(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115719085 A
(43)申请公布日 2023.02.28

(21)申请号 202310030791.1
(22)申请日 2023.01.10
(71)申请人 武汉大学
地址 430072 湖北省武汉市武昌区八一路299号
(72)发明人 曹雨欣 田博为 王骞 龚雪鸾 沈超 李琦
(74)专利代理机构 武汉科皓知识产权代理事务所(特殊普通合伙) 42222
专利代理师 齐晨涵
(51)Int.Cl.
G06N 3/0475(2023.01)
G06N 3/094(2023.01)
G06N 3/084(2023.01)
G06F 18/24(2023.01)
G06F 21/62(2013.01)

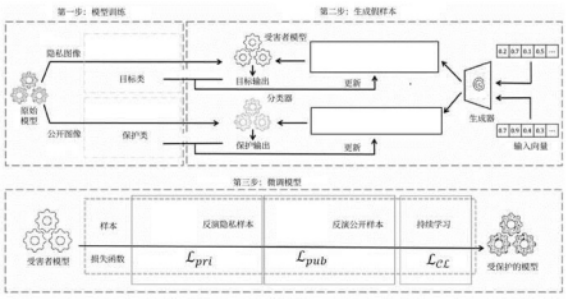
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种深度神经网络模型反演攻击防御方法及设备

(57)摘要

本发明涉及一种深度神经网络模型反演攻击防御方法及设备，提出了一种基于生成对抗网络和假样本的模型反演攻击防御方法，先利用生成对抗网络生成虚假样本，在根据虚假样本在保证目标受害者模型的有效性的基础上，微调目标受害者模型参数，从而实现防御模型反演攻击的目的。本发明可以有效地对抗模型反演攻击，保护数据隐私性，同时确保模型的高可用性。



CN 115719085 A

1.一种深度神经网络模型反演攻击防御方法,其特征在于,包括以下步骤:

步骤1,训练模型;所述模型包括使用隐私数据集训练的目标受害者模型与使用公开数据集的分类器模型,所述目标受害者模型采用深度学习模型;

步骤2,利用生成对抗网络,生成虚假样本;所述生成对抗网络包括由深度神经网络组成的生成器模块和鉴别器模块;所述虚假样本包括反演公开样本和反演隐私样本;所述反演公开样本是指,对基于公开数据集训练的分类器进行反演攻击来重构获得的样本;所述反演隐私样本是指,对目标受害者模型进行反演攻击来重构获得的样本;

步骤3,根据虚假样本在保证目标受害者模型的有效性的基础上,微调目标受害者模型参数,实现防御模型反演攻击的目的,微调过程为最大化反演隐私样本的损失,最小化反演公开样本的损失。

2.根据权利要求1所述的深度神经网络模型反演攻击防御方法,其特征在于:所述目标受害者模型的训练过程具体为:

步骤1.1,根据隐私数据样本的交叉熵构建损失函数;

步骤1.2,利用所述损失函数和随机梯度下降算法训练目标受害者模型。

3.根据权利要求1所述的深度神经网络模型反演攻击防御方法,其特征在于:所述目标受害者模型的训练过程具体为:所述目标受害者模型在resnet、VGGnet、AlexNet、FaceNet、VGG-16中择一使用。

4.根据权利要求1所述的深度神经网络模型反演攻击防御方法,其特征在于:所述目标受害者模型的训练过程具体为:所述目标受害者模型为VGG-16,包括22层和37个深度单元。

5.根据权利要求1所述的深度神经网络模型反演攻击防御方法,其特征在于:所述目标受害者模型的训练过程具体为:所述分类器在resnet、VGGnet、AlexNet、FaceNet中择一使用。

6.根据权利要求1所述的深度神经网络模型反演攻击防御方法,其特征在于:所述反演公开样本的生成过程具体为:

步骤2.1.1,随机选取公开数据集中的一类保护数据l_p;

步骤2.1.2,初始化生成对抗网络的生成器模块输入向量z_{pub},获得输出图像img_{pub};

步骤2.1.3,将步骤2.1.2所述图像img_{pub}输入分类器模块,获得输出向量;

步骤2.1.4,构建损失函数;所述损失函数计算过程如下:

$$Z^* = arg\,min L_{discri} + \lambda L_{iden},\ L_{discri} = -log\left(D\left(G(z_{pub})\right)\right),\ L_{iden} = L_{CE}\left(F_C\left(G(z_{pub})\right), l_p\right)$$

其中Z^*表示优化反演公开样本优化过程中学习后的z_{pub}向量,L_{discri}表示生成对抗网络鉴别器损失,L_{iden}表示可识别性损失,D表示生成对抗网络中的鉴别器模块,F_C表示分类器模块,G表示生成对抗网络中的生成器模块,λ为L_{iden}的参数,L_{CE}表示交叉熵函数;

步骤2.1.5,根据步骤2.1.4的损失函数更新生成对抗网络的输入z_{pub};

步骤2.1.6,循环执行所述步骤2.1.1到步骤2.1.5,此循环过程将在到达预先设置的迭代次数后终止;将最终的z_{pub}输入生成对抗网络的生成器模块,输出即为反演公开样本。

7.根据权利要求1所述的深度神经网络模型反演攻击防御方法,其特征在于:

所述生成反演隐私样本的过程具体为:

步骤2.2.1,选取隐私数据集中的需要保护的一类数据l_t;

步骤2.2.2,初始化生成对抗网络的生成器模块输入向量z_{pri},获得输出图像img_{pri};

步骤2.2.3,将步骤2.2.2所述图像img_{pri}输入目标受害者模型模块,获得输出向量;

步骤2.2.4,构建损失函数;所述损失函数计算过程如下:

$$Z^* = arg\min L_{discri} + \lambda L_{iden}\ ,\ L_{discri} = -log\left(D\left(G(z_{pri})\right)\right),\ L_{iden} = L_{CE}\left(F_C\left(G(z_{pri})\right), l_t\right)$$

其中Z^*表示优化反演隐私样本优化过程中学习后的z_{pri}向量,L_{discri}表示生成对抗网络鉴别器损失,L_{iden}表示可识别性损失,D表示生成对抗网络中的鉴别器模块,F_C表示分类器模块,G表示生成对抗网络中的生成器模块,λ为L_{iden}的参数;

步骤2.2.5,将步骤2.2.4所述的损失向后传递,更新生成对抗网络的输入z_{pri};

步骤2.2.6,循环执行所述步骤2.2.1到步骤2.2.5,此循环过程将在到达预先设置的迭代次数后终止,将最终的z_{pri}输入生成对抗网络的生成器模块,输出即为反演隐私样本。

8.根据权利要求1所述的深度神经网络模型反演攻击防御方法,其特征在于:微调目标模型参数的过程利用了持续学习算法,还包括可塑权重巩固的正则化项,所述正则化项为$\mathcal{L}_{CL} = \sum_i \frac{\lambda_3}{2} F_i(\theta_i - \theta_i^*)^2$,步骤3的全部过程表示为

$\theta = \arg\min -\alpha\mathcal{L}_{\mathbf{pri}} + \beta\mathcal{L}_{\mathbf{pub}} + \omega\mathcal{L}_{\mathbf{CL}}$,其中θ为目标受害者模型的参数集,α、β和ω分别是三个损失项的参数,所述反演隐私样本的损失函数为$\mathcal{L}_{\mathbf{pri}}$,所述反演公开样本的损失函数为$\mathcal{L}_{\mathbf{pub}}$。

9.一种电子设备,其特征在于,包括:

一个或多个处理器;

存储装置,用于存储一个或多个程序;

当一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-8中任一项所述方法所执行的操作。

10.一种计算机可读介质,其上存储有计算机程序,其特征在于:所述程序被处理器执行时实现如权利要求1-8中任一项所述方法所执行的操作。

一种深度神经网络模型反演攻击防御方法及设备

技术领域

[0001] 本发明属于人工智能安全领域，尤其涉及一种深度神经网络模型反演攻击防御方法及设备。

背景技术

[0002] 目前，深度学习在各种现实应用中表现得令人印象深刻，例如人脸识别、自动驾驶和目标检测。训练一个高性能的深度学习模型需要大量的敏感或私有数据，以及大量的计算资源和存储资源。许多模型供应商将训练良好的模型作为web服务提供给用户，用户可以发送输入样例并获得模型的输出。一个流行的例子是人脸识别web APIs，这样的服务被广泛用于各个领域的人脸验证与人脸识别。微软云认知服务和Naver Clova也为用户提供了其他类型的图像分析web API。然而最近的研究表明，通过访问受害者模型的方式，攻击者可能会获取到敏感的训练数据，这带来了极大的安全隐患。近年来，一种被称为模型反演攻击Model Inversion Attacks，简称MIA的新型攻击引起了广泛的关注。

[0003] 模型反演攻击的原理是向深度学习模型发送大量的询问请求，并根据API的输入和输出恢复出任何给定标签的相应训练数据。比如，对于一个人脸识别模型，模型反演攻击可以重建训练数据中任何人的面部。随着模型反演技术的发展，现有的工作甚至可以对在高分辨率数据集上训练的深度神经网络都很有效。这样高准确度的重建面部图像甚至可以通过访问控制系统，带来了严重的安全风险。

[0004] 模型反演攻击可以分为黑盒设置和白盒设置。在黑盒设置中，攻击者只能通过获取预测向量的方式来访问模型，比如谷歌云机器学习引擎)；在白盒设置中，攻击者可以获得商业模型的所有信息，比如一个开源可下载的人脸识别服务模型。

[0005] 迄今为止，很少有专门针对模型反演的防御工作。现有技术一般通过下述方法实现敏感信息的数据的保护：

方法一：基于差分隐私的防御，通过向数据中添加噪声来保护隐私信息。理论上，差分隐私可以用于保护训练数据的安全性。但是，差分隐私无法在保证受害者模型准确率的同时，保护数据的隐私性。此外，有理论分析表明，差分隐私无法防御模型反演攻击。

[0006] 方法二：基于预测纯化的防御，通过纯化模型的预测输出以防御攻击。具体来说，防御方需要训练一个净化器模型，旨在将返回的可信度向量中包含的信息最小化，并保持模型的预测准确性。为了防御模型反演攻击，该方法减少了模型预测置信度向量在成员和非成员数据之间的差别。因此，训练样本与预测向量之间的相关性减弱。于是，攻击者在进行模型反演攻击时，无法获得准确的训练数据。

[0007] 方法三：基于预测扰动的防御，通过向预测输出中加入噪声来干扰攻击者。这种策略要求最大化反演误差，并尽量减少对于目标受害者模型的可用性损失。但是在实际应用中，此方法会损害目标受害者模型的预测准确率。

[0008] 综上，目前针对模型反演的防御存在矛盾：在保证了模型可用性的情况下，无法有效地防御先进的模型反演攻击；相反的，加入过量噪声可以保护训练数据的安全性，却大大

影响了模型的准确率。

发明内容

[0009] 本发明的目的是在于提供一种基于生成对抗网络和假样本的模型反演攻击防御方法,可以在有效地保护数据隐私性的同时,确保模型的高可用性。

[0010] 为实现以上目的,本发明提供了如下方案:

步骤1,训练模型;所述模型包括使用隐私数据集训练的目标受害者模型与使用公开数据集的分类器模型,目标受害者模型为深度学习模型,采用现有的深度学习模型均可,分类器采用现有分类器均可。

[0011] 步骤2,利用生成对抗网络,生成虚假样本。所述生成对抗网络包括由深度神经网络组成的生成器模块和鉴别器模块;所述虚假样本包括反演公开样本和反演隐私样本;所述反演公开样本是指,对基于公开数据集训练的分类器进行反演攻击来重构获得的样本;所述反演隐私样本是指,对目标受害者模型进行反演攻击来重构获得的样本。

[0012] 步骤3,根据虚假样本在保证目标受害者模型的有效性的基础上,微调目标受害者模型参数,从而实现防御模型反演攻击的目的。

[0013] 可选的,如果目标受害者模型已经预训练完成,将直接使用所述目标受害者模型;否则,所述目标受害者模型的训练过程具体为:

步骤1.1,根据隐私数据样本的交叉熵构建损失函数。

[0014] 步骤1.2,利用所述损失函数和随机梯度下降算法训练目标受害者模型。

[0015] 分类器模型的训练过程具体包括:

1)所述分类器模型的训练集为公开数据集;所述的公开数据集是与隐私数据集相同领域的一个常用公开数据集。

[0016] 2)所述分类器模型的训练遵循标准深度神经网络训练过程。

[0017] 可选的,生成反演公开样本包括分类器模块、生成对抗网络模块两个部分,所述生成反演公开样本的过程具体为:

步骤2.1.1,随机选取公开数据集中的一类保护数据l_p;

步骤2.1.2,初始化生成对抗网络的生成器模块输入向量z_{pub},获得输出图像img_{pub};

步骤2.1.3,将步骤2.1.2所述图像img_{pub}输入分类器模块,获得输出向量;

步骤2.1.4,构建损失函数;所述损失函数计算过程如下:

$$Z^* = arg\,min L_{discri} + \lambda L_{iden},\ L_{discri} = -log\left(D\left(G(z_{pub})\right)\right),\ L_{iden} = L_{CE}\left(F_C\left(G(z_{pub})\right), l_p\right)$$

其中Z^*表示优化反演公开样本优化过程中学习后的z_{pub}向量,L_{discri}表示生成对抗网络鉴别器损失,L_{iden}表示可识别性损失,D表示生成对抗网络中的鉴别器模块,F_C表示分类器模块,G表示生成对抗网络中的生成器模块,λ为L_{iden}的参数,L_{CE}表示交叉熵函数。

[0018] 步骤2.1.5,根据步骤2.1.4的损失函数更新生成对抗网络的输入z_{pub};

步骤2.1.6,循环执行所述步骤2.1.1到步骤2.1.5,此循环过程将在到达预先设置的迭代次数后终止;将最终的z_{pub}输入生成对抗网络的生成器模块,输出即为反演公开样

本。

[0019] 可选的，生成反演隐私样本包括目标受害者模型模块、生成对抗网络模块两个部分，所述生成反演隐私样本的过程具体为：

步骤2.2.1，选取隐私数据集中的需要保护的一类数据l_t；

步骤2.2.2，初始化生成对抗网络的生成器模块输入向量z_{pri}，获得输出图像img_{pri}；

步骤2.2.3，将步骤2.2.2所述图像img_{pri}输入目标受害者模型模块，获得输出向量；

步骤2.2.4，构建损失函数；所述损失函数计算过程如下：

$$Z^* = arg\,min\, L_{discri} + \lambda L_{iden}\ ,\ L_{discri} = -log\left(D\left(G(z_{pri})\right)\right),\ L_{iden} = L_{CE}\left(F_C\left(G(z_{pri})\right), l_t\right)$$

其中Z^*表示优化反演隐私样本优化过程中学习后的z_{pri}向量，L_{discri}表示生成对抗网络鉴别器损失，L_{iden}表示可识别性损失，D表示生成对抗网络中的鉴别器模块，F_C表示分类器模块，G表示生成对抗网络中的生成器模块，λ为L_{iden}的参数；

步骤2.2.5，将步骤2.2.4所述的损失向后传递，更新生成对抗网络的输入z_{pri}；

步骤2.2.6，循环执行所述步骤2.2.1到步骤2.2.5，此循环过程将在到达预先设置的迭代次数后终止，将最终的z_{pri}输入生成对抗网络的生成器模块，输出即为反演隐私样。

[0020] 优选地，微调目标模型参数包括最大化反演隐私样本的损失和最小化反演公开样本的损失，所述反演隐私样本的损失函数为

$\mathcal{L}_{pri} = L_{CE}\left(\theta^*, F_V\left(G.generator\left(z_{pri}\right)\right), l_t\right)$；所述反演公开样本的损失函数为

$\mathcal{L}_{pub} = L_{CE}\left(\theta^*, F_C\left(G.generator\left(z_{pub}\right)\right), l_t\right)$；微调目标模型参数的过程利用了持续学习算法，还包括可塑权重巩固的正则化项，所述正则化项为

$\mathcal{L}_{CL} = \sum_i \frac{\lambda_2}{2} F_i(\theta_i - \theta_i^*)^2$，步骤3的全部过程表示为

$\theta = \arg\min -\alpha\mathcal{L}_{\mathbf{pri}} + \beta\mathcal{L}_{\mathbf{pub}} + \omega\mathcal{L}_{\mathbf{CL}}$，其中θ为目标受害者模型的参数集，α、β和ω分别是三个损失项的参数，所述反演隐私样本的损失函数为$\mathcal{L}_{\mathbf{pri}}$，所述反演公开样本的损失函数为$\mathcal{L}_{\mathbf{pub}}$。

[0021] 基于同一发明构思，本发明还设计了一种电子设备，包括：

一个或多个处理器；

存储装置，用于存储一个或多个程序；

当一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现所述深度神经网络模型反演攻击防御方法所执行的操作。

[0022] 基于同一发明构思，本发明还设计了一种计算机可读介质，其上存储有计算机程序，所述程序被处理器执行时实现所述深度神经网络模型反演攻击防御方法所执行的操作。

[0023] 本发明的有益效果是：

本发明可以与通用的网络服务结合，保护数据的安全。在防御对抗样本的性能方面，本发明的防御效果在面对不同种类的模型反演攻击时，均很好地保护了隐私数据，同时不影响目标受害者模型的准确率。

[0024] 本发明可以防御模型反演攻击，保护深度神经网络模型的安全性。本发明可以应用到已有的网络服务上，使得攻击者无法从受保护的网络中获取隐私数据。

[0025] 本发明不会影响原本神经网络模型的功能。与目前的防御方法不同，本发明使用了持续学习算法，在保护了隐私训练数据的同时，维持了目标受害者模型的准确率。

附图说明

[0026] 图1为本发明所提出的防御方法的总流程图。

[0027] 图2为在不同原始训练集下，本发明与其他防御方法在受到不同种类模型反演攻击的对比结果。

具体实施方式

[0028] 为使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例对本发明进行进一步的详细说明。应当理解，此处所描述的具体实施方式仅仅用以解释本发明，并不限定本发明的保护范围。为了实现对模型反演攻击进行防御，本发明提供了一种基于利用生成对抗网络生成假样本的对模型反演攻击的防御方法，该防御方法包括三个阶段，如图1总体架构图所示，分别为训练模型、生成虚假样本、微调目标受害者模型参数。按照本发明内容完整方法实施的实施例如下：

首先进行模型训练，本实例中目标受害者模型的结构为VGG-16，包括22层和37个深度单元。使用公开数据集CelebA作为公开的训练数据集，本发明的目标受害者模型采用现有技术的其他深度学习模型也可，如resnet、VGGnet、AlexNet、FaceNet模型。

[0029] 步骤1，使用隐私数据集训练目标受害者模型，使用公开数据集CelebA训练分类器模型。在训练目标受害者模型时，根据隐私数据样本的交叉熵构建损失函数，并利用所述损失函数和随机梯度下降算法训练目标受害者模型。分类器模型使用CelebA作为训练集，遵循标准深度神经网络训练过程进行训练。本发明的分类器采用现有技术的分类器即可，如resnet、VGGnet、AlexNet。

[0030] 步骤2，利用生成对抗网络，生成虚假样本。本案例构建的生成对抗网络G包括由深度神经网络组成的生成器模块和鉴别器模块。生成器用于生成反演图像，鉴别器用于鉴别生成器生成图像的可信度。本发明使用生成对抗网络生成虚假样本，用于干扰攻击者的攻击模型。生成器模块和鉴别器模块采用现有的GAN结构即可。

[0031] 生成的虚假样本分为反演公开样本和反演隐私样本两种：对基于公开数据集训练的分类器的干扰类别l_p进行反演攻击，重构获得的反演公开样本；对目标受害者模型中受保护的目标类别l_t进行反演攻击，重构获得的反演隐私样本。反演公开样本是用于干扰的图像，反演隐私样本是需要保护的隐私图像。

[0032] 反演公开样本的生成过程具体为：

步骤2.1.1，随机选取公开数据集中的一类保护数据l_p；

步骤2.1.2，初始化生成对抗网络的生成器模块输入向量z_{pub}，获得输出图像

img_{pub}；

步骤2.1.3，将步骤2.1.2所述图像img_{pub}输入分类器模块，获得输出向量；

步骤2.1.4，构建损失函数；所述损失函数计算过程如下：

$$Z^{*}=arg\,minL_{discri}+\lambda L_{iden}\ ,\ L_{discri}=-log\left(D\left(G(z_{pub})\right)\right),\ L_{iden}=L_{CE}\left(F_{C}\left(G(z_{pub})\right),l_{p}\right)$$

其中Z^{*}表示优化反演公开样本优化过程中学习后的z_{pub}向量，L_{discri}表示生成对抗网络鉴别器损失，L_{iden}表示可识别性损失，D表示生成对抗网络中的鉴别器模块，F_{C}表示分类器模块，G表示生成对抗网络中的生成器模块，λ为L_{iden}的参数，L_{CE}表示交叉熵函数；

步骤2.1.5，根据步骤2.1.4的损失函数更新生成对抗网络的输入z_{pub}；

步骤2.1.6，循环执行所述步骤2.1.1到步骤2.1.5，此循环过程将在到达预先设置的迭代次数后终止；将最终的z_{pub}输入生成对抗网络的生成器模块，输出即为反演公开样本。

[0033] 生成反演隐私样本的过程具体为：

步骤2.2.1，选取隐私数据集中的需要保护的一类数据l_{t}；

步骤2.2.2，初始化生成对抗网络的生成器模块输入向量z_{pri}，获得输出图像img_{pri}；

步骤2.2.3，将步骤2.2.2所述图像img_{pri}输入目标受害者模型模块，获得输出向量；

步骤2.2.4，构建损失函数；所述损失函数计算过程如下：

$$Z^{*}=arg\,min\,L_{discri}+\lambda L_{iden}\ ,\ L_{discri}=-log\left(D\left(G(z_{pri})\right)\right),\ L_{iden}=L_{CE}\left(F_{C}\left(G(z_{pri})\right),l_{t}\right)$$

其中Z*表示优化反演隐私样本优化过程中学习后的z_{pri}向量，L_{discri}表示生成对抗网络鉴别器损失，L_{iden}表示可识别性损失，D表示生成对抗网络中的鉴别器模块，F_{C}表示分类器模块，G表示生成对抗网络中的生成器模块，λ为L_{iden}的参数，L_{CE}表示交叉熵函数；

步骤2.2.5，将步骤2.2.4所述的损失向后传递，更新生成对抗网络的输入z_{pri}；

步骤2.2.6，循环执行所述步骤2.2.1到步骤2.2.5，此循环过程将在到达预先设置的迭代次数后终止，将最终的z_{pri}输入生成对抗网络的生成器模块，输出即为反演隐私样本。本实施例中迭代100轮，学习率10的-3次方。

[0034] 在本发明的防御下，攻击者仅能推测出公开数据集中l_{p}的特征，而不能获得隐私数据集中受保护类别l_{t}的样本特征。

[0035] 步骤3，根据虚假样本在保证目标受害者模型的有效性的基础上，微调目标受害者模型参数，从而实现防御模型反演攻击的目的。

[0036] 为了防止反演出的隐私数据集中受保护类别l_{t}，本发明微调目标受害者模型的参数，从而最大化了反演隐私样本的损失$\mathcal{L}_{pri}=L_{CE}\left(\theta^{*},F_{V}\left(G.generator(z_{pri})\right),l_{t}\right)$。为了使得攻击者可以反演出公开数据集的干扰类别$l_{p}$，本发明通过微调目标受害者模型的参数，最小化了反演公开样本的损失

$\mathcal{L}_{pub} = L_{CE}\left(\theta^*, F_V\left(G.generator\left(z_{pub}\right)\right), l_t\right)$。这个过程相当于向隐私数据集受保护类别$l_t$中注入干扰样本，使得目标受害者模型在反演公开样本上达到过拟合的效果同时，为了在持续学习的过程中维持目标受害者模型的高准确率，本发明采用了可塑权重巩固算法，增加正则化项

$$\mathcal{L}_{CL} = \sum_i \frac{\lambda_3}{2} F_i(\theta_i - \theta_i^*)。$$

[0037] 步骤3的全部过程可表示为$\theta = arg\,min - \alpha\mathcal{L}_{pri} + \beta\mathcal{L}_{pub} + \omega\mathcal{L}_{CL}$，其中θ为目标受害者模型的参数集，α、β和ω分别是三个损失项的参数，本实施例中α=1，β=2，ω=5。

[0038] 图2为在不同原始训练集下，本发明与其他防御方法在受到不同种类模型反演攻击的对比结果。可以看到，从整体来看，本方法在各种的模型反演攻击下都表现更好。相对于其他防御，能够更有效地保护隐私数据安全。

[0039] 由此可见，本发明通过利用生成对抗网络生成假样本，可以安全高效地防御模型反演攻击。

[0040] 基于同一发明构思，本发明还设计了一种电子设备，包括：

一个或多个处理器；

存储装置，用于存储一个或多个程序；

当一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现所述深度神经网络模型反演攻击防御方法所执行的操作。

[0041] 基于同一发明构思，本发明还设计了一种计算机可读介质，其上存储有计算机程序，所述程序被处理器执行时实现所述深度神经网络模型反演攻击防御方法所执行的操作。

[0042] 本文中所描述的具体实施例仅仅是对本发明作举例说明。本发明所属技术领域的技术人员可以对所描述的具体实施例做各种各样的修改或补充或采用类似的方式替代，但并不会偏离本发明的精神或者超越所附权利要求书所定。

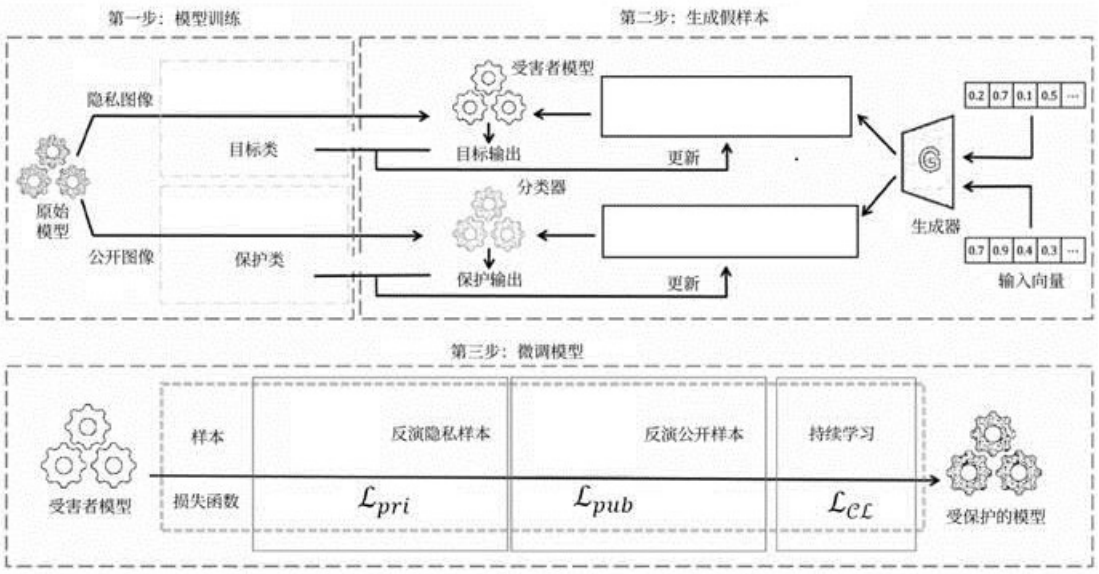

第一步：模型训练
第二步：生成假样本
隐私图像
目标类
原始模型
公开图像
保护类
受害者模型
目标输出
保护输出
分类器
更新
更新
生成器
0.2 0.7 0.1 0.5 ...
0.7 0.9 0.4 0.3 ...
输入向量
第三步：微调模型
受害者模型
样本
损失函数
反演隐私样本
$\mathcal{L}_{pri}$
反演公开样本
$\mathcal{L}_{pub}$
持续学习
$\mathcal{L}_{CL}$
受保护的模型

图1

		Original			DP			Ad-mi			NetGuard		
		MC(l_t)	AA	FID	MC(all)	MC(l_t)	FID	MC(all)	MC(l_t)	FID	MC(all)	MC(l_t)	FID
CelebA	DMI [8]	87.46%	52%	108	40%	56.82%	50%	101	**0%**	87.46%	153	**0%**	86.54%	**100%**	**159**
	Mirror [4]	87.46%	69%	98	45%	56.82%	50%	103	**0%**	87.46%	170	**0%**	85.46%	**100%**	**174**
	Privacy [12]	87.46%	48%	301	**0%**	56.82%	50%	339	1%	87.46%	329	**0%**	85.33%	**100%**	**351**
	Alignment [36]	87.46%	3%	124	3%	56.82%	50%	151	**0%**	87.46%	166	**0%**	85.62%	**100%**	**171**
VGG-Face	DMI [8]	82.41%	21%	162	12%	70.61%	**100%**	154	2%	82.41%	226	**0%**	81.14%	**100%**	**216**
	Mirror [4]	82.41%	49%	163	44%	70.61%	**100%**	173	**0%**	82.41%	219	**0%**	80.01%	**100%**	**273**
	Privacy [12]	82.41%	**0%**	347	**0%**	70.61%	**100%**	320	**0%**	82.41%	399	**0%**	80.75%	**100%**	**401**
	Alignment [36]	82.41%	17%	212	**0%**	70.61%	**100%**	194	**0%**	82.41%	317	**0%**	81.54%	**100%**	**391**
VGG-Face2	DMI [8]	92.03%	27%	142	19%	87.83%	**100%**	144	**0%**	92.03%	211	**0%**	90.77%	**100%**	**261**
	Mirror [4]	92.03%	34%	78	30%	87.83%	**100%**	101	**0%**	92.03%	178	**0%**	88.35%	**100%**	**199**
	Privacy [12]	92.03%	**0%**	410	**0%**	87.83%	**100%**	338	**0%**	92.03%	396	**0%**	91.39%	**100%**	**401**
	Alignment [36]	92.03%	14%	186	14%	87.83%	**100%**	213	**0%**	92.03%	244	**0%**	87.94%	**100%**	**265**

图2