文章编号:1671-5896(2024)01-0059-08

隐私保护的图像替代数据生成方法

李婉莹[a,b], 刘学艳[a,b], 杨 博[a,b]
(吉林大学 a. 计算机科学与技术学院; b. 符号计算与知识工程教育部重点实验室, 长春 130012)

摘要: 针对现有图像数据集存在的隐私保护需求, 提出一种图像数据集隐私保护场景及该场景下隐私保护的图像替代数据生成方法。该场景利用经隐私保护方法处理后的替代图像数据集取代原始图像数据集, 其中替代图像与原始图像一一对应, 人类无法识别替代图像所属类别, 替代图像可训练现有的深度学习图像分类算法, 且具有较好的分类效果。同时针对上述场景, 改进了基于投影梯度下降(PGD: Project Gradient Descent)攻击的数据隐私保护方法, 将原始PGD攻击目标由标签改为图像, 即图像对图像的攻击, 并使用经过对抗训练的鲁棒模型进行图像对图像攻击作为替代数据的生成方法。在标准测试集上, 替代后的CIFAR(Canadian Institute For Advanced Research 10)数据集和CINIC数据集在图像分类任务上分别取得了87.15%和74.04%的测试正确率。实验结果表明, 该方法能在保证替代数据集对人类隐私性的前提下, 生成原始数据集的替代数据集, 并保证现有方法在该数据集上的分类性能。

关键词: 深度学习; 隐私保护; 计算机视觉; 对抗攻击; 对抗样本
中图分类号: TP391 **文献标志码**: A

Alternative Data Generation Method of Privacy-Preserving Image

LI Wanying[a,b], LIU Xueyan[a,b], YANG Bo[a,b]
(a. College of Computer Science and Technology; b. Key Laboratory of Symbolic Computing and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China)

Abstract: Aiming at the privacy protection requirements of existing image datasets, a privacy-preserving scenario of image datasets and a privacy-preserving image alternative data generation method is proposed. The scenario is to replace the original image dataset with an alternative image dataset processed by a privacy-preserving method, where the substitute image is in one-to-one correspondence with the original image. And humans can not identify the category of the substitute image, the substitute image can be used to train existing deep learning images classification algorithm, having a good classification effect. For this scenario, the data privacy protection method based on the PGD (Project Gradient Descent) attack is improved, and the attack target of the original PGD attack is changed from the label to the image, that is the image-to-image attack. A robust model for image-to-image attacks as a method for generating alternative data. On the standard testset, the replaced CIFAR(Canadian Institute For Advanced Research 10) dataset and CINIC dataset achieved 87.15% and 74.04% test accuracy on the image classification task. Experimental results show that the method is able to generate an alternative dataset to the original dataset while guaranteeing the privacy of the alternative dataset to humans, and guarantees the classification performance of existing methods on this dataset.

收稿日期: 2022-12-25
基金项目: 国家自然科学基金资助项目(U22A2098; 62172185; 62202200; 62206105); 国家重点研发计划基金资助项目(2021ZD0112501; 2021ZD0112502); 吉林省重点科技研发基金资助项目(20180201067GX; 20180201044GX); 吉林省自然科学基金资助项目(20200201036JC)
作者简介: 李婉莹(1999—), 女, 黑龙江牡丹江人, 吉林大学硕士研究生, 主要从事深度学习和对抗攻击研究, (Tel)86-18245371367(E-mail)846353603@qq.com; **通讯作者**: 杨博(1974—), 男, 河南新乡人, 吉林大学教授, 博士, 主要从事复杂系统学习和神经符号系统研究, (Tel)86-431-85166892(E-mail)ybo@jlu.edu.cn。

Key words: deep learning; privacy protection; computer vision; adversarial attack; adversarial example

0 引 言

数据集在机器学习的研究发展中起着基础性作用。数据集构成了设计和部署模型的基础，也是进行基准测试和评估的主要媒介。现代机器学习的发展基于大规模和多样化的数据集研究。例如，早期计算机视觉算法只能利用实验室条件下收集的小规模数据集进行相应的训练与测试，因此，不能在现实世界中应用[1]。当人们采用 PASCAL VOC(The PASCAL Visual Object Classes)[2]，MS COCO(Common Objects in COntex)和 ImageNet[3]等数据集时，才使其得到应用。由于这些数据集来源于现实世界，可提供更好的自然图像表示，而且比早期数据集规模更大，有利于测试算法在更接近现实场景下的性能，为算法推广提供了保障。此外，这些数据集还具有基本的科学功能：它们允许对算法进行可重复和定量的比较，并使人们能有效地在彼此工作基础上开展研究。

然而，随着数据集规模的不断增长，越来越多的真实数据集被应用于计算机视觉算法研究，数据的隐私保护问题亟待解决。根据数据集中信息的敏感程度，笔者将数据集分为部分和全隐私数据集。部分隐私图像数据可能包含部分敏感信息，如一张图像中的人脸、地址或车牌等信息为敏感信息，而图像中的环境等为非敏感信息。全隐私图像数据包含的信息均为隐私数据，如医疗和军事数据集等。

针对不同类型的数据集，人们提出了不同的隐私保护方法。目前，部分隐私数据集的隐私保护方法主要为信息模糊技术，即将隐私信息模糊或删除。例如，Yang 等[4]发布了隐私版本的 ImageNet 数据集，主要方法为对人脸等敏感信息定位并手动模糊数据集中定位的敏感信息，该方法将信息模糊后的数据集定义为隐私版本数据集。Asano 等[5]采用人工的方式将人脸与其余隐私信息裁剪，最后得到了隐私版本的 ImageNet 数据集。上述方法在处理部分隐私数据集有较好的效果，然而，全隐私数据集中的信息均为隐私数据，无法采用信息模糊的方法。对全隐私数据集，目前存在的隐私保护方法有密钥加密[6]，基于生成对抗网络(GAN：Generative Adversarial Network)[7-10]生成现实世界中不存在的数据集等。例如，Fernando 等[6]使用密钥将原始数据集转化为受保护的数据集。未授权用户在受保护数据集上训练的模型在测试数据上的性能非常差。相反，授权用户可利用密钥将受保护的数据集恢复为原始数据集，并在恢复的数据集上训练模型，模型在测试数据上具有正常的性能。该方法发布的受保护数据集与原数据集虽然相似，但无法训练图像分类模型。基于 GAN 生成的图像组成的合成数据集主要用于人脸检测任务。Colbois 等[8]通过实验证明了最新的 GAN 可生成高分辨率和逼真的人脸图像。但基于 GAN 的方法存在如下问题：首先，目前基于 GAN 方法生成的图像无法取代真实图像；其次，GAN 方法为人们提供了实现深度伪造(deep-fake)的有效途径，这造成了人脸隐私相关的问题，如生成虚假的视频和音频[11]，篡改选举信息[12]等。

尽管上述工作在数据集隐私保护方面均取得了一些进展，但全隐私数据集的隐私保护任务仍具有挑战性。密钥加密算法及基于 GAN 的生成方法虽然能保护原始数据集，但这两种方法发布的数据集与原始数据集仍存在较大的相似性，暴露了过多的原始图像的类别信息，并且不能很好地保证全隐私数据集的隐私性。

基于以上分析，笔者提出如下场景：利用经隐私保护方法处理后的替代图像数据集取代目标图像数据集，其中替代图像具有以下特征：1）替代图像与目标图像一一对应；2）人类无法识别替代图像所属类别；3）替代图像可训练现有的深度学习图像分类算法，且具有较好的分类效果。针对这种场景，以上的数据集隐私保护方法均不适用。

针对该场景，笔者提出了隐私保护的图像替代数据生成方法。该方法利用图像对图像的对抗攻击算法对数据集进行转化。在攻击过程中，该方法将原始数据集中的原始图像设定为攻击目标，将随机选择的图像作为随机图像开始攻击，生成人类无法识别的替代版本图像训练集。该训练集能较好地训练现有的基于深度学习的图像分类算法，在替代版本训练集中学习到分类信息，即保证在不经过转化的标准测试集上的分类正确率。因此，该方法可发布隐私版本数据集而不影响基于深度神经网络的图像分类算法，从而实现对数据集的隐私保护。本研究的主要工作如下。

1）为了不暴露过多类别信息，笔者设计了隐私保护的图像替代数据生成方法。该方法将原始数据集中的原始图像设定为攻击目标，将从原始数据集或外部数据集中随机选择的图像做为随机图像开始攻击，使生成的替代图像在视觉角度既远离随机图像又远离原始图像，保证了生成的替代图像人类无法识别，即保证替代版本数据集的隐私性。

2）为保证替代版本数据集在分类模型上可替代原始数据集，笔者设计的攻击方法将攻击目标设置为原始图像。该设置可保证生成的替代图像在表示层参数上接近原始图像，从而保证分类模型可从替代版本数据集中学到与原始数据集相似的分类信息，即保证替代版本数据集对原始数据集的替代性。

3）针对笔者提出的数据集隐私保护场景及隐私保护方法，在 CIFAR 数据集[13]及 CINIC 数据集[14]上进行了相应的实验。实验结果表明，笔者提出的方法在保证替代数据集使人类无法识别的前提下，能保证替代版本数据集的替代性。

1　模型构建

1.1　算法框架

为生成替代图像，笔者对现有的对抗攻击方法[15]在图像分类任务上进行了相关实验。实验发现，常见的对抗攻击算法，如 C&W（Carlini & Wagner）攻击[16]，投影梯度下降（PGD：Project Gradient Descent）[17]等，通常将图像向某一类别进行攻击，使原图发生较小的改变，从而使人眼无法察觉改变，但分类算法却将图像划分到错误类别。这与本研究想要达成的目的相反。笔者希望能达到如下效果，即图像发生较大改变使人类无法识别其所属类别，分类算法依旧能学习到类别相关信息。

为此，笔者改进了 PGD[17]攻击方法，设计了隐私保护的替代图像生成算法，该算法整体框架如图 1 所示。具体过程如下。

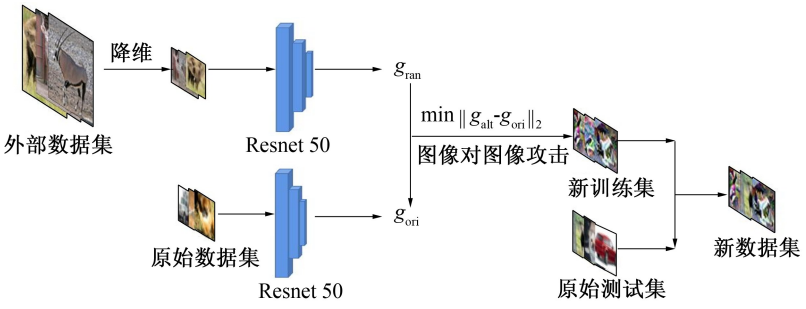


图 1　算法框架

Fig. 1　Algorithm framework

1）从内部数据集（原始数据集）或外部数据集（非原始数据集）中选择随机图像 ran（random）并将随机图像 ran 修改为与原始图像 ori（original）同维度。

2）分别将随机图像 ran 和原始图像 ori 输入到深度神经网络分别得到 ran 和 ori 的表示层参数 g_{ran} 和 g_{ori}。

3）将 g_{ran} 作为攻击的输入，g_{ori} 作为攻击目标，采用

$$g_{ran}^{t+1} = \prod_{g_{ran}+S}(g_{ran}^{t} + \alpha \mathrm{sgn}(\nabla_{g_{ran}} L(\theta, g_{ran}, g_{ori}))), \tag{1}$$

以多步迭代的方式进行攻击，迭代过程为，第 t 步的输入为 g_{ran}^{t}，为其添加扰动，扰动值限制在球面范围 S 内，然后将结果进行投影操作，得到第 t 步的输出 g_{ran}^{t+1}；第 $t+1$ 步的输入 g_{ran}^{t+1}，为其添加扰动并进行投影操作。以此类推，完成对原始图像的攻击。其中 sgn（·）为符号函数，自变量大于等于 0 时取 1，小于 0 时取 −1。g_{ran}^{t} 为输入数据，g_{ran}^{t} 为第 t 步的输入，g_{ran}^{t+1} 为第 $t+1$ 步的输入，g_{ori} 为攻击目标，θ 为模型参数，L 为损失函数，∇_x 为取偏导数。$\prod$ 为投影操作，S 为球面限制，该投影操作是将扰动值限制在球面范围内，g_{ran}^{t} 加上扰动值就是 g_{ran}^{t+1}。α 为超参数，用于控制 g_{ran}^{t} 与 g_{ran}^{t+1} 之间的差异程度。

4）取最后一步（记为第 T 步）的输出 g_{ran}^{T} 作为图像对图像攻击的最终输出，即替代图像

alt(alternative)的表示层参数 g_{alt}。注意，与经典 PGD[17] 攻击方法相比，笔者的攻击方法在攻击过程中无标签信息的参与。最后，该方法的优化目标为

$$\min_{alt} \| g_{alt} - g_{ori} \|_2 , \tag{2}$$

其中 $\| g_{alt} - g_{ori} \|_2$ 为 alt 图像与 ori 图像表示层参数之间的 l_2 距离。

5）攻击完成后，利用攻击生成的替代图像的表示层参数得到替代图像 alt，并将 alt 的标签写为原始图像 ori 的标签。此时，图像 alt 为图像 ori 的替代版本，因为其在表示层角度与 ori 高度接近，且保留了 ori 中基于深度神经网络的图像分类算法需要的大部分分类特征。

1.2 对抗训练

为提高攻击成功率，笔者使用预训练模型进行图像对图像的攻击。常见的有标准和经对抗训练的鲁棒模型 2 种预训练模型，笔者使用经对抗训练的鲁棒模型进行图像对图像的攻击。这是因为，相比标准学习，鲁棒表示具有很多优点[18]，如：具有近似鲁棒特征的图像在语义上相似，鲁棒特征更易于被人类理解等。通过实验发现，使用标准模型进行图像对图像的攻击生成的替代图像无法远离随机图像，因此，不能在视觉上达到人类无法识别所属类别的要求。为此，笔者采用了 Mahloujifar 等[19]提出的对抗训练方法。对抗训练是增强神经网络鲁棒性的重要方法，在对抗训练过程中，训练样本会被混入一些对抗样本，使神经网络适应这种改变，从而具有鲁棒性。笔者使用经过对抗训练的鲁棒模型进行图像对图像的攻击(如图 1 所示)，使生成的替代图像从表示层及视觉角度均偏离随机图像，在一定迭代次数内，替代图像在表示层角度十分接近原始图像，但在视觉角度既偏离随机图像又偏离原始图像。

2 实验与结果分析

为评估模型的有效性，在 CIFAR 和 CINIC 2 个经典数据集上，针对隐私保护的替代数据集任务，对其进行验证。

2.1 实验准备

2.1.1 系统配置与训练参数

系统配置如下：操作系统为 Ubuntu 16.04；中央处理器(CPU：Central Processing Unit)为 Intel(R) Xeon(R)CPU E5-2643v4 @3.4 GHz；图形处理器(GPU：Graphics Processing Unit)为 2080Ti。

笔者提出的基于隐私保护的图像替代数据生成方法，使用 resnet50[20] 网络作为分类模型，对抗性训练超参 α 为 0.25，学习率为 0.1，攻击步长为 1，攻击步数为 200。生成替代图像时，使用迭代器分批次选取训练集图像，每批次向下一批次攻击，最后一批次向第 1 批次攻击，不限制类别。采用此种方式保证了替代版本数据集的完整性以及替代版本数据集与原始数据集图像一一对应。

文中所有数据集训练分类模型时均采用 autoaugment[21] 作为数据增强策略。

2.1.2 数据准备

笔者选用的原始数据集为图像分类任务经典数据集，分别为 CIFAR 数据集[13] 和 CINIC 数据集[14]。CIFAR 是一个用于识别物体的小型数据集，包含 10 个类别的彩色图像。图像的尺寸为32×32像素，数据集中有 50 000 张训练图像和 10 000 张测试图像。CINIC[14] 数据集由 CIFAR 和 ImageNet 的图像组成，CINIC 中有 270 000 张图像，其中包括 3 个子集：训练集、验证集和测试集。除了这 2 个原始数据集，还引入了一个外部数据集 STL-10[22]。该数据集包括 10 个类别，每个类别包含 500 张训练图像，800 张测试图像。此外，其中还包含 10 万张无标签图像用于无监督学习。

笔者采取来自内部和外部数据集 2 种随机图像选取方法。来自内部数据集的选取方法为随机图像取自与原始图像相同的数据集。以 CIFAR[13] 数据集为例，随机图像和原始图像均为 CIFAR 数据集中随机选取的图像。这种选取方式不需要处理外部数据集，实现方便，但存在原始数据集信息暴露过多的问题。来自外部数据集的选取方法为随机图像取自与原始图像不同的数据集。以 CIFAR 数据集为例，外部数据集为除 CIFAR 数据集外的任一数据集。使用外部数据集可在视觉角度有效地淡化原始数据集

的特征，最大限度的降低替代版本数据集对原始数据集的信息暴露，以达到更好的隐私保护效果。笔者使用的外部数据集为 STL-10。

为保证生成的替代图像之间不相似，需要每次选取的随机图像不同。为此，使用 STL-10 中的无标签图像数据集中的前 50 000 张图像，将图像降维为与 CIFAR 同维度，然后分批次向 CIFAR 中的 50 000 张训练集图像发起图像对图像的攻击，生成图像标签写为 CIFAR 图像的标签。

2 个原始数据集 CIFAR 和 CINIC 分别生成了来自内部替代版本数据集(CIFAR-to-CIFAR，CINIC-to-CINIC)以及来自外部版本替代数据集(STL-to-CIFAR，STL-to-CINIC)，两者测试集相同，均为标准测试集。为证明替代数据集对原始数据集的替代性，笔者引入原始数据集作为对比数据集。

2.2　评价标准

生成的替代数据主要用于图像分类任务，其是从给定的类别集合中为图像分配对应标签的任务。该研究中每个样本都只有一个正确的类别，因此，采用图像分类指标 top-1accuracy，即 top-1 正确率，计算方法如下：

$$A_{\text{top-1}} = \sum_{i=0}^{N-1} (f(x_i) == y_i) / N,$$

其中 x_i 为样本，其真实类别为 y_i，类别种类为$(0,1,\cdots,N)$，预测类别函数为f，N 为总样本数。

2.3　替代数据集方法性能测试结果

笔者采用对抗训练方法[18]，对原始数据集进行对抗训练用于训练鲁棒模型，使用鲁棒模型对 CIFAR 和 CINIC 训练集的图像进行图像对图像的攻击，随机图像分别选取内部数据集(CIFAR 数据集/CINIC 数据集)及外部数据集(STL10 数据集)，生成了来自内部替代版本数据集(CIFAR-to-CIFAR，CINIC-to-CINIC)以及来自外部版本替代数据集(STL-to-CIFAR，STL-to-CINIC)，两者测试集相同，均为标准测试集。

在相同的分类模型上评估替代版本和原始数据集性能的结果如表 1 所示。原始数据集训练集为原始 CIFAR/CINIC 训练集，其余训练集为本方法生成的替代数据集。所有测试集均为原始 CIFAR/CINIC 测试集。

表 1　替代数据集测试正确率

Tab. 1　Alternative dataset test accuracy　%

	原始数据集	来自内部	来自外部
CIFAR	94.25	87.15	86.21
CINIC	86.84	74.04	73.67

由表 1 可知，替代数据集训练的神经网络与原始训练集训练的网络在标准测试集上的正确率相比稍有下降，这是由于在生成替代数据集时发生了信息丢失。笔者认为，为保证替代数据集的隐私性，即保证人类无法学到相关信息，该下降必然发生。通过实验发现，来自内部替代数据集与来自外部替代数据集训练的神经网络在标准测试集上的正确率相差不大，因此可认为图像对图像的攻击中随机图像的选取对生成图像在神经网络训练的性能影响很小。

2.4　数据增强结果展示

autoaugment[21]是一种自动搜索合适数据增强策略的方法，该方法创建一个数据增强策略的搜索空间，利用搜索算法选取适合特定数据集的数据增强策略。cutout 是对训练数据进行随机遮挡，该方法激励神经网络在决策时能更多地考虑次要特征，而不是依赖很少的主要特征。由于替代数据集减少了原始数据集的部分信息，并且引入了外部数据集的部分信息，因此选择适合替代数据集的数据增强十分有必要。

表 2 给出了来自 2 种来自内部版本替代数据集：CIFAR-to-CIFAR，CINIC-to-CINIC 分别在无数据增强，A+C(autoaugment+cutout)，A(autoaugment)3 种情况下的性能。

从表 2 可知，使用 autoaugment 作为数据增强策略在两个替代数据集上均有较大提升。这可能是由于替代数据集减少了原始数据集的部分信息，并且引入了外部数据集的部分信息，因此 autoaugment 可自动搜索并选择合适数据增强策略，使测试正确率有较大提升。

表2 数据增强对测试正确率的影响

Tab.2 Impact of data augmentation test accuracy %

数据集	无数据增强	A+C(autoaugment+cutout)	A(autoaugment)
CIFAR-to-CIFAR	83.27	86.74	87.15
CIFAR-to-CINIC	69.23	73.03	74.04

表2中，A+C(autoaugment+cutout)的测试正确率相对A(autoaugment)稍有下降，这可能是由于在生成替代数据集的过程中模糊了主要特征与次要特征，使cutout的效果不明显。

2.5 生成鲁棒模型的对抗性训练参数敏感度分析

以CIFAR-to-CIFAR，CINIC-to-CINIC数据集为例，研究各参数训练鲁棒模型对替代数据集在标准测试集泛化的影响。

表3给出了2种数据集训练鲁棒模型时，式(1)中α的值对生成的2种来自内部版本替代数据集CIFAR-to-CIFAR，CINIC-to-CINIC测试正确率的影响。

表3 不同α生成的替代训练集训练的模型在测试集取得的正确率对比

Tab.3 Comparison of the test accuracy of the models generated by different α to train the alternative trainset on the testset %

数据集	$\alpha=0.25$	$\alpha=0.5$	$\alpha=1$
CIFAR-to-CIFAR	86.74	86.45	80.94
CINIC-to-CINIC	74.03	73.64	70.80

从表3可看出，选取不同α值对替代数据集测试正确率的影响不大。其中$\alpha=0.25$时，使用替代数据集训练的神经网络在标准测试集的正确率最高，因此在生成鲁棒模型时，统一选择$\alpha=0.25$。

2.6 图像对图像攻击参数敏感度分析

为研究PGD攻击中攻击半径R_{eps}(式(1)中S球面限制的半径)对替代数据集在标准测试集泛化的影响，笔者针对CIFAR-to-CIFAR,CINIC-to-CINIC数据集及各攻击半径(R_{eps})的值进行了相关实验。

表4给出了生成替代数据集时，攻击半径(R_{eps})的值对2种来自内部版本替代数据集CIFAR-to-CIFAR，CINIC-to-CINIC测试正确率的影响。

表4 不同eps生成的替代训练集训练的模型在测试集取得的正确率对比

Tab.4 Comparison of the test accuracy of the models generated by differenteps to train the alternative trainset on the testsett %

数据集	eps=500	eps=1 000	eps=5 000
CIFAR-to-CIFAR	86.75	86.76	86.21
CINIC-to-CINIC	73.06	73.67	71.24

从表4可看出，选取不同R_{eps}值对替代数据集测试正确率的影响不大。其中R_{eps}值为1 000时，使用替代数据集训练的神经网络在标准测试集的正确率最高，因此在进行图像对图像攻击时，统一选择攻击半径R_{eps}为1 000。

2.7 对抗训练中鲁棒模型与标准模型对比

为说明笔者采用的经过对抗训练的鲁棒模型相对标准模型生成的替代图像的隐私性，进行了相关的实验，结果如图2所示。

从图2可看出，使用标准模型进行图像对图像的攻击生成的替代图像无法在视觉上达到保护隐私信息的要求，而使用经过对抗训练的鲁棒模型生成的替代图像从视觉角度既偏离随机图像又偏离原始图像，可以更好地保护图像隐私。

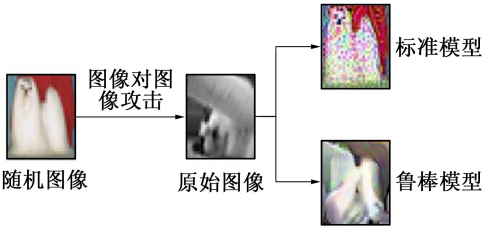


图2 图像对图像的攻击生成图像

Fig.2 Image-to-image attack generates images

2.8　可视化

为进一步验证隐私保护的图像替代数据生成方法的隐私性，对笔者创建的替代数据集进行可视化分析。图 3 分别给出了生成的 CIFAR-to-CIFAR，STL-to-CIFAR，CINIC-to-CINIC，STL-to-CINIC 数据集。从图 3可看出，生成的替代数据集可以在视觉角度远离原始数据集，即可以保证替代数据集的隐私性。



a CIFAR-to-CIFAR数据集　　b CINIC-to-CINIC数据集

c STL-to-CIFAR数据集　　d STL-to-CINIC数据集

图 3　数据集展示

Fig. 3　Dataset display

3　结　语

笔者提出一种图像数据集隐私保护场景及该场景下隐私保护的图像替代数据生成方法，并且使用经对抗训练的鲁棒模型对原始数据集进行图像对图像的攻击实现了该目标。研究了随机图像选取，数据增强，对抗训练参数及 PGD 攻击参数对替代版本数据集训练的分类模型在标准测试集上测试正确率的影响。结果表明，本研究可在该场景下实现较好的隐私保护。笔者认为，对数据集进行替代数据生成并且保证替代版本数据集能正常训练分类模型是一个十分重要的隐私保护场景，如果能实现替代版本数据集的广泛使用，将在不影响研究工作的前提下极大的保护数据集的隐私。

参考文献：

[1] PAULLADA A，RAJI I D，BENDER E M，et al. Data and Its (Dis) Contents：A Survey of Dataset Development and Use in Machine Learning Research [J]. Patterns，2021，2(11)：12-25.

[2] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [C]//Proceedings of CVPR-14. Columbus，USA：IEEE，2014：580-587.

[3] DENG J，DONG W，SOCHER R，et al. Imagenet：A Large-Scale Hierarchical Image Database [C]//Proceedings of CVPR-09. Miami：IEEE，2009：248-255.

[4] YANG K，YAU J H，LI F F，et al. A Study of Face Obfuscation in Imagenet [C]//Proceedings of ICML-22. Baltimore：PMLR，2022：25313-25330.

[5] ASANO Y，RUPPRECHT C，ZISSERMAN A，et al. PASS：An Image Net Replacement for Self-Supervised Pretraining without Humans [EB/OL]. (2021-09-27)[2022-11-07]. https://arxiv.org/pdf/2109.13228.pdf.

[6] FERNANDO P，WEI-KOCSIS J. A Novel Data Encryption Method Inspired by Adversarial Attacks [EB/OL]. (2021-09-14)[2022-11-07]. https://arxiv.org/pdf/2109.06634.pdf.

[7] GOODFELLOW I，POUGET-ABADIE J，MIRZA M，et al. Generative Adversarial Networks [J]. Communications of the

ACM, 2020, 63(11): 139-144.
[8]COLBOIS L, DE FREITAS PEREIRA T, MARCEL S. On the Use of Automatically Generated Synthetic Image Datasets for Benchmarking Face Recognition [C]//Proceedings of IJCB-21. Shenzhen: IEEE, 2021: 1-8.
[9]BROCK A, DONAHUE J, SIMONYAN K. Large Scale GAN Training for High Fidelity Natural Image Synthesis [EB/OL]. (2021-02-25)[2022-11-07]. https://arxiv.org/pdf/1809.11096.pdf.
[10]ODENA A, OLAH C, SHLENS J. Conditional Image Synthesis with Auxiliary Classifier Gans [C]//Proceedings of ICML-17. Sydney: PMLR, 2017: 2642-2651.
[11]LYU S. Deepfake Detection: Current Challenges and Next Steps [C]//2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). London: IEEE, 2020: 1-6.
[12]WESTERLUND M. The Emergence of Deepfake Technology: A Review [J]. Technology Innovation Management Review, 2019, 9(11): 40-53.
[13]KRIZHEVSKY A, HINTON G. Learning Multiple Layers of Features from Tiny Images [EB/OL]. (2009-04-08)[2022-11-07]. https://www.researchgate.net/publication/306218037_Learning_multiple_layers_of_features_from_tiny_images.
[14]DARLOW L N, CROWLEY E J, ANTONIOU A, et al. CINIC-10 is Not Imagenet or CIFAR-10 [EB/OL]. (2018-10-02)[2022-11-07]. https://arxiv.org/pdf/1810.03505.pdf.
[15]SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing Properties of Neural Networks [EB/OL]. (2013-12-21)[2022-11-07]. https://arxiv.org/pdf/1312.6199.pdf.
[16]CARLINI N, WAGNER D. Towards Evaluating the Robustness of Neural Networks [C]//Proceedings of SP-17. San Jose: IEEE, 2017: 39-57.
[17]MADRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks [EB/OL]. (2017-06-19)[2022-11-07]. https://arxiv.org/pdf/1706.06083.pdf.
[18]ENGSTROM L, ILYAS A, SANTURKAR S, et al. Adversarial Robustness as a Prior for Learned Representations [EB/OL]. (2019-06-03)[2022-11-07]. https://arxiv.org/pdf/1906.00945.pdf.
[19]MAHLOUJIFAR S, DIOCHNOS D I, MAHMOODY M. The Curse of Concentration in Robust Learning: Evasion and Poisoning Attacks from Concentration of Measure [C]//Proceedings of the AAAI-19. Palo Alto: [s.n.], 2019: 4536-4543.
[20]HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition [C]//Proceedings of CVPR-16. Las Vegas: IEEE, 2016: 770-778.
[21]CUBUK E D, ZOPH B, MANE D, et al. Autoaugment: Learning Augmentation Policies from Data [EB/OL]. (2018-05-24)[2022-11-07]. https://arxiv.org/pdf/1805.09501.pdf.
[22]COATES A, NG A, LEE H. An Analysis of Single-Layer Networks in Unsupervised Feature Learning [C]//Proceedings of AISTATS-11. Lauderdale: JMLR, 2011: 215-223.

(责任编辑: 张洁)