联邦学习中的模型逆向攻防研究综述

王冬[1]，秦倩倩[1]，郭开天[1]，刘容轲[1]，颜伟鹏[1]，任一支[1]，罗清彩[2]，申延召[3]

（1. 杭州电子科技大学网络空间安全学院，浙江 杭州 310018；
2. 山东浪潮科学研究院有限公司，山东 济南 250000；3. 山东区块链研究院，山东 济南 250000）

摘 要：联邦学习作为一种分布式机器学习技术可以解决数据孤岛问题，但机器学习模型会无意识地记忆训练数据，导致参与方上传的模型参数与全局模型会遭受各种隐私攻击。针对隐私攻击中的模型逆向攻击，对现有的攻击方法进行了系统总结。首先，概括并详细分析了模型逆向攻击的理论框架；其次，从威胁模型的角度对现有的攻击方法进行总结分析与比较；再次，总结与比较了不同技术类型的防御策略；最后，对现有模型逆向攻击常用的评估标准及数据集进行汇总，并对模型逆向攻击现有的主要挑战以及未来研究方向进行总结。

关键词：联邦学习；模型逆向攻击；隐私安全

中图分类号：TP309

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023209

Survey on model inversion attack and defense in federated learning

WANG Dong[1], QIN Qianqian[1], GUO Kaitian[1], LIU Rongke[1], YAN Weipeng[1], REN Yizhi[1], LUO Qingcai[2], SHEN Yanzhao[3]

1. School of Cyberspace Security, Hangzhou Dianzi University, Hangzhou 310018, China
2. Shandong Inspur Science Research Institute Co., Ltd, Jinan 250000, China
3. Shandong Blockchain Research Institute, Jinan 250000, China

Abstract: As a distributed machine learning technology, federated learning can solve the problem of data islands. However, because machine learning models will unconsciously remember training data, model parameters and global models uploaded by participants will suffer various privacy attacks. A systematic summary of existing attack methods was conducted for model inversion attacks in privacy attacks. Firstly, the theoretical framework of model inversion attack was summarized and analyzed in detail. Then, existing attack methods from the perspective of threat models were summarized, analyzed and compared. Then, the defense strategies of different technology types were summarized and compared. Finally, the commonly used evaluation criteria and datasets were summarized for inversion attack of existing models, and the main challenges and future research directions were summarized for inversion attack of models.

Keywords: federated learning, model inversion attack, privacy security

0 引言

海量数据的产生与获取以及计算设备的进步，极大地推动了人工智能技术的发展，各领域皆相继开展人工智能技术的研究，并将其应用投入实际场景中。与此同时，大量的公共、个人信息作为人工智能模型的训练对象，例如，人脸识别模型将个人的面部数据用以训练[1]，医疗辅助诊断模型运用患

收稿日期：2023−07−18；**修回日期：**2023−10−18

通信作者：任一支，renyz@hdu.edu.cn

基金项目：浙江省“尖兵”“领雁”研发基金资助项目（No.2023C03203, No.2023C03180, No.2022C03174）；浙江省属高校基本科研业务费专项资金资助项目（No.GK229909299001-023）

Foundation Items: Zhejiang Province’s “Sharp Blade” and “Leading Goose” Research and Development Project (No.2023C03203, No.2023C03180, No.2022C03174), Zhejiang Province-funded Basic Research Fund for Universities Affiliated with Zhejiang Province (No.GK229909299001-023)

者的病例作为训练数据[2]。但在实际应用中，个人信息通常具有隐私性极强的特点，并且逐渐完善的数据安全和隐私保护要求使这些数据无法聚合，形成数据孤岛[3]。而单一孤岛的数据量又不足以支撑大规模的深度学习模型训练，导致训练后的模型性能不佳，以及难以实现数据价值等问题。

联邦学习[4]作为一种分布式机器学习技术可以解决数据孤岛问题，使机构间可以跨地域协作而数据不出本地，且多方合作构建的全局模型能够更准确地预测各类问题[3]。该技术旨在不直接上传本地数据的前提下，上传通过本地数据训练后的模型参数，包括模型权重、梯度信息等。服务器收到多名参与方的参数后进行聚合，并重复多次得到一个共享的全局模型。然而，由于机器学习模型会无意识地记忆训练数据，并将其编码在模型权重或梯度信息中[5-6]，导致参与方上传的参数遭受各种隐私攻击，使本地的敏感数据泄露[7-8]。因此，针对机器学习模型的隐私攻防研究十分有意义。

目前，联邦学习中的隐私攻击可分为模型隐私攻击、数据隐私攻击两类，其中模型隐私攻击为模型萃取攻击[9]，数据隐私攻击为推理攻击[10]、重构攻击[11-12]。模型萃取攻击旨在创建一个与目标模型任务相同的替代模型，同时表现相似甚至更好；推理攻击并不直接暴露原始数据的隐私信息，而是结合特定的背景信息实现隐私数据的推断并窃取；重构攻击是一种试图通过模型参数生成目标模型训练集数据的技术，相比于前两类攻击，该攻击揭示了训练数据的“细粒度”信息，因此造成的隐私危害更大。

本文重点关注重构攻击中的模型逆向攻击（MIA, model inversion attack）[13]，即通过访问目标模型，重构训练数据、敏感属性或输入数据。该攻击由 Fredrikson 等[13]首次提出，并利用最大后验原则（MAP, maximum a posteriori）成功恢复了训练集敏感属性。此后，随着深度学习模型的广泛应用，研究者探索了针对不同类型和场景的深度学习模型的逆向攻击方法。例如，针对浅层神经网络模型，Fredrikson 等[14]利用梯度下降的方法重构了训练集数据；针对深层卷积神经网络，Zhang 等[15]和 Chen 等[16]利用生成式对抗网络（GAN, generative adversarial network）[17]作为攻击模型，生成了与训练图像高度相似且更具有语义信息的图像。另外，针对敌手能否访问模型内部信息，将攻击场景分为白盒与黑盒场景。在白盒场景下，一类方法通过设计合适的损失函数来提高攻击效果[16-19]，另一类方法通过伪标签指导生成器生成分类更解耦的图像[20]；在黑盒场景下，主要研究如何优化攻击模型以及攻击模型的输入[21-22]。

为了应对模型逆向攻击，多种防御策略被提出，在政务、金融、医疗、联邦学习开源工具等领域均有落地应用的探索，并根据技术类型可分为基于加密技术[23-24]、扰动技术[25-27]、深度学习训练技术[28-29]的防御策略。但上述策略也有相应的局限性和挑战，如加密技术会给本地和服务端带来额外的计算开销，给传输网络增加额外的占用空间。为了解决该问题，基于差分隐私的模型参数扰动技术被提出，但文献[26]证明该技术无法针对最先进的攻击方法成功防御。因此，基于目标模型输出的扰动技术、基于深度学习训练技术的防御策略相继被提出，旨在保证模型可用性的同时，提高模型防御能力。但上述方法同样存在性能损失、可用性与隐私性难以权衡等问题，因此设计更高效、可靠、通用的防御策略，仍然是一个值得持续研究的问题。

综上所述，联邦学习作为解决数据孤岛的隐私计算技术，同样在训练过程中面临着隐私攻击的挑战。如图 1 所示，本文重点关注联邦学习隐私攻击中的模型逆向攻击，首先，概括并详细分析了该攻击的理论框架；然后，根据威胁模型将模型逆向分为白盒、黑盒攻击，并对现有的攻击方法进行分析与比较；并且，根据不同技术类型对现有的防御策略进行分析与比较；最后，对现有模型逆向攻击常用的评估标准及数据集进行汇总，并对模型逆向攻击现有的主要挑战以及未来研究方向进行总结。

1　模型逆向攻击

模型逆向攻击是一类针对机器学习模型的隐私攻击手段，它以模型的输出为依据，逆向地重构隐私数据[13-15]。

模型逆向攻击的目标通常是分类模型，称为目标模型或目标分类器，记作 M_{T}。要重构的隐私数据被称为重构目标，往往是预测值对应的输入样本的部分或全部属性，或是模型训练集中每类的典型特征。攻击重构出的样本称为攻击样本。

根据模型逆向攻击重构目标，其可分为两类：

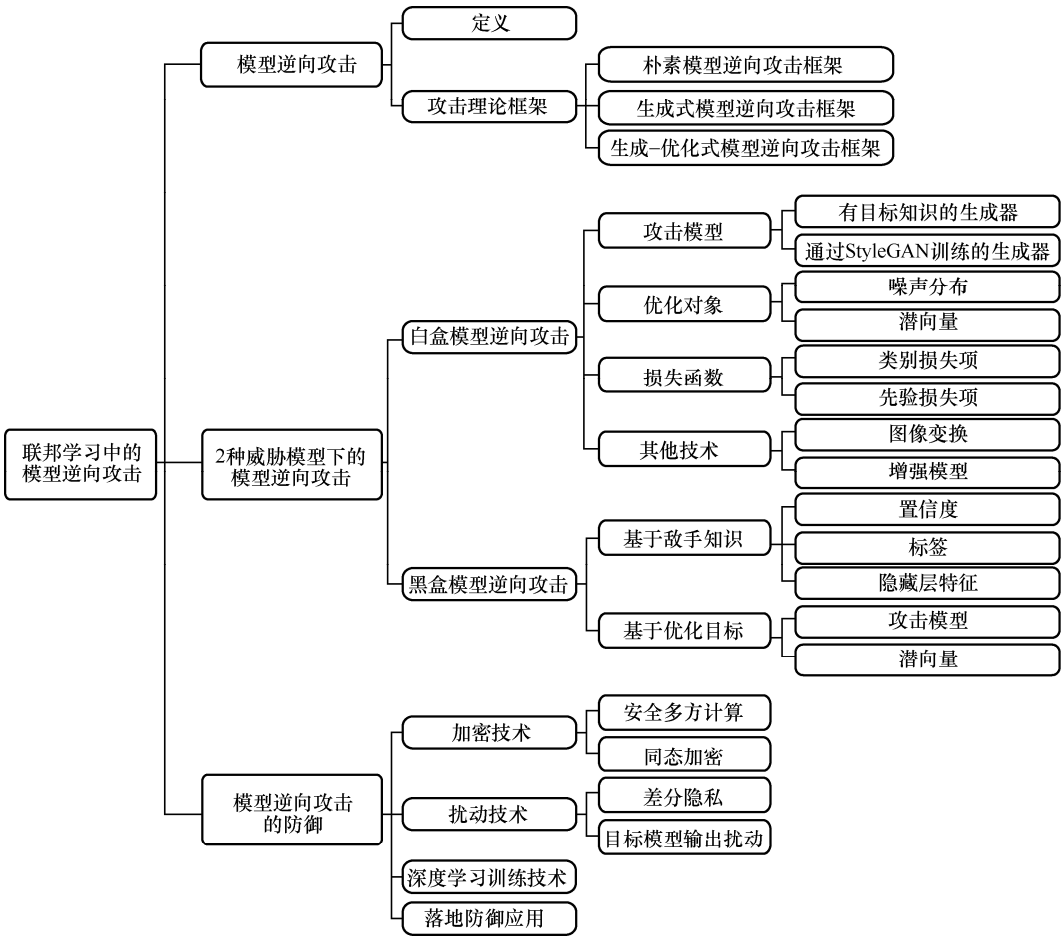


图 1 联邦学习中的模型逆向攻击和防御技术

数据重构攻击和类重构攻击[15]。数据重构攻击通过某个样本在目标模型上的预测结果重构该样本，本文把要重构的样本称为目标样本，把目标样本在目标模型上的预测向量称为目标预测向量。在一些数据重构的攻击场景中，敌手可能不需要重构目标样本的所有属性，而是根据目标样本预测和目标的部分非敏感属性重构敏感属性[13,23]。类重构攻击的目标是重构出具有某个类典型特征的样本，这个类被称为目标类。例如，对人脸识别模型实施类重构攻击可以窃取目标类的典型样本，也就是某个个体的人脸图像。早期的模型逆向攻击研究通常是数据重构攻击[13-14]，而近年来类重构攻击的研究居多[15-16,20]，本文主要介绍类重构攻击。

另外，根据模型逆向攻击的威胁模型，其也分为两类：白盒模型逆向攻击和黑盒模型逆向攻击。前一类攻击中敌手可以访问目标模型的结构和参数，后一类攻击中敌手仅能访问目标模型对某些样本的全部或部分的预测结果。

1.1 朴素模型逆向攻击框架

Fredrikson 等[13]在 2014 年首次提出了模型逆向攻击的概念，并设计了一种基于穷举的模型逆向攻击框架，即通过穷举样本敏感属性的所有可能取值并从中选出可能性最大的取值，进而实现模型逆向数据重构。他们基于此框架设计了一种攻击华法林剂量预测线性回归模型的方法，用样本的非敏感属性和标签推断基因标记这一敏感属性。在逆向攻击前，他们先将目标模型的连续回归输出离散化，从而将原来的回归模型改为分类模型，然后用非敏感属性的边界分布以及模型的误分类概率估计所有敏感属性取值的后验概率，选择其中后验概率最大的取值。这种方法可以攻击线性模型重构表格数据，证明了模型逆向攻击的可行性，但是无法处理具有大量离散特征的输入样本（如基于大量像素点的图像），也无法扩展到非线性模型中运用。

为了从更复杂的模型中重构图像数据，

Fredrikson 等[14]又提出了一种基于随机梯度下降优化的模型逆向攻击框架，如图 2 所示，其旨在从简单的人脸识别模型中重构目标类典型样本。该框架攻击目标模型的过程如下：以随机生成的噪声图像 $\boldsymbol{x}=\boldsymbol{x}^{(0)}$ 作为初始状态，用随机梯度下降方法不断更新优化图像 $\boldsymbol{x}$，最小化图像 $\boldsymbol{x}$ 在目标模型上的预测向量 $\boldsymbol{y}=T(\boldsymbol{x})$ 与目标类向量 $\hat{\boldsymbol{y}}$ 的差异 $L(\boldsymbol{y},\hat{\boldsymbol{y}})$，得到能反映目标类典型特征的攻击图像 $\boldsymbol{x}=\hat{\boldsymbol{x}}$。对于浅层神经网络人脸识别模型，该框架采用人工方式对攻击性能进行判断，最高能达到 75%的攻击准确率。对于更深层的神经网络模型，仅通过随机梯度下降难以生成有意义的图像。

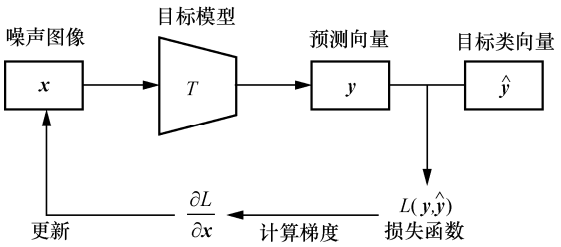


图 2　基于随机梯度下降优化的模型逆向攻击框架

1.2　生成式模型逆向攻击框架

虽然遵循朴素模型逆向攻击框架的模型逆向攻击方法可以在一些对表格数据分类的目标模型上起到比较好的效果，但是这类框架下的方法在面对更复杂、结构性更强、维度更大的目标数据（尤其是图像数据）和更深层的神经网络模型时效果并不好。文献[13]中提出的敏感数据推断方法无法穷举所有可能的高维数据。文献[14]中提出的随机梯度下降方法用深度图像分类目标模型优化图像，由于缺少对目标图像的约束往往会出现过拟合问题，难以生成有意义的图像。

为了解决上述问题，研究者开始考虑使用生成式模型解决朴素模型逆向攻击中存在的问题。Yang 等[21]提出利用辅助数据集来训练攻击模型，以实现对深度神经网络在图像数据上的模型逆向攻击。他们提出了 2 个主要技术。首先，利用对手的背景知识构建了一个辅助数据集，用于训练攻击模型，而无须访问原始的训练数据。具体来说，在一个与目标模型训练集具有相同分布的数据集上训练反向模型，该反向模型学习了人脸图像的公共特征，并将目标模型输出的预测向量 $\hat{\boldsymbol{y}}$ 作为输入，然后通过反向模型 G 重构样本 $\hat{\boldsymbol{x}}$。生成式模型逆向攻击框架如图 3 所示，其中，t 为目标样本，T 为目标模型。其次，他们还设计了一种基于截断的技术，使其能够有效地从对手在受害用户数据上获取的部分预测值逆向目标模型。Yang 等[21]的方法在模型逆向攻击方面取得了显著的进展，该研究为深度神经网络在图像数据上的安全性提供了重要的启示，并为进一步研究和改进提供了基础。

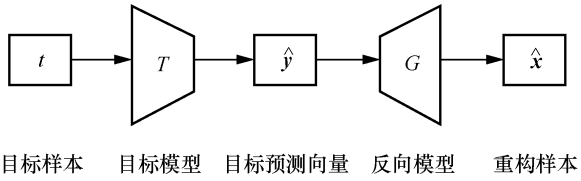


图 3　生成式模型逆向攻击框架

1.3　生成–优化式模型逆向攻击框架

为了进一步从目标模型中挖掘目标类的知识，一些研究者在生成式框架训练攻击模型的基础上继承了文献[14]的优化方法，他们在得到攻击模型的生成样本后，进一步优化该样本，形成了生成–优化式的模型逆向攻击框架，该框架将逆向攻击分为 2 个阶段：训练阶段和优化阶段。

如图 4 所示，在训练阶段中，敌手需要找到一个与目标数据集 D_{tar} 分布接近的辅助数据集 D_{aux}，然后用 D_{aux} 训练一个生成模型 G（也被称为攻击模型）。G 的输入为潜在向量 $\boldsymbol{z}$，输出为一个与辅助数据集样本接近的样本 $\boldsymbol{x}$，例如，如果用人脸图像数据集训练 G，那么生成的样本 $\boldsymbol{x}$ 也是一张人脸图像。潜在向量 $\boldsymbol{z}$ 的取值空间被称为潜在空间。

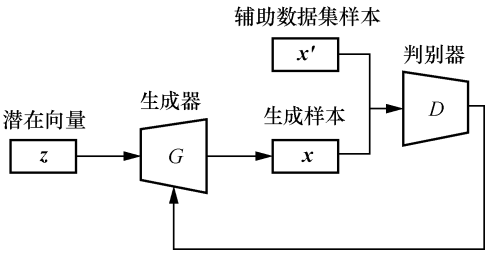


图 4　生成–优化式模型逆向攻击框架训练阶段

如图 5 所示，在优化阶段中，敌手利用生成模型 G 生成图像 $\boldsymbol{x}=G(\boldsymbol{z})$，然后将 $\boldsymbol{x}$ 输入目标模型 T，得到预测向量 $\boldsymbol{y}=T(\boldsymbol{x})$。计算 y 与目标类置信度为 1、其他类均为 0 的目标预测向量 $\hat{\boldsymbol{y}}$ 之间的距离 $L(\boldsymbol{y},\hat{\boldsymbol{y}})$，以此为损失函数，用随机梯度下降等方法优化 $\boldsymbol{z}$，进而优化生成图像，最终得到目标类的攻击样本。

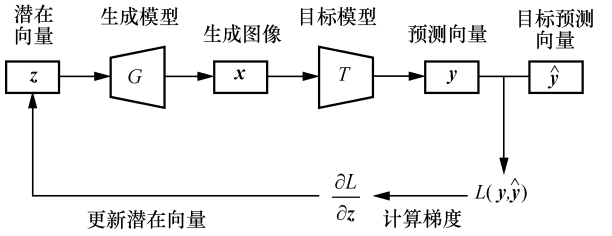


图 5 生成−优化式模型逆向攻击框架优化阶段

Zhang 等[15]提出的生成式模型逆向（GMI, generative model-inversion）攻击是首个使用生成−优化式框架的模型逆向攻击方法，旨在从人脸识别模型中重构出目标类的数据。GMI 用深度卷积生成对抗网络（DCGAN, deep convolutional generative adversarial network）[30]训练生成模型。在优化阶段中，GMI 用交叉熵损失 L_{ce} 计算 $\boldsymbol{y}$ 与 $\hat{\boldsymbol{y}}$ 之间的距离，还在损失函数中加入了判别器 D 的预测值，以保证攻击图像的真实性，即

$$L = \lambda L_{ce}(\boldsymbol{y}, \hat{\boldsymbol{y}}) - D(y) \tag{1}$$

其中，λ 是一个超参数；损失函数的前一项代表生成图像与目标类的差距，称为类别损失项，记作 L_{id}；后一项代表生成图像的不真实程度，称为先验损失项，记作 L_{prior}。

生成−优化式框架通过训练攻击模型，建立了从潜在空间到样本（通常是图像）空间的映射，从而避免在样本空间中直接搜索，从而大大增强了重构复杂样本的真实性。

许多对深度神经网络模型有效的方法[16,18,20,31]都沿用了这种生成−优化式模型逆向攻击框架。而黑盒模型逆向攻击方法[32]由于不能访问模型参数，无法通过目标模型进行反向传播，因此不能直接使用随机梯度下降方法，一些研究者提出了其他的方法来优化潜在向量，详见 2.2 节。

2 2 种威胁模型下的模型逆向攻击

根据威胁模型，模型逆向攻击可以分为 2 种，即白盒模型逆向攻击和黑盒模型逆向攻击。本节主要对这 2 种威胁模型下的模型逆向攻击进行介绍，分类结果如表 1 所示。

表 1 根据威胁模型对模型逆向攻击进行分类

MIA	敌手知识	攻击框架	研究方向	参考文献
白盒	目标模型的网络结构、模型参数等	朴素模型、生成−优化式	攻击模型的选择、优化目标和优化损失项的改进	文献[14-16,18-20,31]
黑盒	样本在目标模型上的全部或部分预测	朴素模型、生成式、生成−优化式	黑盒优化方法的改进、对目标模型查询次数的改进	文献[13-14,21-22,32,33-39]

2.1 白盒模型逆向攻击

在白盒模型逆向攻击场景下，敌手拥有目标模型的全部知识，其中包括目标模型的结构和模型参数等。通常情况下，敌手还可以访问一个与目标数据集分布接近的辅助数据集。白盒模型逆向攻击在一些分布式人工智能应用系统中非常常见。例如，在联邦学习场景下，模型的训练集数据往往是用户提供的隐私数据，中心服务器不具有这些数据的访问权限，却可以访问模型的结构和完整的参数信息，中心服务器运行该模型为用户提供人工智能服务。恶意的中心服务器维护者可以对模型实施白盒模型逆向攻击，窃取训练模型时用到的隐私数据。另外，如果保存不善，敌手可能会从中心服务器窃取模型，实施白盒模型逆向攻击，从而导致隐私数据的泄露。

近年来，白盒模型逆向攻击的研究主要集中在人脸识别模型的类重构攻击。GMI 是最早的白盒模型逆向攻击方法，可以从深度神经网络中重构出有意义的攻击图像，但是准确率和图像质量仍有待提高。Struppek 等[18]指出，模型逆向攻击的效果主要受到以下问题影响。

1) 分布偏移。用于训练生成模型的辅助数据集与目标数据集存在偏差，攻击图像与重构目标差别较大。

2) 局部最小。在优化阶段，优化到一定程度时，可能出现梯度消失现象，导致潜在向量欠拟合或无法跳出局部最小值。

3) 易生成无意义图像。模型逆向攻击可能生成能被目标模型分类正确，但偏离辅助数据集分布的无意义图像。

另外，如何在保证攻击图像真实性的同时提高其在目标模型上的准确率也是白盒模型逆向攻击需要解决的问题之一。研究者在白盒模型逆向攻击的不同方面采用了新的技术和方法，以解决上述问题。

2.1.1 白盒模型逆向攻击的攻击模型

GMI 通过 DCGAN 训练生成器，后续的大多数深度神经网络模型逆向攻击方法也都采用各种 GAN 训练生成模型。DCGAN 在模型逆向攻击中的表现并不完善，其效果存在不足之处。研究者通过训练有目标知识的生成器，或者使用更适合模型逆向攻击的训练框架来训练生成器。

为了提高生成器生成样本 的效果，研究者训练了一些具有目标知识的生成器。Chen 等[16]提出了知识增强分布模型逆向（KED-MI, knowledge-enriched distributional model inversion）攻击方法和一种专用于模型逆向攻击的 GAN，在预处理阶段中用目标模型对辅助数据集的样本进行预测，将预测置信度向量作为它们的“软标签”，再利用软标签辅助训练判别器，间接影响生成器，让生成器可以生成具有各个目标类特征的图像，也就是目标数据集分布的图像。

当辅助数据集与目标数据集取自同一个图像数据集时，KED-MI 能够取得比 GMI 好得多的效果；而当辅助数据集与目标数据集取自不同的人脸数据集时，KED-MI 的效果仍优于 GMI，但比前一情况下的效果稍差。这表明 KED-MI 对分布偏移问题的鲁棒性仍有待改善。

与 KED-MI 相比，Yuan 等[20]提出的伪标签指导的模型逆向（PLG-MI, pseudo label-guided model inversion）攻击方法更直接，这种方法采用了条件 GAN（CGAN, conditional generative adversarial network）[40]架构训练生成模型，不仅可以生成接近目标数据集的样本，还可以生成目标数据集某一类的样本。

PLG-MI 在预处理阶段中采用 top-n 选择策略：用目标模型预测辅助数据集中的样本，然后为每个类置信度最高的 n 个样本赋予该类的“伪标签”，用这些带“伪标签”的样本训练生成模型。PLG-MI 使用 CGAN 训练生成模型。CGAN 在训练和生成图像时可以用标签指导，生成标签对应类的图像。PLG-MI 进一步缩小了生成图像的范围，提高了模型逆向攻击的效果。

PLG-MI 在辅助数据集与目标数据集取自同一数据集和不同数据集的 2 种情况下都有比 KED-MI 更优的效果。

1) 基于 styleGAN 训练的生成器

styleGAN[41]架构中的生成器由一个映射网络 G_{mapping} 和一个合成网络 $G_{\text{synthesis}}$ 组成。其中，映射网络将潜在向量 $\boldsymbol{z}$ 映射到另一个向量 $\boldsymbol{w}$，$\boldsymbol{w}$ 的每个分量所代表的样本特征耦合度较低。例如，对于生成人脸识别模型而言，$\boldsymbol{w}$ 的各个分量可能代表具体的人脸特征，如胡须、发色等。styleGAN 将 $\boldsymbol{w}$ 复制 L 份，然后输入合成网络中，合成网络根据 $\boldsymbol{w}$ 表示的特征生成图像。先前的研究发现，在生成器的中间层做优化比在神经网络的头部或尾部做优化更好[31]，而对 styleGAN 潜在向量的优化刚好发生在中间层。

由于上述原因，Struppek 等[18]提出的即插即用的攻击（PPA, plug and play attack）和 Wang 等[31]提出的变分模型逆向（VMI, variational model inversion）攻击 2 种攻击方法都采用 styleGAN 训练生成器。

2.1.2 白盒模型逆向攻击的优化对象

GMI 攻击在优化阶段中通过优化潜在向量最小化损失函数，进而重构目标类的典型样本。大多数白盒模型逆向攻击在优化阶段的优化对象都是潜在向量，而 KED-MI 与 VMI 这 2 种方法通过优化潜在向量的分布使损失函数的值最小化。

Chen 等[16]认为目标分类器是一个多对一的映射，即将同一个类的多个样本映射到这个类上，所以在实施模型逆向攻击时，攻击的目标应该也是重构出一个目标类的多个图像。KED-MI 选择潜在向量的正态分布作为优化目标。

Wang 等[31]将模型逆向攻击解释为一种变分推理（VI, variational inference）[42]问题：找到一个图像样本分布 $q(\boldsymbol{x})$，使它尽可能接近目标数据集样本 $\boldsymbol{x}$ 在目标类 $\boldsymbol{y}$ 上的条件概率分布 $p_{\text{TAR}}(\boldsymbol{x}|\boldsymbol{y})$，而 $q(\boldsymbol{x})$ 又由潜在向量分布 $q(\boldsymbol{z})$ 确定。因此，VMI 的优化对象就是潜在向量的分布。以分布为优化对象进行的模型逆向攻击往往可以得到更具多样性的攻击图像。

2.1.3 白盒模型逆向攻击的损失函数

白盒模型逆向攻击在优化阶段的损失函数往往由类别损失项与先验损失项组成，即

$$L_{\text{op}} = L_{\text{id}} + L_{\text{prior}} \tag{2}$$

其中，L_{id} 是类别损失项，表示生成图像在目标模型上的预测结果与目标类之间的差异，促使生成模型生成目标类样本；L_{prior} 是先验损失项，对优化的潜在向量或潜在向量分布进行规约，使之生成有意义的图像。

1) 类别损失项

GMI 与 KED-MI、VMI 等其他白盒模型逆向攻击方法都使用交叉熵（CE, cross entropy）损失衡量生成图预测向量与目标类向量之间的差距。但 CE 损失容易出现梯度消失现象和局部最小问题。PLG-MI 使用了一些其他类型敌手攻击[43-44]用到的最大边界损失替代 CE 损失；PPA 在双曲空间中比较 2 个向量的差异，用两者之间的 poincaré 距离作为损失，也可以解决梯度消失问题。Nguyen 等[19]基于对先前模型逆向攻击方法的考察，认为预测向量在目标类上的值比在其他类上的值更加重要，他们提出了一种新的最大对率（LOM, logit maximization）损失，提高了模型逆向攻击的效果。

2) 先验损失项

GMI 与 KED-MI 直接使用 DCGAN 的判别器 D 的输出结果作为损失函数先验损失项，这种先验损失项要求潜在向量能够使生成器生成能被判别器 D 判定为真样本的图像，如果 D 的训练效果较好，这种方法可以产生有意义的图像。VMI 攻击的目标不是重构目标图像，而是找到目标潜在向量分布，这种方法在先验损失项中约束潜在向量分布，使每个潜在向量分布与辅助数据集潜在向量分布（通常是标准高斯分布）之间的 KL（Kullback-Leibler）差异尽可能小，从而生成较真实的生成图像。

在一些模型逆向攻击方法[18-20]的优化阶段，损失函数没有先验损失项。这些方法通过一些其他方法确保生成图像的真实性。

2.1.4 白盒模型逆向攻击的其他技术

1) 图像变换

PPA 使用基于图像变换的方法增强生成图像的真实性和鲁棒性，防止出现局部最小和无意义图像问题。由于有意义的图像即使经过裁剪、水平反转等随机变换，在目标模型中目标类的置信度也会比较高，因此 PPA 在优化阶段得到生成图像后，把生成图像进行有随机性的变换，再通过目标模型进行预测。PLG-MI 方法也在优化阶段沿用了这种图像变换方法，而且由于 PLG-MI 在生成模型的训练中也用到了类别损失项，也需要对生成图像做随机变换。

另外，PPA 还使用基于图像变换的方法为优化阶段筛选初始潜在向量，并且受到仅标签成员推断攻击方法[45-46]的启发，提出了基于图像随机变换的生成图的筛选方法。

2) 增强模型

为了解决过拟合问题，Nguyen 等[19]提出了模型增强（MA, model augmentation）方法。用目标模型 M_{T} 标记辅助数据集，然后利用辅助数据集训练增强模型 M_{aug}。增强模型的输入和输出分别是图像样本和预测向量，与原目标模型的输入和输出相同，功能也相同，但它们具有不同的网络结构，以形成差异性。优化阶段的类别损失函数综合考虑目标模型损失 L_{id} 和增强模型损失 $L_{\mathrm{id}}^{\mathrm{aug}}$。由于原始目标模型与增强模型的差异性，仅通过 L_{id} 得到的过拟合生成图像 x' 的 $L_{\mathrm{id}}^{\mathrm{aug}}$ 往往较大，反之亦然。因此这种方法降低了过拟合的可能性，提高了优化过程的鲁棒性，起到了与损失函数先验损失项相似的作用。

2.2 黑盒模型逆向攻击

白盒模型逆向攻击已知目标系统的网络结构和参数的这种假设并不现实，因为大多数真实图像识别云服务并不向公众公开这些信息。实际上，目标模型经常被打包成黑盒，大部分位于客户端的敌手只能查询目标模型。因此，黑盒模型逆向攻击更具有现实意义。

2.2.1 基于敌手知识的黑盒模型逆向攻击

在黑盒场景中，敌手只能得到目标模型对于给定数据记录的输出，即标签和置信度向量。其中，标签代表目标模型对于给定数据记录所预测的类别，置信度向量代表分类类别的概率分布，且置信度向量中的每个分数代表相应类别的预测置信度。例如，假设目标模型可以分为三类：狗、猫和人类。对于给定的输入数据记录，模型的输出置信度向量可能是(0.2 0.3 0.5)，这表示模型预测给定的记录为狗、猫和人类，置信度分别为 0.2、0.3 和 0.5。在典型的黑盒模型逆向攻击中，攻击者训练一个攻击模型，仅使用置信度向量就可准确恢复样本。然而，在极端情况下，置信度向量会被目标模型的所有者隐藏起来，他们只公开输入数据的预测标签[47]。在上面的示例中，目标模型将“人类”作为标签输出。因此，根据敌手掌握目标模型的知识，黑盒模型逆向攻击可以分为以下两类。

1) 基于置信度的黑盒模型逆向攻击

当目标模型输出为置信度向量时，敌手可利用该向量对生成样本进行更加精准的优化，并且当敌手得到该目标模型与某个数据的预测时，可对该数据进行重构。

Yang 等[21]首次在黑盒场景下利用辅助数据集训练一个攻击模型来恢复样本，他们向攻击模型输入置信度向量，输出重构样本，该方法虽然简单有效，但是也有一定的局限性，即需要对每一个辅助样本进行一次查询，目标模型的查询次数为辅助数据集的大小。Dionysiou 等[33]为降低查询次数，提出了深度黑盒模型逆向（Deep-BMI, deep black-box model inversion）攻击方法，并且成功攻击图像识别模型的训练集数据。该方法假设敌手能够获得目标模型预测的置信度向量，并将其输入黑盒优化器中，以最大化目标类的置信度分数，并且黑盒优化器支持各种优化算法。在 Deep-BMI 中，目标模型的查询次数由所使用的黑盒优化器决定，与辅助数据集大小无关。与上述方法不同的是，Yoshimura 等[35]提出的方法中敌手只能获得最高的 5 个置信分数，而不是全部的置信分数，只需要最高的 5 个置信分数就可以实现模型逆向攻击。除上述方法外，An 等[36]还提出一种特殊的基于置信度向量的模型逆向攻击，他们假设敌手只能获得目标类的置信分数，无法获得其他类的置信度分数，他们将目标类的置信分数作为遗传算法[48]下一代精英选择的凭据。

上述攻击都针对图像数据进行模型逆向攻击。Mehnaz 等[32]针对结构化数据中的敏感属性，提出基于置信度的模型逆向属性推理（MIAI, model inversion attribute inference）攻击推断敏感属性值。他们假设敌手知道除了敏感属性值外的所有信息，包括真实标签。由于目标模型在训练过程中遇到了包含原始敏感属性值的目标记录，MIAI 攻击的关键思想是，当使用包含原始敏感属性值的记录进行查询时，目标模型返回的预测更有可能是正确的，置信度分数可能更高。反之，当使用包含错误敏感属性值的记录进行查询时，目标模型返回的预测更有可能是错误的。

2) 基于标签的黑盒模型逆向攻击

当目标模型的输出为仅标签时，敌手可进行的优化空间相较于置信度会受到很大限制。现阶段针对仅标签的模型逆向攻击可以总结为 3 种：利用预测标签估计真实置信度、利用预测标签推测敏感属性。

Zhu 等[37]提出针对仅标签的模型逆向攻击方法，利用该标签估计真实置信度，攻击流程如图 6 所示。该方法首先在辅助数据集上注入高斯噪声，从而获得目标模型的错误率 σ；然后，利用该错误率计算输入数据到目标模型决策边界的距离 $d=-\sigma\phi^{-1}(\mu)$，其中 $\phi(\mu)$ 是标准正态分布的累积分布函数；接着，敌手训练一个线性回归模型 $h(\boldsymbol{x})=\alpha(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)$ 来逼近目标模型，该模型的输出为置信度向量，线性回归模型中输入数据到决策边界的距离 $d=\dfrac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}{\|\boldsymbol{w}\|_2}=\dfrac{\alpha^{-1}(h(\boldsymbol{x}))}{\|\boldsymbol{w}\|_2}$。利用距离 d 可以得到 $h(\boldsymbol{x})=\dfrac{1}{1+\exp[\sigma\phi^{-1}(\mu)\|\boldsymbol{w}\|_2}$；最后，将标签转化为置信度向量，再利用 Yang 等[21]的方法，训练一个攻击模型完成模型逆向攻击。

同样，Mehnaz 等[32]针对标签也提出了敏感属性的模型逆向属性推理。敌手首先获得辅助数据集 $\mathrm{DS_A}$，该攻击步骤的关键思路是，如果目标模型仅针对敏感属性的一个可能值返回正确的预测，那么该值很可能代表了原始敏感属性的值。敌手收集所有符合上述条件的带标签记录，并得到 $\mathrm{DS_A}$ 数据集。接下来，敌手使用 $\mathrm{DS_A}$ 训练攻击模型，其中输入是目标记录的非敏感属性值集合，输出是对敏感属性值的预测。这一攻击步骤的关键目标是学习目标模型如何将敏感属性值与其他非敏感属性值以及目标模型的预测标签相关联。一旦攻击模型训练

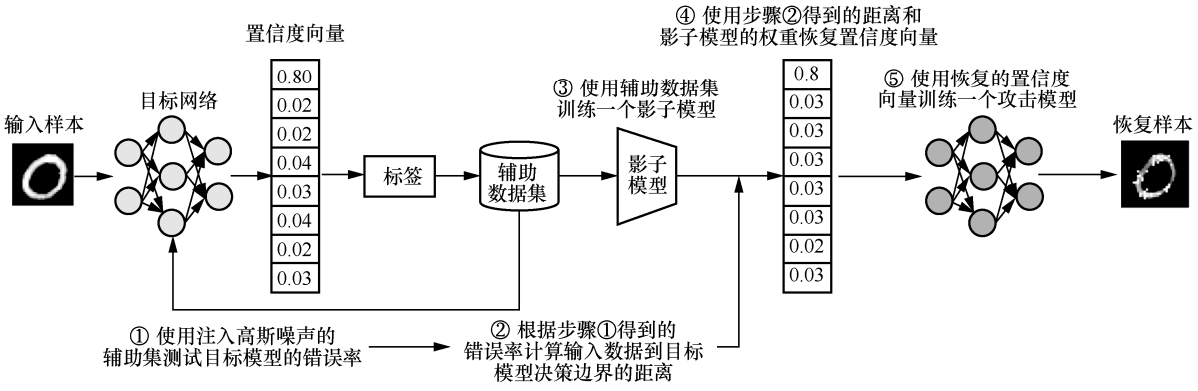


图 6　Zhu 等提出的模型逆向攻击流程

完成，敌手可以简单地使用目标记录的非敏感属性值查询攻击模型，并获取对敏感属性的预测。

3) 基于隐藏层特征的黑盒模型逆向攻击

当敌手在黑盒场景下可以获取隐藏层特征时，（如协作推理场景）可进行此类攻击。该攻击与上述 2 种不同，敌手获取目标模型的前几层特征输出，并基于此进行以下研究。

Yin 等[38]和 He 等[39]在协作推理场景下实现了黑盒模型逆向攻击。协作推理框架如图 7 所示，其中 $f_{\theta A}(\cdot)$ 为客户端的简单模型，$f_{\theta B}(\cdot)$ 为服务端的复杂模型。他们假设云提供者是敌手，敌手的知识就是目标模型中间隐藏层的特征。两者的区别在于攻击模型的训练。He 等的攻击模型是利用辅助数据集训练得到的，而 Yin 等的攻击模型则没有采用辅助数据集，而是采用自然进化策略进行训练。这种策略可以帮助攻击模型逐步优化攻击性能，并且不依赖于额外的训练数据。

2.2.2 基于优化目标的黑盒模型逆向攻击

由于无法获得目标模型的网络结构和参数，黑盒模型逆向攻击重构出的数据通常效果较差，因此，黑盒模型逆向攻击的重点在于优化过程。根据攻击流程，现有的研究大致可以分为两类，一类是优化攻击模型，另一类是优化潜在向量。

1) 优化攻击模型的黑盒模型逆向攻击

在训练攻击模型阶段，Yang 等[21]将生成图像输入目标模型，获得其置信度向量，并利用该向量与隐私图像的置信度向量构建损失函数，通过不断优化攻击模型实现更真实的隐私图像重构。而 Dionysiou 等[33]采用黑盒优化技术对攻击模型编码端进行结构化扰动，使攻击模型生成的图像最大化目标类别的置信度分数。Yin 等[38]和 He 等[39]在协作推理场景下根据中间隐藏层的特征训练攻击模型。He 等[39]将辅助数据集逐一输入目标模型，然后将目标模型的输出再输入到攻击模型，最后计算攻击模型的生成与辅助数据集的损失函数。

由于损失函数不涉及目标模型，则可以直接利用梯度下降法优化攻击模型。Yin 等不借助辅助数据集，因此在最小化目标模型隐私数据 $\boldsymbol{x}$ 的输出 $f_{\theta A}(\boldsymbol{x})$ 和目标模型生成数据的输出 $f_{\theta A}(\hat{\boldsymbol{x}})$ 的过程中，即 $\min\left|\left|f_{\theta A}(\boldsymbol{x})-f_{\theta A}(\hat{\boldsymbol{x}})\right|\right|_2=\min \mathrm{DIS}(\boldsymbol{x},\hat{\boldsymbol{x}})$，$\frac{\partial \mathrm{DIS}(\boldsymbol{x},\hat{\boldsymbol{x}})}{\partial \boldsymbol{x}}$ 不能直接获得，Yin 等借助自然进化策略来估计该梯度。第一步随机采样对称的高斯噪声 $\delta_1,\delta_2,\cdots,\delta_n$；第二步在生成数据上加上噪声和标准差的乘积 $\hat{\boldsymbol{x}}_i=\hat{\boldsymbol{x}}+\sigma\delta_i$；第三步计算损失值 $\mathrm{DIS}(\boldsymbol{x},\hat{\boldsymbol{x}}_1),\mathrm{DIS}(\boldsymbol{x},\hat{\boldsymbol{x}}_2),\cdots,\mathrm{DIS}(\boldsymbol{x},\hat{\boldsymbol{x}}_n)$；最后估计梯度 $\frac{\partial \mathrm{DIS}(\boldsymbol{x},\hat{\boldsymbol{x}})}{\partial \boldsymbol{x}}=\frac{1}{n\sigma}\sum_{i=1}^{n}\delta_i \mathrm{DIS}(\boldsymbol{x},\hat{\boldsymbol{x}}_i)$。这种方法在目标模型只有一层 DNN 时效果最好，实现了在不借助辅助数据集的情况下实现黑盒模型逆向攻击。

2) 优化潜在向量的黑盒模型逆向攻击

为了在有限的知识下进行更精准的黑盒攻击，目前大多数研究集中在优化潜在向量的黑盒模型逆向攻击中。首先训练一个生成式模型，或直接利用预训练的生成式模型作为攻击模型；其次采用优化算法针对潜在向量进行优化，其中的优化算法即研究重点。

Kahla 等[22]和 Yoshimura 等[35]都基于 GAN 训练的生成器作为攻击模型，将攻击模型生成的样本输入目标模型进行预测，并利用该预测进行优化以使生成样本靠近目标类。Kahla 等的目标是找到一个生成图，使该生成图像属于目标类的置信分数与其他类置信分数差别最大，即

$$\max_{\boldsymbol{x}} M_c^*(\boldsymbol{x}), M_c^*(\boldsymbol{x})=$$
$$f_c^*(\boldsymbol{x})-\max_{c\neq c^*} f_c(\boldsymbol{x}), \boldsymbol{x}=G(\boldsymbol{z})$$

其中，c^* 为目标标签，f 为目标模型，G 为 GAN 模型。在黑盒场景下，敌手不知道目标模型的参数，不能直接使用随机梯度下降算法优化损失函数，并且 $\boldsymbol{x}$ 位于高维数据空间，优化此空间容易陷入局部

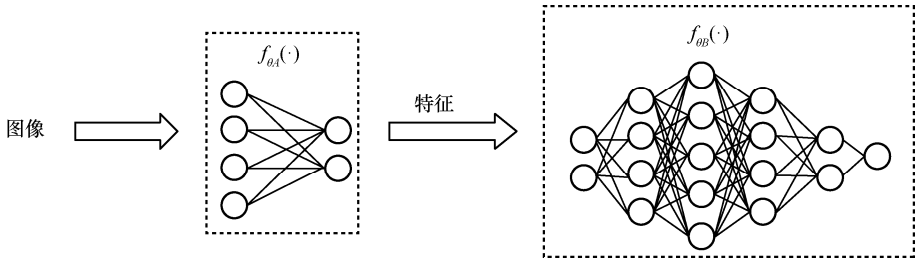


图 7 协作推理框架

最小值。于是，Kahla 等提出梯度估计器，将优化生成图像 $\boldsymbol{x}$ 转换成优化 GAN 的潜在向量，旨在寻找更新潜在向量的方向。具体地，在高维球体上采样点，并查询它们的标签，不属于目标类的点所在的方向是要远离的方向，上取这些点的平均值，朝着平均值相反的方向继续采样，直到所有点都被预测到目标类中，然后再增加采样的半径，重复上述过程。

$$\varphi_{c^*}(z)=\frac{\operatorname{sign}\left(M_c^*(z)\right)-1}{2}=\begin{cases}0, c^*=\arg\max\limits_{c\in C} f_c(G(z))\\ 1, \text{其他}\end{cases}$$

其中，$\varphi_c^*(z)$ 标记不被预测到目标类的点；sign(·) 是一个函数，输入为正则返回 1，输入为负则返回−1。梯度估计器的表达式为

$$\widehat{M_c^*}(z,R)=\frac{1}{N}\sum_{i=1}^{N}\varphi_c^*(z+Ru_n)u_n$$

其中，R 为采样球体的半径，u_n 为球体内的随机点，N 为采样点数目。最终潜在向量更新的表达式为 $z\leftarrow z+\leftarrow\alpha\widehat{M_c^*}(z,R)$，$\alpha$ 为更新的步长。Yoshimura 的目标则是最小化生成图目标模型的预测值 $\boldsymbol{y}$ 与目标预测值 $\hat{\boldsymbol{y}}$ 的损失函数 $L(\boldsymbol{y},\hat{\boldsymbol{y}})$。他们同样将损失函数转化成关于潜在向量的函数，即 $\boldsymbol{z}^{(t+1)}=\boldsymbol{z}^t-\alpha\dfrac{\partial L}{\partial \boldsymbol{z}}(\boldsymbol{z}^t)$，由于损失函数 $L(\boldsymbol{y},\hat{\boldsymbol{y}})$ 涉及目标模型，Yoshimura 等使用一个扰动向量 $\boldsymbol{\epsilon}$ 来近似梯度，即 $g=\dfrac{\partial L}{\partial \boldsymbol{z}}(\boldsymbol{z}^t)=\dfrac{L(\boldsymbol{z}^{(t)}+\boldsymbol{\epsilon})-L(\boldsymbol{z}^{(t)})}{(\boldsymbol{z}^{(t)}+\boldsymbol{\epsilon}-\boldsymbol{z}^{(t)})}$，当 $\boldsymbol{\epsilon}$ 接近 0 时，g 趋近真实值。

Han 等[34]也使用 GAN 作为生成器，但与前面提到的 2 种方法不同的是，他们没有通过估计梯度来优化潜在向量，而是利用马尔可夫决策过程（MDP, Markov decision process）搜索 GAN 潜在空间。在这个过程中，状态和行为都是潜在向量，行为引导初始向量走向高回报的最终向量，即能够生成最逼近目标类图像的向量，其中目标类置信分数越高，回报越高。该方法训练了一个代理，这个代理指导马尔可夫决策过程，并通过状态转变来优化潜在向量。攻击过程如图 8 所示，其中，S_0 为初始样本，S_{term} 为优化后的样本。

3　模型逆向攻击的防御

模型逆向攻击通过本地上传的模型参数可进行白盒攻击，通过本地或全局模型的预测可进行黑盒攻击。这种重构模型训练数据的细粒度攻击不仅使训练数据可见，也威胁到数据拥有者的隐私，因此防御模型逆向攻击是一项未来必要的研究课题。本节将从基于加密技术、基于扰动技术、基于深度学习训练技术 3 个方面进行比较和分析。

3.1　基于加密技术防御

加密技术是一种常用的保护数据隐私和安全的方法，它可以通过对数据或模型进行加密变换，使攻击者无法直接获取原始信息。现有的研究主要应用安全多方计算、同态加密技术防御模型逆向攻击。

在联邦学习训练阶段，模型聚合过程可以视为在不泄露任何参与方私有信息的前提下，安全地计算一个函数的值，其思想与安全多方计算不谋而合。Xu 等[23]提出了 HybridAlpha 算法通过可信第三方认证机构（TPA, third party administrator）生成公钥和私钥，将公钥分发给参与聚合的用户。与其他

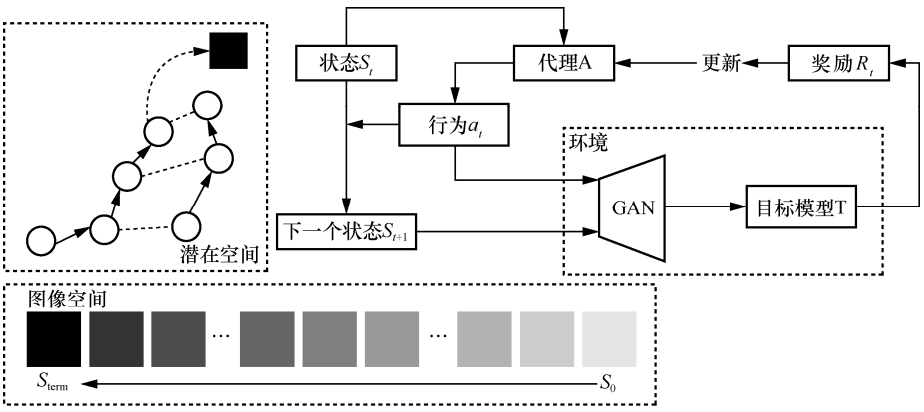


图 8　Han 等[34]的模型逆向攻击过程

的基于加密的安全多方计算解决方案相比，HybridAlpha 在提供同样的模型性能和隐私保证的情况下需要的训练时间和数据传输量更少。

近年来，随着隐私计算能力的提升以及人们对于隐私保护的需要，同态加密技术也逐渐引入联邦学习中。Cheng 等[24]提出了无损隐私保护树提升系统 SecureBoost，并且为了保证数据安全，使用同态加密技术进行处理。该系统为了能够使用同态加密，对数据进行处理以去除无法进行同态加密的运算。

3.2 基于扰动技术防御

通过加密技术保障联邦学习的安全性对传输效率的影响较大，参与方与中心方对加密与解密的操作增加了计算开销；加密后的参数较原始参数占用更多空间，增加了通信开销等。差分隐私技术作为一种量化数据隐私泄露风险的方法，它可以通过在数据或模型中添加一定的噪声，来保证个体数据的不可区分性。相比于基于加密技术的防御策略，差分隐私技术降低了通信开销，提高了传输效率。

Zhang 等[15]通过实验证明差分隐私随机梯度下降（DPSGD, differential privacy stochastic gradient descent）[25]并不能防御他们提出的 GMI 攻击，隐私预算的减小也不影响攻击精确度。在随后的研究中，Wang 等[26]通过理论证明了为什么差分隐私不能防御模型逆向攻击。因此针对目标模型输出的扰动技术相继被提出，例如，Wen 等[27]提出在模型输出中加入对抗性噪声使逆向误差最大化，并给受害者模型带来可忽略的效用损失；Yang 等[49]提出净化器框架来防御模型逆向攻击，该框架减少了模型输出在训练数据集的成员和非成员上的分散性，从而削弱训练样本和预测输出之间的关联性。

3.3 基于深度学习训练技术防御

基于深度学习训练技术的防御策略是一种优化训练过程来抵抗模型逆向攻击的方法。该技术使模型在保持良好的性能的同时，增加敌手攻击结果的不确定性或者误导性。

Peng 等[28]提出双边依赖优化（BiDO, bilateral dependency optimization）策略，最小化潜在特征空间与输入之间的依赖性，同时最大化潜在特征空间与真实标签之间的依赖性。前者限制了输入到潜在特征的冗余信息，提高模型防止隐私泄露的能力；后者促使潜在特征空间判别特征，确保模型的效用。

虽然 BiDO 策略能从本地训练阶段提供理论较好的防御，但是并不能很好地防御最先进的模型逆向攻击，如 MIRROR[36]。于是 Gong 等[29]提出 NetGuard，通过在本地训练阶段插入对抗样本微调目标模型来误导 MIA。

3.4 模型逆向攻击的落地防御应用

目前，联邦学习结合上述防御策略已经开始了在行业领域的落地探索，在不同行业有多样化的应用场景和落地形态[50]。

现有联邦学习开源工具中的隐私计算方法主要由加密技术和扰动技术组成。例如，微众银行 AI 团队开源的工业级联邦学习技术框架 FATE（federated AI technology enabler）[24,51]、OpenMinded 社区开发的 Pysyft[52]、矩阵元开发的 Rosetta[53]等工具采用基于安全多方计算和同态加密的加密技术；Pysyft 与百度开发的 PaddleFL[54]等工具采用 DPSGD 等基于差分隐私的扰动技术。这些防御策略针对训练阶段可能遭受的白盒攻击能起到防御效果。除了上述开源工具中的理论防御措施，在实际工业界也有相应的探索。

在金融领域，保险公司联合科技公司共同搭建隐私计算平台的案例[55]，提供了安全多方计算和联邦学习相结合的隐私计算来抵御模型逆向攻击。

在政务领域，中山市政务服务数据管理局联合科技公司，有应用基于联邦学习、安全多方计算、可信执行环境为核心的防御策略案例；或是科技公司联手，基于洞见数智联邦平台提供支持安全多方计算和联邦学习结合的防御应用[55]，并通过联邦区块链保证过程中免于模型逆向攻击。

在医疗领域，蚂蚁集团基于蚂蚁隐私计算平台“隐语”提供的安全多方计算及联邦学习技术，通过阿里云医疗大数据管理平台[55]，实现模型逆向攻击的防御。

即使已有多种策略已落地应用，但皆存在应用部署的难点，如加密算法对性能的影响、资源因素的影响、多方协同的“木桶效应”、软硬件资源的多样性、部署后的监查审计[55,69]等问题。具有普适性、公平性、隐私与效用权衡的主动防御策略仍是联邦学习中的研究重点。

4 模型逆向攻击的评估指标和数据集

针对不同目标模型、训练数据集、敌手背景知识等，全面的评估指标可以多角度地反映模型逆向

攻击的攻击精准性、恢复相似性以及恢复真实性能力。其中，攻击评估指标可以分为定性评估和定量评估，定性评估依赖于人类的视觉判断，定量评估依赖于数学计算。目前已有多种评估指标和数据集被提出和使用，但其中仍存在一些问题与挑战，如缺乏统一的标准、难以覆盖不同的攻击场景等。本节对模型逆向攻击的定量评估指标以及使用的数据集进行详细的分析与比较。

4.1　模型逆向攻击的评估指标

本节主要介绍模型逆向攻击评估指标。

1) 攻击准确率（Acc, accuracy）。通过训练一个比目标模型框架更复杂、精度更高的评估模型作为人类判断的代理，对攻击模型恢复的样本进行分类。攻击准确率指评估模型能够将恢复样本分类到目标类占所有样本的比率，或置信分数前 5 对应的类中包含目标类占所有样本的比率。

2) 攻击精确率（Pre, attack precision）。在评估模型预测为目标类的样本中，预测正确的目标类样本占所有预测为目标类样本的比率。

3) 攻击召回率（Rec, attack recall）。评估模型预测正确的目标类样本占所有恢复的目标类样本的比率。

4) K 近邻距离（KNN Dist, k-nearest neighbor distance）。给定目标类，并计算恢复样本与真实目标样本的最短特征距离。该距离通过评估模型倒数第二层输出的特征空间进行 L2 范数距离计算。

5) 特征距离（Feat Dist, feature distance）。通过 L2 范数距离，计算恢复样本特征与真实目标样本特征质心间的距离。该特征通过评估模型倒数第二层得到。

6) 弗雷歇距离（FID, Fréchet inception distance）。通过预训练的 Inception-v3 模型，计算恢复样本与真实样本之间的特征相似性。该指标越低说明两组样本越相似。

7) 均方误差/数据逆向误差（MSE, mean squared error）。计算恢复样本与目标样本之间像素值的误差距离。

8) 峰值信噪比（PSNR, peak signal-to-noise ratio）。图像的最大平方像素波动与目标和恢复样本之间的均方误差之比。该指标评估像素级的恢复质量。

9) 结构相似性（SSIM, structural similarity）。衡量两组图像的结构相似度，该指标考虑了两幅图像的亮度、对比度和结构。SSIM 是一个 0~1 的单一数值，其中 0 代表最不相似，1 表示最相似。

4.2　数据集

数据集是模型逆向攻击的重要影响因素，因为作为敌手可以利用的辅助数据集，其数据分布、类型直接影响攻击结果。本节将总结现有攻击常用的数据集及其特性，包括数据类别、数据任务、数据集名称、样本个数、类别个数、特征维度，如表 2 所示。

表 2　　**模型逆向攻击的常用数据集**

数据类别	数据任务	数据集名称	样本个数	类别个数	特征维度	参考文献
图像	分类	MNIST[56]	70 000	10	28×28×1	文献[15-16,19,21,32-33,37,39]
		Fashion-MNIST[57]	70 000	10	28×28×1	文献[19,33]
		CIFAR-10[58]	60 000	10	32×32×3	文献[16,18-19,21,37-39]
		CelebA[59]	202 599	10 177	218×178×3	文献[15-16,19-22,25,33-34,37]
		FaceScrub[60]	100 000	530	—	文献[16,18,20-22,34,37-38]
		ChestX-Ray8[61]	108 948	32 717	1 024×1 024×1	文献[15-16,31]
		Stanford dogs[62]	20 580	120	—	文献[18]
		AT&T Face[63]	400	40	92×112×1	文献[14,33]
		Pubfig83[64]	13 600	83	—	文献[22,34]
		VGGFace2[65]	3 310 000	9 131	—	文献[35-36]
	生成	Flickr-Faces-HQ[41]	70 000	—	1 024×1 024×3	文献[16,18-20,34,36]
		MetFaces[66]	1 336	—	1 024×1 024×3	文献[18]
		Animal Faces-HQ Dogs[67]	15 000	—	512×512×3	文献[18]
结构化	分类	IWPC[68]	5 700	—	—	文献[13]

5 主要挑战和未来研究方向

在模型逆向攻击研究领域，尽管已经取得了一些成果，但仍然存在许多挑战和问题。本节将探讨这些主要挑战，并提出未来研究的可能方向，如表 3 所示。

5.1 模型逆向攻击的主要挑战和研究方向

现有的模型逆向攻击方法可以在多种场景上取得较好的攻击效果，但这些方法在普适性、高效性和准确率等方面还存在诸多不足。本文将模型逆向攻击的一些研究方向列举如下。

5.1.1 研究普适性的跨领域模型逆向攻击方案

目前大部分研究仅涉及对一种或几种相同领域的目标模型攻击效果实验和评估，往往无法在不同领域的目标模型上展开相同的攻击，如图像识别领域与自然语言处理领域的攻击方法无法平移。因为不同领域的模型存在着不同的结构和规模，这使设计出一种既具有普适性的模型逆向攻击方法变得极其困难。因此，如何克服这个挑战，设计出能够适应跨领域的模型逆向攻击方法，将是未来研究的一个重要方向。

5.1.2 研究高效性和可迁移性更强的白盒模型逆向攻击方案

一些模型逆向攻击方法通过访问目标模型，用辅助数据集训练逆向攻击的专用生成模型，生成了具有很高攻击准确率的优质图像。但是这种利用专用生成模型的攻击无法针对同领域但不同结构和规模的目标模型进行模型迁移。此外，专用生成模型训练过程耗时较长，在对多个目标模型实施攻击时，需要花费更多的时间。一些方法利用在公开数据集上预训练的非专用生成模型实施攻击，效率较高但准确率偏低。因此，对攻击优化阶段进行改进，使预训练攻击模型也能达到较好的攻击效果；或者研究一种可迁移的训练框架，降低多次攻击开支的方法将是未来的研究方向。

5.1.3 研究准确率高和访问次数少的黑盒模型逆向攻击方案

白盒模型逆向攻击虽然准确率更高，但威胁模型更加严苛。黑盒访问目标模型时，有必要提升黑盒模型逆向攻击的准确率。另一方面，在实际的攻击情境中，敌手往往也不能无限次地访问黑盒目标模型，因此需要研究如何在黑盒模型逆向攻击中减少对目标模型的访问次数。针对性地优化黑盒模型逆向攻击的框架，或提出新的潜在向量优化算法将会是未来的研究方向。

5.2 模型逆向攻击防御的主要挑战和研究方向

模型逆向攻击防御的方法尚不成熟，可以从以下方面开展研究。

5.2.1 研究普适性更高的本地模型训练策略

已有的基于深度学习训练技术的防御策略不具有普适性，需要参与方根据本地数据设计攻击模型以生成对抗样本，再对其模型进行微调；而其他防御策略不能抵御最先进的模型逆向攻击。因此，需要研究和设计一种普适性更高的本地模型训练策略以抵御攻击。

5.2.2 研究隐私性与可用性均衡的防御策略

目前，大部分模型逆向攻击的防御策略无法做到隐私性与可用性的权衡，即使做到权衡也很难应用于更深层的神经网络模型中，并且很少研究不同防御设置对隐私性与可用性的影响。因此，未来可以根据现如今发展迅猛的深度技术，挖掘更多可能的防御策略，以做到隐私性与可用性的权衡，并探讨不同防御设置的影响。

表 3 联邦学习中模型逆向攻击和防御的主要挑战和未来研究方向

研究内容	主要挑战	未来研究方向
模型逆向攻击	相同的模型逆向攻击方法无法针对不同领域的目标模型展开攻击	研究并探索具有普适性、适应性的跨领域模型逆向攻击方法
	白盒模型逆向攻击方法训练攻击模型耗时较长，且无法针对同领域但不同结构和规模的目标模型进行迁移	研究可迁移的方法训练白盒模型逆向攻击的攻击模型
	黑盒模型逆向攻击方法在一些情况下的准确率较差，且无法在有限访问次数下进行有效的攻击	研究黑盒模型逆向攻击的框架，提出新的潜在向量优化方法，提升准确率，减少访问次数
模型逆向攻击防御	基于深度学习训练技术的防御策略无法运用到各参与方的本地训练中	研究和设计一种普适性更强的本地模型训练策略
	已有研究忽视防御策略的设置对于可用性和隐私性的影响，如差分隐私中隐私预算、模型参数大小	研究和设计一种隐私性与可用性均衡的防御策略，并探讨不同设置的影响

6　结束语

随着人工智能技术的发展和应用，数据安全和隐私保护问题日益突出。联邦学习作为一种解决数据孤岛问题的分布式机器学习技术，在训练过程中也面临着隐私攻击的挑战。本文重点关注联邦学习隐私攻击中的模型逆向攻击，概括并详细分析了该攻击的理论框架，并根据威胁模型将模型逆向分为白盒、黑盒攻击，对现有的攻击方法进行分析与比较。此外，本文还根据不同技术类型对防御策略进行分析与比较，对现有模型逆向攻击常用的评估标准及数据集进行汇总。最后，本文总结了模型逆向攻击现有的主要挑战以及未来研究方向。

参考文献：

[1] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

[2] WANG S, KANG B, MA J L, et al. A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19)[J]. European Radiology, 2021, 31(8): 6096-6104.

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

[4] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[J]. arXiv Preprint, arXiv: 1602.05629, 2016.

[5] SONG C Z, RISTENPART T, SHMATIKOV V. Machine learning models that remember too much[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2017: 587-601.

[6] 牛俊, 马骁骥, 陈颖, 等. 机器学习中成员推理攻击和防御研究综述[J]. 信息安全学报, 2022, 7(6): 1-30.
NIU J, MA X J, CHEN Y, et al. A survey on membership inference attacks and defenses in Machine Learning[J]. Journal of Cyber Security, 2022, 7(6): 1-30.

[7] SUN L C, QIAN J W, CHEN X. LDP-FL: practical private aggregation in federated learning with local differential privacy[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2021: 1571-1578.

[8] PAPERNOT N, ABADI M, ERLINGSSON Ú, et al. Semi-supervised knowledge transfer for deep learning from private training data[J]. arXiv Preprint, arXiv: 1610.05755, 2016.

[9] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction APIs[C]//Proceedings of the 25th USENIX Conference on Security Symposium. Berkeley: USENIX Association, 2016: 601-618.

[10] SHOKRI R, STRONATI M, SONG C Z, et al. Membership inference attacks against machine learning models[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 3-18.

[11] 刘艺璇, 陈红, 刘宇涵, 等. 联邦学习中的隐私保护技术[J]. 软件学报, 2022, 33(3): 1057-1092.
LIU Y X, CHEN H, LIU Y H, et al. Privacy-preserving techniques in federated learning[J]. Journal of Software, 2022, 33(3): 1057-1092.

[12] 陈明鑫, 张钧波, 李天瑞. 联邦学习攻防研究综述[J]. 计算机科学, 2022, 49(7): 310-323.
CHEN M X, ZHANG J B, LI T R. Survey on attacks and defenses in federated learning[J]. Computer Science, 2022, 49(7): 310-323.

[13] FREDRIKSON M, LANTZ E, JHA S, et al. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing[C]//Proceedings of the USENIX Security Symposium. Berkeley: USENIX Association, 2014: 17-32.

[14] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2015: 1322-1333.

[15] ZHANG Y H, JIA R X, PEI H Z, et al. The secret revealer: generative model-inversion attacks against deep neural networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 250-258.

[16] CHEN S, KAHLA M, JIA R X, et al. Knowledge-enriched distributional model inversion attacks[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 16158-16167.

[17] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[18] STRUPPEK L, HINTERSDORF D, CORREIA A D A, et al. Plug & play attacks: towards robust and flexible model inversion attacks[J]. arXiv Preprint, arXiv: 2201.12179, 2022.

[19] NGUYEN N B, CHANDRASEGARAN K, ABDOLLAHZADEH M, et al. Re-thinking model inversion attacks against deep neural networks[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 16384-16393.

[20] YUAN X, CHEN K, ZHANG J, et al. Pseudo label-guided model inversion attack via conditional generative adversarial network[J]. arXiv Preprint, arXiv: 2302.09814, 2023.

[21] YANG Z Q, ZHANG J Y, CHANG E C, et al. Neural network inversion in adversarial setting via background knowledge alignment[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2019: 225-240.

[22] KAHLA M, CHEN S, JUST H A, et al. Label-only model inversion attacks via boundary repulsion[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 15025-15033.

[23] XU R H, BARACALDO N, ZHOU Y, et al. HybridAlpha: an efficient approach for privacy-preserving federated learning[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2019: 13-23.

[24] CHENG K W, FAN T, JIN Y L, et al. SecureBoost: a lossless federated learning framework[J]. IEEE Intelligent Systems, 2021, 36(6): 87-98.

[25] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM

Press, 2016: 308-318.

[26] WANG T H, ZHANG Y H, JIA R X. Improving robustness to model inversion attacks via mutual information regularization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(13): 11666-11673.

[27] WEN J, YIU S M, HUI L C K. Defending against model inversion attack by adversarial examples[C]//Proceedings of 2021 IEEE International Conference on Cyber Security and Resilience (CSR). Piscataway: IEEE Press, 2021: 551-556.

[28] PENG X, LIU F, ZHANG J F, et al. Bilateral dependency optimization: defending against model-inversion attacks[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2022: 1358-1367.

[29] GONG X L, WANG Z Y, CHEN Y J, et al. NetGuard: protecting commercial Web APIs from model inversion attacks using GAN-generated fake samples[C]//Proceedings of the ACM Web Conference 2023. New York: ACM Press, 2023: 2045-2053.

[30] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv Preprint, arXiv: 1511.06434, 2015.

[31] WANG K C, FU Y, LI K, et al. Variational model inversion attacks[J]. Advances in Neural Information Processing Systems, 2021, 34: 9706-9719.

[32] MEHNAZ S, DIBBO S V, DE-VITI R, et al. Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models[J]. arXiv Preprint, arXiv: 2201.09370, 2022.

[33] DIONYSIOU A, VASSILIADES V, ATHANASOPOULOS E. Exploring model inversion attacks in the black-box setting[J]. Proceedings on Privacy Enhancing Technologies, 2023(1): 190-206.

[34] HAN G, CHOI J, LEE H, et al. Reinforcement learning-based black-box model inversion attacks[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 20504-20513.

[35] YOSHIMURA S, NAKAMURA K, NITTA N, et al. Model inversion attack against a face recognition system in a black-box setting[C]//Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Piscataway: IEEE Press, 2022: 1800-1807.

[36] AN S W, TAO G H, XU Q L, et al. MIRROR: model inversion for deep learning network with high fidelity[C]//Proceedings of 2022 Network and Distributed System Security Symposium. Reston: Internet Society, 2022: 1-18.

[37] ZHU T Q, YE D Y, ZHOU S, et al. Label-only model inversion attacks: attack with the least information[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 991-1005.

[38] YIN Y P, ZHANG X L, ZHANG H L, et al. Ginver: generative model inversion attacks against collaborative inference[C]//Proceedings of the ACM Web Conference. New York: ACM Press, 2023: 2122-2131.

[39] HE Z C, ZHANG T W, LEE R B. Model inversion attacks against collaborative inference[C]//Proceedings of the 35th Annual Computer Security Applications Conference. New York: ACM Press, 2019: 148-162.

[40] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks[J]. arXiv Preprint, arXiv: 1802.05957, 2018.

[41] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 4396-4405.

[42] BLEI D M, KUCUKELBIR A, MCAULIFFE J D. Variational inference: a review for statisticians[J]. Journal of the American Statistical Association, 2017, 112(518): 859-877.

[43] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 39-57.

[44] SRIRAMANAN G, ADDEPALLI S, BABURAJ A, et al. Guided adversarial attack for evaluating and enhancing adversarial defenses[J]. arXiv Preprint, arXiv: 2011.14969, 2020.

[45] CHOQUETTE-CHOO C A, DULLERUD N, DZIEDZIC A, et al. CaPC learning: confidential and private collaborative learning[J]. arXiv Preprint, arXiv: 2102.05188, 2021.

[46] LI Z, ZHANG Y. Membership leakage in label-only exposures[C]//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2021: 880-895.

[47] CHOQUETTE-CHOO C A, TRAMER F, CARLINI N, et al. Label-only membership inference attacks[J]. arXiv Preprint, arXiv: 2007.14321, 2020.

[48] BHANDARI D, MURTHY C A, PAL S K. Genetic algorithm with elitist model and its convergence[J]. International Journal of Pattern Recognition and Artificial Intelligence, 1996, 10(6): 731-747.

[49] YANG Z, WANG L, YANG D, et al. Purifier: defending data inference attacks via transforming confidence scores[J]. arXiv Preprint, arXiv: 2212.00612, 2022.

[50] 微众银行 AI 项目组. 联邦学习白皮书 V1.0[R]. 2019.
WeBank AI Project Team. Federated learning white paper V1.0[R]. 2019.

[51] 微众银行 AI 项目组. 联邦学习开源平台 FATE [R]. 2019.
WeBank AI Project Team. Federated learning open source platform FATE [R]. 2019.

[52] ZILLER A, TRASK A, LOPARDO A, et al. Pysyft: a library for easy federated learning[J]. Federated Learning Systems: Towards Next-Generation AI, 2021: 111-139.

[53] CHEN Y A, HUANG G, SHI J, et al. Rosetta: a privacy-preserving framework based on TensorFlow[E]. 2020.

[54] MA Y, YU D, WU T, et al. PaddleFL: an open-source deep learning platform from industrial practice[J]. Frontiers of Data and Domputing, 2019, 1(1): 105-115.

[55] 隐私计算联盟, 中国信息通信研究院云计算与大数据研究院. 隐私计算应用研究报告[R]. 2022.
Privacy Computing Alliance, Cloud Computing and Big Data Research Institute of China Academy of Information and Communications Technology. Privacy computing application research report[R]. 2022.

[56] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[57] XIAO H, RASUL K, VOLLGRAF R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms[J]. arXiv Preprint, arXiv: 1708.07747, 2017.

[58] KRIZHEVSKY A. Learning multiple layers of features from tiny images[D]. Tront: University of Tront, 2009.

[59] LIU Z W, LUO P, WANG X G, et al. Deep learning face attributes in the wild[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2016: 3730-3738.

[60] NG H W, WINKLER S. A data-driven approach to cleaning large face datasets[C]//Proceedings of IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2015: 343-347.

[61] WANG X S, PENG Y F, LU L, et al. ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 3462-3471.

[62] KHOSLA A, JAYADEVAPRAKASH N, YAO B, et al. Novel dataset for fine-grained image categorization: Stanford dogs[C]//Workshop on Fine-grained Visual Categorization (FGVC). Piscataway: IEEE Press, 2011: 1-2.

[63] ADDLESEE M, CURWEN R, HODGES S, et al. Implementing a sentient computing system[J]. Computer, 2001, 34(8): 50-56.

[64] PINTO N, STONE Z, ZICKLER T, et al. Scaling up biologically-inspired computer vision: a case study in unconstrained face recognition on facebook[C]//Proceedings of CVPR 2011 WORKSHOPS. Piscataway: IEEE Press, 2011: 35-42.

[65] CAO Q, SHEN L, XIE W D, et al. VGGFace2: a dataset for recognising faces across pose and age[C]//Proceedings of 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Piscataway: IEEE Press, 2018: 67-74.

[66] KARRAS T, AITTALA M, HELLSTEN J, et al. Training generative adversarial networks with limited data[J]. arXiv Preprint, arXiv: 2006.06676, 2020.

[67] CHOI Y, UH Y, YOO J, et al. StarGAN v2: diverse image synthesis for multiple domains[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 8185-8194.

[68] The International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data[J]. New England Journal of Medicine, 2009, 360(8): 753-764.

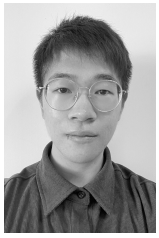
[69] 贾轩, 白玉真, 马智华. 隐私计算应用场景综述[J]. 信息通信技术与政策, 2022(5): 45-52.
JIA X, BAI Y Z, MA Z H. Overview of privacy preserving computing application scenarios[J]. Information and Communications Technology and Policy, 2022(5): 45-52.

[作者简介]

王冬（1990− ），女，山东泰安人，博士，杭州电子科技大学讲师，主要研究方向为人工智能安全、隐私计算等。

秦倩倩（2000− ），女，湖北随州人，杭州电子科技大学博士生，主要研究方向为人工智能安全、隐私计算等。

郭开天（2000− ），男，山东菏泽人，杭州电子科技大学博士生，主要研究方向为人工智能安全、隐私计算等。



刘容轲（1999− ），男，安徽安庆人，杭州电子科技大学博士生，主要研究方向为人工智能安全、隐私计算等。

颜伟鹏（2001− ），男，福建泉州人，杭州电子科技大学博士生，主要研究方向为人工智能安全、数据安全等。

任一支（1981− ），男，安徽枞阳人，博士，杭州电子科技大学教授，主要研究方向为大数据安全、人工智能、区块链、知识图谱。

罗清彩（1978− ），男，山东青岛人，山东浪潮科学院有限公司高级工程师，主要研究方向为隐私计算、人工智能安全等。

申延召（1984− ），男，河南汝州人，博士，山东区块链研究院高级工程师，主要研究方向为隐私计算、人工智能安全、密码学等。