基于注意力特征解耦的跨年龄身份成员推理

刘宇璐, 武淑红, 于丹, 马垚, 陈永乐

引用本文

刘宇璐, 武淑红, 于丹, 马垚, 陈永乐. 基于注意力特征解耦的跨年龄身份成员推理[J]. 计算机科学, 2024, 51(9): 401-407.

LIU Yulu, WU Shuhong, YU Dan, MA Yao, CHEN Yongle. Cross-age Identity Membership Inference Based on Attention Feature Decomposition [J]. Computer Science, 2024, 51(9): 401-407.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600112

基于注意力特征解耦的跨年龄身份成员推理

刘宇璐　武淑红　于　丹　马　垚　陈永乐
太原理工大学计算机科学与技术学院(大数据学院)　山西 晋中 030600
(2893066290@qq.com)

摘　要　生成对抗网络(GANs)模型可以生成高分辨率的“不存在”的物体真实图像,近期被广泛应用于各种人工合成数据,尤其是人脸图像生成领域。然而,由于基于该模型的人脸生成器通常需要根据不同身份高度敏感的面部图像进行训练,其中存在潜在数据泄露使得攻击者能够对身份成员关系进行推断的问题。为此,首先设计对查询身份所获取样本与其实际参与训练样本之间存在巨大差异时的身份成员推理攻击,这些差异会导致基于样本推理身份成员关系的性能急剧下降;其次,在此基础上设计基于各身份解耦表征的重建误差攻击方案,在最大化消除不同样本间背景姿势等因素影响的同时,消除巨大年龄跨度导致的表征差异,进一步提高了攻击性能;最后,基于3个代表性的人脸数据集在3个主流GAN架构上训练生成模型并进行攻击,实验结果表明,在各种攻击场景下,此攻击方案较对比方法AUCROC值平均提高0.2。

关键词:身份成员推理;人脸嵌入;注意力特征解耦;生成对抗网络;人脸生成
中图分类号　TP309;TP181

Cross-age Identity Membership Inference Based on Attention Feature Decomposition

LIU Yulu,WU Shuhong,YU Dan,MA Yao and CHEN Yongle
College of Computer Science and Technology(College of Data Science),Taiyuan University of Technology,Jinzhong,Shanxi 030600,China

Abstract　Generative adversarial networks(GANs) can generate high-resolution “non-existent” realistic images,so they are widely used in various artificial data synthesis scenarios,especially in the field of face image generation. However,the face generators based on these models typically require highly sensitive facial images of different identities for training,which may lead to potential data leakage enabling attackers to infer identity membership relationships. To address this issue,this study proposes an identity membership inference attack when significant difference exist between the obtained samples and the actual training samples for the queried identity,resulting in a drastic decline in the performance of identity membership inference based on samples. Subsequently,a reconstruction error attack scheme is designed based on attention feature decomposition to further enhance the attack performance. This scheme maximizes the elimination of influences from factors such as background poses between different samples,as well as mitigates the representation difference caused by a large age span. Extensive experiments are conducted on three representative face datasets,training generative models with three mainstream GAN architectures and performing the proposed attacks. Experimental results demonstrate that the proposed attack scheme achieves an average increase of 0.2 in AUCROC value compared to previous researches.

Keywords　Identity membership inference,Face embedding,Attention feature decomposition,Generative adversarial networks,Face generation

1　引言

机器学习主要包括鉴别模型和生成模型,在过去几年里这两类技术都取得了巨大进展。然而,上述模型的训练数据集包含医疗记录、收入等敏感属性,因此恶意用户可以对目标模型的原始训练数据进行推断,窃取隐私信息。其中最常见的就是成员推理攻击(Membership Inference Attack MIA),其目的是确定给定的数据样本是否参与了目标模型的训练。目前关于MIA的研究主要集中在鉴别模型上[1-2],生成模型常被用于联邦学习[3]或无数据[4]场景下,当敌手缺乏数据知识时,在攻击过程中使用生成模型可以提高数据多样性。然而,针对生成模型进行攻击却因存在很大挑战性而很少受到关注,主要有以下两点原因:1)与鉴别模型不同,生成模型并不会提供暴露模型过拟合的任何置信度相关的信息,因此很难获得可以进行成员推断的线索;2)由于某些生成模型自身架构的问题,不可避免地会低估某些样本数据进而出现模式

到稿日期:2023-06-13　返修日期:2024-03-03
基金项目:山西省基础研究计划(20210302123131,20210302124395)
This work was supported by the Basic Research Program of Shanxi Province(20210302123131,20210302124395).
通信作者:武淑红(wushuhong@tyut.edu.cn)

丢失和模式坍塌现象。

但是目前生成模型已经被应用于健康记录、医学图像、人脸生成等多个敏感领域，相关训练数据集的隐私泄露将会暴露患者疾病史、面部特征等信息，因此这些模型的训练数据信息同样敏感。由于缺乏足够的训练数据，许多公司会收集处理客户或者网络上的用户图像，将其用于训练具有商业用途的人脸生成模型。因此，隐私保护单位判断目标身份是否被用于人脸生成器的训练，进而对网络用户的个人照片进行保护十分必要。此外，从恶意用户的角度出发，对目标生成器的训练数据所属身份进行恶意推理将会产生额外的收益。例如，某人脸生成器使用罪犯相关图像训练，若经过推断得知某身份为其成员身份，则泄露了罪犯的相关信息。因此，判断给定身份是否参与了目标生成器的训练具有重要研究价值。

基于人脸生成器的使用范围广泛以及对人脸身份成员关系推理的必要性，本文重点考虑人脸生成器，主要贡献包括以下 3 个方面：

1）提出面向 GAN 的身份成员推理攻击，设计了各身份参与训练样本与所获取样本之间年龄跨度过大的情况，进一步提高了攻击的难度。

2）基于注意力特征解耦，面向黑盒场景从原理上提出了可行的攻击方案。

3）使用代表性的跨年龄人脸数据集在主流 GAN 模型上进行了实验，证明了该攻击的有效性。

2 相关工作

本章将对面向的生成模型以及身份成员推理攻击进行简要介绍。

2.1 生成模型

生成模型用于逼近真实数据的概率分布并基于此生成新数据。目前主流的生成模型是生成对抗网络（Generative Adversarial Network，GAN）。GAN 通过最小化生成数据和真实训练数据分布之间的差异来训练，利用深度神经网络强大的表示能力来构建异常丰富的参数族，进而在生成高维数据分布上取得了巨大的成功。自 Goodfellow 等[5]首次提出 GAN 的思想后，陆续出现了许多工作来提高原始 GAN 的性能，但 Lucic 等[6]证实了所有 GAN 的改进只涉及计算预算和超参数的调整，并没有产生实际的科学突破。本文基于计算成本，将重点研究 LSGAN[7]，WGANGP[8]和 DCGAN[9]。此外，虽然本文着重在这些 GAN 上进行实验，但攻击方法是通用的，也可以应用到其他类型的 GAN 中。

2.2 身份成员推理

成员推理指判断某特定样本是否参与目标机器学习模型的训练。Shokri 等[1]在 2017 年首次面向鉴别模型的黑盒场景提出了成员推理攻击这一概念，所提出的方法主要利用了模型的过拟合特征，通过训练影子模型模拟目标模型的行为，进而获取训练攻击模型的数据。

随后 Hayes 等[10]提出了面向 GAN 的成员推理攻击，通过训练目标 GAN 的影子模型，检查影子鉴别器对查询样本的置信度分数来判断样本成员关系。Hilprecht 等[11]则通过统计查询样本 ε-*ball* 中生成样本的数量来作为成员关系的判断依据。Chen 等[12]根据敌手对目标模型的访问程度将面向 GAN 的攻击分为完全黑盒（只可以盲目地获取生成样本）、部分黑盒（可以操控目标模型的潜在代码并获取生成样本）、白盒（可以完全访问生成器内部架构和参数）和可访问鉴别器（鉴别器和生成器内部架构及参数）4 类，并提出了基于样本重建误差设计的通用攻击方法。本文面向其中最困难的完全黑盒场景进行设计。

前文所述的攻击都是面向样本级别的成员推理，Tinsley 等[13]证明在建立或使用现有生成模型时，人脸图像中的身份信息可以从训练数据集中流入生成样本，间接证明了身份成员推理攻击的可行性。本文将该攻击视为一个二分类任务，对其进行如下明确定义：

$$\mathcal{A}:(y_t, G_t)\rightarrow\{0,1\} \tag{1}$$

给定目标身份 y_t 的查询样本集，判断其是否被用于 G_t 的训练。如果攻击者 $\mathcal{A}$ 推断该目标身份存在样本参与了 G_t 的训练，则输出为 1，否则输出为 0。在本文工作之前，Webster 等[14]率先提出了针对生成模型的身份成员推理攻击，即判断给定身份是否参与了生成模型的训练。该攻击通过统计生成样本被识别为给定身份的次数来进行身份成员关系的推断，并在人脸生成模型上展开了实验。然而该攻击并未考虑各身份在不同年龄段面部的极大差异性，当查询身份获取到的样本与其实际参与训练的样本年龄跨度过大时，该方法不能表现出良好的推断性能。本文则通过引入注意力特征解耦的方法解决这一难题。

3 攻击方案

本章将介绍对查询样本所属身份成员关系推理的攻击方案。如图 1 所示，该方案基于生成模型对所查询身份的特征重建误差进行设计，其特点是采用注意力特征解耦的方法获取对年龄变化不敏感的身份特征，进而使得攻击方案更适合于实际场景中使用。整个方案主要包括注意力特征解耦、攻击模型的构建以及重建特征的获取三大模块，下面将详细介绍每个模块。

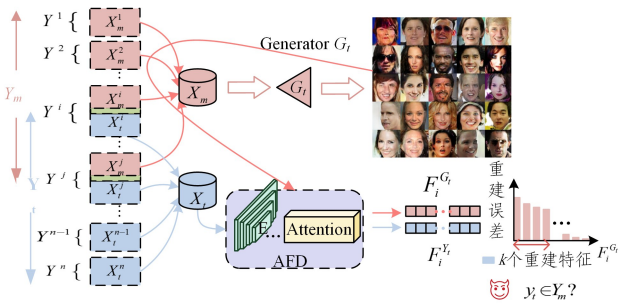

图 1 攻击框架

Fig. 1 Attack framework

3.1 注意力特征解耦

先前的研究工作[15]已经证实，将查询身份的面部图像 $\boldsymbol{x}$ 送入编码器 $\boldsymbol{E}(\cdot)$ 所生成的高维特征向量 $\boldsymbol{E}(\boldsymbol{x})$，包含该身份面部的身份、年龄、性别等多种不同信息。虽然在各身份标识监督下训练的编码器可以去除背景、姿势等与人脸标识无关的大量信息，但是随着年龄增长各身份面部的巨大变化难以

去除。为了获取年龄不变人脸特征，本文参考 Huang 等[16]的工作，借助注意力机制将混合人脸特征分解为身份特征和年龄特征两个不相关的分量。因为一些与年龄变化相关的特征，如胡须、皱纹等会在一维特征中丢失，但是会保留在特征图中。Huang 等的工作正是利用注意力机制操作特征图来进行特征分解。如图 2 所示，本文首先使用类似 ResNet 的骨干网络作为编码器从输入图像 $\boldsymbol{x}$ 中提取混合特征图 $\boldsymbol{F}\in\mathbb{R}^{C\times H'\times W'}$，即 $\boldsymbol{F}=\boldsymbol{E}(\boldsymbol{x})$；之后采用通道注意力和空间注意力之和的平均值来抓取通道和空间上的身份相关信息。则注意力解耦(Attention Feature Decomposition，AFD)的过程可以描述为：

$$\boldsymbol{F}=\underbrace{\boldsymbol{F}\circ\boldsymbol{\sigma}(\boldsymbol{F})}_{\boldsymbol{F}_i}+\underbrace{\boldsymbol{F}\circ(1-\boldsymbol{\sigma}(\boldsymbol{F}))}_{\boldsymbol{F}_a} \tag{2}$$

其中，$\circ$ 表示元素级别的乘法，$\boldsymbol{\sigma}$ 表示注意力模块。

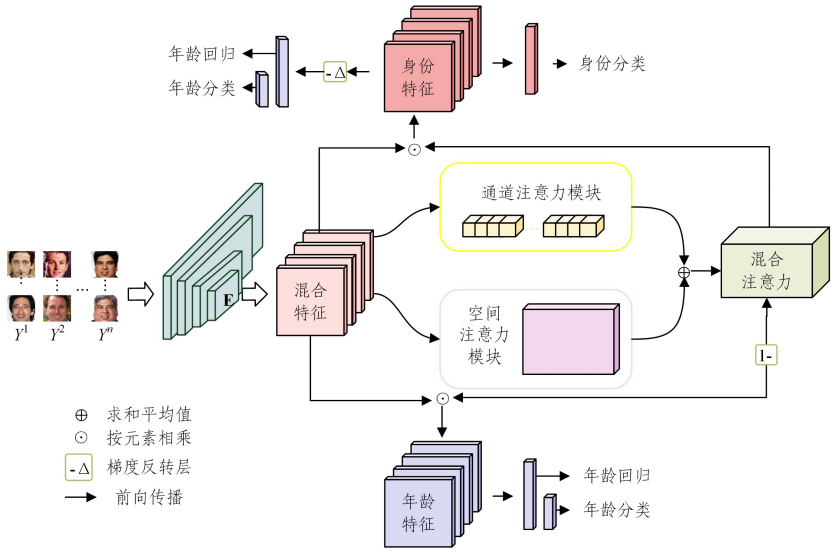

图 2 注意力特征解耦

Fig. 2 Attention feature decomposition(AFD)

为了使注意力模块稳健地分解特征，本文通过人脸识别任务监督的注意力模块来分离特征图中的身份相关信息，通过年龄估计任务来监督剩余的年龄相关信息。具体地，$\boldsymbol{F}_a$ 通过年龄估计任务对年龄相关信息编码，而 $\boldsymbol{F}_i$ 对身份相关信息进行编码。本文设计一个分别具有 512 和 101 个神经元的两个线性层的年龄估计网络 $\boldsymbol{A}$，通过计算 softmax 值实现年龄回归，学习年龄分布特征。最后添加线性层 $\boldsymbol{W}\in R^{101\times N}$ 用于年龄组分类，使学习到的分布正则化，其中 N 表示年龄组的数量。用于优化年龄估计的损失函数定义如下：

$$\ell_{\mathrm{AE}}(\boldsymbol{F}_a)=\mathbb{E}[\ell_{\mathrm{MSE}}(\mathrm{softmax}(\boldsymbol{A}(\boldsymbol{F}_a)),y_a)+\ell_{\mathrm{CE}}(\boldsymbol{A}(\boldsymbol{F}_a)\boldsymbol{W},c_a)] \tag{3}$$

其中，y_a 表示真实年龄，c_a 表示真实年龄组，ℓ_{MSE} 表示年龄回归的均方误差，ℓ_{CE} 表示年龄组分类的交叉熵损失。

此外，本文使用含有 512 个神经元的线性层 L 进行身份估计，并使用目前被证明为人脸识别领域最有效的 CosFace 损失[17]来监督身份特征的训练。此外，本文引入跨年龄域对抗性学习，使用具有梯度反转层(GRL)的连续域自适应，使得 $\boldsymbol{F}_i$ 保持年龄不变。最终整个注意力解耦模块损失函数如式(4)所示：

$$\mathcal{L}=\lambda_1\ell_{\mathrm{CosFace}}(L(\boldsymbol{F}_i),y_i)+\lambda_2\ell_{\mathrm{AE}}(\boldsymbol{F}_a)+\lambda_3\ell_{\mathrm{AE}}(GRL(\boldsymbol{F}_i)) \tag{4}$$

其中，第一项是 CosFace 损失，第二项是年龄估计损失，最后一项是域自适应损失。y_i 是身份标签，λ_* 用于控制不同损失项的平衡。

3.2 攻击模型

本节将介绍如何利用注意力特征解耦所得的年龄不变身份特征构建攻击模型。生成模型用于学习训练数据集的分布，因此生成模型生成训练数据集中样本的概率将会更大。但在实际攻击的过程中，往往不能准确获取训练数据集中的样本。由于背景、光照等多种因素的影响，通过重建样本误差[12]来判断给定身份是否参与训练是不可靠的。而针对同一身份而言，其面部特征如五官、脸型等会随着年龄的增长产生巨大变化。因此当目标身份所获取样本与实际参与训练样本的年龄跨度过大时，仅仅通过生成样本被识别为目标身份的数量[14]进行身份成员关系的判断会产生巨大误差。敌手的攻击目标即计算身份 y_t 是目标人脸生成模型成员的概率 $P(m_t=1|y_t,\boldsymbol{G}_t)$。已有研究证实[10-11]，生成模型生成样本中不可避免地包含了参与训练身份的相关特征信息，因此本文假设这个概率与生成器生成具有查询身份年龄不变身份特征的概率成正比。简言之，如果目标生成模型生成样本具有查询身份年龄不变身份特征的概率越大，则该身份参与训练的可能性越大。数学表示如下：

$$P(m_t=1|y_t,\boldsymbol{G}_t)\propto P_{\mathrm{AFD}}(\boldsymbol{F}_i^t|\boldsymbol{G}_t) \tag{5}$$

然而，由于生成样本的身份特征数据复杂且维度高，难以使用显式密度函数来表示，因此对该概率进行精确计算十分困难。因此本文采用核密度估计，概率密度表示如下：

$$\begin{aligned}P_{\mathrm{AFD}}(\boldsymbol{F}_i^t|\boldsymbol{G}_t)&=\frac{1}{k}\sum_{m=1}^{k}\phi(\boldsymbol{F}_i^t,\boldsymbol{F}_i^{G_t(z_m)})\\&\approx\frac{1}{k}\sum_{m=1}^{k}\exp(-L_2(\boldsymbol{F}_i^t,\boldsymbol{F}_i^{G_t(z_m)}))\end{aligned} \tag{6}$$

其中，$\boldsymbol{F}_i^t$ 表示查询身份特征；$\boldsymbol{F}_i^{G_t(z_m)}$ 表示生成身份特征，即目标人脸生成模型的重建特征；z_m 是目标人脸生成器输入的潜在代码，在本文所面向的黑盒场景下对攻击者不可见；k 表示

生成样本的数量；$\phi(\cdot,\cdot)$表示核函数；$L_2(\cdot,\cdot)$表示欧氏距离，用于衡量身份特征之间的差距。重建误差本质上计算的是两个身份特征之间的差异，由于这里使用欧氏距离衡量，因此随着二者之间距离的增大，式中的对应项会呈指数减小。在该方法中可以通过寻找年龄不变身份特征距离最近的少数个重建特征简化计算。在4.3节中将详细描述重建特征的构建过程。

攻击模型是一个基于阈值的二元分类器。整个攻击过程主要有两个步骤：1)敌手利用预先训练好的注意力特征解耦模型，分别获取查询身份特征和重建特征；2)将获取的特征带入成员概率计算式后计算出二者之间的相似程度，通过与给定阈值进行比较即可得到身份的成员关系。数学表达式如下：

$$m_t=\begin{cases}0, & P_{\mathrm{AFD}}=(\boldsymbol{F}_i^t\,|\,\boldsymbol{G}_t)<\varepsilon\\ 1, & P_{\mathrm{AFD}}=(\boldsymbol{F}_i^t\,|\,\boldsymbol{G}_t)\geqslant\varepsilon\end{cases} \tag{7}$$

3.3 重建估计

为了达到有效攻击的目的，获取有效的重建身份特征十分重要。本文面向的是完全黑盒场景，此时攻击者只可以随机地从目标生成器中获取生成样本。首先攻击者访问目标生成器获取生成样本$\boldsymbol{X}_G=\{\boldsymbol{G}_t(\boldsymbol{z}_1),\boldsymbol{G}_t(\boldsymbol{z}_2),\cdots,\boldsymbol{G}_t(\boldsymbol{z}_{|X|})\}$，其中$\boldsymbol{G}_t(z_i)$表示由目标生成模型通过潜在代码$z_i$生成的相应样本，潜在代码$z_i$由$\boldsymbol{G}_t$随机生成，处于攻击者不可控的范围内。之后将身份查询样本$\boldsymbol{x}$和生成样本集$\boldsymbol{X}_G$输入预训练的模型后分别获取身份特征。本文使用$\mathscr{R}$来表示攻击者基于$\boldsymbol{G}_t$构建的y_t重建特征，数学表示如下：

$$\mathscr{R}(y_t\,|\,\boldsymbol{G}_t)=\underset{\boldsymbol{F}_i^{G_t(\cdot)}\in\{AFD(\boldsymbol{G}_t(\cdot)_m)\}_{m=1}^k}{\arg\min} L_2(\boldsymbol{F}_i^y,\boldsymbol{F}_i^{G_t(\cdot)}) \tag{8}$$

如图1所示，本文针对目标身份的查询样本，计算距离其身份特征最近的k个生成样本特征作为重建特征。

4 实验

本章对所提攻击展开实验并进行性能分析。4.1节主要对数据集的处理以及目标GAN进行介绍，并对后续实验中攻击性能分析时使用的评估指标进行分析。4.2—4.4节主要分析影响攻击性能的3个因素，设置了多组实验进行对比分析。4.5节将本文所提攻击与文献[14]所提攻击进行比较。

4.1 实验设置

4.1.1 数据集处理

在目前主流的Celeba，FGNET和CALFW人脸数据集上进行测试，具体如下。

Celeba数据集[18]是一个拥有10000以上身份、共计20万张RGB图像的大规模人脸属性数据集。其人脸图像根据面部标志相互对齐，这有利于GAN的人脸训练过程。本文根据身份将数据集进行划分，并根据需要选择身份数以及每个身份的图片样本数进行目标GAN模型的训练。

FGNET数据集[19]包含82个身份、不同年龄段的多张照片，年龄跨度相对较大，是跨年龄人脸数据集的典型代表。本文利用该数据集同一身份不同样本间变化大的特性，查看所提攻击方案的攻击效果。

CALFW(Cross-Age LFW)数据集[20]为跨年龄野生标记人脸数据集，包含3000对年龄跨度较大的正面人脸图像。本文利用该数据集同一身份图片样本年龄跨度大的特性，查看所提攻击方案的攻击效果。

4.1.2 目标模型

本文着重考虑WGANGP，DCGAN，LSGAN作为目标模型进行攻击，因为它们是实际应用中的典型和代表性模型。在训练目标GAN前，首先将图像按人脸面部图像进行居中裁剪，并将该图像大小调整为112 * 112。所有GAN均采用100维的潜在代码并生成112 * 112 * 3的人脸图像。在实际应用中使用FID(Fréchet Inception Distance)衡量生成图像的多样性和质量，FID越小，则图像多样性越好，质量越高。

LSGAN[7]使用最小二乘损失函数代替了GAN的损失函数，缓解了GAN训练不稳定和生成图像质量差、多样性不足的问题。这里对实验所设计的每一个生成器均采用64的批次大小，并选择迭代4000 epoch。

WGANGP[8]的目的是改进WarrersteinGAN(WGAN)的训练过程，它通过加入梯度惩罚实现了GAN的稳定训练。为了实现精准对比，实验所设计的每一个生成器均采用64的批次大小，并迭代2000 epoch。

DCGAN[9]使用卷积层替换全连接层，使得其在大多数环境下可以稳定地训练。同样对实验所设计的每一个生成器均采用64的批次大小。经过实验观察，为了生成更高质量的人脸图像，选择迭代4000 epoch。

4.1.3 实验评估

本文所面向的黑盒场景在实际应用中往往是模型所有者发布在线闭源API以供使用，本文通过模拟这种API实现攻击场景的设计。此外，本文定义有查询身份集$S=\{(y_i,m_i)\}_{i=1}^N$，其中m_i表示成员关系指示变量。该集合既包含成员身份($y_t^i\in Y_m,m_i=1$)，也包含非成员身份($y_t^i\notin Y_m,m_i=0$)。借此可以分别计算攻击的真阳性率$\mathbb{E}_{y_i}[P(\mathscr{A}(y_i,\boldsymbol{G}_t)=1\,|\,m_i=1)]$和真阴性率$\mathbb{E}_{y_i}[P(\mathscr{A}(y_i,\boldsymbol{G}_t)=0\,|\,m_i=0)]$用于对攻击性能的评估。

本文方法用于判断目标身份是否是成员，因此从攻击者的角度，将成员身份视为正例样本，将非成员身份视为负例样本。所提出的攻击方法通过改变等式中的给定阈值实现二分类。因此，本文通过改变阈值、绘制ROC曲线、测量ROC曲线下的面积AUCROC值来评估攻击的性能。此外，通过实验观察到，当k值越大时，本文方法会自觉优化特征的重建过程，但会导致高额的查询和计算代价。因此在实际攻击过程中应选择合适的k值来平衡二者。在综合考虑硬件设备以及攻击性能的情况下，本文在整个实验过程中固定式(5)中参数k值为5。

4.2 参与训练的身份数目

训练数据集的大小与生成模型的过拟合程度密切相关。使用较小训练数据集的生成模型更容易对单个样本产生记忆，因此攻击成功的概率越大。本文通过分析实验要求，根据各数据集身份总数的不同，针对性地设计了各数据集参与训练的多组身份数目，以评估不同身份数下攻击效果的差异。需要注意的是，由于FGNET数据集只包含82个身份的数据，因此本文设计其身份数分别为20，41，64，82。当身份数超过该数据集身份总数的一半时，使用Celeba中与其不相交的身份样本作为负例进行实验。

图 3 展示了针对不同 GAN 模型在 3 个数据集不同身份数上的攻击性能。如图 3 所示，当身份数较少时，该攻击展现了良好的性能。例如，在 Celeba 上，当只有 64 个身份参与训练时，3 个目标 GAN 模型的 AUCROC 都在 0.8 左右浮动。这代表着成员身份集存在着严重的隐私泄露。但是，随着身份集数量增加，目标 GAN 模型的过拟合程度也会相应降低，导致攻击变得不那么有效。因此，收集更多的身份样本参与 GAN 的训练可以减少个别身份的隐私泄露。经过实验观察，WGANGP 在各数据集上都更易于受到身份成员推理的攻击，经过计算 FID 发现，这是由于 WGGAN 生成的样本质量更高，更好地拟合了训练数据，因此更多地暴露了各身份的相关特征。

4.3 各身份参与训练的样本数目

直觉上讲，目标身份参与训练的样本越多，则生成模型对该身份的记忆程度越高，使得对该身份攻击成功的概率越大。本文根据 Celeba 和 FGNET 数据集中各身份的样本数量，在各个身份数目下对 Celeba 数据集分别选择 2，5，8，对 FGNET 分别选择 2，5，10 个数据样本，对攻击效果差异进行观察。

如图 3(a)－3(f)所示，整体看来，随着参与训练样本数目的增多，攻击性能会更优。例如图 3(e)，当参与训练身份数为 64，各身份参与训练样本数由 2 增加至 10 时，攻击的 AUCROC 值提高 0.112。因此，可以通过控制单个身份参与训练的样本数有效防御攻击。但是可以发现，当所获取的样本与参与训练的样本差异过大时，即使增加训练样本的数量，攻击准确率的提高十分微小甚至会表现出更差的攻击性能。并且发现随着身份总数的增多，各成员身份参与训练的样本数目增多带来的优势也会降低。如图 3(a)－3(c)所示，当参与训练身份数为 4096 时，随着各身份样本数的增多，攻击性能并不会发生明显的改变。此外，本文观察到在 FGNET 和 CALFW 数据集下训练的 LSGAN 对该攻击呈现出更优的抵抗能力，经过计算 FID 发现，这是由于 LSGAN 的生成样本质量较差，进而导致特征重建误差难以准确的估计。

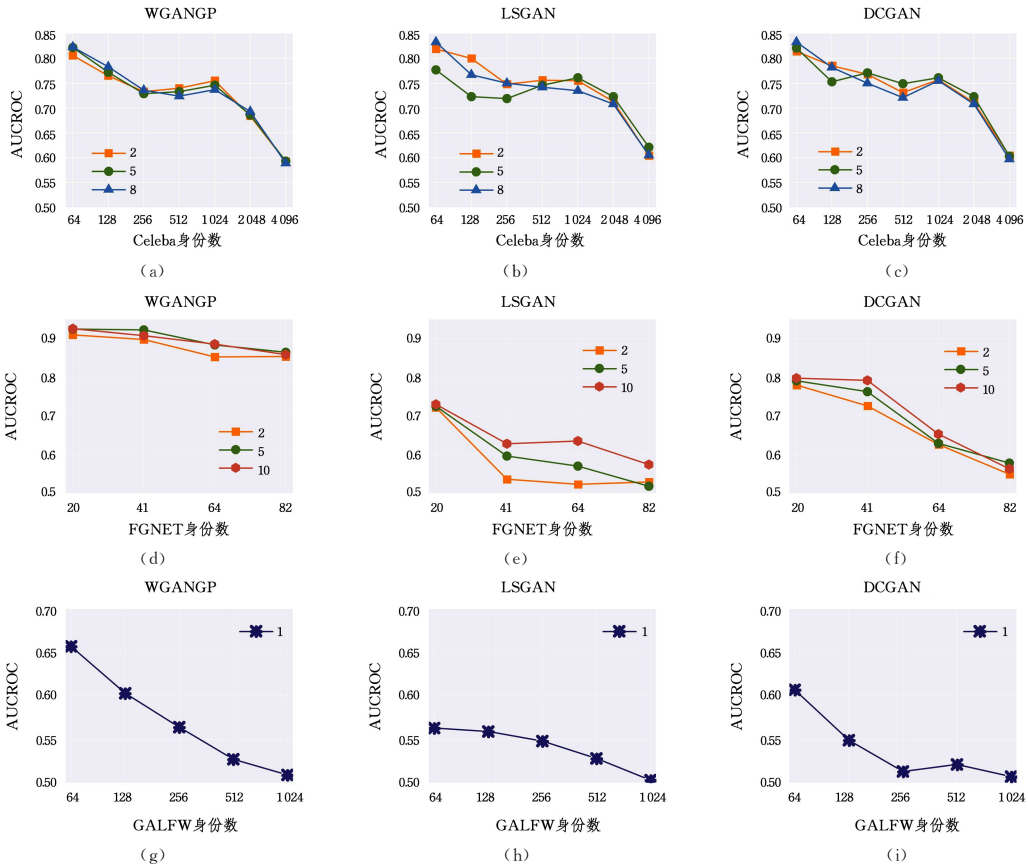

图 3 攻击性能 & 身份数和样本数

Fig. 3 Attack performance & number of identity and samples

4.4 目标身份所获样本成员关系

直觉上讲，当目标身份为成员时，所获取的样本参与训练，即对目标身份所获取的查询样本被包含在训练数据集之内时，攻击成功的概率将会变大。而当所获取样本与成员样本之间的差距越大，年龄变化带来的影响越显著，身份特征之间的差异越大，攻击难度随之增大。因此，本文通过设计目标身份的查询样本是否包含在训练数据集之内进行分组实验，观察当查询样本与成员样本差异过大时的攻击表现。

图 4 展示了几组 GAN 当查询样本是否包含在训练数据集中的实验效果。Celeba 以及 FGNET 所选实验组为数据集参与训练身份最多，且各身份参与样本最少。可以观察到，当目标 GAN 模型在 Celeba 数据集上训练时，所获取样本若参与训练，将会显著提高攻击性能。通过分析 Celeba 数据集发现，由于 Celeba 数据集除显著的年龄变化以外，更多的是面部表情以及眼镜、帽子、发型等不可控因素的变化，这些变化也会间接影响攻击效果。在未来的研究中可以考虑更优的解耦表征方案以分离出这些因素的影响。此外，此发现也可以应用到防御策略中，对训练数据集进行预处理，通过技术手段

给面部图像添加复杂的要素使得面部图像产生明显变化，从而有效防御攻击。

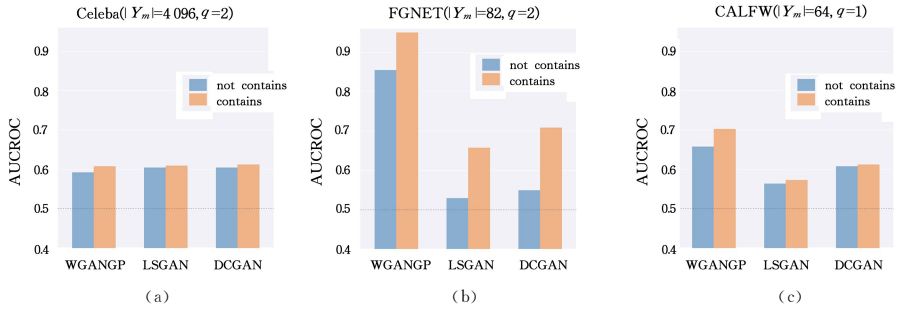

图 4 攻击性能 & 所获样本成员关系

Fig. 4 Attack performance & sample membership attribute

4.5 对比实验

将本文的攻击方法与文献[14]的身份成员推理攻击方法进行了比较。文献[14]的攻击也是面向人脸生成模型，首先使用目标身份训练人脸分类器，之后通过统计目标生成模型生成样本被识别为目标身份的次数与给定阈值进行比较来判断成员关系。因此该方法也可以使用 AUCROC 来进行评估。为公平起见，在本文实验所涉及的目标模型以及数据集上使用文献[14]的方法再次进行实验以进行对比。由于该方法的基础思想使用的是人脸分类器，因此本文使用 FC(Face Classifier)来指代该攻击方法。

图 5 展示了不同数据集、不同 GAN、不同身份数目以及所获样本是否参与训练下本文方法与 FC 攻击效果的对比。可以明显看出，本文攻击方法呈现出了更好的性能，如图 5(a)—5(c)所示，在各种设置下本文方法的 AUCROC 值比 FC 高出 0.2 左右。主要有以下两点原因：1)本文使用的特征去除了年龄变化所带来的影响；2)统计生成样本被识别为给定身份的数量本质上相当于统计给定身份一定范围内的样本数量，相较而言距离误差可以综合各个特征的差异，更精确地反映出重建误差的大小。但如图 5(e)所示，当所获样本未参与训练时本文方法呈现劣势，与 4.3 节分析所得的原因相同，这是 LSGAN 生成样本的质量不高所导致的特征重建误差难以准确评估所致。

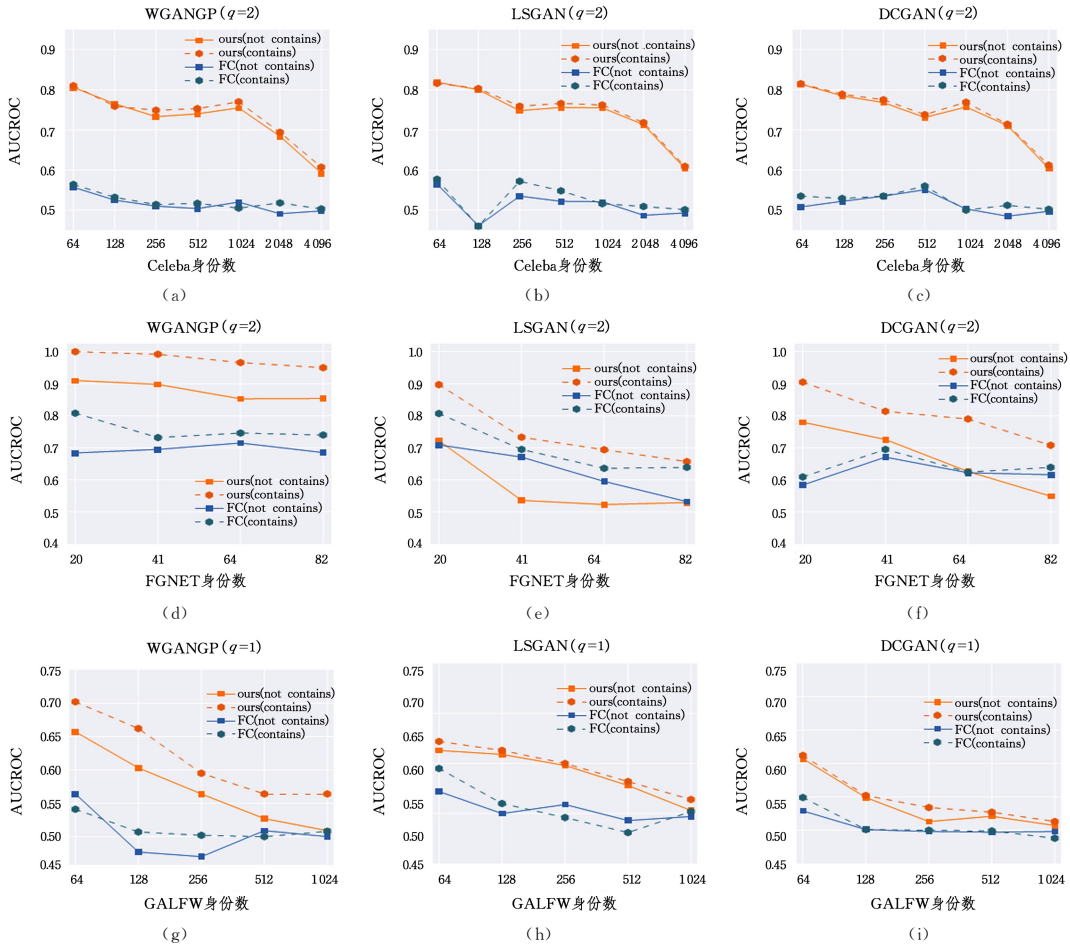

图 5 对比实验

Fig. 5 Contrast experiment

结束语 本文面向GAN进行身份成员推理攻击,并通过改变设置进一步提升攻击难度,即对查询身份所获取样本与实际参与训练样本存在巨大年龄变化等显著差异时进行身份成员推理。在此基础之上,基于注意力特征解耦提出了一种攻击方法,综合实验表明,该方法在人脸生成模型身份成员关系的推理上展现了良好的性能。但考虑到本文所提出的解耦模型只分离了与身份特征不相关的年龄特征,并且难以保证所剔除的年龄特征中不含有身份相关信息,故未来的研究方向将探索更优的解耦表征方法,用于提取出表情、姿势、遮挡物等所有除身份外的相关信息。此外,对于提取出的上述信息,考虑如何进一步用于优化样本成员关系的推断。

参 考 文 献

[1] SHOKRI R,STRONATI M,SONG C,et al. Membership Inference Attacks Against Machine Learning Models[C]// 2017 IEEE Symposium on Security and Privacy(SP). San Jose,CA,USA:IEEE,2017:3-18.

[2] PENG C G,GAO T,LIU H L,et al. PCA-based membership inference attack for machine learning models[J]. Journal on Communications,2022,43(1):149-160.

[3] ZHANG J L,ZHU C C,SUN X B,et al. Membership inference attack and defense method in federated learning based on GAN[J]. Journal on Communications,2023,44(5):193-205.

[4] YANG P P,ZHANG X M. Label-based data-free membership inference attack[J]. Cyber Security And Data Governance,2023,42(5):44-49.

[5] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial nets[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2. 2014,2672-2680.

[6] LUCIC M,KURACH K,MICHALSKI M,et al. Are gans created equal? a large-scale study[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. 2018:698-707.

[7] MAO X,LI Q,XIE H,et al. Least squares generative adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:2794-2802.

[8] GULRAJANI I,AHMED F,ARJOVSKY M,et al. Improved training of wasserstein gans[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. 2017:2234-2242.

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

[10] HAYES J,MELIS L,DANEZIS G,et al. Logan:Membership inference attacks against generative models[J]. arXiv:1705.07663,2017.

[11] HILPRECHT B,HÄRTERICH M,BERNAU D. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models[J]. Proceedings Priv. Enhancing Technol.,2019,2019(4):232-249.

[12] CHEN D,YU N,ZHANG Y,et al. Gan-leaks:A taxonomy of membership inference attacks against generative models[C]// Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. 2020:343-362.

[13] TINSLEY P,CZAJKA A,FLYNN P. This face does not exist… but it might be yours! identity leakage in generative models[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021:1320-1328.

[14] WEBSTER R,RABIN J,SIMON L,et al. This person(probably) exists. identity membership attacks against gan generated faces[J]. arXiv:2107.06018,2021.

[15] SUN Y,WANG X,TANG X. Deeply learned face representations are sparse,selective,and robust[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:2892-2900.

[16] HUANG Z,ZHANG J,SHAN H. When age-invariant face recognition meets face age synthesis:A multi-task learning framework[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:7282-7291.

[17] WANG H,WANG Y,ZHOU Z,et al. Cosface:Large margin cosine loss for deep face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:5265-5274.

[18] LIU Z,LUO P,WANG X,et al. Deep learning face attributes in the wild[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:3730-3738.

[19] PANIS G,LANITIS A,TSAPATSOULIS N,et al. Overview of research on facial ageing using the FG-NET ageing database[J]. Iet Biometrics,2016,5(2):37-46.

[20] ZHENG T,DENG W,HU J. Cross-age lfw:A database for studying cross-age face recognition in unconstrained environments[J]. arXiv:1708.08197,2017.

LIU Yulu, born in 1998, postgraduate, is a student member of CCF (No. P2665G). Her main research interests include artificial intelligence security and information security.

WU Shuhong, born in 1969, Ph.D, associate professor, master supervisor. Her main research interests include embedded systems, intelligent information processing, brain informatics and information security.

(责任编辑:何杨)