面向正常拟合迁移学习模型的成员推理攻击

陈晋音[1,2]，上官文昌[2]，张京京[3]，郑海斌[2]，郑雅羽[2]，张旭鸿[4]

（1. 浙江工业大学网络空间安全研究院，浙江 杭州 310012；2. 浙江工业大学信息工程学院，浙江 杭州 310012；
3. 军事科学院系统工程研究院信息系统安全技术国防科技重点实验室，北京 100039；4. 浙江大学控制科学与工程学院，浙江 杭州 310007）

摘 要：针对现有成员推理攻击（MIA）在面向正常拟合迁移学习模型时性能较差的问题，对迁移学习模型在正常拟合情况下的 MIA 进行了系统的研究，设计异常样本检测获取容易受攻击的数据样本，实现对单个样本的成员推理攻击。最终，将提出的攻击方法在 4 种图像数据集上展开攻击验证，结果表明，所提 MIA 有较好的攻击性能。例如，从 VGG16（用 Caltech101 预训练）迁移的 Flowers102 分类器上，所提 MIA 实现了 83.15%的成员推理精确率，揭示了在迁移学习环境下，即使不访问教师模型，通过访问学生模型依然能实现对教师模型的 MIA。

关键词：成员推理攻击；深度学习；迁移学习；隐私风险；正常拟合模型

中图分类号：TN92

文献标识码：A

DOI: 10.11959/j.issn.1000−436x.2021209

Membership inference attacks against transfer learning for generalized model

CHEN Jinyin[1,2], SHANGGUAN Wenchang[2], ZHANG Jingjing[3],
ZHENG Haibin[2], ZHENG Yayu[2], ZHANG Xuhong[4]

1. Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310012, China
2. School of Information Engineering, Zhejiang University of Technology, Hangzhou 310012, China
3. National Key Laboratory of Science and Technology on Information System Security, Institute of System Engineering, Chinese Academy of Military Science, Beijing 100039, China
4. School of Control Science and Engineering, Zhejiang University, Hangzhou 310007, China

Abstract: For the problem of poor performance of exciting membership inference attack (MIA) when facing the transfer learning model that is generalized, the MIA for the transfer learning model that is generalized was first systematically studied, the anomaly detection was designed to obtain vulnerable data samples, and MIA was carried out against individual samples. Finally, the proposed method was tested on four image data sets, which shows that the proposed MIA has great attack performance. For example, on the Flowers102 classifier migrated from VGG16 (pretraining with Caltech101), the proposed MIA achieves 83.15% precision, which reveals that in the environment of transfer learning, even without access to the teacher model, the MIA for the teacher model can be achieved by visiting the student model.

Keywords: membership inference attack, deep learning, transfer learning, privacy risk, generalized model

1 引言

随着深度学习技术的飞速发展，深度学习模型已成功应用于多种任务，包括图像分类[1-4]、文本识别[5-6]、语音识别[7-8]、社交网络挖掘[9-10]、电磁信号处理[11-12]等，且均取得了令人满意的性能。

收稿日期：2021−07−20；**修回日期**：2021−09−27

基金项目：国家重点研发计划基金资助项目（No.2018AAA0100801）；国家自然科学基金资助项目（No.62072406）；浙江省自然科学基金资助项目（No.LY19F020025）；宁波市“科技创新 2025”重大专项基金资助项目（No.2018B10063）

Foundation Items: The National Key Research and Development Program of China (No.2018AAA0100801), The National Natural Science Foundation of China (No.62072406), The Natural Science Foundation of Zhejiang Province (No.LY19F020025), The Major Special Funding for “Science and Technology Innovation 2025” in Ningbo (No.2018B10063)

伴随应用任务的复杂化，以及对任务性能的需求日益增高，深度学习模型日趋复杂化[13-14]，通过本地独立完成这些复杂模型的训练需要大量的训练数据与计算资源的支持。例如，OpenAI 公司花费了将近 1.4 TB 的训练数据和 460 万美元来训练 GPT-3 模型[15]。通常情况下，个体研究人员和小公司负担不了这么多的资源。针对这一问题，近期研究提出了迁移学习，即通过在一个或多个源领域训练获得模型，总结有用的知识并将其应用于新的目标任务。常用的迁移学习方法之一是在已有的预训练模型（教师模型）的基础上进行微调训练，获得性能较好的学生模型。这种方式使个体研究人员不需要大量训练数据和训练资源也能获得性能良好的模型，提高模型的利用效率，降低训练成本。例如，应用于文本处理预训练模型 Transformer[16]，可以通过微调训练的方法应用于众多不同任务（如情感分类、文本识别等）中，且取得较好性能。

深度学习在现实商业中的应用日益广泛，其数据的误用和不充足的法律基础所导致的数据隐私问题频繁发生。例如，DeepMind 项目中存在滥用国家健康服务数据的问题。在众多深度学习技术的安全问题中，较严重的一个是数据隐私问题，即模型的恶意使用者通过成员推理攻击（MIA, membership inference attack）实现对模型训练数据的窃取。具体而言，成员推理攻击是指给定数据样本和模型的访问权限，判定该样本是否存在于模型的训练数据集中。至今为止，针对成员推理攻击的研究[17-28]已引起学术界的广泛关注。成员推理攻击根据攻击的方式可以分为 2 种类型：1)基于模型的成员推理攻击[21]，通过攻击者训练攻击模型，利用攻击模型判断待测样本是否为目标模型的成员样本；2)基于指标的成员推理攻击[20,23]，不需要训练攻击模型，通过计算预测向量的指标并与预设阈值进行比较来给出成员关系的推理。

通常假设成员推理攻击的攻击方具有目标模型的数据知识、训练知识和输出知识，获取数据知识表明攻击者已知训练数据的分布特征，训练知识意味着攻击者知道目标模型的训练方法，输出知识表示攻击者可以得到目标模型的输出。根据攻击者是否能够访问模型参数的模型知识，MIA 可分为黑盒推理攻击[20-21,23]和白盒推理攻击[24-25]。然而，上述工作都是在所有样本中不加选择地进行攻击，这种场景下的攻击成功率在所有目标样本上平均，而不考虑误判的代价。文献[26]首次研究了针对单个样本点的 MIA，从另一个角度清晰地阐明了隐私风险，但是该攻击需要获取目标模型的置信度信息，在目标模型只输出标签信息的情况下无法正常工作。

鉴于迁移学习的优势，即利用较少训练资源获得较高性能的深度学习模型，通过迁移学习的方式获取深度模型成为主流方式之一，随之而来的是迁移学习的隐私安全问题，例如，面向迁移学习模型的成员推理攻击。迁移学习中主要包含 2 种模型：教师模型和学生模型。与现有的针对单独模型的成员推理攻击不同，本文面向迁移学习的 MIA 根据攻击对象不同和攻击者的访问权限不同，提出了分别窃取教师模型和学生模型的数据隐私，判断目标样本是否为目标模型的训练数据。例如，当攻击者攻击教师模型且可以访问教师模型时，可以判断目标样本是否为教师模型的训练数据。文献[27]首次研究了面向迁移学习的成员推理攻击，该方法在模型处于过拟合状态下取得较好攻击效果，当模型处于正常拟合状态时，攻击性能有明显的下降。这一现象也普遍存在于针对非迁移学习的 MIA 方法中，包括 Salem[20]、Yeom[23]、Nasr[24]和 Lenio[25]等。为了表明这些方法只能在过拟合的深度学习模型上取得较好的攻击效果，而当模型是正常拟合的情况下攻击性能大幅下降。本文在 VGG16 模型、Caltech101 数据集上复现了上述攻击方法，攻击结果如图 1 所示。在过拟合与正常拟合情况下，各种 MIA 方法的攻击性能均存在明显下降。具体而言，当模型处于过拟合时，攻击有较高的精确率；当模型处于正常拟合时，攻击性能明显降低，而实际应用中的大部分深度学习模型都是正常训练且处于正常拟合的情况。针对这一问题，文献[26]提出了模型处于正常拟合环境下的成员推理攻击，然而该攻击需要获取置信度信息，在目标模型只输出标签的情况下则无法展开攻击，限制了其实际应用的可操作性。

综上所述，本文提出了针对迁移学习的深度学习模型在正常拟合情况下的成员推理攻击方法，通过搜索对目标模型预测产生特殊影响的异常样本，利用异常样本在目标模型的训练集中存在与否对预测结果产生较大差异，通过异常样本展开成员推理攻击，实现正常拟合模型的成员推理攻击。此外，针对现有成员推理攻击需要获取置信度才能实现攻击的问题，本文提出了一种只需要输出标签不需要置信度的更高效的 MIA 方法，采用置

信度分数表示样本与模型决策边界的距离，并使用对抗噪声进行衡量，从而实现置信度重构，通过对抗攻击和回归分析获取攻击样本所需对抗噪声的大小与样本在模型下的置信度关系，仅获取模型输出标签的情况下，实现与置信度攻击相当的攻击性能。

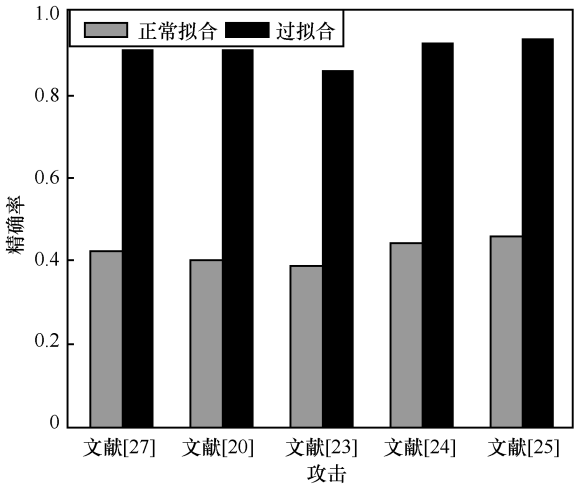


图 1　训练集为 Caltech101 的 VGG16 模型的攻击样本精确率

本文的主要创新点如下。

1) 研究了迁移学习的模型处于正常拟合时的成员推理攻击。设计了 3 种攻击模式，分别实现对教师模型和学生模型的成员推理攻击，提出了异常样本检测和置信度重构方法，实现了面向正常拟合迁移学习模型的成员推理攻击（TMIA, transfer membership inference attack）。

2) 在目标模型只输出标签的情况下，通过置信度重构，寻找对抗攻击成功时对抗噪声的大小与置信度的逻辑关系，实现了成员推理攻击，即在最小信息量的情况下，依然达到了与拥有置信度的 MIA 相当的攻击性能。

3) 展开对不同数据集的实验验证，证明了本文所提出的成员推理攻击在不同迁移学习方式下的攻击有效性，并与现有的多种 MIA 方法展开对比，本文方法在大部分情况下达到了最优攻击性能（SOTA, state-of-the-art）。另外，在特征提取器的迁移方式下，揭示了冻结层数对攻击性能的影响。

4) 为了进一步验证本文提出的 TMIA 攻击的有效性，假设实际应用中存在对 TMIA 的防御方法，对防御模型展开适应性攻击，实验结果表明，即使存在防御，本文的 TMIA 依然具有较高的攻击精确率。

2　相关工作

本节主要介绍面向机器学习模型的成员推理攻击方法，以及迁移学习安全性研究。

2.1　成员推理攻击

成员推理攻击主要是为了窃取模型的数据隐私，当数据包含大量敏感信息时，如生物医学数据[28-29]和移动跟踪数据[30]，将造成严重的隐私风险，因此，成员推理攻击引起广泛关注。

文献[21]首次提出了针对机器学习模型的成员推理攻击，利用影子模型模拟目标模型的行为，为攻击模型生成训练数据，通过攻击模型判定样本是否为成员样本。然而该攻击的前提是需要获取目标模型的结构和训练数据的分布，而实际应用中大部分情况下，目标模型的结构与训练数据分布获取异常困难，限制了其实际应用。因此，文献[20]提出目标模型结构和训练数据分布未知情况下的成员推理攻击，在训练攻击模型时不使用所有的置信度分数，只从中选取前 3 个最大的值进行训练。另外，文献[20]也提出了基于阈值的成员推理攻击，通过比较阈值和置信度分数的最大值进行成员推理，当置信度大于设定阈值，则判定为成员样本。文献[23]提出了 2 种成员推理攻击：第一种只利用标签信息，将样本的真实标签与预测的标签相比，如果相同则认为是成员样本；第二种攻击计算样本的交叉熵损失，并将计算出的损失与所有训练样本的平均损失相比，从而判断是否是成员样本。文献[24]评估了针对深度学习算法的白盒成员推理攻击，认为白盒场景是黑盒场景的拓展，不同于黑盒环境下只能使用模型最后一层的输出，白盒环境则可利用任意层的输出进行攻击，但攻击性能并没有明显的提升。随后，他们又提出了一种预测损失对模型训练参数求导的方法，利用得到的梯度信息进行白盒攻击，并表明该攻击的性能优于黑盒攻击。但这种攻击需要得到模型的训练数据，在实际应用中面临较大困难。文献[25]针对这一问题，提出了一种不需要模型训练数据的白盒成员推理攻击。

总结上述攻击的有效性保证是模型处于过拟合状态，而当模型处于正常拟合状态时，攻击性能会大大降低。

除了针对批量成员数据的推理攻击，文献[26]首次提出了针对单个样本点的成员推理攻击。该方

法只对部分样本点进行攻击，即使在模型处于正常拟合状态下，依然有较高的攻击准确率。然而，该方法需要获取模型输出的置信度信息，在模型输出标签的环境下无法正常工作。

综上，现有工作尚未对面向正常拟合迁移学习模型的成员推理攻击进行研究，且在目标模型只输出标签的情况下无法达到较好的攻击效果。

2.2 迁移学习安全性研究

面向深度学习的迁移学习方法在计算机视觉[31-34]、语音分析[35-38]和文本处理[39-40]等领域均取得了较好的性能。但已有研究表明，迁移学习存在安全隐患，包括对抗攻击[41]、中毒攻击[42]和成员推理攻击[27]。

文献[41]提出了一种针对迁移学习的对抗攻击。常用的对抗攻击[38-40]主要是优化图像，使其被预测为目标标签，与已有方法的不同之处是，文献[41]提出的方法的核心思想是优化图像来模仿目标图像的内部表现。文献[42]提出了一种针对词嵌入的数据中毒攻击，基于嵌入的自然语言处理任务遵循迁移学习模式，其中嵌入模型和下游模型分别被视为教师模型和学生模型。目标可以是使目标单词在单词中排名更高，也可以将目标单词与特定的单词集的距离进行移近或者移远。论文进行了大量的实验，表明对嵌入模型（教师模型）进行攻击可以严重影响多个下游模型（学生模型）。文献[27]利用影子模型模仿目标模型，通过影子模型的输出训练攻击模型，最后使用攻击模型判断样本是否为成员样本，首次研究了面向迁移学习的成员推理攻击，但是该攻击只能在目标模型处于过拟合状态时有较好的攻击效果。

综上，现有的面向迁移学习成员隐私的研究只在模型处于过拟合的状态下进行，所提方法无法在模型正常拟合时有较好的攻击效果。

3 方法

本节首先介绍了攻击模式和威胁模型，随后对提出的攻击方法展开详细描述。

3.1 攻击模式

与成员推理攻击[22-25]不同，迁移学习场景中包含教师模型和学生模型 2 种模型，微调和特征提取器 2 种迁移方式。微调是指不冻结教师模型，直接用学生数据集训练教师模型得到学生模型。特征提取器是指假设教师模型共 n 层，冻结其前 k 层，只用学生数据集训练教师模型的 $n-k$ 层。另外，从攻击者能获得的权限来看，攻击者在某些情况下可能获得教师模型的访问权限，在某些情况下可能获得学生模型的访问权限。从攻击者的目标来看，攻击者可能想要推断教师模型的训练数据，也可能想要推断学生模型的训练数据。根据上述迁移方式的不同和攻击者的能力及需求，本文将攻击分为以下 3 种模式。

攻击Ⅰ：微调模式下，攻击者攻击教师模型，能且仅能访问教师模型。

攻击Ⅱ：特征提取器模式下，攻击者攻击教师模型，能且仅能访问学生模型。

攻击Ⅲ：微调模式下，攻击者攻击学生模型，能且仅能访问学生模型。

3.2 威胁模型

与现有成员推理攻击[25]相似，本文假设攻击者可以获得目标模型的结构和数据分布，并且可以访问目标模型，获得目标模型的输入输出对。

攻击Ⅰ模式下，攻击者 A 攻击教师模型 f_{t}，能且仅能访问教师模型。攻击目标是判断一个数据样本点 (x,y) 是否是教师模型的训练数据 $D_{\mathrm{t}}^{\mathrm{train}}$，计算式为

$$A(f_{\mathrm{t}}(x)) \in D_{\mathrm{t}}^{\mathrm{train}} \tag{1}$$

该模式下，本文默认攻击者 A 可以获得以下内容。

1) 教师模型结构和训练方式。

2) 教师模型训练集的特征分布和其同分布的数据集。

3) 教师模型的黑盒访问权限。

攻击Ⅱ模式下，攻击者 A 攻击教师模型 f_{t}，能且仅能访问学生模型 f_{s}。攻击目标是判断一个数据样本点 (x,y) 是否是教师模型的训练数据 $D_{\mathrm{t}}^{\mathrm{train}}$，计算式为

$$A(f_{\mathrm{s}}(x)) \in D_{\mathrm{t}}^{\mathrm{train}} \tag{2}$$

该模式下，A 可以获得以下内容。

1) 教师模型和学生模型的结构和训练方式。

2) 教师模型和学生模型的训练集的特征分布和其同分布的数据集。

3) 学生模型的黑盒访问权限。

攻击Ⅲ模式下，攻击者 A 攻击学生模型 f_{s}，能且仅能访问学生模型 f_{s}。攻击目标是判断一个数据样本点 (x,y) 是否是学生模型的训练数据 $D_{\mathrm{s}}^{\mathrm{train}}$，计算式为

$$A(f_{\mathrm{s}}(x)) \in D_{\mathrm{s}}^{\mathrm{train}} \tag{3}$$

该模式下，本文默认攻击者 A 可以获得以下内容。

1) 教师模型和学生模型的结构和训练方式。

2) 教师模型和学生模型训练集的特征分布和其同分布的数据集。

3) 学生模型的黑盒访问权限。

3.3 攻击框架

本节对攻击方法进行整体概述。本文方法的整体框架如图 2 所示，主要分为 3 种攻击模式。

1) 攻击 Ⅰ

攻击 Ⅰ 模式下，攻击者攻击教师模型，判断待测样本是否为教师模型的训练数据，且仅能访问教师模型。为实现这一目标，本文建立了对比模型。对比模型的作用有 2 个，首先是构建样本特征，然后是生成输出特征累计概率分布图。对比模型的结构与目标模型相同，对比模型的训练集与目标模型的训练集特征分布一致。为构建样本特征，本文构建 k 个对比模型，考虑到攻击者存在获得的数据集样本数量不足的问题，使用 bootstrap 采样[27]来生成对比数据集，bootstrap 采样减少了对比训练集之间的重叠，使对比模型之间的相似性降低。对比模型的训练方法与目标模型一致。

随后，将待测样本输入 k 个对比模型，获取其中间层输出并将其合并构建样本特征，通过异常样本检测得到异常样本，本文只针对异常样本进行成

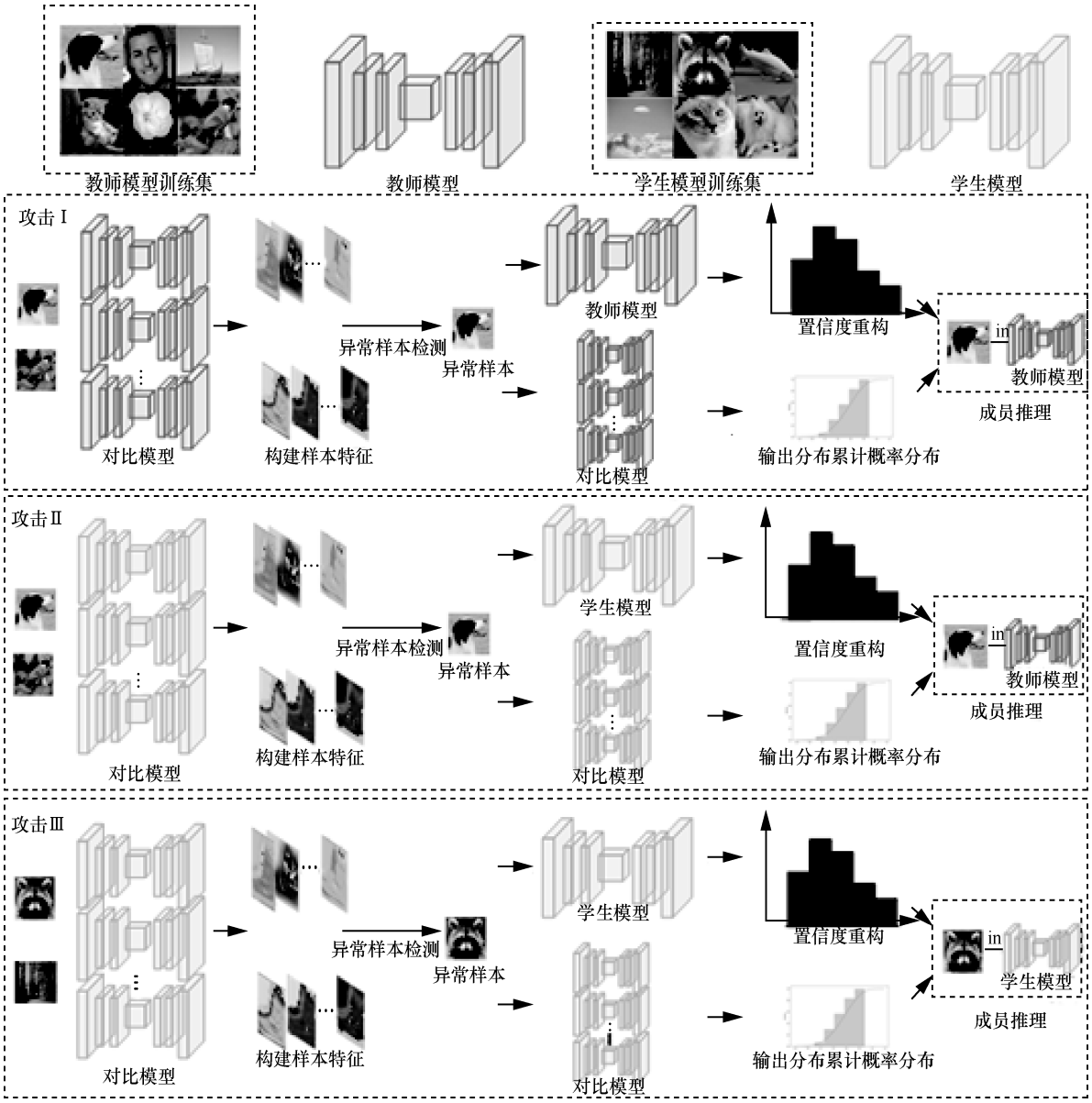


图 2　面向正常拟合模型的成员推理攻击方法整体框架

员推理攻击。

将异常样本输入对比模型，绘制其输出特征累计概率分布图，对数损失函数在训练模型时常用作标准函数，故本文采用对数损失函数构建输出特征分布图，定义为

$$L(M,x)=-\log p_{y_x} \tag{4}$$

其中，M 表示分类器，x 表示输入样本，y_x 表示输入样本的标签，P_{y_x} 表示分类器 M 将样本 x 分类为 y_x 的置信度。

具体步骤如下。首先将目标样本输入对比模型获取其输出 L，利用 L 构建累积分布函数（CDF, cumulative distribution function）图 $D(L)$，函数形式表示为 $F(L)$。然后将异常样本输入教师模型，使用置信度重构方法获取教师模型预测该样本的置信度。最后是成员推理阶段，本文根据假设检验评估样本 x 是目标模型训练数据的置信度。零假设 H_0：样本 x 不是目标模型的训练数据。备择假设 H_1：样本 x 是目标模型的训练数据。根据假设检验，存在 p 值和显著性水平 β，当 $p>\beta$ 时，零假设 H_0 正确；反之，备择假设 H_1 正确。显著性水平 β 由人为设置，p 值计算式为

$$p=F(L) \tag{5}$$

将重构的置信度输入式(4)计算得到对数损失，再将其输入式(5)计算，获取 p 值，若 $p>\beta$，则认为该样本不是成员样本，反之，则是成员样本。

2) 攻击Ⅱ

攻击Ⅱ模式下，攻击者攻击教师模型，判断待测样本是否为教师模型的训练数据，且仅能访问学生模型。与攻击Ⅰ不同，攻击Ⅱ建立了学生模型的对比模型，其训练集分布与学生模型训练集分布一致，训练方式相同。

构建样本特征时将异常样本输入对比模型获取其中间层输出并合并，通过异常样本检测得到异常样本。随后将异常样本分别输入对比模型绘制输出特征累计概率分布图，输入学生模型利用置信度重构得到置信度。与攻击Ⅰ不同，攻击Ⅱ绘制输出特征累计概率分布图时，将目标模型输出的最大置信度代入式(4)计算对数损失。最后通过假设检验，推理该样本是否为成员样本。

3) 攻击Ⅲ

攻击者攻击学生模型，判断待测样本是否是学生模型的训练数据，能且仅能访问学生模型。与上述攻击不同，攻击Ⅲ攻击目标是学生模型，待测样本与学生模型训练集相同。攻击者建立对比模型，模型的结构与学生模型相同，其训练数据分布与学生模型的训练数据分布一致，训练方式与学生模型相同。

随后，将待测样本输入对比模型，提取中间层输出并将其合并得到样本特征，通过异常样本检测获取异常样本，并只对异常样本进行成员推理攻击。将异常样本输入对比模型绘制输出特征累计概率分布图，与攻击Ⅱ不同之处在于绘制输出特征累计概率分布图时，将目标模型输出的预测类对应的置信度代入式(4)计算对数损失，后将异常样本输入学生模型，利用置信度重构方法获取异常样本在目标模型下的预测置信度。最后利用假设检验，推理异常样本是否为学生模型的成员样本。

3.4 异常样本检测

本文只对检测到的异常样本进行成员推理攻击，这些异常样本在特征分布上与其他待测样本存在较大差异，故在训练模型时，异常样本会对模型产生特殊的影响。在模型训练集包含与不包含异常样本时，模型对异常样本的预测会有明显的差别，故能达到较好的攻击效果，异常样本检测算法如算法 1 所示。

算法 1 异常样本检测算法

输入 待测样本与对比模型训练样本 n，类别数 k，最大迭代次数 max_iter，当前迭代次数 num_iter，距离阈值 α

输出 待测样本中的异常样本集合 Q

1)从数据中随机挑选 k 个样本点作为原始的簇中心 $u_1,u_2,\cdots,u_k$。

2)while num_iter<max_iter, do

3) for i=0: n do

4) 根据式(6)计算第 i 个样本的类别

5) 根据式(7)计算每个类别的簇中心

6) if Δu_k=0 then

7) break

8) end if

9) end for

10)end while

11) for i=0:n do

12)根据式(8)计算第 i 个样本到其簇中心的距离 dis

13)if dis > α then

14)　$Q=Q+i$

15)end if

16)Return Q

样本类别计算式为

$$c^i = \arg\min_j \left\| x_f^i - u_j \right\|^2 \tag{6}$$

其中，c^i 表示第 i 个样本的类，j 表示第 j 个类，u_j 表示第 j 个类的中心，x_f^i 表示第 i 个样本特征，即样本 x^i 在 k 个对比模型中间层输出的组合。

簇中心计算式为

$$u_j = \frac{\sum_{i=1}^{n} 1\{c^i = j\} x_f^i}{\sum_{i=1}^{n} 1\{c^i = j\}} \tag{7}$$

其中，u_j 表示第 j 个类的中心，n 表示样本特征的个数，c^i 表示第 i 个样本的类，j 表示第 j 个类，x_f^i 表示第 i 个样本特征。当 c^i 为 j 时，$1\{c^i = j\}$ 的值为 1，否则为 0。

样本间距离计算式为

$$\text{dis} = \left\| x_f^i - x_f^j \right\|^2 \tag{8}$$

其中，x_f^i 表示第 i 个样本特征，x_f^j 表示第 j 个样本特征。

3.5　置信度重构

本文提出置信度重构技术，即使模型只输出预测标签，也能使攻击有较好的攻击性能。

置信度重构基于的思想是将一个样本输入深度模型，模型输出的置信度越大，则该样本越难被对抗攻击，即攻击成功所需要的对抗噪声越大。本文提出的置信度重构主要分为 2 个部分：首先通过对抗攻击，获取攻击成功所需要的对抗噪声大小；然后利用回归分析，获取对抗噪声和置信度的逻辑关系。“HopSkipJump”攻击[41]是最近提出的攻击效率最高的对抗攻击，具有查询次数少、添加噪声少的特点，本文选用该攻击作为攻击方法。第一步，将样本输入对比模型，获取其置信度，随后将样本输入目标模型进行对抗攻击，获取对抗噪声大小。第二步，将第一步中获取的置信度–噪声大小对进行回归分析，获取其对应关系。回归分析采用最小二乘法，具体步骤如下。

1)根据样本点分布特征，初始化近似函数 $y = f(w,x)$。

2)计算残差函数

$$L(y, f(w,x)) = \sum_{i=1}^{m} [y_i - f(w_i, x_i)]^2 \tag{9}$$

3)更新 w，取残差函数最小时的 w 为近似函数的最终参数。

因为对比模型的训练数据分布与目标模型的训练数据分布一致，本文认为在对比模型上得到的置信度和噪声的大小关系与目标模型的基本一致。

4　实验

本节在多个真实数据集和模型上进行实验验证正常拟合迁移学习模型的 3 种成员推理攻击有效性。首先，在 4 种攻击模式下评估了攻击的性能，分别分析了成员推理攻击在访问教师模型时对教师模型造成的成员隐私风险、访问学生模型时对教师模型造成的成员隐私风险和访问学生模型时对学生模型造成的成员隐私风险。其次，对本文方法的有效性进行分析，解释了本文方法在正常拟合模型下有效的原因，随后解释了相比于其他攻击需要获得置信度信息，而本文方法仅需获得标签信息就能有效的原因。再次，进行了参数敏感性分析，评估了异常样本检测阶段不同参数对攻击性能造成的影响。最后，进行了适应性攻击实验，对添加了防御的模型进行攻击，以说明本文所提方法的普适性。

4.1　实验设置

本节主要介绍了实验环境、数据集、模型和评价指标、对比算法等信息。

实验硬件及软件平台：i7-7700K 4.20GHzx8 (CPU), TITAN Xp 12GiBx2 (GPU), 16GBx4 memory (DDR4), Ubuntu16.04 (OS), Python(3.6), tensorflow-gpu (1.12.0), keras (2.2.4), torch (0.4.1) 和 torchvision (0.2.1)。

数据集：本文实验采用 4 个公共数据集。

Caltech101[43]。该数据集包含 5 486 个训练图像和 3 658 张测试图像，分为 101 个不同的物体类别（如人脸、手表、蚂蚁、钢琴等）和一个背景类别。每个类别大约有 40~800 张图片，大多数类别大约有 50 张图片。

CIFAR100[44]。该数据集是广泛用于评价图像识别算法的基准数据集，由彩色图像组成，这些图像被平均分为 100 类，如食物、人、昆虫等。每个

类别有 500 张训练图片和 100 张测试图片。

Flowers102[45]。该数据集包含 102 种常见的花卉类别，包含 6 149 张训练图像和 1 020 张测试图像。

PubFig83[46]。该数据集由 8 300 张裁剪面部图像组成，这些图像来自 83 张公共人脸图像，每一张人脸图像包含 100 个变体。PubFig83 中的图片是从网上获取的，并不是在可控的环境中收集的。

本文选用 4 个常用的深度模型，分别是 VGG16 模型[47]、VGG19 模型[47]、ResNet50 模型[48]和 Inception_v3 模型[49]。模型训练阶段，优化算法采用 Adam 方法，batch_size 设置为 64，epoch 设置为 100。训练完成后，模型均处于正常拟合状态，训练准确率与测试准确率较高且无明显差异。

精确率是衡量成员推理攻击[10]的常用指标，精确率越大表示攻击性能越高，定义为

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

其中，TP 表示实际为成员样本预测为成员样本的样本个数，FP 表示实际为非成员样本预测为成员样本的样本个数。

另外，本文引入覆盖率衡量成员推理攻击性能，覆盖率越大，表示攻击性能越好。

$$\text{coverage}=\frac{\text{TP}}{N} \tag{11}$$

其中，TP 表示实际为成员样本预测为成员样本的样本个数，N 表示成员样本总数。

本文采取 Zou 等[27]（FMIA）、Salem 等[20]（GMIA）和 Long 等[26]（PMIA）这 3 种攻击方法作为本文方法的对比算法。FMIA 和 GMIA 在攻击过程中都建立了攻击模型，区别是 FMIA 针对每一类样本建立了一个攻击模型，GMIA 只需要建立一个攻击模型。攻击模型由两层全连接层组成，第一层包含 64 个神经元，激活函数选用 ReLU，输出层选用 Softmax。PMIA 不建立攻击模型，通过建立参考模型获取样本在不同模型下的输出差异进行攻击。为评估攻击方法的性能，本文建立 100 个目标模型进行测试，其中 50 个包含待测样本，另外 50 个不包含待测样本。

4.2 攻击 I：访教−攻教

本节在微调的迁移方式下评估了本文提出的成员推理攻击性能。攻击 I 模式下，攻击者攻击教师模型，判断输入样本是否为教师模型的成员样本，且攻击者能且仅能访问教师模型。本文教师模型分别在 4 种数据集和 3 种常见的深度模型上训练。实验结果如表 1 所示。本文用精确率和覆盖率来衡量不同攻击方法之间的攻击性能。

首先，本文比较了 PMIA 和 TMIA 检测的异常样本数量。TMIA 检测到的异常样本比 PMIA 多，这主要是因为 PMIA 基于密度检测异常样本，只能在样本分布稀疏时检测到较多异常样本，而 TMIA 基于距离检测异常样本，更具普适性。FMIA 和 GMIA 本身无异常检测步骤，为与本文方法 TMIA

表 1 攻击 I：访教−攻教模式下不同攻击的攻击性能比较

比较项	方法	Caltech101			Flowers102			CIFAR100			PubFig83		
		VGG16	VGG19	Resnet50	VGG16	VGG19	Resnet50	VGG16	VGG19	Resnet50	VGG16	VGG19	Resnet50
异常样本	FMIA	62	60	59	42	43	40	76	73	77	35	31	32
	GMIA	62	60	59	42	43	40	76	73	77	35	31	32
	PMIA	51	53	50	29	27	28	58	58	54	23	22	20
	TMIA	62	60	59	42	43	40	76	73	77	35	31	32
精确率	FMIA	42.41%	48.23%	45.12%	40.17%	46.66%	44.31%	48.03%	47.64%	49.46%	42.49%	46.31%	49.15%
	GMIA	51.39%	52.87%	51.01%	48.35%	46.12%	47.16%	43.93%	45.63%	46.29%	41.27%	51.50%	41.30%
	PMIA	92.64%	93.83%	91.64%	94.22%	93.99%	93.23%	94.33%	91.38%	94.09%	90.66%	92.28%	94.86%
	TMIA	**93.98**%	93.60%	**91.92**%	93.49%	93.54%	93.01%	90.08%	90.20%	92.34%	**91.98**%	**93.36**%	90.74%
覆盖率	FMIA	54.42%	52.59%	53.37%	48.05%	47.33%	45.27%	40.26%	41.45%	40.48%	44.02%	41.63%	45.48%
	GMIA	54.69%	53.60%	53.81%	43.68%	44.00%	44.92%	41.30%	41.27%	43.18%	48.61%	46.94%	46.77%
	PMIA	70.61%	74.97%	71.30%	72.57%	71.08%	72.86%	71.66%	72.20%	72.50%	73.58%	73.49%	71.14%
	TMIA	**72.05**%	70.79%	70.39%	**73.40**%	**72.51**%	**73.10**%	71.61%	70.35%	71.13%	72.62%	73.12%	**72.28**%

对比，测试时攻击 TMIA 检测到的异常样本，故其异常样本数量与 TMIA 相同。

其次，本文比较了不同攻击方法在不同数据集和不同模型下的精确率。在任意模型和任意数据集中，TMIA 和 PMIA 的精确率均高于 FMIA 和 GMIA，FMIA 和 GMIA 在 Caltech101 数据集的 Resnet50 模型下的精确率分别为 45.12%和 51.01%，这主要是因为 FMIA 和 GMIA 是针对过拟合模型的成员推理攻击，它们基于成员样本和非成员样本在目标模型下的输出差异进行攻击，然而，在攻击正常拟合模型时，成员样本和非成员样本在目标模型下的输出差异较小，FMIA 和 GMIA 攻击性能大大降低。

本文所提方法 PMIA 和 TMIA 的攻击性能相近，均有较好的攻击性能，例如在 Flowers102 数据集的 VGG16 模型下，精确率分别为 94.22%和 93.49%，这是因为 TMIA 和 PMIA 利用异常样本检测找到了容易受到攻击的样本，这些样本对模型的预测输出有特殊的影响，有较高的概率被攻击成功。

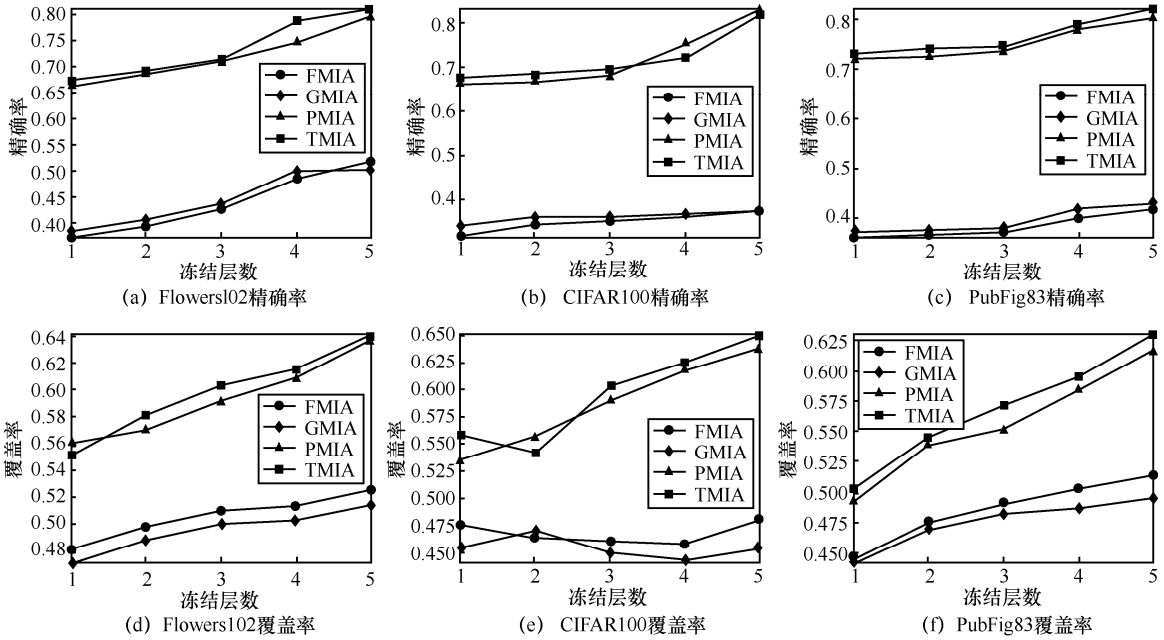
与 PMIA 需要获取置信度不同，本文所提方法只需要获取样本在目标模型下输出的标签信息，获得的信息更少，但是攻击性能与 PMIA 相比并没有明显的降低，表明了 TMIA 的优越性。

最后，本文比较了不同攻击方法在不同模型和不同数据集下的覆盖率。在任意模型和数据集下，TMIA 的覆盖率明显高于 FMIA 和 GMIA，这显示了 TMIA 较好的攻击性能。与 PMIA 需要置信度相比，TMIA 只需要获取标签信息，在获得信息较少的情况下，性能并没有明显的降低，再次表现了 TMIA 的优越性。

4.3　攻击 II：访学–攻教

本节在特征提取器的迁移方式下评估了本文提出的成员推理攻击性能。攻击 II 模式下，攻击者攻击教师模型，判断输入样本是否为教师模型的成员样本，且攻击者能且仅能访问学生模型。本节教师模型均由 Caltech101 数据集训练，学生模型在另外 3 种数据集上训练，教师模型和学生模型都采用 VGG16。实验结果如图 3 所示，其中横坐标表示冻结教师模型的层数，纵坐标表示攻击的性能指标。本文用精确率和覆盖率来衡量不同攻击方法之间的攻击性能。

由图 3 可知，随着冻结层数的增加，攻击的性能也会上升。这是因为冻结的层数越多，学生模型会更多地保留教师模型训练集的特征，增加了攻击的成功率。上述结果表明，即使在不访问教师模型的情况下，只访问学生模型，也会造成教师模型训练数据的成员隐私泄露。这主要是因为学生模型也包含教师模型训练数据的特征，故存在泄露其数据隐私的可能。



(a) Flowersl02精确率　(b) CIFAR100精确率　(c) PubFig83精确率

(d) Flowers102覆盖率　(e) CIFAR100覆盖率　(f) PubFig83覆盖率

图 3　不同攻击方法在不同冻结层数下的性能比较

其次，在任意数据集下，TMIA 的精确率和覆盖率均大于 FMIA 和 GMIA，表明了本文方法有较好的攻击性能。这主要因为 FMIA 和 GMIA 基于成员样本和非成员样本在模型下的输出差异进行攻击，而模型处于正常拟合状态下，输出几乎无差异，而 TMIA 只攻击异常样本，这些异常样本对目标模型的预测产生特殊影响，当模型训练集中存在和不存在异常样本时，模型对异常样本的预测会有较大的差异，对比模型训练集中不包含异常样本，在推理阶段，利用假设检验，若异常样本在目标模型下的输出特征不符合异常样本在对比模型下的输出特征分布，则认为该样本为成员样本，有较高的精确率推理成功。

最后，TMIA 在只获得标签信息的情况下，获得的信息更少，但是和 PMIA 性能几乎无差异，再次表明了 TMIA 方法的优越性。

4.4 攻击 III：访学−攻学

本节在微调的迁移模式下评估了本文提出的成员推理攻击性能。攻击 III 模式下，攻击者攻击学生模型，判断输入样本是否为学生模型的成员样本，且攻击者能且仅能访问学生模型。本节教师模型均由 Caltech101 数据集训练，学生模型在另外 3 种数据集上训练，分别在 3 种常见的深度模型上进行成员推理攻击。本文用精确率和覆盖率来衡量不同攻击方法之间的攻击性能。

如表 2 所示，在任意模型和任意数据集中，TMIA 和 PMIA 的精确率和覆盖率均高于 FMIA 和 GMIA，例如在 Flowers102 上训练的 VGG19 的精确率分别为 53.55%和 53.06%，PMIA 和 TMIA 的精确率分别为 94.37%和 93.53%。这是因为 FMIA 和 GMIA 是基于成员样本和非成员样本在模型输出下的置信度差异进行攻击，当模型处于正常拟合时，成员样本和非成员样本的置信度差异很小，导致 FMIA 和 GMIA 攻击性能大大降低。本文所提方法 TMIA 和 PMIA 的攻击性能更强，因为 TMIA 和 PMIA 挑选对模型输出有特殊影响的样本，这些样本更容易被攻击。

与 PMIA 需要获取置信度不同，本文所提方法 TMIA 只需要获取样本在目标模型下输出的标签信息，获得的信息更少，但是攻击性能与 PMIA 相比并没有明显的降低，这也表明了本文置信度重构的有效性。

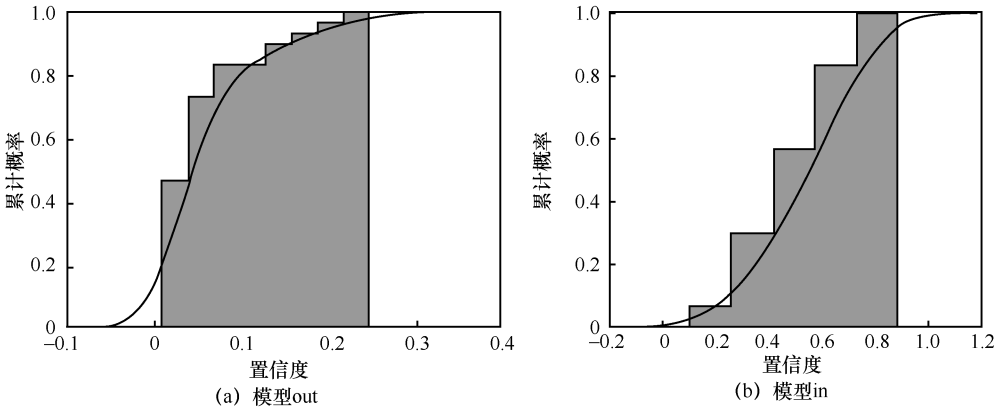
4.5 有效性分析

本节分析了 TMIA 有较强攻击性能的原因。为此，本文给出了异常样本在模型 in 和模型 out 下输出的置信度概率累计分布，模型 in 表示该模型的训练数据包含异常样本，模型 out 表示该模型的训练数据不包含异常样本。

如图 4 所示，异常样本在模型 in 和模型 out 下的输出分布有着明显差异。异常样本在模型 in 下的输出置信度明显大于在模型 out 下的输出置信度，这说明了本文方法的攻击有效性，解释了本文方法

表 2 攻击 III：访学−攻学模式下不同攻击的攻击性能比较

比较项	方法	Flowers102			CIFAR100			PubFig83		
		VGG16	VGG19	Resnet50	VGG16	VGG19	Resnet50	VGG16	VGG19	Resnet50
异常样本	FMIA	46	43	40	76	73	77	35	31	32
	GMIA	46	43	40	76	73	77	35	31	32
	PMIA	29	27	28	58	58	54	23	22	20
	TMIA	**46**	**43**	**40**	**76**	**73**	**77**	**35**	**31**	**32**
精确率	FMIA	52.34%	53.55%	53.91%	41.59%	44.28%	42.51%	43.82%	45.70%	46.15%
	GMIA	52.54%	53.06%	53.65%	41.72%	40.79%	41.49%	40.87%	44.08%	44.13%
	PMIA	94.10%	94.37%	94.76%	92.36%	91.89%	92.56%	92.87%	93.12%	90.29%
	TMIA	93.29%	93.53%	93.97%	92.00%	91.55%	**93.43%**	92.34%	**94.65%**	**90.47%**
覆盖率	FMIA	51.42%	52.99%	52.76%	42.11%	50.21%	53.57%	48.36%	45.50%	45.63%
	GMIA	50.30%	50.09%	50.32%	47.01%	47.84%	49.64%	47.21%	48.48%	49.80%
	PMIA	73.51%	71.89%	74.30%	73.42%	72.24%	74.12%	73.43%	73.63%	72.84%
	TMIA	71.15%	71.50%	72.36%	**74.48%**	**73.29%**	72.97%	**73.80%**	72.95%	72.12%



(a) 模型out　　(b) 模型in

图 4　异常样本累计概率分布

可以推断出样本成员状态的原因。

最后，本节给出了置信度重构，如图 5 所示。构成对抗样本的噪声大小和模型对样本预测的置信度有明显的逻辑关系，置信度越大，攻击该样本所需要的噪声就越大，说明了本文所提置信度重构方法的有效性，解释了即使在模型只输出标签信息的情况下 TMIA 依然能有较好攻击性能的原因。

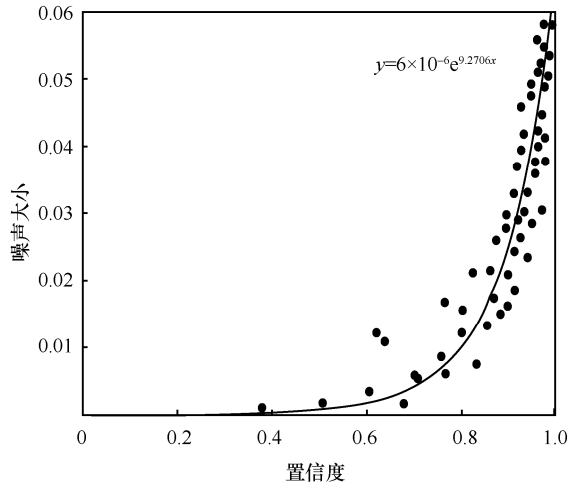


图 5　置信度重构

4.6　参数敏感性分析

本节主要对异常样本检测过程中距离阈值 α 进行敏感性分析，评估了不同阈值 α 对攻击性能的影响。实验结果如表 3 所示，随着阈值 α 的增加，检测到的异常样本数量会减少，精确率和覆盖率有一定增加，这表明阈值的增高会让更少的样本被认为是异常样本，这些异常样本离聚类中心更远，特征差异越大，对模型预测造成的影响也越大，更容易被攻击成功。

表 3　　**参数敏感性分析**

数据集	α	异常样本	精确率	覆盖率
Caltech101	0.1	212	56.61%	35.28%
	0.2	103	77.39%	59.50%
	0.3	57	91.66%	78.03%
Flowers102	0.2	207	58.91%	37.66%
	0.3	125	71.30%	58.62%
	0.4	75	92.25%	79.10%
CIFAR100	0.1	243	58.89%	34.72%
	0.2	157	76.24%	55.17%
	0.3	68	91.22%	75.97%
PubFig83	0.1	154	57.74%	35.71%
	0.2	80	73.15%	54.74%
	0.3	33	93.16%	75.29%

4.7　适应性攻击

本节主要对 TMIA 方法在施加了防御后的模型的攻击效果进行分析。现有研究[19]表明，Dropout 对成员推理攻击有较好的防御性能。本节在 Caltech101 数据集上训练的目标模型上应用了 Dropout，随后用 TMIA 进行攻击。

表 4 给出了应用 Dropout 前后，模型的准确率和攻击性能的差异。结果显示，Dropout 方法降低了异常样本检测环节检测到的异常样本数量，但是检测出的异常样本仍然以较高的精确率被攻击成功。例如，当 Dropout 的参数被设置为 0.1 时，检测到了 6 个异常样本，这些异常样本以高达 96.15% 的精确率被推理成功。

综上，Dropout 在一定程度上缓解了成员推理攻击，但是并没有完全解决成员推理攻击的隐私威胁，防御效果有限，进一步说明了本文方法对成员

隐私的危害。

表 4 不同 Dropout 下 TMIA 的攻击性能

Dropout	训练	测试	异常样本/个	精确率	覆盖率
0	0.98	0.97	33	91.61%	55.40%
0.1	0.98	0.98	6	96.15%	63.55%
0.3	0.96	0.96	6	95.44%	56.08%

4.8 累计分布图

本节在微调的迁移方式下展示了 Flowers102 数据集在 VGG16、VGG19 和 Resnet50 模型上的对数损失 L 累计分布。攻击 I 模式下，攻击者攻击教师模型，判断输入样本是否为教师模型的成员样本，且攻击者能且仅能访问教师模型。TMIA 在 VGG16、VGG19 和 Resnet50 模型上分别检测到了 42、43 和 40 个异常样本。本文将检测到的异常样本输入对比模型，得到输出标签后，利用回归分析得到其置信度，最后通过式(4)计算其对数损失。根据获得的对数损失，绘制累计分布图。

累计概率分布如图 6 所示，其中横坐标表示对数损失 L，纵坐标表示累计概率。判别输入样本是否为成员样本时，将输入样本输入目标模型，得到输入样本在目标模型下真实标签类的置信度，随后利用式(4)计算其对数损失，最后根据假设检验判别输入样本是否为成员样本。

5 结束语

本文对不同迁移学习下，正常拟合模型的数据成员隐私风险进行了系统的研究。针对过去的工作主要面向过拟合模型，本文考虑的是正常拟合这一更加符合现实条件的环境，通过异常点检测选择容易受到成员推理攻击的目标数据并根据假设检验保守地做出成员关系预测，使攻击失败成本降至最低。针对过去的工作主要面向个人独自训练的模型，本文在迁移学习环境中设置了 2 种不同迁移方式，并设计了 3 种不同的攻击模式。本文系统地设计了攻击框架，并根据实验结果评估了 3 种攻击对 4 个真实数据集的攻击性能。针对模型只能在标签信息情况下过去攻击无法正常工作的问题，本文提出了置信度重构方法，在获得信息更少的情况下，达到了与基于置信度攻击几乎一致的性能。

此外，本文 TMIA 方法存在异常样本检测数量少的问题，这是由于本文提出的异常样本检测技术可能无法找到所有对模型预测产生特殊影响的样本。另外，本文方法需要获取目标模型的训练集分布，当攻击者无法获取目标模型训练集分布时，攻击性能有一定降低。因此，在未来的研究中，作者将继续研究异常样本的检测方法，找到更多的异常样本，并找到一种不需要获取目标模型训练集的通用方法。

参考文献：

[1] 高红民, 曹雪莹, 陈忠昊, 等. 基于多尺度近端特征拼接网络的高光谱图像分类方法[J]. 通信学报, 2021, 42(2): 92-102.
GAO H M, CAO X Y, CHEN Z H, et al. Hyperspectral image classification method based on multi-scale proximal feature concatenate network[J]. Journal on Communications, 2021, 42(2): 92-102.

[2] 崔颖, 徐凯, 陆忠军, 等. 主动学习策略融合算法在高光谱图像分类中的应用[J]. 通信学报, 2018, 39(4): 91-99.
CUI Y, XU K, LU Z J, et al. Combination strategy of active learning for hyperspectral images classification[J]. Journal on Communications, 2018, 39(4): 91-99.

[3] KIM I, BAEK W, KIM S. Spatially attentive output layer for image classification[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9530-9539.

[4] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 1-9.

[5] WANG T W, ZHU Y Z, JIN L W, et al. Decoupled attention network for text recognition[C]// The Thirty-Second Innovative Applications of

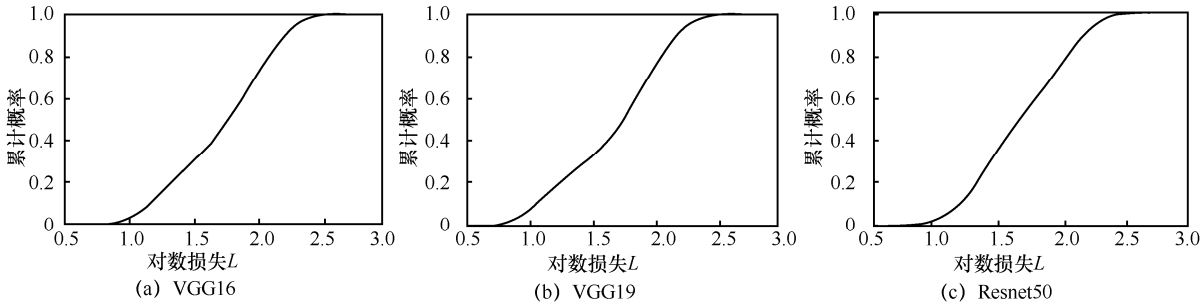


图 6 累计概率分布

Artificial Intelligence Conference. Palo Alto: AAAI Press, 2020: 12216-12224.

[6] YU D L, LI X, ZHANG C Q, et al. Towards accurate scene text recognition with semantic reasoning networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 12110-12119.

[7] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2013: 6645-6649.

[8] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[9] SEN P, NAMATA G, BILGIC M, et al. Collective classification in network data[J]. AI Magazine, 2008, 29(3): 93-106.

[10] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.

[11] 张思成, 林云, 涂涯, 等. 基于轻量级深度神经网络的电磁信号调制识别技术[J]. 通信学报, 2020, 41(11): 12-21.
ZHANG S C, LIN Y, TU Y, et al. Electromagnetic signal modulation recognition technology based on lightweight deep neural network[J]. Journal on Communications, 2020, 41(11): 12-21.

[12] WANG Q, DU P F, YANG J Y, et al. Transferred deep learning based waveform recognition for cognitive passive radar[J]. Signal Processing, 2019, 155: 259-267.

[13] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of 3rd International Conference on Learning Representations. [S.n.:s.l.], 2015: 803-807.

[14] 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.

[15] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of 2020 Advances in Neural Information Processing Systems (NIPS). [S.n.:s.l.], 2020:6-12.

[16] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21(1): 1-67.

[17] OLATUNJI I E, NEJDL W, KHOSLA M. Membership inference attack on graph neural networks[J]. arXiv Preprint, arXiv:2101.06570, 2021.

[18] HUI B, YANG Y C, YUAN H L, et al. Practical blind membership inference attack via differential comparisons[C]//Proceedings of 2021 Network and Distributed System Security Symposium. Reston: Internet Society, 2021: 21-25.

[19] LI J C, LI N H, Ribeiro B. Membership inference attacks and defenses in supervised learning via generalization gap[J]. arXiv Preprint, arXiv:2002.12062, 2020.

[20] SALEM A, ZHANG Y, HUMBERT M, et al. ML-leaks: model and data independent membership inference attacks and defenses on machine learning models[C]//Proceedings of 2019 Network and Distributed System Security Symposium. Reston: Internet Society, 2019: 24-27.

[21] 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.

[22] SONG L W, SHOKRI R, MITTAL P. Privacy risks of securing machine learning models against adversarial examples[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2019: 241-257.

[23] YEOM S, GIACOMELLI I, FREDRIKSON M, et al. Privacy risk in machine learning: analyzing the connection to overfitting[C]//Proceedings of 2018 IEEE 31st Computer Security Foundations Symposium (CSF). Piscataway: IEEE Press, 2018: 268-282.

[24] 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 (SP). Piscataway: IEEE Press, 2019: 739-753.

[25] LEINO K, FREDRIKSON M. Stolen memories: leveraging model memorization for calibrated white-box membership inference[C]//Proceedings of 2020 USENIX Security Symposium (USENIX Security 20). Berkeley: USENIX Association, 2020: 1605-1622.

[26] LONG Y H, WANG L, BU D Y, et al. A pragmatic approach to membership inferences on machine learning models[C]//Proceedings of 2020 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE Press, 2020: 521-534.

[27] ZOU Y, ZHANG Z K, BACKES M, et al. Privacy analysis of deep learning in the wild: membership inference attacks against transfer learning[J]. arXiv Preprint, arXiv:2009.04872, 2020.

[28] BACKES M, BERRANG P, HUMBERT M, et al. Membership privacy in MicroRNA-based studies[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2016: 319-330.

[29] HAGESTEDT I, ZHANG Y, HUMBERT M, et al. MBeacon: privacy-preserving beacons for DNA methylation data[C]//Proceedings of 2019 Network and Distributed System Security Symposium. Reston: Internet Society, 2019: 21-27.

[30] PYRGELIS A, TRONCOSO C, DE CRISTOFARO E. Knock knock, who's there? membership inference on aggregate location data[C]//Proceedings of 2018 Network and Distributed System Security Symposium. Reston: Internet Society, 2018: 35-42.

[31] CHEN J C, RANJAN R, KUMAR A, et al. An end-to-end system for unconstrained face verification with deep convolutional neural networks[C]//Proceedings of 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). Piscataway: IEEE Press, 2015: 360-368.

[32] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway: IEEE Press, 2015: 1137-1149.

[33] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 779-788.

[34] CAELLES S, MANINIS K K, PONT-TUSET J, et al. One-shot video object segmentation[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 5320-5329.

[35] KUNZE J, KIRSCH L, KURENKOV I, et al. Transfer learning for speech recognition on a budget[C]//Proceedings of the 2nd Workshop on Representation Learning for NLP. Stroudsburg: Association for Computational Linguistics, 2017: 168-177.

[36] WANG D, ZHENG T F. Transfer learning for speech and language processing[C]//Proceedings of 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Piscataway: IEEE Press, 2015: 1225-1237.

[37] HEIGOLD G, VANHOUCKE V, SENIOR A, et al. Multilingual acoustic models using distributed deep neural networks[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2013: 8619-8623.

[38] CIREŞAN D C, MEIER U, SCHMIDHUBER J. Transfer learning for Latin and Chinese characters with deep neural networks[C]//Proceedings of 2012 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2012: 1-6.

[39] JOHNSON M, SCHUSTER M, LE Q V, et al. Google's multilingual neural machine translation system: enabling zero-shot translation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 339-351.

[40] MIKOLOV T, LE Q V, SUTSKEVER I. Exploiting similarities among languages for machine translation[J]. Computer Science, 2014, 17(4): 45-52.

[41] WANG B, YAO U, CHICAGO U O, et al. With great training comes great vulnerability: practical attacks against transfer learning[C]//Proceedings of 2018 USENIX Security Symposium (USENIX Security). Berkeley: USENIX Association, 2018: 1281-1297.

[42] SCHUSTER R, SCHUSTER T, MERI Y, et al. Humpty dumpty: controlling word meanings via corpus poisoning[C]//Proceedings of 2020 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2020: 1295-1313.

[43] LI F F, FERGUS R, PERONA P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories[J]. Computer Vision and Image Understanding, 2007, 106(1): 59-70.

[44] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[J]. Handbook of Systemic Autoimmune Diseases, 2009, 1(4): 130-138.

[45] NILSBACK M E, ZISSERMAN A. Automated flower classification over a large number of classes[C]//Proceedings of 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. Piscataway: IEEE Press, 2008: 722-729.

[46] 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.

[47] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014, 8(2): 475-483.

[48] 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.

[49] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 2818-2826.

[作者简介]

陈晋音（1982− ），女，浙江宁波人，博士，浙江工业大学教授，主要研究方向为智能计算、数据挖掘、网络安全等。

上官文昌（1996− ），男，湖北十堰人，浙江工业大学硕士生，主要研究方向为深度学习、人工智能、深度学习、隐私攻防等。

张京京（1988− ），男，北京人，博士，军事科学院系统工程研究院工程师，主要研究方向为深度学习、人工智能和对抗性攻击和防御等。

郑海斌（1995− ），男，浙江台州人，浙江工业大学博士生，主要研究方向为深度学习、人工智能和对抗性攻击和防御等。

郑雅羽（1978− ），男，浙江温州人，博士，浙江工业大学副教授，主要研究方向为嵌入式软硬件应用开发、视频图像处理算法、服务器网络技术等。

张旭鸿（1988− ），男，河北石家庄人，博士，浙江大学助理教授，主要研究方向为分布式大数据与人工智能系统、大数据挖掘与分析、数据驱动安全、人工智能与安全等。