

可逆神经网络的隐私泄露风险评估

何毅凡^{1,2}, 张杰^{2,3}, 张卫明^{2,3}, 俞能海^{2,3}

(1. 中国科学技术大学信息科学技术学院, 安徽 合肥 230027; 2. 中国科学院电磁空间信息重点实验室, 安徽 合肥 230027
3. 中国科学技术大学网络空间安全学院, 安徽 合肥 230027)

摘要: 近年来, 深度学习已经成为多方领域的核心技术, 而深度学习模型的训练过程中往往需要大量的数据, 这些数据中可能含有隐私信息, 包括个人身份信息(如电话号码、身份证号等)和敏感信息(如金融财务、医疗健康等)。因此, 人工智能模型的隐私风险问题成为学术界的研究热点。深度学习模型的隐私研究仅局限于传统神经网络, 而很少针对特殊网络结构的新兴网络(如可逆神经网络)。可逆神经网络的上层信息输入可以由下层输出直接得到, 直观上讲, 该结构保留了更多有关训练数据的信息, 相比传统网络具有更大的隐私泄露风险。为此, 提出从数据隐私泄露和模型功能隐私泄露两个层面来探讨深度网络的隐私问题, 并将该风险评估策略应用到可逆神经网络。具体来说, 选取了两种经典的可逆神经网络(RevNet 和 i-RevNet), 并使用了成员推理攻击、模型逆向攻击、属性推理攻击和模型窃取攻击 4 种攻击手段进行隐私泄露分析。实验结果表明, 可逆神经网络在面对数据层面的隐私攻击时存在相比传统神经网络更严重的隐私泄露风险, 而在面对模型层面的隐私攻击时存在相似的隐私泄露风险。由于可逆神经网络研究越来越多, 目前被广泛应用于各种任务, 这些任务也涉及敏感数据, 在实验结果分析基础上提出了一些潜在的解决方法, 希望能够应用于未来可逆神经网络的发展。

关键词: 可逆神经网络; 隐私保护; 成员推理攻击; 隐私威胁

中图分类号: TP309.7

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023051

Privacy leakage risk assessment for reversible neural network

HE Yifan^{1,2}, ZHANG Jie^{2,3}, ZHANG Weiming^{2,3}, YU Nenghai^{2,3}

1. School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China
2. Key Laboratory of Electro-magnetic Space Information, Chinese Academy of Sciences, Hefei 230027, China
3. School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230027, China

Abstract: In recent years, deep learning has emerged as a crucial technology in various fields. However, the training

收稿日期: 2022-10-26; **修回日期:** 2023-02-06

通信作者: 张卫明, zhangwm@ustc.edu.cn

基金项目: 国家自然科学基金(U20B2047, 62072421, 62002334, 62102386, 62121002); 中国科学技术大学探索基金(YD3480002001); 中央高校基本科研业务费专项基金(WK2100000011)

Foundation Items: The National Natural Science Foundation of China (U20B2047, 62072421, 62002334, 62102386, 62121002), Exploration Fund of University of Science and Technology of China (YD3480002001), Fundamental Research Funds for the Central Universities (WK2100000011)

引用格式: 何毅凡, 张杰, 张卫明, 等. 可逆神经网络的隐私泄露风险评估[J]. 网络与信息安全学报, 2023, 9(4): 29-39.

Citation Format: HE Y F, ZHANG J, ZHANG W M, et al. Privacy leakage risk assessment for reversible neural network[J]. Chinese Journal of Network and Information Security, 2023, 9(4): 29-39.

process of deep learning models often requires a substantial amount of data, which may contain private and sensitive information such as personal identities and financial or medical details. Consequently, research on the privacy risk associated with artificial intelligence models has garnered significant attention in academia. However, privacy research in deep learning models has mainly focused on traditional neural networks, with limited exploration of emerging networks like reversible networks. Reversible neural networks have a distinct structure where the upper information input can be directly obtained from the lower output. Intuitively, this structure retains more information about the training data, potentially resulting in a higher risk of privacy leakage compared to traditional networks. Therefore, the privacy of reversible networks was discussed from two aspects: data privacy leakage and model function privacy leakage. The risk assessment strategy was applied to reversible networks. Two classical reversible networks were selected, namely RevNet and i-RevNet. And four attack methods were used accordingly, namely membership inference attack, model inversion attack, attribute inference attack, and model extraction attack, to analyze privacy leakage. The experimental results demonstrate that reversible networks exhibit more serious privacy risks than traditional neural networks when subjected to membership inference attacks, model inversion attacks, and attribute inference attacks. And reversible networks have similar privacy risks to traditional neural networks when subjected to model extraction attack. Considering the increasing popularity of reversible neural networks in various tasks, including those involving sensitive data, it becomes imperative to address these privacy risks. Based on the analysis of the experimental results, potential solutions were proposed which can be applied to the development of reversible networks in the future.

Keywords: reversible neural network, privacy protection, membership inference attack, privacy threat

0 引言

近年来,深度学习的快速发展使其成为多个领域的核心技术。尽管带来了可观的商业价值,但是深度学习模型以及训练这些模型所使用的用户数据面临着诸多隐私安全问题。例如,现有的深度模型往往部署在第三方的云服务平台上,攻击者可以通过访问模型获得相应的输出。通过对这些输出进行分析,攻击者可以方便地提取有关训练数据的隐私信息,甚至可以窃取整个模型功能。随着云计算和大数据的普遍使用,深度学习模型的隐私安全问题逐渐成为学术界的研究热点。按照敌手策略划分,目前深度模型面临的常见隐私威胁包括成员推理攻击^[1]、属性推理攻击、模型逆向攻击^[2]和模型窃取攻击^[3]。其中,成员推理攻击、属性推理攻击和模型逆向攻击为数据隐私层面的攻击策略,而模型窃取攻击为模型功能隐私层面的攻击策略^[4]。

当前针对深度模型隐私安全风险的问题研究仅局限于传统神经网络,而忽略了近些年来取得飞速发展的可逆神经网络。相比传统神经网络,可逆神经网络采用了可逆模块作为模型结构主

体,可以使用更少的参数量实现信息的无损传输。在可逆神经网络提出之前,关于流模型的研究就指出:当模型的雅可比行列式不为零时,模型可以实现可逆。为了使得雅可比矩阵易于求得,Laurent等利用耦合层的想法提出了NICE^[5],即将输入 x 通道上分割成两个部分 x_1 和 x_2 进行前向传播。Gomez等^[6]借鉴了耦合层的思想提出RevNet,从而不需要存储中间结果,解决了深层网络发展受制于显存瓶颈的问题。之后,i-RevNet^[7]进一步将网络中的不可逆结构(如池化层等)进行改进从而使得模型彻底可逆,保证了信息传输的双向性,同时验证了信息瓶颈理论的不适用性。此外,可逆神经网络还可用于数据压缩、信息隐藏等其他领域,并取得了令人称赞的性能表现。然而,可逆神经网络的无损传输并保留更多有关数据信息的特性是否会造成更大的隐私风险,目前在学术界少有人关注。

为了解决上述重要问题,本文专注于对可逆神经网络隐私泄露风险进行全面评估。具体来说,本文选取了两种经典的可逆神经网络,即RevNet和i-RevNet作为研究对象;从数据隐私和模型功

能隐私两方面入手,分别采用成员推理攻击、属性推理攻击^[8]、模型逆向攻击和模型窃取攻击作为评估方法,并在不同数据集上进行相应的实验。通过大量实验,本文验证了可逆神经网络相比传统神经网络存在更为严重的隐私泄露问题。具体来说,在成员推理攻击场景下,RevNet和i-RevNet的隐私泄露风险分别提升了5.19%和8.56%;在面对模型逆向攻击和属性推理攻击时,隐私风险有着不同程度的提升;而面对模型功能隐私攻击即模型窃取攻击时,可逆神经网络相比传统神经网络,隐私安全风险没有明显提升。本文主要贡献点如下。

1) 首次关注了可逆神经网络的隐私泄露问题。

2) 提出了从基于数据隐私和模型功能隐私两个角度进行隐私泄露风险评估的策略,采用相应的隐私攻击方法进行实现,并提出了一种针对可逆神经网络的基于中间层信息的成员推理攻击策略来评估。

3) 通过实验表明,相比传统神经网络,可逆神经网络在数据隐私层面面临着更严重的隐私泄露风险。

4) 基于以上可逆神经网络的隐私泄露风险,提出了一些潜在的解决方法,希望可以促进该领域将隐私安全问题纳入可逆神经网络的设计需求中。

1 相关工作

1.1 机器学习典型隐私威胁

现有的深度学习模型在部署之后,主要面临的隐私攻击包括成员推理攻击、模型逆向攻击、属性推理攻击和模型窃取攻击。

(1) 成员推理攻击

成员推理攻击主要指攻击者通过访问模型预测API,通过预测结果可以推理得到某个数据是否在模型训练集中^[5]。这种攻击方法攻击者只需要得到预测的分类输出置信度,并不需要模型的结构、训练方法、模型参数、训练数据集分布等信息。成员推理攻击对于过拟合的模型攻击尤其有效,即模型的泛化能力越差,其训练集成员与非训练集成员越容易区分。

(2) 模型逆向攻击

模型逆向攻击是指攻击者从模型预测的结果中提取和训练数据有关的信息。攻击者可以在不接触隐私数据的情况之下,利用模型输出的结果反推用户的隐私数据。有学者提出利用模型输出中的后验概率分布为目标模型的每一类重构一个具有代表性的样本,将模型逆向攻击问题转变为一个优化问题,优化目标为使逆向数据的输出向量和原始输入的输出向量差异尽可能缩小,从而泄露训练数据的隐私信息^[2]。最近工作通过引入生成对抗网络,使得模型逆向攻击可以成功应用于人脸等高维隐私数据。

(3) 属性推理攻击

Melis等^[8]证明了攻击者不仅可以准确推理出准确数据点的存在,还可以推断属性在训练期间何时出现和消失在数据中。例如,确定特定的人何时首次出现在用于训练通用性别分类器的照片中。机器学习模型可以学习与原始任务无关的关于训练数据的额外信息,如一个从照片预测年龄的模型也可以学习到预测种族所需的信息,属性推理旨在利用这种意外的信息泄露。

(4) 模型窃取攻击

模型窃取攻击是指攻击者对目标模型的功能和使用价值进行窃取^[9]。模型是一种有非常高价值的知识产权,利用了最有价值的数据去训练,如金融交易、医疗信息、用户交易信息等等,因此训练模型花费大量的时间、金钱、人力等。模型窃取攻击的威胁是双重的:攻击者可能会使用窃取的模型来获得金钱收益,或作为发动进一步攻击的前提步骤。与知识蒸馏相似,都是得到了原模型的性能,但不同之处在于攻击者在这种场景中无法访问受害者的专有训练集。实际情况中,攻击者与目标模型之间仅存在黑盒的交互作用,即输入图像并得到输出的预测结果,除此之外并不知道任何信息。

1.2 可逆神经网络模型

可逆神经网络指的是一类可以通过双向映射以相反的过程得到最初输入数据的网络。近年来,出现了很多与网络可逆结构研究的相关论文,其中NICE和RVNP将可逆神经网络与最大似然估计相结合提出了基于流的生成模型,通过最大化

生成过程的逆过程, 得到潜变量的对数似然函数值来优化模型生成能力与拟合真实数据分布的能力。NICE 提出了一种耦合层结构来实现可逆, RNVP 引入了卷积操作, 提出了更为一般的仿射耦合层, 二者都采用了将输入 x 通道分割成两个部分 x_1 和 x_2 进行传播的办法。Gomez 等受到二者的启发, 在其基础之上提出 RevNet^[6]。

RevNet 的可逆块结构如图 1 所示, 正常的残差块输入 x 到 y 的映射 $y=x+F(x)$ 变为

$$y_1 = x_1 + F(x_2) \quad (1)$$

$$y_2 = x_2 + G(y_1) \quad (2)$$

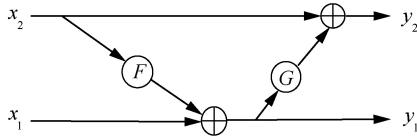


图 1 RevNet 的可逆块结构
Figure 1 Reversible block structure of RevNet

每一层的激活可以从下一层的激活中重建, 如式(3)、式(4)所示。

$$x_1 = y_1 - F(x_2) \quad (3)$$

$$x_2 = y_2 - G(y_1) \quad (4)$$

与残差网络不同, 可逆块的步长必须为 1, 否则步长为 2 的卷积层或者池化层会丢弃信息, 这样的丢弃操作是不可逆的。在逆向过程中, 已知激活 (y_1, y_2) 和它们的总导数 $(\bar{y}_1, \bar{y}_2)$, 希望得到上层输入 (x_1, x_2) , 输入的导数 $(\bar{x}_1, \bar{x}_2)$ 和与 F 、 G 相关的所有参数的总导数。

算法 1 可逆残差网络反向传播算法

输入 下层的输出 (y_1, y_2) 、梯度 $(\bar{y}_1, \bar{y}_2)$

输出 上层的输入 (x_1, x_2) 、梯度 $(\bar{x}_1, \bar{x}_2)$ 、权重梯度 $(\bar{w}_F, \bar{w}_G)$

$$1) z_1 \leftarrow y_1;$$

$$2) x_2 \leftarrow y_2 - G(z_1);$$

$$3) x_1 \leftarrow z_1 - F(x_2);$$

$$4) \bar{z}_1 \leftarrow \bar{y}_1 + \left(\frac{\partial G}{\partial z_1} \right)^T \bar{y}_2;$$

$$5) \bar{x}_2 \leftarrow \bar{y}_2 + \left(\frac{\partial F}{\partial x_2} \right)^T \bar{z}_1;$$

$$6) \bar{x}_1 \leftarrow \bar{z}_1;$$

$$7) \bar{w}_F \leftarrow \left(\frac{\partial F}{\partial w_F} \right)^T \bar{z}_1;$$

$$8) \bar{w}_G \leftarrow \left(\frac{\partial G}{\partial w_G} \right)^T \bar{y}_2;$$

$$9) \text{return } (x_1, x_2) (\bar{x}_1, \bar{x}_2) (\bar{w}_F, \bar{w}_G)$$

反向传播过程主要分为三步, 这里引入一个中间变量 z_1 : ① 利用式(3)、式(4)还原输入 (x_1, x_2) ; ② 计算得到 y_1, y_2 关于损失函数的梯度; ③ 利用链式法则求得输入特征以及输入特征的权值梯度。通过反复应用算法 1^[2], 只要简单给出序列中顶层的激活及其导数, 就可以对可逆块序列进行反向传播。通常, 一个典型的 ResNet 体系结构包含很长的残差块序列和少量的下采样层, 这些层的输入需要在反向传播期间显式存储。相反, RevNet 中需要显式存储的激活值数量并不会随着网络深度而增加, 这样, 激活的存储成本将会很小并且与网络的深度没有关系。

后来, Jacobsen 等^[7]在 RevNet 的基础上通过同胚层级联 (cascade of homeomorphic layer) 构建了 i-RevNet, 用来证明通过信息丢弃实现分类网络的泛化能力是非必要性的。信息瓶颈理论认为提升网络的泛化能力必须要丢失一些无关特征 (如噪声等) 来获得最优特征, 而这种可逆神经网络用事实粉碎了该理论。它舍弃了 RevNet 中的部分不可逆结构 (如最大池化和下采样层), 这是一个可以完全可逆、从下层恢复上层输入的网络, 即没有信息被丢弃, 但依然能够训练并收敛。

i-RevNet 的可逆块结构如图 2 所示, 它的初始输入被分成两个大小相等的子层 $(x_0, \tilde{x}_0)$, 卷积块 F_j 交替连接, 并且通过 S_j 运算进行洗牌, 以保证架构的可逆性和计算效率, 输出结果通过 $\tilde{M}$ 处理进行合并。证明了在学习可以泛化到陌生数据的表征时, 信息损失并不是必要条件, 在逐步收缩的过程中, 信息损失是可以避免的。

2 本文方法与设置

本节主要根据不同的隐私攻击策略, 验证不同攻击下 3 种模型结构受到的隐私威胁。本节对

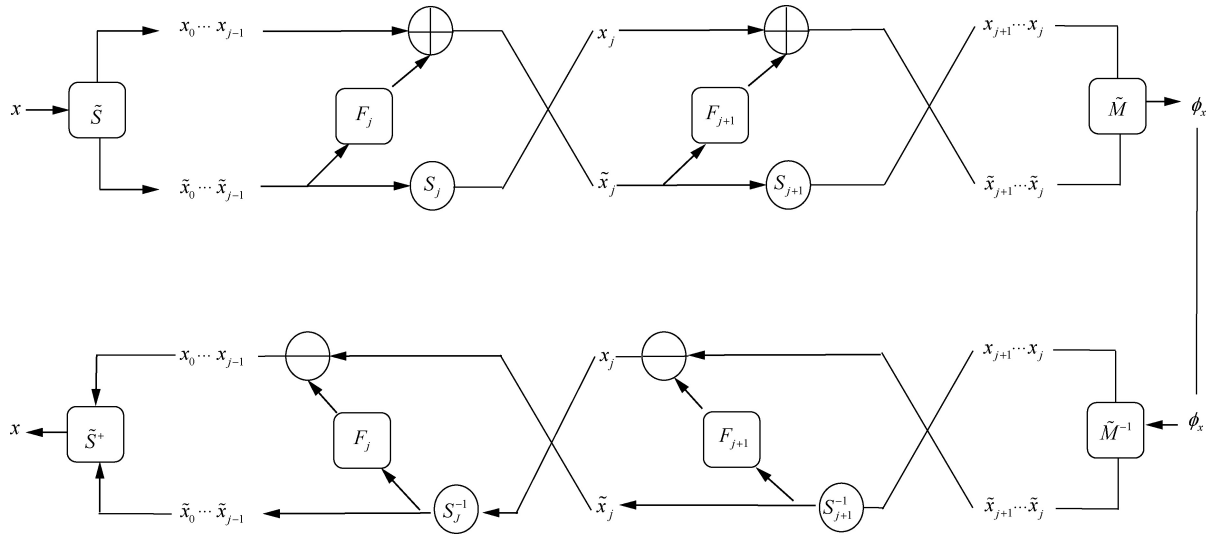


图 2 i-RevNet 的可逆块结构
Figure 2 Reversible block structure of i-RevNet

可逆神经网络做隐私泄露风险评估，专注于探究如下 3 个问题。

1) 可逆块在不同隐私攻击策略下受到的影响程度是否有所不同。

2) 可逆结构有更大隐私泄露风险的背后原因是什么，这样的问题会不会阻止可逆神经网络的发展。

3) 可逆块的数量和隐私泄露影响程度的关系。

为了公平，本文保证不同模型的参数量维持在同一个数量级上，网络模型分别采用 ResNet-32、RevNet-38、i-RevNet-31，3 种模型的网络结构参数如表 1 所示，本文首先从成员推理攻击开始验证，接着引申到其他的隐私攻击策略上，采用不同的对手策略来评估可逆神经网络的隐私安全问题。

表 1 模型的网络结构参数

Table 1 Network structure parameters of the model

网络模型	单位	通道数	参数量/MB
ResNet-32	5-5-5	16-16-32-64	0.46
RevNet-38	3-3-3	32-32-64-112	0.46
i-RevNet-31	1-3-6	8-32-128	0.48

2.1 成员推理攻击

Salem 等^[10]关于成员推理攻击可行性有很多关于攻击方法的假设，最终提出了 3 种攻击方法：面向模型无关的成员推理攻击方法、面向独立于数据的成员推理攻击方法、独立于模型和数据的不经训练的成员推理攻击方法。这里采用经典的面向模型无关的成员推理攻击方法。

这种攻击方法需要训练影子模型。为了训练影子模型，假设攻击者有一个数据集，用 D_s 表示，它与目标模型训练数据有着相同的底层分布。攻击方法可以分为 3 个阶段，即影子模型训练、攻击模型训练和成员推理。首先将数据集 D_s 分割成两个不相交的集合 D_s^{Train} 和 D_s^{Out} ，用 D_s^{Train} 来训练影子模型 S 。使用训练好的影子模型 S 对两个集合 D_s^{Train} 和 D_s^{Out} 中的数据进行预测并得到相应的后验概率，将前者生成的特征向量标记为 1，后者标记为 0。接着将生成的所有特征向量和标签用于二分类训练攻击模型 A 。对于成员推理阶段，将目标数据访问目标模型得到的相应的后验概率分布作为特征进行二分类即可实现成员推理攻击。这里将整个数据集的训练集二等分，分别作为 D_s 和 D_T 。根据策略，把 D_s 划分为 D_s^{Train} 和 D_s^{Out} 。 D_T 也一分为二，一部分 D_T^{Train} 作为目标模型训练数据的成员，另一部分 D_T^{Out} 作为非成员数据点。

成员推理攻击的本质是对输入数据在模型上的表现进行二分类的结果，故对于可逆神经网络，如果采用中间层的特征信息作为攻击模型的二分类的输入特征，可逆神经网络和传统神经网络或许会产生不一样的结果。假设一种攻击场景，攻击者拿到了半白盒的模型，获取到了分类网络在全连接层之前的输出特征，对于传统神经网络，攻击者只能获得此特征，而对于可逆神经网络，攻击者可以逆向恢复上一层的中间信息。本文提

出一种以该中间层的信息作为特征向量训练攻击模型的成员推理攻击方法, 对可逆神经网络做进一步的隐私评估。

2.2 模型逆向攻击

模型逆向攻击旨在从目标模型中重建数据样本, 即让攻击者获得关于训练数据集的信息。本文首先考虑攻击者拥有白盒访问权限, 并且不需要任何辅助数据集。对于目标模型的每一类, 攻击者首先创建一个高斯噪声样本, 将该样本提供给模型, 然后获取后验输出。攻击者使用梯度下降来优化输入样本进行修正, 使类对应的后验可以超过预先设置的阈值, 从而获得高置信度的样本。

为了更加直观展示模型逆向攻击的实验效果, 本文在 CelebA 人脸数据集上做模型逆向可视化实验, 由于 CelebA 数据集较大, 使用基于梯度的攻击方法效果较差, 无法成功反演。因此尝试了基于 GAN 的模型逆向攻击方法^[11], 该方法分为两个步骤完成: ① 构建一个反演特异性 GAN 从公开数据中提取人脸图像的分布信息; ② 恢复私有域分布, 其中图 3 为反演特异性 GAN 训练过程, 假设目标网络将一个样本分类为 K 个可能类别中的一个。这里的判别器 (D) 是一个 $K+1$ 分类器, 其中前 K 个类对应目标网络的标签, 第 $K+1$ 个类代表假样本。为了训练这种判别器, 使用目标模型为公共数据集中的每幅图像生成一个软标签 $T(x)$ 。判别器训练损失由自监督 (supervised) 损失和无监督 (unsupervised) 损失两部分组成。

$$L_D = L_{\text{supervised}} + L_{\text{unsupervised}} \quad (5)$$

$$L_{\text{supervised}} = -\mathbb{E}_{x \sim p_{\text{data}}(x)} \sum_{k=1}^K T_k(x) \log p_{\text{disc}}(y=k|x) \quad (6)$$

$$L_{\text{unsupervised}} = -\{\mathbb{E}_{x \sim p_{\text{data}}(x)} \log D(x) + \mathbb{E}_{z \sim \text{noise}} \log(1 - D(G(z)))\} \quad (7)$$

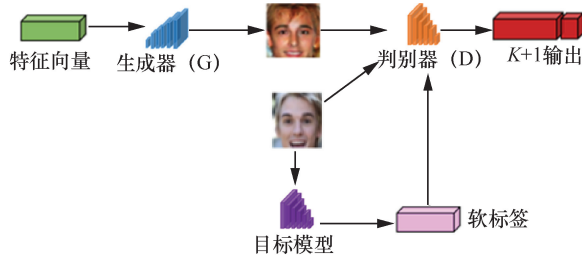


图 3 反演特异性 GAN 训练过程
Figure 3 Inverse specific GAN training process

这里 p_{data} 是公共数据集的分布, $p_{\text{disc}}(y|x)$ 是判别器预测 x 为 y 的概率, 随机噪声 z 从 $N(0,1)$ 中采样, $D(x)$ 输出 x 为真实图像的概率, 对于训练生成器 (G), 使用特征匹配损失来根据鉴别器的中间层学习到的特征 $f(x)$, 将生成的图像与真实的对应图像对齐比较。

$$L_G = \|\mathbb{E}_{x \sim p_{\text{data}}(x)} f(x) - \mathbb{E}_{z \sim \text{noise}} f(G(z))\|_2^2 + \lambda_h L_{\text{entropy}} \quad (8)$$

其中, L_{entropy} 是一个熵正则化矩阵。图 4 是恢复私有域分布流程, 给定一个标签 k , 在两个可学习参数的 $p_{\text{gen}} = N(\mu, \sigma^2)$ 中, 根据参数化策略采样得到 $z' = \sigma \varepsilon + \mu, \varepsilon \sim N(0,1)$ 作为特征向量。训练过程的损失由 identity loss (身份损失) 和 prior loss (先验损失) 两部分组成, 具体表达式如下。

$$L = L_{\text{prior}} + \lambda_i L_{\text{id}} \quad (9)$$

$$L_{\text{prior}} = -\frac{1}{L} \sum_{l=1}^L \log D(G(\sigma \varepsilon_l + \mu)) \quad (10)$$

$$L_{\text{id}} = -\frac{1}{L} \sum_{l=1}^L \log T_k(G(\sigma \varepsilon_l + \mu)) \quad (11)$$

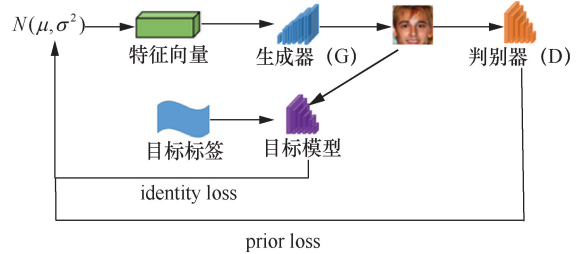


图 4 恢复私有域分布流程
Figure 4 Recover the distribution of private domain process

这里仍然采用前文提到的 3 种模型作为目标模型进行实验, 并建立了一个评估分类器, 根据输入的重建图像预测身份。如果评估分类器在重建图像上达到较高的精度, 则认为重建图像公开了目标标签的私有信息, 将识别的准确度作为攻击成功率来评估。

2.3 属性推理攻击

属性推理攻击的工作与成员推理攻击类似, 每个数据点都有两个标签, 以第一个标签训练目标模型, 第二个标签作为额外属性, 将目标模型的输出和第二个标签作为输入训练攻击模型, 即可对目标数据在目标模型上的输出进行预测判断

其是否有第二个标签的属性。

2.4 模型窃取攻击

模型窃取攻击指的是攻击者通过黑盒查询，还原模型或者获取模型训练集信息的攻击^[7]。也就是说，需要找到学生模型的参数 θ_s ，使学生模型的预测 $S(x)$ 和受害者模型预测 $V(x)$ 之间的误差概率最小化

$$\arg \min_{\theta_s} P_{x \sim D_v} \left(\arg \max_i V_i(x) \neq \arg \max_i S_i(x) \right) \quad (12)$$

受害者模型的训练集 D_v 是不可公开的，无数据模型提取攻击通过优化损失函数 L 使分类误差最小化，这个误差函数用来衡量受害者模型和学生模型之间的差异。

$$\arg \min_{\theta_s} E_{x \sim D_S} [L(V(x), S(x))] \quad (13)$$

这里采用数据无关的基于 GAN 的模型窃取攻击方法^[9]。攻击流程如图 5 中的黑色箭头所示：从一个标准正态分布中采样一个随机噪声向量 z ，并将其输入生成器中，生成图像 x 。受害者模型 V 和学生模型 S 分别以 x 作为输入得到后验输出，计算出损失函数 L 。在反向传播阶段需要计算两个不同来源的梯度： L 对学生模型的参数 θ_s 的梯度和对生成器的参数 θ_g 的梯度。因为受害者模型只能作为一个黑盒来访问，所以不能通过它传播梯度。虚线箭头表示需要正向差分进行梯度近似。每次迭代交替训练生成器和学生模型 S 。用窃取模型的性能来评估模型泄露风险。

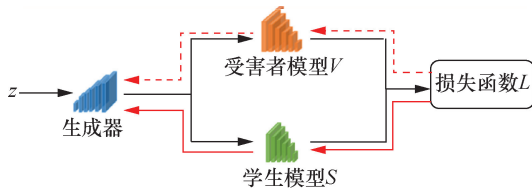


图5 无数据模型窃取攻击流程
Figure 5 Data-free model extraction attack process

3 实验结果与分析

3.1 成员推理攻击

本文实验在 CIFAR-10、CIFAR-100、GTSRB、CelebA 等数据集上进行，采用的服务器为 NVIDIA RTX2080，操作系统为 Ubuntu16.04。为了尽可能避免随机性，本文重复实验 10 次，

并报告平均值和标准差值。CIFAR10 数据集共有 60 000 张彩色图像，这些图像尺寸为 $32 \times 32 \times 3$ ，分为 10 个类，每个类 6 000 张，CIFAR100 也是 60 000 张彩色图像，图像尺寸为 $32 \times 32 \times 3$ ，GTSRB 为交通标志识别图像，包括 43 种交通信号，STL10 图像从 ImageNet 中获取，包含 10 个类别（飞机、鸟、汽车、猫、狗、马、猴子、鹿、船、卡车），每个类含有 800 张测试样本， 96×96 大小的 RGB 彩色图像。运用前文所述攻击方法，得到面向模型无关的成员推理攻击成功率如表 2 所示，在这样的成员推理攻击的敌手策略下，RevNet 和 i-RevNet 的攻击成功率都高于 ResNet，这也说明可逆神经网络在成员推理攻击的攻击策略下比 ResNet 有着更高的隐私泄露风险。

表2 面向模型无关的成员推理攻击成功率
Table 2 Accuracy of model-oriented independent member inference attacks

网络模型	CIFAR10	CIFAR100	GTSRB	STL10
ResNet-32	63.88%	70.48%	71.36%	75.96%
RevNet-38	69.07% (↑ 5.19%)	73.54% (↑ 3.06%)	72.58% (↑ 1.22%)	81.02% (↑ 5.06%)
i-RevNet-31	72.44% (↑ 8.56%)	74.58% (↑ 4.10%)	74.67% (↑ 3.31%)	81.53% (↑ 5.57%)

假设一种场景：攻击者获取了模型全连接层之前的输入特征向量 x 和最靠近全连接层模块的网络结构与参数信息。此时，传统神经网络只能采用 x 作为二分类特征训练攻击模型，可逆神经网络则可以逆向恢复上一层输入 x_1 作为特征训练攻击模型。因此，3 种网络的训练特征分别为： $N \times$ 分类类别数， $N \times 112 \times 8 \times 8$ ， $N \times 256 \times 8 \times 8$ ，中间层信息的成员推理攻击如图 6 所示。攻击模型分别设计为 MLP 和上层堆叠预处理多维度数据卷积核的卷积神经网络模型，二者线性层部分结构相同，数据使用 CIFAR10，得到中间层信息与面向模型无关的成员推理攻击成功率对比如表 3 所示。

从表 3 可知，ResNet 在该条件下的攻击成功率很低，几乎失效，但是两种可逆神经网络仍然可以有一定的攻击成功率。成员推理攻击是模型的过拟合导致训练集和测试集在模型上的表现差异而产生的。通常分类模型过拟合的关键因素在于最后的全连接层上，而可逆神经网络很可能记住了过多训练数据的信息，使得模型过拟合的位置提前，在前面

可逆的层中就已经造成训练集和测试集表现差异较大从而攻击成功,泄露了更多的数据隐私。

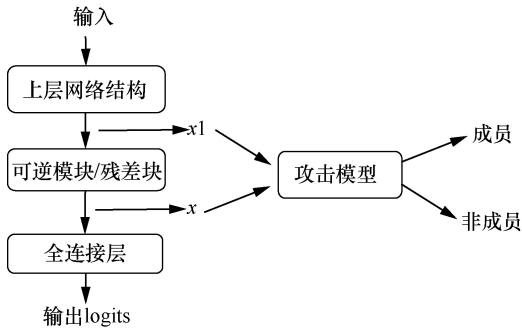


图 6 中间层信息的成员推理攻击
Figure 6 Membership inference attacks against middle layers

表 3 中间层信息与面向模型无关的
成员推理攻击成功率对比

Table 3 Comparison of accuracy between membership inference attack against middle layers and model independent membership inference attack

网络模型	中间层信息的 成员推理攻击	面向模型无关的 成员推理攻击
ResNet-32	53.20%	63.88%
RevNet-38	68.72%	69.07%
i-RevNet-31	72.32%	72.44%

3.2 模型逆向攻击

模型逆向攻击是为目标模型的每一类生成一个具有代表性的样本。这里设置阈值为 0.999 999, 初始学习率 0.1, 最大迭代次数为 500 000, 在达到迭代次数之后, 将反演出来的数据访问原始模型得到后验概率, 损失越小说明反演效果越好, 模型逆向攻击越成功。模型逆向攻击损失如表 4 所示, 可见 RevNet 比 ResNet 的损失小了一个数量级, 反演效果 i-RevNet 和 RevNet 均好于 ResNet。为了更直观地感受重建图像的效果, 对 3 种模型逆向攻击结果可视化来进行隐私风险评估。本文对 CelebA 的 1 000 个类别的 30 027 张图片作为要逆向恢复私有数据集来训练目标模型, 选取了另外 1 000 个类别的 30 000 张图片来训练 GAN, 两个 1 000 类别之间没有任何重叠, 并建立了一个评估分类器, 根据输入的重建图像预测身份。建立了一个评估分类器, 根据输入的重建图像预测身份。如果评价分类器达到较高的精度, 则认为重建图像公开了目标标签的私有信息, 将识别的准确度作为攻击成功率, 反演出来的模型逆向攻击结果可视化如图 7 所示。

表 4 模型逆向攻击损失

Table 4 Loss of model inversion attacks

网络模型	CIFAR100	CIFAR10
ResNet-32	3.47×10^{-6}	4.05×10^{-6}
RevNet-38	6.38×10^{-7}	5.96×10^{-7}
i-RevNet-31	2.27×10^{-7}	2.45×10^{-6}

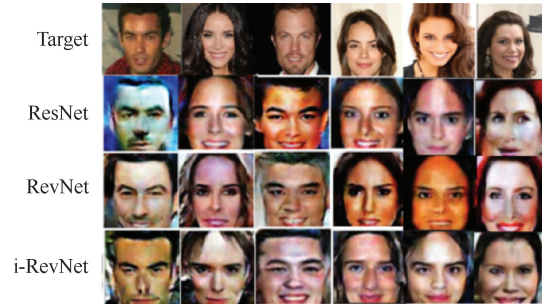


图 7 模型逆向攻击结果可视化
Figure 7 Model inversion attack results visualization

从表 5 可以看到, 对于 RevNet 和 i-RevNet, 采用基于 GAN 的模型逆向攻击策略均产生了更高的攻击成功率。

表 5 模型逆向攻击成功率

Table 5 Accuracy of model inversion attack

网络模型	CelebA
ResNet-32	39.28%
RevNet-38	43.25%(↑1.97%)
i-RevNet-31	50.25%(↑10.97%)

3.3 属性推理攻击

本文使用数据集 UTKFace 来评估属性推理攻击, 因为该数据集为样本提供了额外的属性, 可以用作目标属性。对于该数据集, 采用“人种”作为目标属性进行推理。攻击模型是一个两层 MLP, 它的输入是目标样本从目标模型的第一层到最后一层的输入, 属性推理攻击成功率如表 6 所示。

表 6 属性推理攻击成功率

Table 6 Accuracy of attribute inference attacks

网络模型	UTKFace
ResNet-32	46.29%
RevNet-38	48.02% (↑1.73%)
i-RevNet-31	48.96% (↑2.67%)

3.4 模型窃取攻击

模型窃取攻击实验仍然以 CIFAR10 作为数

据集，这里分别采用 ResNet-18、RevNet-38、i-RevNet-31 作为学生模型的结构进行窃取实验，结果如表 7 所示，在学生模型和被攻击模型结构类似时效果较好。但是窃取 RevNet 和 i-RevNet 模型得到的性能并没有比 ResNet 高。接下来，针对 3 种网络同样采用 InceptionV3^[12] 作为学生模型结构又进行了一样的对比实验。从表 8 可以看出，可逆神经网络较传统神经网络反而没有那么容易被模型窃取攻击。这是由于可逆神经网络保留了输入数据的信息，在数据隐私层面会出现隐私泄露而在模型隐私层面并不一定会出现隐私泄露。

表 7 模型窃取攻击成功率

Table 7 Accuracy of model extraction attack

网络模型	ResNet-18	RevNet-38	i-RevNet-31
ResNet-32	86.97%	80.22%	55.23%
RevNet-38	82.03%	83.38%	38.56%
i-RevNet-31	81.76%	72.10%	66.41%

表 8 模型窃取攻击成功率（攻击模型统一为 Inception V3）
Table 8 Accuracy of model extraction attack (The attack model is unified as Inception V3)

网络模型	InceptionV3
ResNet-32	84.76%
RevNet-38	79.85%
i-RevNet-31	80.79%

3.5 消融实验

从上述几种不同的隐私攻击策略中可以看出，可逆块会在不同的数据层面隐私攻击策略下造成不同程度的隐私泄露。本文在不改变整体 RevNet 和 i-RevNet 结构的前提下，将部分可逆模块修改为残差模块，分别将 RevNet 和 i-RevNet 的最后 3 个和 6 个可逆模块修改还原为残差模块，消融实验所用的网络结构参数如表 9 所示，A、B 分别表示还原部分可逆块为不可逆的残差块模型。通过这样的控制变量实验，能够更好地评估可逆模块造成的隐私风险。

所有实验均在 CIFAR10 数据集上进行，其他实验设置与前文实验保持一致，RevNet 和 i-RevNet 的消融实验结果如表 10 和表 11 所示。

表 9 消融实验所用的网络结构参数（加粗部分为残差块）
Table 9 Network structure used in ablation experiment parameter (the bold part is residual block)

网络模型	单位	通道数
RevNet-38	3-3-3	32-32-64-112
RevNet-38-A	3-3- 3	32-32-64-112
RevNet-38-B	3-3-3	32-32-64-112
i-RevNet-31	1-3-6	8-32-128
i-RevNet-31-A	1-3- 6	8-32-128
i-RevNet-31-B	1-3-6	8-32-128

表 10 RevNet 消融实验结果

Table 10 Results of RevNet ablation experiment

网络模型	成员推理攻击成功率	模型逆向攻击损失	属性推理攻击成功率
RevNet-38	69.07%	5.96×10^{-7}	48.02%
RevNet-38-A	67.15%	5.96×10^{-7}	47.97%
RevNet-38-B	66.37%	1.66×10^{-6}	47.64%

表 11 i-RevNet 消融实验结果

Table 11 Results of i-RevNet ablation experiment

网络模型	成员推理攻击成功率	模型逆向攻击损失	属性推理攻击成功率
i-RevNet-31	72.44%	2.45×10^{-6}	48.96%
i-RevNet-31-A	71.76%	2.62×10^{-6}	48.06%
i-RevNet-31-B	70.18%	3.33×10^{-6}	46.67%

实验表明，当部分可逆块为残差块时，隐私攻击成功率有所下降，可逆块的个数和模型的隐私泄露风险是正相关的，随着神经网络可逆结构数量的增加，网络保留的数据信息更多更全面，隐私泄露问题也更严重。

4 防御策略

Liu 等^[13]曾在对机器学习模型推理攻击的全局性研究中通过大量实验提出两大观点。其一，深度模型的过拟合程度能对隐私攻击的效果造成显著影响。针对成员推理攻击，过拟合的模型更容易关注于训练集的特征而对训练集的预测比非成员的预测更有信心，往往更容易受到影响^[14]。针对模型窃取攻击，过度拟合的模型在很大程度上记忆了它的训练数据集，因此被盗模型很可能无法较好地逼近目标模型的输出，从而导致攻击效果变差。其二，复杂的训练集可以导致成员推理攻击成功率更高，而模型窃取攻击成功率较低，

这是因为模型对复杂的数据集的特征很难进行归纳。学术界防御隐私攻击的主要手段可分为知识蒸馏和差分隐私两大类,对于所有的隐私攻击暂时还没有一种通用的防御方法^[15]。知识蒸馏可以缩小成员数据与非成员数据之间的差距,从而保护成员数据隐私。但知识蒸馏仅能有效防御成员推理攻击,对其他类型的攻击没有太好的效果。Abadi^[16]基于差分隐私策略提出了 DP-SGD,即在模型训练反向传播过程中对目标模型的梯度进行高斯噪声处理,该方法可以防御多种隐私攻击方式。除此之外,利用模型压缩使模型信息泄露风险降低也是一种可行的防御策略。

通过实验可知,可逆神经网络的隐私泄露风险与网络的可逆程度正相关,说明保留的信息越多,可逆神经网络的隐私泄露风险越大。若对可逆神经网络中间层保留的数据集信息采用一定的方法进行信息约减,也许能够在维持网络可逆性的同时降低其隐私泄露的风险。

5 结束语

本文通过将4种隐私攻击应用于3种不同的模型,提出了模型可逆性与隐私泄露风险之间存在的关系,证明了可逆神经网络模型具有更大的隐私风险。通过运用3种数据隐私层面的隐私攻击策略,发现攻击模型 RevNet 和 i-RevNet 有更高的攻击成功率。对于模型隐私层面的模型窃取攻击,结果并非如此,没有明显的隐私泄露现象。

本文使用的这4种攻击中,成员推理攻击和属性推理攻击属于数据隐私层面的直接隐私攻击方式,仅仅推理出了数据是否属于训练集和部分隐私属性^[17],模型逆向攻击反演出了更多有关训练数据集的隐私信息。而模型窃取攻击属于模型隐私层面的间接攻击策略,关注点在于模型自身的隐私^[18]。可逆神经网络结构保留了更多的数据信息,因此在数据层面的直接隐私攻击更容易。本文假设只获取了分类网络全连接层的输入信息的半白盒攻击场景,设计了一种针对可逆神经网络的中间层信息的成员推理攻击,该种攻击场景下,成员推理攻击对 ResNet 几乎无效,而对两种可逆神经网络仍然有一定的成功率,也从侧面说明,可逆神经网络受到隐私攻击的场景更加多样

化,中间层对输入数据信息的保留导致了更大的隐私泄露风险。

通过消融实验发现,当参数数量和网络层数接近,可逆模块的数量增加,数据隐私攻击成功率也有所提高。对于特殊结构的网络,如果能够使用额外的隐式输出变量来捕获正向过程中可能丢失的信息,通过观测值逆向得到中间层的隐藏信息,则可能更容易受到数据层面的 AI 隐私攻击,攻击成功率的提升与模型保留信息的程度呈正相关。未来,建立隐私保护机制多维评估体系十分必要,在不同模型、不同攻击方式下对三者之间的关系进行建模,实现三者在不同应用场景下的权衡最优化。目前,学术界的隐私泄露风险评估基本是针对某一特定领域,评估手段的应用范围受到限制,深度学习的隐私问题可以从多方面进行考量^[19],隐私攻击的评估手段只是其中一种具有局限性的方式,建立一套系统的隐私评估方式是未来人工智能安全领域研究的发展趋势。

随着可逆神经网络的应用领域越来越广泛,在处理隐私数据时必须重新考虑可逆神经网络的隐私泄露风险问题,至少模型应该对数据隐私泄露进行定性分析^[20]。目前还需要做更多的工作来量化评估模型的隐私泄露度,以实现可逆神经网络的优势和隐私的权衡,有利于开发更好的攻击和防御策略^[21]。

参考文献:

- [1] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models[C]//IEEE Symposium on Security and Privacy. 2017
- [2] KHOSRAVY M, NAKAMURA K, HIROSE Y, et al. Model inversion attack: analysis under gray-box scenario on deep learning based face recognition system[J]. KSII Transactions on Internet and Information Systems, 2021, 15(3): 1100-1119.
- [3] TRUONG J B, MAINI P, WALLS R, et al. Data-free model extraction[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021
- [4] 谭作文, 张连福. 机器学习隐私保护研究综述[J]. 软件学报, 2020, 31(7): 30.
TAN Z W, ZHANG L F. Survey on privacy preserving techniques for machine learning [J], Journal of Software. 2020, 31(7): 30.
- [5] DINH L, KRUEGER D, BENGIO Y. NICE: non-linear independent components estimation[J]. Computer Science, 2014.
- [6] GOMEZ A N, REN M, URTASUN R, et al. The reversible residual network: backpropagation without storing activations[C]//International Conference on Neural Information Processing Systems(NIPS).

- 2017: 2211-2221
- [7] JACOBSEN J H, SMEULDERS A, OYALLON E. i-RevNet: deep invertible networks[C]//International Conference on Learning Representations. 2018.
- [8] MELIS L, SONG C Z, DE-CRISTOFARO E, et al. Exploiting unintended feature leakage in collaborative learning[C]//Proc of the IEEE Symp on Security and Privacy. 2019
- [9] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction APIs[C]//Proc of the USENIX Security Symposium. 2016: 601-618.
- [10] SALEM A, ZHANG Y, HUMBERT M, et al. ML-Leaks: model and data independent membership inference attacks and defenses on machine learning models[C]//Network and Distributed Systems Security (NDSS) Symposium. 2019.
- [11] CHEN S I, KAHLA M, JIA R X, et al. Knowledge-enriched distributional model inversion attacks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 16178-16187.
- [12] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2818-2826.
- [13] LIU Y, WEN R, HE X, et al. ML-doctor: holistic risk assessment of inference attacks against machine learning models[C]//USENIX Security Symposium. 2022.
- [14] MEJIA F A, GAMBLE P, HAMPEL-ARIAS Z, et al. Robust or private? adversarial training makes models more vulnerable to privacy attacks[R]. 2019.
- [15] AL-RUBAIE M, CHANG J M. Privacy-preserving machine learning: threats and solutions[J]. IEEE Security & Privacy, 2019, 17(2): 49-58.
- [16] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//ACM SIGSAC Conference on Computer and Communications Security (CCS). 2016: 308-318.
- [17] CHOO C A C, TRAMER F, CARLINI N, et al. Label-only membership inference attacks[J]. arXiv Preprint arXiv: 2007. 14321, 2020.
- [18] CHENG Z, LI Z, ZHANG J, et al. Differentially private machine learning model against model extraction attack[C]//2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics). 2020.
- [19] JIA J Y, SALEM A, BACKES M, et al. MemGuard: defending against black-box membership inference attacks via adversarial examples[C]//Proceedings of 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 259-274.
- [20] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning[C]//Proceedings of 2019 IEEE Symposium on Security and Privacy. 2019: 739-753.
- [21] ZHANG G, LIU B, TIAN H, et al. How does a deep learning model architecture impact its privacy[J]. arXiv Preprint arXiv: 2210. 11049, 2022.

[作者简介]



何毅凡（1998—），男，湖北武汉人，中国科学技术大学硕士生，主要研究方向为人工智能安全与隐私保护。



张杰（1995—），男，河南驻马店人，中国科学技术大学博士生，主要研究方向为人工智能安全、隐私及其版权保护。



张卫明（1976—），男，河北定州人，中国科学技术大学教授、博士生导师，主要研究方向为信息隐藏、多媒体内容安全、人工智能安全。



俞能海（1964—），男，安徽无为，人，中国科学技术大学教授、博士生导师，主要研究方向为多媒体信息检索、图像处理与视频通信、数字媒体内容安全。