的生成式图像先验。相关工作包括Yin等人<sup>[34]</sup>提出的GI(Grad Inversion)及面向ViT模型的GradViT<sup>[35]</sup>，Jeon等人<sup>[36]</sup>提出的GIAS(Gradient Inversion in Alternative Spaces)和Li等人<sup>[37]</sup>提出的GGL(Generative Gradient Leakage)。

GI<sup>[34]</sup>的优化目标在梯度匹配的基础上还增加了保真度正则和组一致性正则。保真度正则用于将待解数据在迭代优化中导向自然图像，既包括了惩罚待解数据的总方差及L2范数的全变分正则和L2正则，更关键的是对批归一化统计信息的正则约束 $R_{\text{BN}}$ 如式(5)所示，其中 $\mu_l(\mathbf{x}_{\text{rec}})$ 和 $\sigma_l^2(\mathbf{x}_{\text{rec}})$ 分别为待解数据在第 $l$ 个卷积层的特征图的均值和方差。GI还同时利用多个随机种子联合优化，在优化过程中多路径联合探索以扩大搜索空间，并利用组一致性正则惩罚其中偏离“共识”的图像。

$$R_{\text{BN}}(\mathbf{x}_{\text{rec}}) = \sum_l \|\mu_l(\mathbf{x}_{\text{rec}}) - \text{BN}_l(\text{mean})\|^2 + \sum_l \|\sigma_l^2(\mathbf{x}_{\text{rec}}) - \text{BN}_l(\text{variance})\|^2 \quad (5)$$

GI<sup>[34]</sup>将数据重建批次大小的上限提升至48，然而由于并非所有的联邦学习框架都支持向服务器报告批归一化统计信息，其适用性大大受限。之后，GradViT<sup>[35]</sup>将GI方法迁移至ViT模型上，取得了相比CNN模型更好的攻击效果。然而，文献<sup>[38]</sup>指出GI中批归一化层统计信息和标签已知的假设太强，放松这些假设则会极大地削弱攻击效果。

基于GAN先验的GIAS<sup>[36]</sup>，GGL<sup>[37]</sup>和HCGLA<sup>[39]</sup>则将待解数据的搜索范围由像素空间缩小到低维的GAN隐空间 $\mathbf{z}$ ，从而显著减小了优化难度。其中GGL直接在隐空间 $\mathbf{z}$ 上进行优化，并且评估了在梯度裁剪、梯度稀疏、Soterial<sup>[40]</sup>特征扰动4种防御防御方法下的重建效果。而GIAS在完成隐变量的迭代优化后，还进一步优化生成模型的参数以减少生成模型本身带来的数据重建误差。HCGLA则通过GAN模型初始化待解数据，并利用基于CNN的降噪模型进一步提高待解数据的图像质量。然而此类基于GAN的方法需动用全部训练集训练GAN的生成器，而攻击效果往往仅能恢复同类样本，存在如图3最后一行末尾3个案例所示的严重失真。

在文本数据重建方面，DLG<sup>[20]</sup>在掩码语言模型

(Masked Language Model)上首次对文本数据重建进行了探索,其采用欧氏距离作为梯度匹配目标并在嵌入空间上进行迭代优化,再根据优化完成后的嵌入在嵌入矩阵中找到最接近的原始词。Deng等人<sup>[41]</sup>提出的TAG则将L2范数与L1范数结合作为梯度匹配目标。其在基于Transformer的语言模型上可实现比DLG更好的重建效果。然而DLG和TAG两种方法仅简单地将图像领域的的数据重建方法应用至文本数据,未针对文本数据进行改进,并且均局限于批次大小为1的单文本数据重建。

Balunovic等人<sup>[42]</sup>提出的LAMP交替进行嵌入空间上的连续优化和文本顺序上的离散优化,首次在批次大小大于1的情况下对多文本进行重建。在连续优化阶段,为解决Transformer自注意力层的点乘操作引起梯度匹配目标优化过程中嵌入向量长度持续增长的问题,LAMP在优化目标中引入了对嵌入长度约束的正则化项 $L_{\text{reg}}(\mathbf{x})$ 如式(6)所示,使

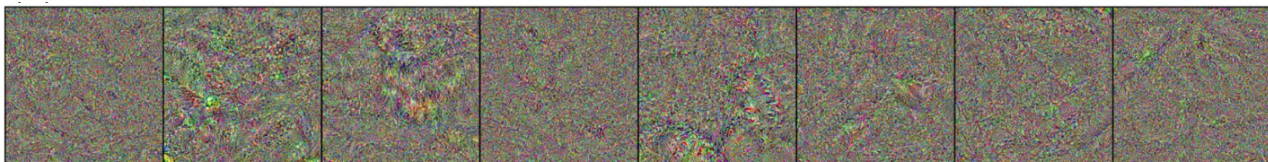
得重建序列的嵌入平均长度接近于词汇表中嵌入的平均长度。虽然连续优化通常可成功恢复词嵌入,但其顺序可能有误。因此在离散优化阶段,LAMP通过词交换、词移动、子序列移动、前缀移动等离散序列变换方法产生候选序列,并选择其中低重建损失低困惑度(可通过GPT-2或其它语言模型计算)的序列,以避免优化过程陷入局部最优。

$$L_{\text{reg}}(\mathbf{x}) = \left( \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i\|^2 - \frac{1}{V} \sum_{j=1}^V \|e_j\|^2 \right)^2 \quad (6)$$

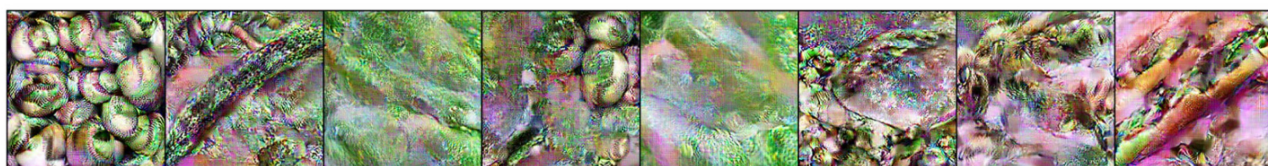
在语音处理领域,深度学习模型通常是在原始语音信号中提取出的梅尔频谱(Mel-spectrogram)、梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients, MFCC)等特征上进行训练。因此,Li等人<sup>[43]</sup>提出了一种两阶段的语音波形数据重建方法。在特征重建阶段,首先通过优化梯度匹配目标及对特征的各向异性全变分正则重建语音特征。而在波形重



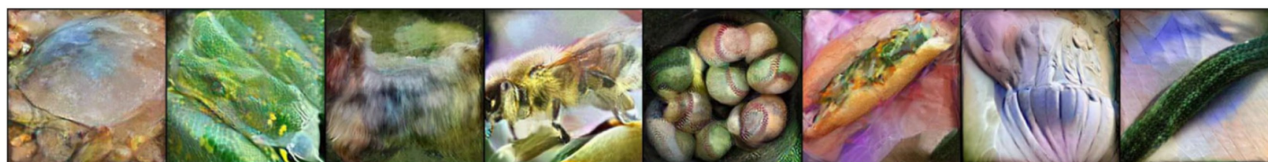
(a) 大批量原始数据



(b) DLG重建结果



(c) IG重建结果



(d) GI重建结果



(e) GGL重建结果

图3 基于迭代优化的典型数据重建方法对大批量图像数据的反演效果

建阶段，则根据第一阶段提取出的梅尔频谱或MFCC特征重建语音波形。梅尔频谱可在转换回时域信号后，使用Griffin-Lim算法重建语音波形。而对于MFCC，则需要先应用逆离散余弦变换将其转换为梅尔频谱，再通过梅尔频谱重建语音波形。

在表格数据重建方面，Vero等人<sup>[44]</sup>首次提出了对表格数据的重建攻击TabLeak。由于表格中既包含离散类别数据也包括连续的数值数据，TabLeak通过对类别数据真值进行独热编码，并利用softmax函数将重建的类别数据限制为0~1之间的实数，从而将其完全转化为连续优化问题。为避免优化过程陷入局部最优，TabLeak通过运行具有不同初始化的独立优化过程并对它们的优化结果进行特征粒度的中位数池化以得到最终重建结果。注意到平均训练梯度是表格中多行数据产生的梯度的均值并且与顺序无关，因此在进行中位数池化前还需利用匈牙利算法对每个独立优化结果中的数据进行重新排序。即使在批次大小达到128的情况下，TabLeak也可以90%以上的准确率恢复其中25%的数据。

## 3.2 基于解析的梯度反演攻击

### 3.2.1 基于被动解析的数据重建

Phong等人<sup>[21]</sup>首先发现对于包含偏置项的全连接层，其输入可通过计算输出神经元与输入的连接权重的梯度与偏置梯度的商直接恢复，即 $\mathbf{x} = \nabla \mathbf{W} / \nabla \mathbf{b}$ 。但该方法只适用于批次大小为1的单样本情况，且不适用于卷积网络或无偏置项的全连接网络。

Zhu等人<sup>[45]</sup>提出的R-GAP以及Chen等人<sup>[46]</sup>提出的COPA结合网络的前向过程和反向传播，将神经网络输入数据的重建问题分解为自后向前逐层依次联立权重约束、梯度约束方程组求解的子问题，权重约束和梯度约束分别如式(7)、式(8)所示，其中 $\mathbf{Z}_i$ 和 $\nabla \mathbf{Z}_i$ 分别为特征图及其梯度。R-GAP和COPA将解析法数据重建推广到卷积层网络，但仍然只适用于单样本情况。

$$\mathbf{W}_i \cdot \mathbf{X}_i + \mathbf{B}_i = \mathbf{Z}_i \quad (7)$$

$$\nabla \mathbf{Z}_i \cdot \nabla \mathbf{X}_i = \nabla \mathbf{W}_i \quad (8)$$

Kariyappa等人<sup>[47]</sup>提出的CPA(Cocktail Party Attack)将迭代优化法和解析法相结合。观察到全连接层的训练梯度是输入的线性组合，即 $\nabla \mathbf{W}' = \mathbf{A}\mathbf{X}$ ，CPA攻击将从全连接层的平均梯度重建单样本输入问题转换为信号处理领域的盲源分离问题，从而利用独立成分分析(Independent Component Analysis, ICA)算法重建全连接层输入。而对于在全连接层之前还包含卷积层等其它结构的模型，则还需对重建得到的全连接层输入进行特征反演，与梯度

反演重建网络输入相似，但特征反演迭代优化过程中以模型的输出作为匹配目标。

在文本数据重建方面，Gupta等人<sup>[48]</sup>提出的FILM攻击则进一步展示了对多达128个句子的大量文本重建的可行性。与之前通过梯度匹配目标对词嵌入进行迭代优化的方法不同，FILM首先根据非零的平均梯度的位置恢复句子中出现的所有词语，再通过波束搜索(beam search)重建句子，最后基于句子的困惑度和梯度范数对句子重新排序。

### 3.2.2 基于主动解析的数据重建

Lam等人<sup>[49]</sup>在采用安全聚合的跨设备联邦学习场景下，提出了一种将聚合梯度分解为各个客户端产生的局部梯度的方法，从而破坏安全聚合。其假设恶意服务器已知来自设备分析的总结信息(用户参与训练的频率)，可在多轮训练过程中将模型的权重固定并收集聚合梯度，之后将聚合梯度的分解问题表示为一个约束二元矩阵分解问题，如式(9)所示，其中 $\mathbf{P}$ 为各用户在多轮训练中的参与矩阵， $\mathbf{noise}$ 为用户训练数据改变等引起的梯度变化。

$$\nabla \mathbf{W}'_{\text{aggregated}} = \mathbf{P} \nabla \mathbf{W}'_{\text{individual\_avg}} + \mathbf{noise} \quad (9)$$

Boenisch等人<sup>[50]</sup>进一步利用了采用ReLU激活的全连接层输入与梯度间的线性关系。由于ReLU激活函数的单侧抑制性质，一个训练样本输入至ReLU激活的全连接层时仅能通过其中的一部分神经元(即输出大于0，称为激活态)，而其余神经元输出被ReLU抑制为0(称为休眠态)，该训练样本在休眠态神经元上产生零梯度。因此，当某个神经元只被训练批次中的一个样本独占激活时，该神经元产生的梯度只来自于这个样本，可通过计算该独占神经元的权重梯度与偏置梯度的商完美恢复该样本，即 $\mathbf{x} = \nabla \mathbf{W} / \nabla \mathbf{b}$ 。然而，当批次大小较大时，独占神经元的数量会迅速减少，只有少数训练样本可通过以上方法恢复。因此，Boenisch等人<sup>[50]</sup>通过恶意构造模型权重，诱导出更多的独占神经元，使得可重建的样本数量提升至随机模型权重的两倍以上。其从高斯正态分布中采样，将模型权重行中的一半分量选择为负值，另一半选择为正值，并且负权重采样的绝对值略大于正权重，以减少多个训练样本激活同一个神经元的可能性。当全连接层前还包含卷积层时，则还需修改卷积核参数，使卷积层输入恒等地传递到输出。

Wen等人<sup>[51]</sup>提出了基于恶意的分类层模型权重的两种“钓鱼”策略，使得平均梯度退化到在单个目标样本上训练产生的梯度。其中，类钓鱼策略通过增大最后一个分类层中非目标类神经元的偏置项，降低了网络对目标类的预测置信度，从而显著

增加了目标类训练数据对平均训练梯度的贡献。当训练批次包含多个目标类别的训练样本时, 还需采用特征钓鱼策略调整分类层中目标特征连接到目标类别神经元输出的权重和偏置项的大小, 改变目标特征的决策边界, 降低网络对目标图像的预测的置信度, 放大目标图像的梯度。

Pasquini等人<sup>[52]</sup>提出了一种基于模型不一致的梯度抑制攻击, 使得聚合梯度退化为目标用户的梯度, 从而破坏安全聚合。具体来说, 其向目标用户发送正常的模型权重, 使其产生正常的局部梯度。对于非目标用户, 则利用ReLU神经元未被激活时产生零梯度的性质, 通过发送的恶意模型权重(例如将全连接层的连接权重设置为0, 偏置项设置为负数)使其产生零梯度。

Fowl等人<sup>[53]</sup>提出了一种基于恶意模型权重的线性层梯度分离方法, 首先对训练图像数据集的一个线性统计量(以平均亮度 $h(\mathbf{x})$ 为例)的累积分布函数 $\Phi(\cdot)$ 进行估计, 向模型前插入一个采用ReLU激活的包含 $k$ 个输出神经元的全连接层(印记模块), 并利用输入与梯度间的线性关系。但与Boenisch等人<sup>[50]</sup>通过恶意的模型权重暴露出更多的独占神经元不同, Fowl等人<sup>[53]</sup>的恶意模型权重使得线性层第 $i$ 个神经元的输出值为输入图像的亮度加上一个通过偏置项控制的偏移量 $b^i = -\Phi^{-1}(i/k) = -c_i$ 。那么对于一个亮度满足不等式 $c_l \leq h(\mathbf{x}_t) \leq c_{l+1}$ 的图像 $\mathbf{x}_t$ , 如果训练批次中没有其它图片满足该亮度不等式, 那么该图像可通过式(10)完美重建, 其中 $\nabla \mathbf{W}^l$ 和 $\nabla b^l$ 分别为第 $l$ 个输出神经元的权重梯度和偏置梯度。

$$\mathbf{x}_t = \frac{\nabla \mathbf{W}^l - \nabla \mathbf{W}^{l+1}}{\nabla b^l - \nabla b^{l+1}} \quad (10)$$

Zhao等人<sup>[54]</sup>指出了Fowl等人<sup>[53]</sup>工作中的一个错误观点, 即对聚合梯度的攻击与攻击单个用户的局部梯度方法相同, 因此当参与训练的用户数增加时, 为从聚合梯度中重建原始训练数据, 所需插入的印记模块的尺寸也随之增加, 造成额外开销。为解决上述问题, Zhao等人<sup>[54]</sup>在印记模块前额外插入了一个卷积层并为不同用户分配不同的恶意卷积核参数, 使得不同用户的训练数据恒等映射到卷积层输出的不同位置, 并且仅将印记模块连接到当前用户训练数据卷积层输出的连接权重设置为非零值, 从而将不同用户在印记模块产生的权重梯度分离。至此, 印记模块的尺寸只与批次大小有关, 而与用户数无关, 大大减少了参数量和计算量。

Pan等人<sup>[55]</sup>提出的NEA(Neuron Exclusivity Analysis)则是首先利用ReLU激活的全连接层的梯度特点求解训练批次中每个样本在采用ReLU激活函

数的全连接网络中的神经元激活状态, 从而将ReLU激活函数等价于激活矩阵乘法, 之后为全连接网络的输入与梯度建立线性方程组并求解。由于该攻击方法需满足充分独占性条件, 即训练批次中的每个样本在最后一个ReLU层上至少有2个独占神经元, 并且在其他ReLU层上至少有1个独占神经元, 而该条件在批次大小较大时很难满足。为此, NEA假设恶意服务器可修改用户的数据预处理算法, 通过向训练批次添加特定扰动, 使其满足充分独占性条件。

Fowl等人<sup>[56]</sup>将主动解析攻击应用至基于Transformer的语言模型中, 提出了一种基于恶意模型权重的文本数据重建方法Decepticons。对于单样本重建, 首先禁用Transformer模型中所有的注意力层和大部分的前馈模块输出(只保留1个输出), 使得网络不会混合嵌入信息, 即每一层的输入都是相同的令牌嵌入加上位置嵌入。由于前馈模块为全连接层, 之后可通过基于恶意模型权重的线性层梯度分离方法<sup>[53]</sup>将各个词嵌入在前馈模块的梯度分离并重建输入。而当批次大小大于1时, 简单地使用上述策略仅能重建词嵌入及其在句子中的位置, 但无法恢复其具体在哪个句子中。为此, 他们进一步提出利用注意力机制, 通过构造注意力层的权重将序列信息编码到嵌入中, 从而消除序列歧义。

### 3.2.3 标签恢复

Zhao等人<sup>[57]</sup>率先提出一种单样本的标签解析算法, 将数据重建与标签恢复分离以改进DLG攻击的数据重建效果, 称作iDLG。iDLG<sup>[57]</sup>观察到在单样本情形下分类网络最后一个全连接层的梯度符号揭示了样本标签的存在, 即输出层中对应真实标签的神经元与前层神经元连接的权重梯度与其它标签对应神经元的梯度是相反的, 因此可通过模型输出层权重的梯度符号确定真实标签, 如式(11)所示, 其中 $\nabla \mathbf{W}_L^i$ 为模型输出层第 $i$ 个神经元与前层神经元连接权重的梯度向量。特别地, 当模型输出层前为非负激活函数时, 该方法可简化为寻找梯度小于0的输出层神经元。

$$\mathbf{y}_{\text{rec}} = i, \text{ s.t. } (\nabla \mathbf{W}_L^i)^T \cdot \nabla \mathbf{W}_L^j \leq 0, \forall j \neq i \quad (11)$$

虽然在批次大小为1并且模型输出层之前采用了非负激活函数时, 模型的输出层神经元中有且仅有对应真实标签的神经元与前层神经元连接权重的梯度为负值, 但当批次大小大于1时, 模型权重的梯度是所有训练样本产生的梯度的叠加, 这些信息可能会在梯度叠加的过程中丢失。但是, Yin等人<sup>[32]</sup>注意到单个训练样本在模型输出层中对应真实标签的神经元与前层连接权重的梯度绝对值大小通常会远大于其它神经元, 这使得在各训练样本产生的梯

度在叠加的过程中，各样本在真实标签的输出层神经元上产生的负梯度在叠加后基本仍能保持负号。因此，GI<sup>[34]</sup>将输出层各个神经元与前层神经元连接权重中的最小梯度进行排序，并取其中梯度最小的 $K$ 个神经元的位置作为恢复出的真实标签，如式(12)所示，其中 $K$ 为批次大小。GI<sup>[34]</sup>在模型输出层前采用非负激活函数的条件下，对iDLG<sup>[57]</sup>中的标签恢复方法进行了改进，使其能扩展到批次大小大于1的情况，但要求训练批次中无重复标签。

$$\mathbf{y}_{\text{rec}} = \arg \text{sort}(\min_i \nabla \mathbf{W}_L^i)[:K] \quad (12)$$

Dang等人<sup>[58]</sup>提出的RLG (Revealing Labels from Gradients)算法提供了标签恢复的另一思路。RLG通过奇异值分解，将模型输出层权重的梯度分解为 $\nabla \mathbf{W} = \mathbf{P}\Sigma\mathbf{Q}$ ，其中 $\mathbf{P}$ 与 $\mathbf{Q}$ 均为正交矩阵， $\Sigma \in \mathbf{R}^{S \times S}$ 是一个在对角线上有非负元素的对角矩阵。RLG证明如果待恢复的批量标签中包含标签 $c$ ，则存在向量 $\mathbf{r} \in \mathbf{R}^S$ 使得 $\mathbf{r} \cdot \mathbf{q}^c < 0$ 而 $\mathbf{r} \cdot \mathbf{q}^{j \neq c} > 0$ ，其中 $\mathbf{q}^j$ 为 $\mathbf{Q}$ 的第 $j$ 列，也就是存在超平面 $\mathbf{r} \cdot \mathbf{x} = 0$ 可以将点 $\mathbf{q}^c$ 与其它点 $\mathbf{q}^{j \neq c}$ 分开。因此，标签恢复问题转化为通过线性规划为 $\mathbf{Q}$ 的每一列寻找可以将其与其它列分开的分类器，如式(13)所示。与GI<sup>[34]</sup>相比，RLG的批量标签恢复方法不再要求非负激活函数，但仍有一定限制。首先，RLG要求批次大小小于模型输出类别数。其次，无法恢复重复标签样本的具体数量。最后，RLG依赖于矩阵分解和求解线性规划问题，标签恢复速度慢于GI，并且在训练的后阶段平均梯度趋于0时标签恢复的精度会下降。

$$\begin{aligned} \text{LP}(c) : \min_{\mathbf{r}} \mathbf{r} \cdot \mathbf{q}^c \\ \text{s.t. } \mathbf{r} \cdot \mathbf{q}^c \leq 0 \text{ and } \mathbf{r} \cdot \mathbf{q}^j \geq 0, \forall j \neq c \end{aligned} \quad (13)$$

然而，GI<sup>[34]</sup>和RLG<sup>[58]</sup>的标签恢复技术也仅限于初步揭示训练批次中是否存在某个标签的样本，当训练批次中有多个相同标签的样本时则无法恢复其具体数量。为此，Ma等人<sup>[59]</sup>针对更符合实际场景的设置，即训练批次中每类样本均包含多个实例，提出一种批量标签高精度解析恢复算法iLRG (instance-wise Label Restoration from Gradients)，率先将攻击粒度提升至实例数目，为当前从平均梯度反演数据标签的最先进水平。iLRG利用分类层梯度恢复类平均嵌入及归一化输出概率(Post-Softmax)，并建立梯度、概率与实例数目的线性方程组，最后通过 Moore-Penrose 伪逆算法求得实例数目的最小二乘解。

### 3.3 小结

根据攻击范式的不同，梯度反演攻击分为迭代优化法与解析法。解析法通过建立梯度与训练数据

间的线性关系、分类层梯度与标签间的关系等方法求解训练数据及标签。虽然解析法在特定条件下具有更好的重建速度和重建质量，但其适用条件较为严格，例如只适用于单样本数据重建或全连接网络等。并且，训练梯度的完整性和数值精度对解析法的数据重建效果影响较大，一旦梯度被扰动，利用有扰梯度得到的重建结果往往很不理想。迭代优化法则以梯度匹配为目标并结合一定先验正则约束，对待解数据进行迭代优化。虽然迭代优化法对于通过梯度匹配反演原始输入的条件和有效性等方面的理论性和可解释性较弱，并且优化过程中可能陷入局部最优，但该方法与解析法相比适用的范围更广，在复杂神经网络、有扰梯度等更为复杂的条件下拥有更好的攻击效果。

根据攻击能力的不同，梯度反演攻击可分为被动攻击和主动攻击，并且目前已有的主动攻击均基于解析法。随着批次大小的增加，各个训练样本产生的梯度在平均的过程中会损失大量信息，因此被动攻击的重建效果会迅速下降。而主动攻击通过构造恶意模型结构及权重等方法诱导用户数据泄露，虽然增强了对大批量数据的梯度反演效果，但恶意的模型结构及权重也极易被用户检测到。

根据攻击对象的不同，梯度反演攻击分为数据重建与标签恢复，分别旨在恢复原始数据和标签。自iDLG<sup>[57]</sup>发现模型输出层梯度中隐含着样本标签信息，并提出了单样本的标签恢复方法后，后续的梯度反演攻击开始将标签恢复与数据重建分离为两个攻击任务，并且数据重建任务通常建立在标签已知的假设上展开。由于正确恢复样本标签是进行数据重建攻击的前提，因此能力不断增强的标签恢复方法也为进一步的数据重建攻击提供了基础，弥补了数据重建攻击中假设标签已知的不足。在数据重建方面，敌手攻击能力从低分辨率的单样本灰度图像反演发展到了高分辨率的大批量彩色图像反演，从单文本数据重建发展到大批量文本的数据重建，并且开始探索对语音、表格等其它类型数据的重建。在标签恢复方面，敌手攻击能力也已从单样本标签恢复发展至可恢复包含重复标签样本的实例级批量标签高精度解析恢复算法。

从攻击范式、攻击能力、攻击对象及其数据类型、攻击的前提假设、特点、不足，对梯度反演攻击进行总结，如表1所示。

## 4 深度梯度反演防御研究进展

随着深度梯度反演攻击方兴未艾，针对性的防御研究近年来纷纷涌现。而基于密码学的PPFL方案的威胁模型中只考虑了半诚实服务器，只对上传

表1 梯度反演攻击总结

攻击方法	攻击范式	攻击能力	攻击对象	数据类型	前提假设	特点	不足
DLG <sup>[20]</sup>	迭代优化	被动	数据重建+ 标签恢复	图像、 标签	—	采用欧氏距离作为梯度匹配目标；同时优化数据与标签。	仅适用于小批量低分辨率图像，重建效果对模型权重初始化方法、训练阶段敏感。
SAPAG <sup>[31]</sup>	迭代优化	被动	数据重建	图像	—	采用基于加权高斯核的梯度匹配目标。	仅适用于小批量低分辨率图像。
CPL <sup>[32]</sup>	迭代优化	被动	数据重建	图像	—	引入标签正则化项。	仅适用于小批量低分辨率图像。
IG <sup>[33]</sup>	迭代优化	被动	数据重建	图像	标签已知。	基于余弦相似度的梯度匹配目标，引入全变分正则项，可重建高分辨率图像。	仅适用于小批量图像。
GI <sup>[34]</sup>	解析+ 迭代优化	被动	标签恢复 +数据重建	标签、 图像	采用非负激活函数，训练批次中无重复标签样本；已知BN统计信息。	引入保真度正则和组一致性正则。	无法恢复重复标签样本的具体数量；BN统计信息已知的假设太强。
GIAS <sup>[36]</sup>	迭代优化	被动	数据重建	图像	预训练GAN模型，标签已知。	依次在GAN隐空间和参数空间上优化。	重建图像失真。
GGL <sup>[37]</sup>	迭代优化	被动	数据重建	图像	预训练GAN模型，标签已知。	评估多种防御方法下的梯度反演效果。	重建图像失真。
HCGLA <sup>[39]</sup>	迭代优化	被动	数据重建	图像	预训练GAN模型、去噪模型。	对稀疏至0.1%的梯度成功反演。	仅针对单样本。
TAG <sup>[41]</sup>	迭代优化	被动	数据重建	文本	—	L2、L1范数结合的梯度匹配目标，在嵌入空间上迭代优化。	仅针对单文本。
LAMP <sup>[42]</sup>	迭代优化	被动	数据重建	文本	标签已知，辅助语言模型(如GPT-2)。	交替进行嵌入空间上的连续优化和文本顺序上的离散优化。	仅适用于小批量文本。
Li等人 <sup>[43]</sup>	迭代优化	被动	数据重建	语音	—	首个两阶段语音波形重建方法。	仅针对单样本。
TabLeak <sup>[44]</sup>	迭代优化	被动	数据重建	表格	—	首个表格数据重建方法。	—
Phong 等人 <sup>[21]</sup>	解析	被动	数据重建	图像	针对全连接网络。	利用全连接网络输入与梯度的线性关系。	仅针对单样本且依赖梯度的完整性。
R-GAP <sup>[45]</sup>	解析	被动	数据重建	图像	针对全连接网络、卷积神经网络。	通过求解梯度约束、权重约束方程组重建输入。	仅针对单样本且依赖梯度的完整性。
COPA <sup>[46]</sup>	解析	被动	数据重建	图像	针对全连接网络、卷积神经网络。	将全连接网络梯度反演建模为盲源分离问题。	批次大小需小于首个全连接层神经元数，依赖梯度的完整性。
CPA <sup>[47]</sup>	解析+ 迭代优化	被动	数据重建	图像	针对全连接网络、卷积神经网络。	将全连接网络梯度反演建模为盲源分离问题。	批次大小需小于首个全连接层神经元数，依赖梯度的完整性。
FILM <sup>[48]</sup>	解析	被动	数据重建	文本	—	从梯度中恢复所有词，再基于波束搜索和重排序重建句子。	依赖梯度的完整性。
Lam 等人 <sup>[49]</sup>	解析	主动	聚合梯度 分解	—	固定模型权重。	将聚合梯度的分解问题表示为约束二元矩阵分解问题。	依赖梯度的完整性，易被用户检测出。
Boenisch 等人 <sup>[50]</sup>	解析	主动	数据重建	图像	篡改模型权重，针对ReLU激活的全连接网络。	通过恶意权重暴露更多的独占神经元。	依赖梯度完整性，攻击成功率提升有限。
Wen 等人 <sup>[51]</sup>	解析	主动	平均梯度 退化	—	篡改模型最后一个分类层的权重。	通过钓鱼策略使平均梯度退化至单个目标样本产生的梯度。	一次攻击仅能恢复一个样本，易被用户检测出。
Pasquini 等人 <sup>[52]</sup>	解析	主动	聚合梯度 退化	—	篡改模型权重，采用ReLU激活函数。	通过恶意权重使聚合梯度退化至单个目标用户产生的梯度。	易被用户检测出。
Fowl 等人 <sup>[53]</sup>	解析	主动	数据重建	图像	篡改模型结构(插入印记模块)及权重。	通过恶意模型权重分离线性层梯度。	依赖梯度的完整性，易被用户检测出。
Zhao 等人 <sup>[54]</sup>	解析	主动	数据重建	图像	篡改模型结构及权重。	向印记模块前插入卷积层以分离不同用户在印记模块的梯度。	依赖梯度的完整性，易被用户检测出。
NEA <sup>[55]</sup>	解析+ 迭代优化	主动	数据重建	图像	篡改用户数据预处理算法，采用ReLU激活函数。	为全连接网络的输入与梯度建立线性方程组。	依赖梯度的完整性，易被用户检测出。
Fowl 等人 <sup>[56]</sup>	解析	主动	数据重建	文本	篡改模型权重。	通过恶意模型权重禁用注意力层，分离线性层梯度。	依赖梯度的完整性，易被用户检测出。
iDLG <sup>[57]</sup>	解析	被动	标签恢复	标签	—	根据最后一层的梯度恢复样本标签。	仅针对单样本。
RLG <sup>[58]</sup>	解析	被动	标签恢复	标签	批次大小小于模型输出类别数。	将标签恢复转化为矩阵分解和线性规划问题求解。	无法恢复重复标签样本的数量，训练后期的标签恢复精度下降。
iLRG <sup>[59]</sup>	解析	被动	标签恢复	标签	—	将标签恢复粒度提升至实例数目，可恢复相同标签样本的具体数量。	训练后期的恢复精度下降。

的局部梯度进行保护，而没有考虑半诚实用户的深度梯度反演能力，无法抵抗半诚实用户对全局梯度的反演攻击。因此本节主要针对基于扰动变换的防御方法展开介绍，并根据扰动对象的不同，将扰动变换分为梯度扰动、输入扰动、特征扰动，分别旨在混淆梯度、输入或特征以隐藏敏感信息。

#### 4.1 梯度扰动

由于梯度反演敌手通常通过梯度匹配或根据梯度建立线性方程组以重建原始数据，因此最直观的防御方法即为直接对梯度进行扰动，具体可分为梯度加噪与梯度压缩两类。虽然梯度扰动对解析攻击的防御效果较好，但对于迭代优化法的梯度反演攻击，梯度扰动在提高隐私性的同时，却是以降低模型性能、增加收敛时间为代价。

Abadi等人<sup>[60]</sup>提出的差分隐私(Differential Privacy, DP)是典型的梯度加噪算法，即向原始梯度中注入一定的噪声，通过添加随机性来避免隐私泄露，对梯度反演攻击有一定的防御效果。DLG<sup>[20]</sup>探究了不同强度的高斯噪声和拉普拉斯加性噪声对攻击的影响，发现梯度加噪方案在隐私性和模型可用性之间存在无法解决的冲突，当增加噪声强度至其可抗梯度反演时模型性能会严重下降。此外，Wei等人<sup>[61]</sup>揭示了传统的服务器协调差分隐私方法Fed-SDP无法保障数据隐私，据此提出一种逐样本的客户端差分隐私算法Fed-CDP，并通过动态衰减噪声注入策略实现了一定程度的隐私-效用权衡。后续他们进一步研究基于动态灵敏度机制、动态噪声尺度机制和动态参数策略的不同组合，提出一套具有动态隐私参数的算法DP-dynS<sup>[62]</sup>。Wang等人<sup>[63]</sup>观察到梯度相对训练数据变化的敏感度是信息泄露风险的一个重要指标，从而提出LRP算法，其通过干扰梯度以匹配信息泄露风险实现了轻量级的防御开销，并借助梯度全局相关性补偿保障了训练精度。

梯度压缩方法包括梯度量化与梯度稀疏。其中，DLG<sup>[20]</sup>验证了16 bit半精度梯度量化无法抗梯度反演，而8比特梯度量化虽成功防御，但也导致了高达20%的精度损失。DLG<sup>[20]</sup>也测试了不同梯度稀疏率的抗梯度反演效果，梯度稀疏率达到20%以上时即可有效抵抗梯度反演，但随着梯度反演敌手能力的不断增强，90%的梯度稀疏率也已无法抵抗GGL<sup>[37]</sup>等更强大的梯度反演攻击。

#### 4.2 输入扰动

对神经网络的原始输入数据进行扰动的工作主要包括Huang等人<sup>[64]</sup>提出的Instahide和Gao等人<sup>[65]</sup>提出的ATS(Automatic Transformation Search)算法，核心思路均为利用数据增强技术保护隐私性。

InstaHide<sup>[64]</sup>首先对目标训练图片与从训练集和公共数据集中随机选取的图片分别在图片像素上和标签上进行线性组合，得到复合图片及复合标签，并对复合图片的像素值进行随机符号翻转，最后将复合图片及复合标签输入网络进行训练。该方法随机选取组合的图片及符号翻转掩码，作为一次性密钥保护了训练数据的隐私性。InstaHide根据其进行线性组合的图片数据集的不同，可分为内部数据集InstaHide(目标训练图片仅与同一训练集中的其它图片随机组合)和跨数据集InstaHide(目标训练图片与同一训练集中的图片和公共数据集的图片同时随机组合)。其中跨数据集的线性组合可解决训练集中不同图片存在相似背景的问题，并且增加复合图片的不可预测性，增强了训练数据的隐私性。

现有的数据增强方法主要用于改进模型性能和泛化能力，缺少其在减少信息泄露方面的效果评估。因此为了在增强隐私和减少建模精度损失两方面取得平衡，ATS<sup>[65]</sup>定义了度量数据增强方法引起隐私泄露情况的隐私分数(privacy score)和度量数据增强方法对模型精度影响情况的精度分数(accuracy score)。其中，隐私分数为变换后图片与原始图片在不同比例下线性组合后得到的复合图片与原始图片产生的梯度的余弦相似性曲线下方的面积，精度分数则采用Joseph等人<sup>[66]</sup>提出的神经架构搜索技术中的计算方法。之后在AutoAugment<sup>[67]</sup>的50种图像变换函数中随机选取若干种组成图像变换策略，计算其隐私分数及精度分数，并从中选取若干个满足精度分数且隐私分数最高的图像变换策略。另外，ATS并不采用单一的某种变换策略，而是对每个数据样本随机采用一种变换策略以保证低隐私泄露和高模型精度。

#### 4.3 特征扰动

对神经网络的中间层特征进行扰动的工作主要包括Sun等人<sup>[40]</sup>提出的Soteria、Huang等人<sup>[68]</sup>提出的TextHide以及Scheliga等人<sup>[69]</sup>提出的PRECODE(PPrivacy Enhancing mODule)。

Soteria<sup>[40]</sup>指出梯度反演引起的隐私泄露问题主要来自于特征泄露，因此其通过对网络的中间特征进行裁剪，在对特征添加尽可能小的扰动的条件下，使得利用扰动后特征重构得到的输入尽可能与真实输入不同，从而既能保持模型性能，又能减少隐私信息泄露。

Huang等人<sup>[68]</sup>还将InstaHide推广至自然语言处理任务中并提出了TextHide，将目标文本与其它文本的标签及在Transformer编码器的输出特征进行线性组合，再根据从掩码池中随机选取的掩码对复

合特征进行随机符号反转,最后在“加密”的复合特征上训练分类器。

PRECODE<sup>[69]</sup>则是首次通过修改神经网络结构,扰动输出层的输入特征来减少隐私泄露。其在原神经网络的输出层前插入了基于变分建模(Variational Modeling)的隐私增强模块以隐藏原始特征,该模块包含一个将特征近似编码至多元正态分布的概率编码器和一个随机解码器。

#### 4.4 小结

从扰动对象、特点、模型准确率损失及不足对梯度反演防御方法进行总结如表2所示,典型深度梯度反演攻防能力矩阵如表3所示。基于线性层泄露的解析攻击对梯度的完整性较为敏感,对有扰梯度的攻击效果较差,此时基于梯度扰动的防御方法

往往能有很好的防御效果。但面对基于迭代优化法的梯度反演攻击时,梯度扰动算法往往难以在隐私性与模型准确率间取得平衡。而基于输入扰动算法的InstaHide,ATS以及基于特征扰动的Soteria,TextHide,PRECODE虽然对模型准确率影响较小,但也很快被指出并不安全。Balunovic等人<sup>[70]</sup>提出的BFGL(Bayesian Framework for Gradient Leakage)可在模型训练初期有效攻击Soteria,ATS和PRECODE等多种防御手段。Carlini等人<sup>[71]</sup>可从实例编码后的数据恢复原始数据,证明了实例编码方案InstaHide和TextHide的不安全性。与梯度反演攻击相比,对梯度反演防御方案的研究有一定滞后性。另外,表3中大量的‘-’也说明当前对梯度反演攻防能力的讨论尚不充分。

表2 梯度反演防御方法总结

防御方法	扰动对象	特点	模型准确率损失	不足
梯度加噪 <sup>[20]</sup>	梯度扰动	直接向梯度注入固定方差的高斯或拉普拉斯噪声。	-30.5%(CIFAR100)	
Fed-CDP <sup>[61]</sup>	梯度扰动	逐样本的客户端侧差分隐私。	-6.2%(MNIST)	难以在隐私性与模型准确率间取得平衡。
DP-dynS <sup>[62]</sup>	梯度扰动	动态隐私参数的差分隐私。	-6.6%(CIFAR10)	
LRP <sup>[63]</sup>	梯度扰动	根据梯度随输入变化的敏感度向梯度注入针对性噪声。	-1.58%(CIFAR10)	
8比特梯度量化 <sup>[20]</sup>	梯度扰动	将高精度梯度量化至低精度。	-22.6%(CIFAR100)	
InstaHide <sup>[64]</sup>	输入扰动	对训练图像进行实例编码。	-5.8%(CIFAR100)	
ATS <sup>[65]</sup>	输入扰动	搜索隐私性强、对模型准确率影响小的数据增强策略。	+1.04%(CIFAR100)	已无法抵抗攻击能力不断增强的梯度反演。
Soteria <sup>[40]</sup>	特征扰动	对网络中间特征进行扰动。	-0.5%(CIFAR100)	
TextHide <sup>[68]</sup>	特征扰动	对训练文本在网络的中间特征实例编码。	-1.9%(GLUE)	
PRECODE <sup>[69]</sup>	特征扰动	向网络中插入隐私增强模块。	-2.2%(CIFAR10)	

表3 典型深度梯度反演攻防能力矩阵

防御	DLG <sup>[20]</sup>	IG <sup>[33]</sup>	R-GAP <sup>[45]</sup>	GI <sup>[34]</sup>	NEA <sup>[55]</sup>	GGL <sup>[37]</sup>	BFGL <sup>[70]</sup>	HCGLA <sup>[39]</sup>	Carlini等人 <sup>[71]</sup>
无防御措施									-
增大批次大小									-
增大输入尺寸			-						-
Fed-CDP <sup>[61]</sup>			-	-	-	-	-	-	-
DP-dynS <sup>[62]</sup>		-	-	-	-	-	-	-	-
LRP <sup>[63]</sup>			-	-	-	-	-	-	-
加性噪声 <sup>[20]</sup>		-	-	-	-		-	-	-
梯度量化 <sup>[20]</sup>		-	-	-	-		-	-	-
梯度稀疏 <sup>[20]</sup>		-	-	-	-		-		-
Soteria <sup>[40]</sup>			-	-	-			-	-
PRECODE <sup>[69]</sup>			-	-	-	-		-	-
ATS <sup>[65]</sup>			-	-	-	-		-	-
InstaHide <sup>[64]</sup>		-	-	-	-	-	-	-	
TextHide <sup>[68]</sup>		-	-	-	-	-	-	-	

## 5 结论与展望

联邦学习作为一种“保留数据所有权, 释放数据使用权”的机器学习方法, 打破了阻碍大数据建模的数据孤岛。而自2019年出现的深度梯度反演攻击利用联邦学习训练过程中的共享梯度中重建训练数据, 严重威胁了联邦学习的私密性, 深度梯度反演攻防因此成为新的研究热点。本文首先给出了梯度反演威胁模型, 根据敌手能力将其分为被动攻击者与主动攻击者两类, 之后对深度梯度反演攻击与防御方法进行了详细梳理与总结, 包括梯度反演攻击的攻击范式、攻击能力、攻击对象、假设、特定、不足, 以及梯度反演防御的特点、模型准确率损失、不足。对梯度反演攻防未来的研究方向展望如下。

第一, 梯度反演攻击在其它场景下的应用研究。当前的梯度反演攻击研究主要集中于采用交叉熵损失的图像分类任务, 少量研究工作对文本等其它类型的数据重建进行了探索。因此, 在其它任务领域和数据类型下的梯度反演攻击是重要的研究方向。例如在智慧安防、智慧交通等领域, 采用的目标检测及语义分割等模型均采用了混合损失函数, 使得当前大多数的梯度反演攻击均无法直接应用, 在这些更为复杂的任务下的梯度反演方法仍亟待研究。

第二, 对主动攻击的隐藏与检测的研究。恶意服务器可通过篡改模型结构及权重破坏安全聚合或诱导大批量平均梯度的隐私泄露。然而, 部分攻击方法构造的恶意权重过于结构化, 严重偏离正常的模型权重, 极易被用户检测出来。因此对于主动攻击的隐藏与检测的对抗研究也是下一步的研究重点。

第三, 对训练梯度的理论安全边界的研究。当前的梯度反演防御方法大多仅通过设计实验对已有的攻击方法展开经验性地防御效果评估, 缺少对训练梯度在何种条件下会泄露训练数据的理论分析, 因此往往提出的防御算法很快便被能力更加强大的梯度反演攻击攻破。并且由于缺乏对安全边界的理论分析, 现有的防御方法难以精准平衡隐私增益与性能损失, 导致现有防御方法为对抗深度梯度反演而过度牺牲模型准确率, 如表2所示。因此, 训练梯度的理论安全边界的研究是一个未来的研究方向。例如, 针对线性层梯度泄露问题<sup>[50,53-55]</sup>, 可根据ReLU神经元的激活状态判断训练梯度是否会泄露训练数据。

第四, 对复合梯度反演防御方法的研究。单一采用安全聚合技术无法抵抗对聚合梯度的反演攻击, 仅进行梯度扰动对基于线性层泄露的解析攻击的防御效果较好, 面对迭代优化法攻击时却难以平

衡隐私性与模型准确率, 现有的输入扰动和特征扰动也已无法抵抗攻击能力不断增强的梯度反演攻击。因此, 同时利用多种防御手段也是一个潜在的研究方向。

## 参考文献

- [1] JORDAN M I and MITCHELL T M. Machine learning: Trends, perspectives, and prospects[J]. *Science*, 2015, 349(6245): 255-260. doi: 10.1126/science.aaa8415.
- [2] LECUN Y, BENGIO Y, and HINTON G. Deep learning[J]. *Nature*, 2015, 521(7553): 436-444. doi: 10.1038/nature14539.
- [3] FANG Binxing. Breaking the conflict between data element flows and privacy protection[EB/OL]. <http://event.chinaet.com/huodong/cite2022/>, 2022.
- [4] MCMAHAN B, MOORE E, RAMAGE D, *et al*. Communication-efficient learning of deep networks from decentralized data[C]. The 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, USA, 2017: 1273-1282.
- [5] YANG Qiang, LIU Yang, CHENG Yong, *et al*. Federated Learning[M]. San Rafael: Morgan & Claypool, 2020: 1-207.
- [6] 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): 12. doi: 10.1145/3298981.
- [7] LIU Yang, FAN Tao, CHEN Tianjian, *et al*. FATE: An industrial grade platform for collaborative learning with data protection[J]. *Journal of Machine Learning Research*, 2021, 22: 1-1.
- [8] 马艳军, 于佃海, 吴甜, 等. 飞桨: 源于产业实践的开源深度学习平台[J]. *数据与计算发展前沿*, 2019, 1(1): 105-115. doi: 10.11871/jfdc.issn.2096.742X.2019.01.011.
- [9] MA Yanjun, YU Dianhai, WU Tian, *et al*. Paddlepaddle: An open-source deep learning platform from industrial practice[J]. *Frontiers of Data and Computing*, 2019, 1(1): 105-115. doi: 10.11871/jfdc.issn.2096.742X.2019.01.011.
- [10] BONAWITZ K A, EICHNER H, GRIESKAMP W, *et al*. Towards federated learning at scale: System design[C]. *Machine Learning and Systems 2019*, Stanford, USA, 2019: 374-388. doi: 10.48550/arXiv.1902.01046.
- [11] RYFFEL T, TRASK A, DAHL M, *et al*. A generic framework for privacy preserving deep learning[EB/OL]. <https://arxiv.org/pdf/1811.04017v2.pdf>, 2018.
- [12] HAO Meng, LI Hongwei, LUO Xizhao, *et al*. Efficient and privacy-enhanced federated learning for industrial artificial intelligence[J]. *IEEE Transactions on Industrial Informatics*, 2020, 16(10): 6532-6542. doi: 10.1109/TII.2019.2945367.
- [13] RIEKE N, HANCOX J, LI Wenqi, *et al*. The future of digital health with federated learning[J]. *NPJ Digital Medicine*, 2020, 3: 119. doi: 10.1038/s41746-020-00323-1.

- [13] XU Jie, GLICKSBERG B S, SU Chang, *et al.* Federated learning for healthcare informatics[J]. *Journal of Healthcare Informatics Research*, 2021, 5(1): 1–19. doi: [10.1007/s41666-020-00082-4](https://doi.org/10.1007/s41666-020-00082-4).
- [14] MILLS J, HU Jia, and MIN Geyong. Communication-efficient federated learning for wireless edge intelligence in iot[J]. *IEEE Internet of Things Journal*, 2020, 7(7): 5986–5994. doi: [10.1109/JIOT.2019.2956615](https://doi.org/10.1109/JIOT.2019.2956615).
- [15] YANG Wensi, ZHANG Yuhang, YE Kejiang, *et al.* FFD: A federated learning based method for credit card fraud detection[C]. Proceedings of the 8th International Conference on Big Data, San Diego, USA, 2019: 18–32. doi: [10.1007/978-3-030-23551-2\\_2](https://doi.org/10.1007/978-3-030-23551-2_2).
- [16] LONG Guodong, TAN Yue, JIANG Jing, *et al.* Federated learning for open banking[M]. YANG Qiang, FAN Lixin, and YU Han. Federated Learning: Privacy and Incentive. Cham: Springer, 2020: 240–254. doi: [10.1007/978-3-030-63076-8\\_17](https://doi.org/10.1007/978-3-030-63076-8_17).
- [17] NASR M, SHOKRI R, and HOUMANSADR A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning[C]. 2019 IEEE Symposium on Security and Privacy, San Francisco, USA, 2019: 739–753. doi: [10.1109/SP.2019.00065](https://doi.org/10.1109/SP.2019.00065).
- [18] MELIS L, SONG Congzheng, DE CRISTOFARO E, *et al.* Exploiting unintended feature leakage in collaborative learning[C]. 2019 IEEE Symposium on Security and Privacy, San Francisco, USA, 2019: 691–706. doi: [10.1109/SP.2019.00029](https://doi.org/10.1109/SP.2019.00029).
- [19] WANG Zhibo, SONG Mengkai, ZHANG Zhifei, *et al.* Beyond inferring class representatives: User-level privacy leakage from federated learning[C]. Proceedings of 2019 IEEE Conference on Computer Communications, Paris, France, 2019: 2512–2520. doi: [10.1109/INFOCOM.2019.8737416](https://doi.org/10.1109/INFOCOM.2019.8737416).
- [20] ZHU Ligeng, LIU Zhijian, and HAN Song. Deep leakage from gradients[C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 2019: 1323. doi: [10.5555/3454287.3455610](https://doi.org/10.5555/3454287.3455610).
- [21] PHONG L T, AONO Y, HAYASHI T, *et al.* Privacy-preserving deep learning via additively homomorphic encryption[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(5): 1333–1345. doi: [10.1109/TIFS.2017.2787987](https://doi.org/10.1109/TIFS.2017.2787987).
- [22] DONG Ye, CHEN Xiaojun, SHEN Liyan, *et al.* Eastfly: Efficient and secure ternary federated learning[J]. *Computers & Security*, 2020, 94: 101824. doi: [10.1016/j.cose.2020.101824](https://doi.org/10.1016/j.cose.2020.101824).
- [23] ZHANG Chengliang, LI Suyi, XIA Junzhe, *et al.* Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning[C/OL]. 2020 USENIX Annual Technical Conference, 2020: 493–506.
- [24] ZHU Hangyu, WANG Rui, JIN Yaochu, *et al.* Distributed additive encryption and quantization for privacy preserving federated deep learning[J]. *Neurocomputing*, 2021, 463: 309–327. doi: [10.1016/j.neucom.2021.08.062](https://doi.org/10.1016/j.neucom.2021.08.062).
- [25] ZHANG Jiale, CHEN Bing, YU Shui, *et al.* PEFL: A privacy-enhanced federated learning scheme for big data analytics[C]. 2019 IEEE Global Communications Conference, Waikoloa, USA, 2019: 1–6. doi: [10.1109/GLOBECOM38437.2019.9014272](https://doi.org/10.1109/GLOBECOM38437.2019.9014272).
- [26] BONAWITZ K, IVANOV V, KREUTER B, *et al.* Practical secure aggregation for privacy-preserving machine learning[C]. 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, USA, 2017: 1175–1191. doi: [10.1145/3133956.3133982](https://doi.org/10.1145/3133956.3133982).
- [27] XU Guowen, LI Hongwei, LIU Sen, *et al.* Verifynet: Secure and verifiable federated learning[J]. *IEEE Transactions on Information Forensics and Security*, 2019, 15: 911–926. doi: [10.1109/TIFS.2019.2929409](https://doi.org/10.1109/TIFS.2019.2929409).
- [28] GUO Xiaojie, LIU Zheli, LI Jin, *et al.* VeriFL: Communication-efficient and fast verifiable aggregation for federated learning[J]. *IEEE Transactions on Information Forensics and Security*, 2021, 16: 1736–1751. doi: [10.1109/TIFS.2020.3043139](https://doi.org/10.1109/TIFS.2020.3043139).
- [29] LUO Fucui, AL-KUWARI S, and DING Yong. SVFL: Efficient secure aggregation and verification for cross-silo federated learning[J]. *IEEE Transactions on Mobile Computing*, 2024, 23(1): 850–864. doi: [10.1109/TMC.2022.3219485](https://doi.org/10.1109/TMC.2022.3219485).
- [30] HAHN C, KIM H, KIM M, *et al.* VerSA: Verifiable secure aggregation for cross-device federated learning[J]. *IEEE Transactions on Dependable and Secure Computing*, 2023, 20(1): 36–52. doi: [10.1109/TDSC.2021.3126323](https://doi.org/10.1109/TDSC.2021.3126323).
- [31] WANG Yijue, DENG Jieren, GUO Dan, *et al.* SAPAG: A self-adaptive privacy attack from gradients[EB/OL]. <https://arxiv.org/pdf/2009.06228.pdf>, 2020.
- [32] WEI Wenqi, LIU Ling, LOPER M, *et al.* A framework for evaluating gradient leakage attacks in federated learning[EB/OL]. <https://arxiv.org/pdf/2004.10397v2.pdf>, 2020.
- [33] GEIPING Jonas, BAUERMEISTER H, DRÖGE H, *et al.* Inverting gradients - how easy is it to break privacy in federated learning?[C]. The 34th Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 16937–16947. doi: [10.48550/arXiv.2003.14053](https://doi.org/10.48550/arXiv.2003.14053).
- [34] YIN Hongxu, MALLYA A, VAHDAT A, *et al.* See through gradients: Image batch recovery via gradinversion[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 16332–16341. doi: [10.1109/CVPR46437.2021.01607](https://doi.org/10.1109/CVPR46437.2021.01607).

- [35] HATAMIZADEH A, YIN Hongxu, ROTH H, *et al.* GradViT: Gradient inversion of vision transformers[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 10011–10020. doi: [10.1109/CVPR52688.2022.00978](https://doi.org/10.1109/CVPR52688.2022.00978).
- [36] JEON J, KIM J, LEE K, *et al.* Gradient inversion with generative image prior[C/OL]. The 35th Conference on Neural Information Processing Systems, 2021: 29898–29908.
- [37] LI Zhuohang, ZHANG Jiaxin, LIU Luyang, *et al.* Auditing privacy defenses in federated learning via generative gradient leakage[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 10122–10132. doi: [10.1109/CVPR52688.2022.00989](https://doi.org/10.1109/CVPR52688.2022.00989).
- [38] HUANG Yangsibo, GUPTA S, SONG Zhao, *et al.* Evaluating gradient inversion attacks and defenses in federated learning[C/OL]. The 35th Conference on Neural Information Processing Systems, 2021: 7232–7241.
- [39] YANG Haomiao, GE Mengyu, XIANG Kunlan, *et al.* Using highly compressed gradients in federated learning for data reconstruction attacks[J]. *IEEE Transactions on Information Forensics and Security*, 2022, 18: 818–830. doi: [10.1109/TIFS.2022.3227761](https://doi.org/10.1109/TIFS.2022.3227761).
- [40] SUN Jingwei, LI Ang, WANG Binghui, *et al.* Soteria: Provable defense against privacy leakage in federated learning from representation perspective[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 9307–9315. doi: [10.1109/CVPR46437.2021.00919](https://doi.org/10.1109/CVPR46437.2021.00919).
- [41] DENG Jieren, WANG Yijue, LI Ji, *et al.* TAG: Gradient attack on transformer-based language models[C]. Findings of the Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021: 3600–3610. doi: [10.18653/v1/2021.findings-emnlp.305](https://doi.org/10.18653/v1/2021.findings-emnlp.305).
- [42] BALUNOVIĆ M, DIMITROV D I, JOVANOVIĆ N, *et al.* LAMP: Extracting text from gradients with language model priors[C]. The 36th Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 7641–7654. doi: [10.48550/arXiv.2202.08827](https://doi.org/10.48550/arXiv.2202.08827).
- [43] LI Zhuohang, ZHANG Jiaxin, and LIU Jian. Speech privacy leakage from shared gradients in distributed learning[C]. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 2023: 1–5. doi: [10.1109/ICASSP49357.2023.10095443](https://doi.org/10.1109/ICASSP49357.2023.10095443).
- [44] VERO M, BALUNOVIĆ M, DIMITROV D I, *et al.* TabLeak: Tabular data leakage in federated learning[C]. The 40th International Conference on Machine Learning, Hawaii, USA, 2023: 1460. doi: [10.5555/3618408.3619868](https://doi.org/10.5555/3618408.3619868).
- [45] ZHU Junyi and BLASCHKO M B. R-Gap: Recursive gradient attack on privacy[C/OL]. The 9th International Conference on Learning Representations, 2021: 1–17.
- [46] CHEN Cangxiong and CAMPBELL N D F. Understanding training-data leakage from gradients in neural networks for image classification[EB/OL]. <https://arxiv.org/pdf/2111.10178.pdf>, 2021.
- [47] KARIYAPPA S, GUO Chuan, MAENG K, *et al.* Cocktail party attack: Breaking aggregation-based privacy in federated learning using independent component analysis[C]. The 40th International Conference on Machine Learning, Honolulu, USA, 2023: 651.
- [48] GUPTA S, HUANG Yangsibo, ZHONG Zexuan, *et al.* Recovering private text in federated learning of language models[C]. The 36th Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 8130–8143.
- [49] LAM M, WEI G Y, BROOKS D, *et al.* Gradient disaggregation: Breaking privacy in federated learning by reconstructing the user participant matrix[C/OL]. The 38th International Conference on Machine Learning, 2021: 5959–5968.
- [50] BOENISCH F, DZIEDZIC A, SCHUSTER R, *et al.* When the curious abandon honesty: Federated learning is not private[C]. The 2023 IEEE 8th European Symposium on Security and Privacy, Delft, Netherlands, 2021: 175–199. doi: [10.1109/EuroSP57164.2023.00020](https://doi.org/10.1109/EuroSP57164.2023.00020).
- [51] WEN Yuxin, GEIPING J A, FOWL L, *et al.* Fishing for user data in large-batch federated learning via gradient magnification[C]. The 39th International Conference on Machine Learning, Baltimore, USA, 2022: 23668–23684. doi: [10.48550/arXiv.2202.00580](https://doi.org/10.48550/arXiv.2202.00580).
- [52] PASQUINI D, FRANCATI D, and ATENIESE G. Eluding secure aggregation in federated learning via model inconsistency[C]. 2022 ACM SIGSAC Conference on Computer and Communications Security, Los Angeles, USA, 2022: 2429–2443. doi: [10.1145/3548606.3560557](https://doi.org/10.1145/3548606.3560557).
- [53] FOWL L, GEIPING J, CZAJA W, *et al.* Robbing the fed: Directly obtaining private data in federated learning with modified models[C/OL]. The 10th International Conference on Learning Representations, 2021: 1–25. doi: [10.48550/arXiv.2110.13057](https://doi.org/10.48550/arXiv.2110.13057).
- [54] ZHAO J C, ELKORDY A R, SHARMA A, *et al.* The resource problem of using linear layer leakage attack in federated learning[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 3974–3983. doi: [10.1109/CVPR52729.2023.00387](https://doi.org/10.1109/CVPR52729.2023.00387).
- [55] PAN Xudong, ZHANG Mi, YAN Yifan, *et al.* Exploring the security boundary of data reconstruction via neuron exclusivity analysis[C]. The 31st USENIX Security Symposium, Boston, USA, 2020: 3989–4006. doi: [10.48550/arXiv.2010.13356](https://doi.org/10.48550/arXiv.2010.13356).
- [56] FOWL L H, GEIPING J, REICH S, *et al.* Decepticons: Corrupted transformers breach privacy in federated learning

- for language models[C]. The 11th International Conference on Learning Representations, Kigali, Rwanda, 2022: 1–23. doi: [10.48550/arXiv.2201.12675](https://arxiv.org/abs/10.48550/arXiv.2201.12675).
- [57] ZHAO Bo, MOPURI K R, and BILEN H. iDLG: Improved deep leakage from gradients[EB/OL]. <https://arxiv.org/pdf/2001.02610.pdf>, 2020.
- [58] DANG T, THAKKAR O, RAMASWAMY S, *et al.* Revealing and protecting labels in distributed training[C]. The 35th Conference on Neural Information Processing Systems, Sydney, Australia, 2021: 1727–1738. doi: [10.48550/arXiv.2111.00556](https://arxiv.org/abs/10.48550/arXiv.2111.00556).
- [59] MA Kailang, SUN Yu, CUI Jian, *et al.* Instance-wise batch label restoration via gradients in federated learning[C]. The 11th International Conference on Learning Representations, Kigali, Rwanda, 2023: 1–15.
- [60] ABADI M, CHU A, GOODFELLOW I, *et al.* Deep learning with differential privacy[C]. 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 2016: 308–318. doi: [10.1145/2976749.2978318](https://doi.org/10.1145/2976749.2978318).
- [61] WEI Wenqi, LIU Ling, WU Yanzhao, *et al.* Gradient-leakage resilient federated learning[C]. The 2021 IEEE 41st International Conference on Distributed Computing Systems, Washington, USA, 2021: 797–807. doi: [10.1109/ICDCS51616.2021.00081](https://doi.org/10.1109/ICDCS51616.2021.00081).
- [62] WEI Wenqi and LIU Ling. Gradient leakage attack resilient deep learning[J]. *IEEE Transactions on Information Forensics and Security*, 2021, 17: 303–316. doi: [10.1109/TIFS.2021.3139777](https://doi.org/10.1109/TIFS.2021.3139777).
- [63] WANG Junxiao, GUO Song, XIE Xin, *et al.* Protect privacy from gradient leakage attack in federated learning[C]. 2022 IEEE Conference on Computer Communications, London, UK, 2022: 580–589. doi: [10.1109/INFOCOM48880.2022.9796841](https://doi.org/10.1109/INFOCOM48880.2022.9796841).
- [64] HUANG Yangsibo, SONG Zhao, LI Kai, *et al.* InstaHide: Instance-hiding schemes for private distributed learning[C/OL]. The 37th International Conference on Machine Learning, 2020: 419. doi: [10.5555/3524938.3525357](https://doi.org/10.5555/3524938.3525357).
- [65] GAO Wei, GUO Shangwei, ZHANG Tianwei, *et al.* Privacy-preserving collaborative learning with automatic transformation search[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 114–123. doi: [10.1109/CVPR46437.2021.00018](https://doi.org/10.1109/CVPR46437.2021.00018).
- [66] MELLOR J, TURNER J, STORKEY A, *et al.* Neural architecture search without training[C/OL]. The 38th International Conference on Machine Learning, 2021: 7588–7598.
- [67] CUBUK E D, ZOPH B, MANÉ D, *et al.* Autoaugment: Learning augmentation strategies from data[C]. The 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 113–123. doi: [10.1109/CVPR.2019.00020](https://doi.org/10.1109/CVPR.2019.00020).
- [68] HUANG Yangsibo, SONG Zhao, CHEN Danqi, *et al.* TextHide: Tackling data privacy in language understanding tasks[C/OL]. Findings of the Association for Computational Linguistics, 2020: 1368–1382. doi: [10.18653/v1/2020.findings-emnlp.123](https://doi.org/10.18653/v1/2020.findings-emnlp.123).
- [69] SCHELIGA D, MÄDER P, and SEELAND M. PRECODE - a generic model extension to prevent deep gradient leakage[C]. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, USA, 2022: 3605–3614. doi: [10.1109/WACV51458.2022.00366](https://doi.org/10.1109/WACV51458.2022.00366).
- [70] BALUNOVIĆ M, DIMITROV D I, STAAB R, *et al.* Bayesian framework for gradient leakage[C/OL]. The 10th International Conference on Learning Representations, 2021: 1–16. doi: [10.48550/arXiv.2111.04706](https://doi.org/10.48550/arXiv.2111.04706).
- [71] CARLINI N, DENG S, GARG S, *et al.* Is private learning possible with instance encoding?[C]. Proceedings of 2021 IEEE Symposium on Security and Privacy, San Francisco, USA, 2021: 410–427. doi: [10.1109/SP40001.2021.00099](https://doi.org/10.1109/SP40001.2021.00099).
- 孙 钰: 男, 副教授, 研究方向为无线网络安全、智能系统安全等。  
 严 宇: 男, 硕士生, 研究方向为联邦学习隐私保护。  
 崔 剑: 男, 讲师, 研究方向为工业互联网安全、硬件系统安全。  
 熊高剑: 男, 博士生, 研究方向为联邦学习隐私保护。  
 刘建华: 男, 博士生, 研究方向为联邦学习隐私保护。

责任编辑: 马秀强