软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.007009]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向开集识别的稳健测试时适应方法 *

周植[1], 张丁楚[1], 李宇峰[1], 张敏灵[2]

[1](南京大学 软件新技术国家重点实验室,江苏 南京 210023)

[2](东南大学 计算机科学与工程学院,江苏 南京 210096)

通讯作者: 李宇峰, E-mail: liyf@nju.edu.cn

摘 要: 开集识别旨在研究测试阶段突现未见类别对于机器学习模型的挑战,以期学习模型既能分类已见类别又可识别/拒绝未见类别, 是确保机器学习模型能够在开放世界中高效稳健部署的重要技术.既有开集识别技术通常假设已见类别的协变量分布在训练与测试阶段维持不变.然而在实际场景中,类别的协变量分布常不断变化. 直接利用既有技术不再奏效,其性能甚至劣于基线方案.因此,亟需研究新型开集识别方法,使其能不断适应协变量分布偏移,以期模型在测试阶段既能稳健分类已见类别又可识别未见类别.我们将此新问题设置命名为开放世界适应问题(简称 AOW)并提出了一种开放测试时适应方法(简称 OTA).本文方法基于无标注测试数据优化自适应熵损失与开集熵损失更新模型,维持对已见类的既有判别能力,同时增强了识别未见类的能力.大量实验分析表明,本文方法在多组基准数据集、多组不同协变量偏移程度下均稳健地优于现有先进的开集识别方法.

关键词: 开集识别;测试时适应;分布偏移;图像识别;流数据

中图法分类号: TP311

中文引用格式: 周植,张丁楚,李宇峰,张敏灵. 面向开集识别的稳健测试时适应方法.软件学报. http://www.jos.org.cn/1000-9825/7009.htm

英文引用格式: Zhou Z, Zhang DC, Li YF, Zhang ML. Towards robust test-time adaptation for open-set recognition. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7009.htm

Towards Robust Test-Time Adaptation for Open-Set Recognition

ZHOU Zhi[1], ZHANG Ding-Chu[1], LI Yu-Feng[1], ZHANG Min-Ling[2]

[1](National Key Laboratory for Novel Software Technology, Nanjing Univeristy, Nanjing 210023, China)

[2](School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: Open-set recognition is an important issue for ensuring the efficient and robust deployment of machine learning models in the open world. It aims to address the challenge of encountering samples from unseen classes that emerge during testing, i.e., to accurately classify the seen classes while identifying and rejecting the unseen ones. Current open-set recognition studies assume that the covariate distribution of the seen classes remains constant during both training and testing. However, in practical scenarios, the covariate distribution is constantly shifting, which can cause previous methods to fail, and their performance may even be worse than the baseline method. Therefore, it is urgent to study novel open-set recognition methods that can adapt to the constantly changing covariate distribution so that they can robustly classify seen categories and identify unseen categories during testing. We name this novel problem Adaptation in the Open World (AOW) and propose a test-time adaptation method for open-set recognition called Open-set Test-time Adaptation (OTA). OTA method only utilizes unlabeled test data to update the model with adaptive entropy loss and open-set entropy loss, maintaining the model's ability to discriminate seen classes while further enhancing its ability to recognize unseen classes. We conduct comprehensive experiments on multiple benchmark datasets with different covariate shift levels. The results show that our proposal is robust to covariate shift and gives superior performance compared to many state-of-the-art methods.

* 周植与张丁楚为共同第一作者

基金项目:科技创新 2030-“新一代人工智能”重大项目课题(2022ZD0114803); 国家自然科学基金(62176118)

收稿时间: 2023-05-11; 修改时间: 2023-07-07; 采用时间: 2023-08-24; jos 在线出版时间: 2023-09-11

Key words: open-set recognition; test-time adaptation; distribution shift; image classification; streaming data

近年来,随着机器学习技术的发展,深度学习方法在诸多领域取得显著成效,在图像分类[1]、语音识别[2]、文本翻译[3]、商品推荐[4]等任务中得到广泛应用.传统深度神经网络(deep neural network, DNN)依赖于封闭世界假设,即训练数据与测试数据的类别空间相同.然而,在开放环境中,测试数据可能包含训练数据里从未见过类别的样本,使得上述假设难以成立.现有研究[5-9]发现,模型面对未见类别样本时,往往以高置信度将其错误分类为已见类别,为深度模型部署于真实应用带来潜在风险.例如,在自动驾驶任务[10]中,有限的训练数据难以覆盖自动驾驶汽车在开放世界中所面对的全部情形,模型对未知情形做出的错误预测可能会导致严重的车祸,危及人民生命财产安全.因此,在测试阶段赋予模型识别未见类别样本的能力, 是机器学习模型能够在开放环境中安全、稳定部署的重要问题.

开集识别(Open-set recognition, OSR)[6,11,12]旨在正确地分类已见类别样本,同时,准确地识别并拒绝未见类别样本.近年来,此领域涌现出大量的基于深度学习模型的研究工作,可以分为判别式方法与生成式方法两种.判别式方法利用 SoftMax 层输出的概率分布[13]、OpenMax 层输出的概率分布[14]、样本在表示空间内的距离[15,16]等信息显式地建模样本属于未见类别的概率.生成式方法通常利用自编码器(auto-encoder, AE)[17]、生成对抗网络(generative adversarial network, GAN)[5]、迪利克雷过程(dirichlet process, DP)[18]等技术对样本进行重构,并利用重构误差来度量样本属于为未见类别的概率.这些工作利用了已见类别的数据分布在训练数据与测试数据之间保持不变的假设,有效地提升了深度学习模型的开集识别能力.

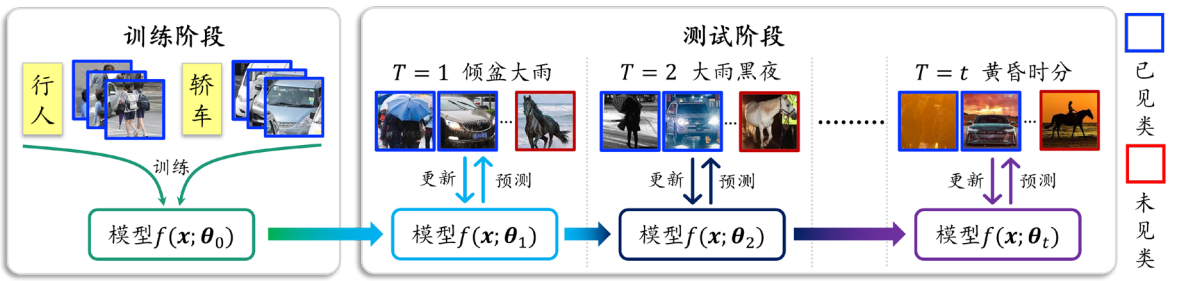

图 1 开放世界适应问题设置示意图

尽管大量研究有效地提升了深度学习模型的开集识别能力,但是他们未考虑到已见类别的协变量分布在实际场景中将连续变化,限制了这些算法在更多现实任务中的应用.例如,在自动驾驶任务[10]中,摄像头捕捉到图像数据分布(即,协变量分布)将随着汽车所处环境的时间、天气、地理位置等因素改变[19,20].在这种情况下,理想的开集识别模型即使面对协变量分布偏移也应稳健地分类已见类别并识别未见类别.据此,我们提出一个新颖的问题设置:开放世界适应问题(Adaptation in the Open World, 简称 AOW).在此问题中,利用已见类别样本训练得到的模型将被部署到持续变化的测试环境中.与开集识别问题相同,在开放世界适应问题中,模型在测试阶段将遇到未见类样本,模型准确分类已见类样本的同时,还需识别未见类样本;与开集识别问题不同,开放世界适应问题在测试阶段面临已见类样本协变量分布变化的问题,使模型分类已见类别、识别未见类别的性能均严重退化.因此,适用于开放世界适应问题的模型一方面需要正确分类已见类别样本,同时,准确识别并拒绝未见类别样本,防止模型对未见类别样本产生错误的预测.另一方面,此模型还需要不断适应于变化的协变量分布,防止模型性能显著退化.图 1 直观地展示了开放世界适应问题的具体设置.图 2 所展示的结果证明,一旦协变量分布发生偏移,既有方法分类已见类别、识别未见类别的能力都会显著下降,其性能表现甚至劣于基线方法.因此,如何使开集识别模型适应动态变化的协变量分布是一个有价值且极具挑战的问题.

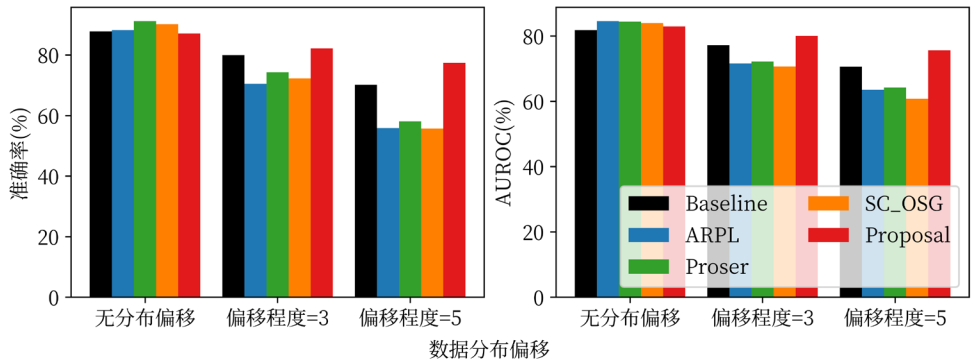

图 2 协变量分布偏移时,开集识别方法与本文方法分类已见类与识别未见类的性能表现. 左图展示了模型分类已见类别的准确率;右图展示了模型识别了未见类样本的 AUROC

测试时适应(Test-time Adaptation, TTA)是一类解决协变量分布偏移的有效技术.其在测试阶段仅利用无标注测试样本使模型的预测结果对协变量分布偏移更稳健.现有工作可以分为优化模型参数[21−23]与调整模型预测结果[24,25]两种类别.近期,研究人员开始考虑真实场景下的测试时适应,例如:测试数据同时存在混合协变量偏移[26]、测试数据的标记分布存在偏移[27]等.这些工作均假设训练数据与测试数据的类别空间完全相同,然而,当测试时适应直接应用于类别空间存在变化的开放世界时,既有方法将受到未见类别样本的影响,导致模型性能显著退化.因此,本文设计了一种针对开放世界适应问题的测试时适应方法,称为开放测试时适应(Open-set Test-time Adaptation,简称 OTA).首先,OTA 方法利用自适应熵损失消除了未见类样本在模型更新中的负面影响,有效地维持了开集识别模型对已见类别的判别能力.进一步, OTA 结合轻量级的未见类别检测模块与开集熵损失,有效地利用了未见类样本,进一步提升开集识别模型区分已见类别与未见类别的能力.最终,OTA 方法引入参数正则化损失,防止模型在持续更新的过程中出现灾难性遗忘问题.本文在包含不同程度协变量分布偏移的基线数据集上进行实验,实验结果证明了本文所提的 OTA 方法的有效性.

综上所述,本文的贡献有如下三点:

1. 本文研究了一个新颖的开放世界适应问题设置 AOW,即,开集识别模型在测试阶段面临协变量分布偏移的问题.开集识别模型需要不断适应于变化的协变量分布,以保证其稳健地分类已见类别样本并识别未见类别样本.
2. 本文提出了一种针对开放世界适应问题的测试时适应方法 OTA.在 OTA 方法中,我们提出自适应熵损失与开集熵损失,一方面消除更新过程中未见类样本的负面影响,有效地维持了模型分类已见类别的能力;另一方面,充分地利用未见类样本,进一步提升模型识别未见类别样本的能力.
3. 本文在包含多种协变量分布偏移的基准数据集上测试了所提的 OTA 方法.实验结果证明,OTA 方法能稳健地适应于变化的协变量分布.其不仅击败了最先进的开集识别方法,同时,也显著优于对分布偏移稳健的开集识别方法、组合开集识别与测试时适应的混合方法.

本文第 1 节将介绍开放世界适应问题的相关工作.第 2 节将介绍本文所提的开放世界适应问题,并对此问题展开分析.第 3 节介绍本文提出的开放测试时适应方法 OTA.第 4 节通过对比实验验证了所提方法的有效性.最后总结全文.

1 开放世界适应问题的相关工作

本章节将介绍与开放世界适应问题相关的两类工作:开集识别与测试时适应.

1.1 开集识别

开集识别研究训练数据与测试数据的类别空间存在差异的问题,旨在准确分类训练数据中的已见类别,同

时识别并拒绝训练数据中从未见过的类别.现有开集识别方法可以分为统计方法与深度方法两种类别.针对统计开集识别方法,Scheirer 等人[28]首先形式化了开集识别问题,并提出一种基于 SVM 模型的开集识别方法.进而,Jain 等人[29]将极值理论(Extreme Value Theory)应用于开集识别 SVM 模型中并获得了更好的性能.近期,基于深度模型的开集识别方法发展迅速,其又可以分别判别式方法与生成式方法两类.Bendale 等人[14]提出了第一个基于深度学习模型的开集识别方法,将深度神经网络中的 SoftMax 模块替换为 OpenMax 模块.继而,Ge 等人[30]将生成对抗网络与 OpenMax 模块相结合,提出了 G-OpenMax 模块.Neal 等人[5]首先利用数据增广技术生成虚拟的未见类样本,从而使已见类别与未见类别之间的决策边界更准确.Oza 等人[31]则利用条件自编码器来解决开集识别问题,通过利用极值理论建模样本的重建误差来区分已见类别与未见类别.此外,Shao 等人[32]首先考虑了开集识别问题中训练数据与测试数据间可能存在协变量偏移,并基于因果理论提出了一种利用不变表征的稳健开集识别方法.然而,既有方法要么未考虑到协变量分布偏移问题,要么依赖于严苛的假设难以在实际场景中奏效.当这些方法被用于协变量分布连续变化的真实测试环境中时,往往性能会严重退化,甚至不如开集识别基线方法.

1.2 测试时适应

测试时适应旨在仅利用无标注测试数据,使源模型不断适应于测试阶段变化的数据分布.Sun 等人[33]首先提出可以在测试阶段更新模型来解决数据分布偏移的问题.早期测试时训练工作[23,34]需要同时介入模型的训练与测试过程.这些工作在训练阶段将额外优化一个自监督学习目标,并在测试阶段继续优化这个目标来更新模型参数.Nado 等人[35]发现在测试阶段动态更新批标准化层中的统计信息,有利于提升模型对于数据分布偏移的稳健性.在此基础上,Wang 等人[21]提出了测试时适应方法,将熵最小化损失作为测试阶段模型的优化目标更新模型参数.Niu 等人[22]提出了一种基于样本选择的测试时适应方法,旨在提升测试时适应方法的计算效率.此外,Wang 等人[24]首次考虑了测试时适应算法持续地在测试环境中更新模型,导致模型性能退化的问题,并提出了一种能够持续更新模型的稳健算法.Gong[27]等人考虑了真实场景下测试时适应算法面对非独立同分布的测试数据可能遇到的稳健性问题,提出了一种基于缓冲区的稳健测试时适应算法.然而,既有方法均假设训练数据与测试数据的类别空间相同,一旦测试数据中出现训练数据从未见过类别的样本,这些方法将无法有效地使模型适应于变化的测试分布.

2 问题与分析

本章节首先介绍本文研究的开放世界适应问题的形式化.然后,针对开集识别问题在协变量分布偏移场景下应用所遇到的问题进行深入分析.

2.1 问题形式化

本文考虑输入空间为 $\mathsf{X} \in \mathbb{R}^{\mathbf{d}}$,标记空间为 $\mathsf{Y} = \{0,1\}^K$ 的多分类开集识别问题.其中,$\mathbf{d}$ 是一个向量表示输入空间的维度,K 表示类别的数量. X,Y 分别表示样本与标记的随机变量,$\mathsf{D}_t(X)$,$\mathsf{D}_t^{OS}(X)$ 分别表示在 t 时刻下的已见类样本与未见类样本的协变量分布.在开集识别问题中,模型不仅要准确分类已见类样本,同时还要识别并拒绝未见类样本.因此,开集识别模型 $f(\mathbf{x};\theta):\mathsf{X} \to [0,1]^{K+1}$ 的输出空间比标记空间略大,其中,前 K 个维度代表已见类别,第 $K+1$ 个维度代表未见类别.

本文研究的开放世界适应问题分为训练与测试两个阶段.在训练阶段,我们给定算法一个包含 N 个有标注数据的训练集 $\mathsf{D}_{train} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$.其中,$\mathbf{x}_i \in \mathsf{X}$ 表示从分布 $\mathsf{D}_0(X)$ 中采样得到的训练样本,$\mathrm{y}_i \in \{0,1\}^K$ 表示训练标注.开放世界适应算法需要利用有标注数据集 D_{train} 训练一个具有开集识别能力的模型 $f(\mathbf{x};\theta_0)$,使其能正确分类已见类别,同时准确识别未见类别.在测试阶段,测试样本的协变量分布 $\mathsf{D}_t(X)$ 在不同时刻 t 间连续变化.开放世界适应算法需要在线地执行开集识别任务,并利用无标注测试样本不断更新模型,使其适应于当前的数据分布.具体来说,在任意时刻 t 均有一批包含 N_t 个测试样本的集合 $D_t = \{\mathbf{x}_i\}_{i=1}^{N_t}$ 到达.其中,$\mathbf{x}_i$ 采样于已见类别与未见类别的组合分布 $\mathsf{D}_t(X) \cup \mathsf{D}_t^{OS}(X)$.开放世界适应算法需要首先给出测试样本集合 D_t 的开集预测结果,然后利

用 D_t 将模型参数 θ_{t-1} 更新为 θ_t,使其适应于当前协变量数据分布 $\mathsf{D}_t(X)$,以便模型 $f(\mathbf{x};\theta_t)$ 在后续时刻能给出更准确的预测结果.

2.2 问题分析

本章节详细分析了解决开放世界适应问题所遇到的关键问题:(1)协变量分布偏移导致模型性能退化;(2)测试时更新模型又受到未见类样本的影响.具体来说,2.2.1 节分析了既有开集识别方法面对协变量分布偏移时性能下降的问题,并介绍了能够在测试时更新模型使其适应于协变量分布偏移的测试时适应方法;2.2.2 节分析了测试时适应方法受到未见类样本影响,性能依旧退化的问题.

2.2.1 协变量分布偏移

图 2 分别在无分布偏移、分布偏移程度为 3 与分布偏移程度为 5 的 CIFAR10 数据集上进行实验,将基线开集识别方法、两种先进的开集识别方法、具有较强域泛化能力的开集识别方法和本文所提方法进行了对比.其中,左图展示了模型分类已见类别的性能,右图展示了模型识别未见类别的性能.实验结果证明了在开放世界适应问题中,一旦协变量分布偏移,无论是经典开集识别方法,还是基于域不变特征的稳健开集识别方法,都面临严重的性能下降问题.因此,仅凭静态模型是难以有效应对连续变化的协变量分布 $\mathsf{D}_t(X)$.测试时适应,可以在测试阶段利用无标注测试样本持续更新模型,使其适应于连续变化的协变量分布.既有研究[21,24]表明测试时适应技术能够有效地解决协变量分布偏移的问题.因此,本文考虑使用测试时适应来解决本文所提的开放世界适应问题.

Tent 是一种代表性的测试时适应方法,由于其使用方式简单、性能提升显著,取得了研究者们的广泛关注.既有研究[21]指出 Tent 在多个基准数据集上均取得了显著的性能提升.例如,在图像分类任务 CIFAR10 与 CIFAR100 的协变量分布偏移测试集上,Tent 方法对比基线方法,错误率相对降低 64.95%与 44.49%.Tent 方法在测试阶段利用无标注样本更新深度模型的批标准化层[36](Batch Normalization Layer,简称 BN Layer).批标准化层是深度学习广为使用的技术,它能解决数据内部的协变量偏移问题.定义 $\mathrm{r}\in\mathbb{R}^{B\times C\times L}$ 为一批样本的特征表示,其中, B,C,L 分别表示这批样本的数量、图片的通道数量、特征表示的维度.批标准化层的输出为:

$$BN\left(r_{:,c,:};\mu_c,\sigma_c^2\right)=\gamma\cdot\frac{r_{:,c,:}-\mu_c}{\sqrt{\sigma_c^2+\varepsilon}}+\beta \tag{1}$$

其中, γ,β 是批标准化层中的可学习参数,$\varepsilon>0$ 是一个小常数来保证批准化层中数值计算的稳定性, $\mu_c,\sigma_c^2\in\mathbb{R}^C$ 分别表示测试数据分布的均值与方差.在传统机器学习问题中,训练数据分布 $\mathsf{D}_0(X)$ 与测试数据分布 $\mathsf{D}_t(X)$ 相同.因此, μ_c,σ_c^2 可以直接从训练数据集 D_{train} 中估计得到.当测试数据的协变量偏移时 $\mathsf{D}_0(X)\neq\mathsf{D}_t(X)$,从训练数据集 D_{train} 中估计得到的均值与方差无法近似测试数据分布的均值与方差.Tent 方法假设测试数据分布 $\mathsf{D}_t(X)$ 连续变化,用上一时刻的测试数据分布 $\mathsf{D}_{t-1}(X)$ 近似当前时刻的测试数据分布 $\mathsf{D}_t(X)$.然后，其基于如下公式(2)利用最近一批测试样本动态估计 μ_c,σ_c^2：

$$\begin{cases}\mu_c=\dfrac{1}{BL}\displaystyle\sum_{b,l}r_{b,c,l},\\ \sigma_c^2=\dfrac{1}{BL}\displaystyle\sum_{b,l}\left(r_{b,c,l}-\mu_c\right)^2.\end{cases} \tag{2}$$

既有工作[35]证明这种简单有效估计方式能够有效地缓解测试数据分布中的协变量偏移问题.在此基础上,Tent 方法发现样本的预测正确率与熵值大小呈正相关.因此,其利用熵最小化损失 $\mathsf{L}_{ent}(\mathbf{x})$ 来优化批标准化层中的可学习参数 γ,β：

$$\mathsf{L}_{ent}(\mathbf{x})=-\sum_{k=1}^{K}f(\mathbf{x})^k\log f(\mathbf{x})^k \tag{3}$$

Tent 及一系列方法[21,22,27]均利用熵最小化损失在测试阶段对模型参数进行更新,显著地提升了深度学习模型面

对协变量分布偏移时的性能表现.

2.2.2 未见类样本影响

然而,既有测试时适应方法与开集识别问题不适配. 图 3 分别在无分布偏移、分布偏移程度为 3 与分布偏移程度为 5 的 CIFAR10 数据集上进行实验,将基线开集识别方法、将测试时适应方法与开集识别方法相结合的方案和本文所提方法进行了对比.其中,左图展示了模型分类已见类别的性能,右图展示了模型识别未见类别的性能.实验结果证明将测试时适应方法直接应用于现有开集识别模型反而会导致模型性能退化.一方面,由于测试数据中存在未见类样本,$t-1$ 时刻的测试数据分布为 $\lambda \mathsf{D}_t(X)+(1-\lambda)\mathsf{D}_t^{OS}(X)$.其中,λ 是未知常数,表示数据分布的混合比例. $t-1$ 时刻的测试数据分布无法近似 t 时刻已见类的数据分布 $\mathsf{D}_t(X)$.这使得基于公式(2)动态估计 μ_c,σ_c^2 的方式错误地将未见类别数据分布信息引入模型的批标准化层中,导致模型对已见类别的分类能力显著下降.另一方面,公式(3)中的熵最小化损失没有考虑测试阶段可能出现的未见类别.当模型使用熵最小化损失更新参数 γ,β 时,会错误地将未见类别样本归类于某个已见类别并更新模型.这将导致模型区分已见类与未见类样本的能力下降.基于上述分析,本文发现开集识别问题中的未见类样本将导致测试时适应技术中动态估计的统计信息 μ_c,σ_c^2 有偏、动态更新的可学习参数 γ,β 错误,分别损害开集识别模型分类已见类别、识别并拒绝未见类别的能力.因此,亟待开发适用于开集识别问题的测试时适应技术,解决上述两个关键难点,从而解决本文所提的开放世界适应问题.

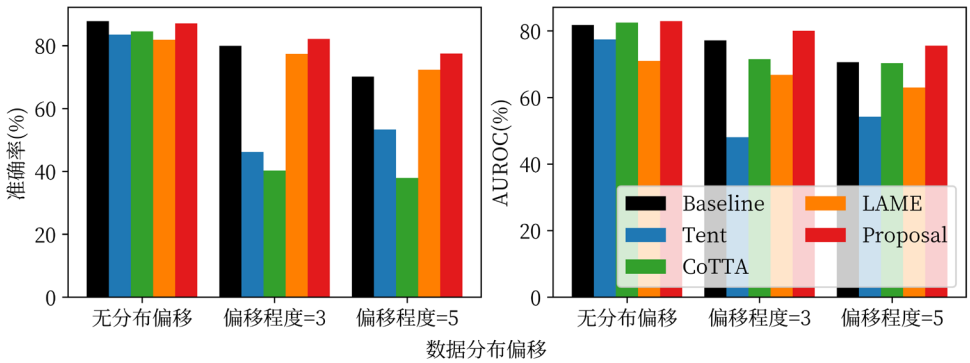

图 3 协变量分布偏移时,测试时适应方法与本文方法分类已见类与识别未见类的性能表现. 左图展示了模型分类已见类别的准确率;右图展示了模型识别了未见类样本的 AUROC

3 开放测试时适应方法 OTA

本章节针对开放世界适应问题中未见类样本的两个难点: (1)将导致测试时适应技术中动态估计的统计信息 μ_c,σ_c^2 有偏; (2)动态更新的可学习参数 γ,β 错误,提出了一种新颖的开放测试时适应方法 OTA,有效地提升了开集识别算法在协变量偏移情形下分类已见类别、识别并拒绝未见类别的性能表现.具体来说,我们提出一种自适应熵损失,在动态估计 μ_c,σ_c^2 与更新 γ,β 的过程中消除未见类样本对已见类分类的不利影响.进一步,我们结合轻量级的开集识别模块,提出了一种开集熵损失更新 γ,β ,帮助模型在测试阶段更准确地识别未见类样本.最终,我们利用模型参数正则化损失,防止模型在更新过程中出现灾难性遗忘现象.接下来,本文将分三个子章节分别介绍 OTA 算法中的三个关键技术.

3.1 自适应熵损失

开放世界适应问题中,测试阶段出现的未见类样本是导致测试时适应技术无法有效估计批标准化层中统计信息 μ_c,σ_c^2 并正确更新批标准化层中可学习参数 γ,β 的核心原因.因此,如何有效地找出测试数据中的已见类样本并将它们合理地用于模型更新是解决开放世界适应问题的关键.基于此动机,我们提出了一种自适应熵损失.首先,OTA 方法结合测试时增广技术与不确定性度量,从任意时刻 t 的测试数据 D_t 中筛选出较高置信度的

已见类样本 D_t^{Kn}.然后,OTA 方法使用高置信度的已见类样本集合 D_t^{Kn} 对统计信息 μ_c,σ_c^2 与可学习参数 γ,β 进行更新,有效地规避了未见类样本对开集识别模型分类已见类别能力的负面影响.

具体来说,我们首先基于测试时增广技术与基于熵的不确定性度量,提出了一种对协变量分布偏移更稳健的样本不确定性度量指标 $Unc(\mathbf{x})$:

$$Unc(\mathbf{x})=\max_{1\le k\le K}\frac{1}{P}\sum_{\mathrm{p}=1}^{P}f(A(\mathbf{x});\theta)^k \tag{4}$$

其中,超参数 P 控制测试时增广次数.增加数据增广次数 P,$Unc(\mathbf{x})$ 的稳健性提升,但 OTA 方法的时间开销也随之增加.当 P 设置为 1 时,不确定性度量指标 $Unc(\mathbf{x})$ 近似退化为不使用测试时增广技术的版本.在本文实验中,OTA 方法的超参数 P 被设置为 6,以兼顾 $Unc(\mathbf{x})$ 的稳健性与模型的时间开销.函数 $A(\mathbf{x})$ 是数据增广函数,为样本 $\mathbf{x}$ 生成一个相似却不同的增广版本 $\mathbf{x}'$.既有工作[37]表明测试时增广技术能够显著增加模型对于协变量分布偏移的稳健性.本文所提不确定性度量 $Unc(\mathbf{x})$ 首先将模型对样本 $\mathbf{x}$ 的 P 种不同增广版本的预测结果以平均的方式集成[38,39],然后利用最大 Logit 分数[13](Max Logit Score, 简称 MLS)度量样本 $\mathbf{x}$ 的不确定性.

基于 $Unc(\mathbf{x})$,OTA 方法进一步筛选得到的已见类样本集合 D_t^{Kn} 为:

$$D_t^{Kn}=\left\{\mathbf{x}_i\mid \mathbf{x}_i\in D_t\wedge Unc(\mathbf{x}_i)\ge\alpha\right\} \tag{5}$$

其中,α 为筛选置信已见类别样本的阈值.α 越大,集合 D_t^{Kn} 内的已见类样本越纯净,但已见类样本的数量随之降低.得益于本文所提不确定性度量 $Unc(\mathbf{x})$ 的稳健性,当超参数 α 在一定范围内变化时,OTA 方法的性能相对稳健.本文在 4.5 节的实验结果也证明了这一现象.

得到置信的已见类样本集合 D_t^{Kn} 后,我们利用 D_t^{Kn} 中的样本更新批标准化层中统计信息 μ_c,σ_c^2 与可学习参数 γ,β.具体来说,针对统计信息 μ_c,σ_c^2,我们使用基于动量的更新方式:

$$\begin{cases}\mu_c=\eta\dfrac{1}{BL}\displaystyle\sum_{b,l}r_{b,c,l}+(1-\eta)\mu_{old}\ ,\\ \sigma_c^2=\eta\dfrac{1}{BL}\displaystyle\sum_{b,l}\left(r_{b,c,l}-\hat{\mu}_c\right)^2+(1-\eta)\sigma_{old}^2.\end{cases} \tag{6}$$

其中,μ_{old},σ_{old}^2 分别表示特征表示均值与方差在更新前的值,η 是控制动量更新时新旧参数值的混合比例的超参数,r 表示集合 D_t^{Kn} 内样本的特征表示向量.在 OTA 方法中,η 根据批标准化层[36]内的默认值被设置为 0.1.针对可学习参数 γ,β,我们仅利用 D_t^{Kn} 内的样本计算公式(3)中的熵最小化损失,并利用梯度下降算法更新 γ,β.

3.2 开集熵损失

OTA 方法使用自适应熵损失,有效地防止未见类样本对估计统计信息 μ_c,σ_c^2 与更新参数 γ,β 的负面影响,提升了开集识别模型面对协变量偏移时分类已见类别的能力.然而,D_t^{Kn} 中仅包含测试样本集合 D_t 中有限的置信已见类样本,忽略了大量未见类样本,使得模型的开集识别性能无法达到最优.因此,我们引入一个轻量级的开集识别模块,帮助 OTA 方法筛选置信的未见类样本,并结合本文所提的开集熵损失进一步强化模型识别未见类样本的能力.

具体来说,轻量级的开集识别模块为 $g(\mathbf{x};\theta):\mathsf{X}\to[0,1]^{2K}$,其输出一个长度为 K 的二维向量 $[g(\mathbf{x};\theta)^i,1-g(\mathbf{x};\theta)^i]$,表示当前样本属于或不属于第 i 个已见类别的概率.由于开集识别模块 $g(\mathbf{x};\theta)$ 与模型 $f(\mathbf{x};\theta)$ 共享特征表示提取器,$g(\mathbf{x};\theta)$ 不会为模型训练与推理带来沉重负担,其参数量仅为 $2FK$.其中,F 为共享特征表示的维度.在训练阶段,OTA 方法使用经典的一对多损失优化 $\mathsf{L}_{ova}(\mathbf{x},y)$ [40]开集识别模块 $g(\mathbf{x};\theta)$:

$$\mathsf{L}_{ova}(\mathbf{x},y)=-\log\left(g(\mathbf{x};\theta)^y\right)-\min_{j\ne y}\log\left(1-g(\mathbf{x};\theta)^j\right) \tag{7}$$

$\mathsf{L}_{ova}(\mathbf{x},y)$ 使样本 $\mathbf{x}$ 属于真实类别 y 的概率 $g(\mathbf{x};\theta)^y$ 升高,使样本 $\mathbf{x}$ 不属于其他类别 $j,j\ne y$ 概率的最小值也升高.在测试阶段,如果样本 $\mathbf{x}$ 属于已见类别 y,那么其对应类别概率 $g(\mathbf{x};\theta)^y$ 较大;如果样本 $\mathbf{x}$ 属于未见类别,那么

属于所有已见类别的概率 $g(\mathbf{x};\theta)^k, 1 \le k \le K$ 都较小.

在测试阶段,OTA 方法首先利用所提开集识别模块 $g(\mathbf{x};\theta)$ 找到置信的未见类样本.我们定义样本属于未见类别的度量指标 $Osr(\mathbf{x})$ 为:

$$Osr(\mathbf{x}) = \frac{1}{P}\max_{1\le k\le K}\sum_{p=1}^{P} g(A(\mathbf{x});\theta)^k \tag{8}$$

其中,超参数 P 控制测试时增广次数,函数 $A(\mathbf{x})$ 为数据增广函数.继而,OTA 方法基于 $Osr(\mathbf{x})$ 指标从当前时刻 t 的测试样本集合 D_t 中筛选得到置信的未见类样本集合 D_t^{OS}:

$$D_t^{OS} = \{\mathbf{x}_i \mid \mathbf{x}_i \in D_t \wedge Osr(\mathbf{x}_i) \ge \delta\} \tag{9}$$

其中,δ 为筛选置信已见类别样本的阈值.最终,OTA 方法使用熵最大化损失 $\mathrm{L}_{r-ent}(\mathbf{x})$ 强化已见类别样本与未见类别样本在特征表示空间内的差异:

$$\mathrm{L}_{r-ent}(\mathbf{x}) = \sum_{k=1}^{K} f(\mathbf{x};\theta)^k \log f(\mathbf{x};\theta)^k \tag{10}$$

3.3 参数正则化损失

OTA 方法利用自适应熵损失与开集熵损失,从每个时刻 t 内测试数据 D_t 中筛选置信的已见类样本集合 D_t^{Kn} 与未见类样本集合 D_t^{OS},并分别利用公式(6)、(10)对模型进行更新.在实际情况中,集合 D_t^{Kn} 与集合 D_t^{OS} 仍存在少量错分样本,导致模型的更新过程不稳健.同时,近期研究[22,24]也发现仅利用无标注测试数据持续更新模型,会引发灾难性遗忘问题,导致模型性能逐渐退化.因此,OTA 方法在损失函数中引入参数正则化损失 $\mathrm{L}_r(\theta,\theta^s)$ 来约束模型不遗忘源模型的知识:

$$\mathrm{L}_r(\theta,\theta^s) = \| \theta - \theta^s \|_2^2 \tag{11}$$

其中,θ 是模型的参数,θ^s 是源模型的参数.

算法 1 开放测试时适应算法 OTA 伪代码

算法 1:开放测试时适应算法 OTA

输入:t 时刻模型参数 θ_t,源模型参数 θ^s,t 时刻测试数据 D_t

输出:下一时刻模型参数 θ_{t+1}

$D_t^{Kn} \leftarrow \{\mathbf{x}_i \mid \mathbf{x}_i \in D_t \wedge Unc(\mathbf{x}_i) \ge \alpha\}$ //选择置信已见类样本集合

$D_t^{OS} \leftarrow \{\mathbf{x}_i \mid \mathbf{x}_i \in D_t \wedge Osr(\mathbf{x}_i) \ge \delta\}$ //选择置信未见类样本集合

$\mathrm{L} \leftarrow 0$

for $\mathbf{x}_i \in D_t^{Kn}$ **do**

  $\mathrm{L} \leftarrow \mathrm{L} + \mathrm{L}_{ent}(\mathbf{x}_i)$ //计算自适应熵损失

end for

for $\mathbf{x}_i \in D_t^{OS}$ **do**

  $\mathrm{L} \leftarrow \mathrm{L} + \mathrm{L}_{r-ent}(\mathbf{x}_i)$ //计算开集熵损失

end for

$\mathrm{L} \leftarrow \mathrm{L} + \mathrm{L}_r(\theta,\theta^s)$ //计算参数正则化损失

$\theta_{t+1} \leftarrow$ 优化总体损失 L 并返回新参数

return θ_{t+1}

3.4 OTA方法总结

OTA 方法在测试阶段更新模型的总体损失是针对已见类样本 D_t^{Kn} 的自适应熵损失、针对未见类样本 D_t^{OS}

的开集熵损失和缓解灾难性遗忘的参数正则化损失之和,即 $L = L_{ent} + L_{r-ent} + L_r$ .OTA 的伪代码如算法 1 所示.

4 实验验证

本节将通过多个基准数据集中不同协变量偏移程度下的实验验证所提开放测试时适应算法 OTA 的有效性,并回答以下问题:

RQ1:OTA 方法是否对协变量分布偏移稳健,并给出优于既有方法的性能表现?

RQ2:OTA 方法提出的自适应熵损失与开集熵损失是否分别有效地提升了已见类别的分类性能与未见类别的识别性能?

4.1 实验数据集

我们选取两个测试时适应数据集 CIFAR10-C 与 CIFAR100-C[41],用于评估各种开集识别方法、测试时适应方法与所提 OTA 方法在不同协变量偏移程度下的性能表现.对于所有的实验,我们将在不存在协变量分布偏移的 CIFAR10 与 CIFAR100 数据集上训练源模型,然后将源模型部署于存在协变量偏移的环境中进行测试.其中,CIFAR10 数据集包含 10 个类别,每个类别包含 5000 张 32×32 的训练样本.在本文实验中,我们将 CIFAR10 中的 6 个动物类别作为已见类别,其他 4 个类别作为未见类别.CIFAR100 数据集包含 100 个类别,每个类别包含 500 张 32×32 的训练样本.在本文实验中,我们将 CIFAR100 中随机的 80 个类别作为已见类别,其他 20 个类别作为未见类别.CIFAR10-C 是 CIFAR10 数据集包含协变量偏移的版本,包含与 CIFAR10 相同的 10 个类别.CIFAR10-C 中包含 15 种不同的协变量偏移场景,每种场景中又存在 5 种不同等级的偏移程度.其中,协变量偏移程度等级由加入样本中的自然噪声强度决定,由弱至强分别对应 1 至 5 五个等级.类似的,CIFAR100-C 是 CIFAR100 数据集包含协变量偏移的版本.本文选取偏移程度为 3 与 5 这两种情况分别进行实验.在实验中,所测试的模型将会依次预测 15 种不同的协变量偏移场景,并评估其已见类别的分类性能与未见类别的检测性能.

4.2 对比方法

为了证明本文所提 OTA 方法的先进性,我们选取了代表性的开集识别方法、代表性的测试时适应方法作为对比方法:

- MLS[13]: Max Logit Score 是一类经典的开集识别方法,其利用模型输出的 Logit 最大值来判断样本是否属于已见类别.MLS 是一个后处理方法,具有较强的通用性,可以应用于任意深度学习模型.既有工作[13]证明 MLS 方法相比经典的 MSP 方法[42]具有更好的开集识别性能.因此,在本文在实验中选取 MLS 方法作为对比方法.在本文中,我们也将 MLS 称为做基线方法,也称为 Baseline.
- APRL[43]: APRL 方法在所学习的特征空间中定义“互补点”的概念.样本属于某个某个类别的概率正比于其与所学互补点的距离.未见类样本由于与所有已见类别均不同,其距离所有互补点的距离更大.基于这个假设,APRL 用测试样本距离互补点的最大距离来度量其属于未见类的程度.
- ARPL+cs[16]: APRL+cs 方法在 APRL 方法的基础上,利用生成对抗网络在训练过程中生成虚拟的未见类样本,从而帮助模型学得一个更容易区分已见类别与未见类别的特征表示空间.
- Proser[15]: Proser 方法在训练过程中利用 MixUP 技术基于已见类别生成虚拟的未见类别样本,在学习的过程中,利用虚拟的未见类别样本使得已见类别的决策边界更加紧致,从而使模型获得更优的开集识别性能.
- SC-OSG[32]: SC-OSG 方法是第一个考虑开集识别模型可能在测试阶段中遭遇协变量偏移问题的工作.其结合因果学习技术,利用域不变特征完成开集识别任务,提升了模型对协变量偏移的稳健性.然而,SC-OSG 由于无法在连续演变的测试环境中不断更新模型,因此,在实际情况中性能仍有提升空间.
- BN Stats[35]: BN Stats 是经典的测试时适应方法,其在测试环境中依旧动态更新批标准化层中的统计信息,来适应测试环境中的协变量偏移问题.
- Tent[21]: Tent 是利用熵最小化损失更新模型参数的测试时适应方法.其在 BN Stats 更新批标准化层统

计信息的基础上,利用熵最小化损失同时更新批标准化层中的可学习参数,使模型进一步适应测试数据中的协变量偏移.

- EATA[22]: EATA 是一种高效的测试时适应方法.EATA 方法在 Tent 方法的基础上,引入了自适应的样本选择技术,剔除测试数据中对更新存在负面影响的样本.同时,其结合防止灾难性遗忘的技术,缓解模型在测试环境中逐渐遗忘源模型知识的问题.在本文的实验中,我们通过对比 EATA 方法来验证既有的自适应样本选择技术是否能够有效的消除未见类样本带来的负面影响.
- LAME[25]: LAME 是一个无需更新模型参数的测试时适应方法.LAME 方法使用半监督学习中标记传播算法的目标式,直接对模型输出概率进行优化.由于无需更新模型参数,LAME 方法解决了既有测试时适应方法在测试环境中连续更新导致性能退化的问题.
- CoTTA[24] :CoTTA 考虑模型在测试环境中连续适应到不同协变量分布的情形,结合模型指数集成技术、伪标记修正技术与防止灾难性遗忘的技术.目前,CoTTA 在测试时适应领域取得了稳健且先进的性能表现.

4.3 实验细节

本文采用残差神经网络[44](Residual Network,简称 ResNet)作为分类器的主干网络,网络的深度设置为 50.对于所有的算法,我们均采用原始论文中推荐的超参数对模型进行训练与测试.测试时适应方法所使用的源模型,使用传统监督学习训练 200 轮次得到.在训练过程中,图像批大小设置为 256,神经网络的学习率设置为 0.1 并在学习的过程中使用余弦退火的方式动态调整.针对实验中使用的测试时适应方法与本文提出的 OTA 方法,我们均使用后处理 MLS 方法识别未见类别.本文的所有实验均使用{0,1,2,3,4}五个随机种子重复运行五次,并汇报性能的均值与标准差. 我们使用 Close-set Accuracy、AUROC 分别来评估模型分类已见类别的能力、区分已见类别与未见类别的能力.进一步,我们还汇报了使用 OSCR[45](Open-set Classification Rate)指标评估的结果,综合性地权衡了模型分类已见类别与识别未见类别的能力.

4.4 实验结果与分析

RQ1:OTA 方法是否对协变量分布偏移稳健,并给出优于既有方法的性能表现?

为了回答这个问题,我们在协变量偏移等级为 3 和 5 的 CIFAR10 与 CIFAR100 数据集上进行实验,并汇报了本文所提 OTA 方法与对比方法的性能表现.CIFAR10 数据集上的结果如表 1 所示,CIFAR100 数据集上的结果如表 2 所示.表格中第一行展示了,使用监督学习在无偏数据集上训练得到的深度学习模型,再结合 MLS 基线方法的性能表现.由于 MLS 方法的在训练过程中并未主动考虑未见类别并在测试过程中也未主动适应偏移的协变量分布.因此,MLS 方法可以作为本文研究问题的基线方法.然而,实验结果表明,当测试数据存在协变量分布偏移的情况下,既有开集识别方法与测试时适应方法在多数情况下都劣于基线 MLS 方法.这说明既有的开集识别方法与测试时适应方法均无法有效地处理协变量分布偏移的问题,导致性能显著退化并不如基线方法.表 1 与表 2 中的结果表明,本文提出的 OTA 方法在所有情形下都显著优于基线 MLS 方法,这说明 OTA 方法能不断地适应于变化的协变量分布,有效地解决了测试数据中协变量分布偏移的问题.同时,OTA 方法的性能也显著优于全部的对比方法,这证明了本文所提方法的优越性并有力地回答了 RQ1.

此外,我们在图 4 中展示了本文所提 OTA 方法与对比方法在协变量分布偏移程度为 3 的 CIFAR10 数据集上的详细性能表现.图 4 左侧展示了 OTA 方法与既有开集识别方法在 15 种偏移上的性能表现,结果表明 OTA 方法在绝大多数情况下都好于基线 MLS 方法与其他开集识别方法.图 4 右侧展示了 OTA 方法与既有测试时适应方法在 15 种偏移上的性能表现.不更新模型参数的 BN Stats 与 LAME 方法虽然性能仍比不过基线 MLS 方法,但是未出现严重的性能退化情况.更新模型参数的 Tent 方法与 CoTTA 方法性能严重下降,这说明既有测试时适应方法在开集识别问题设定下无法有效地更新模型参数.EATA 方法通过样本选择技术,一定程度上缓解了性能退化问题,但其性能仍比不过基线 MLS 方案.这说明样本选择技术有利于剔除未见类样本在参数更新过程中的负面影响,但在开集识别问题设定下,EATA 的选择技术仍有极大的优化空间.OTA 方法性能优于全部的

测试时适应方法,证明了本文所提方法的先进性.

表 1 存在不同协变量偏移程度的 CIFAR10 数据集上的性能对比.最优的结果加粗标注,对比基线 MLS 方法性能退化的结果用下划线标注.实验结果显示 OTA 方法显著优于对比方法.

对比方法	偏移程度=3			偏移程度=5		
	Close-set Acc	AUROC	OSCR	Close-set Acc	AUROC	OSCR
MLS	79.92±0.00	77.17±0.00	66.82±0.00	70.17±0.00	70.58±0.00	55.96±0.00
ARPL	70.45±0.22	71.60±0.31	56.70±0.32	55.86±0.30	63.52±0.75	42.10±0.49
ARPL+cs	70.57±0.21	71.77±0.56	56.97±0.38	56.18±0.46	64.03±0.41	42.59±0.40
Proser	74.27±0.93	72.17±1.37	59.55±1.60	58.10±1.11	64.26±1.18	43.39±1.25
SC-OSG	72.27±0.75	70.67±1.81	57.77±0.72	55.68±1.38	60.79±1.72	41.05±0.66
BN Stats	79.55±0.04	77.46±0.06	66.77±0.07	75.72±0.04	74.58±0.03	62.12±0.03
Tent	46.18±11.29	48.04±10.25	21.95±3.09	53.34±9.48	54.21±7.03	30.29±7.41
EATA	79.89±0.07	77.18±0.13	66.80±0.13	75.99±0.08	74.44±0.08	62.23±0.05
LAME	77.37±0.09	66.77±0.12	57.40±0.15	72.34±0.06	62.99±0.07	51.33±0.10
CoTTA	40.31±0.53	71.48±1.17	29.98±0.34	37.95±0.83	70.30±1.11	28.00±0.63
OTA	**82.14±0.02**	**80.02±0.03**	**70.77±0.02**	**77.45±0.07**	**75.58±0.07**	**64.21±0.07**

表 2 存在不同协变量偏移程度的 CIFAR100 数据集上的性能对比.最优的结果加粗标注,对比基线 MLS 方法性能退化的结果用下划线标注.实验结果显示 OTA 方法显著优于对比方法.

对比方法	偏移程度=3			偏移程度=5		
	Close-set Acc	AUROC	OSCR	Close-set Acc	AUROC	OSCR
MLS	51.30±0.00	62.08±0.00	39.09±0.00	38.08±0.00	58.54±0.00	28.60±0.00
ARPL	44.81±0.53	58.77±0.28	32.59±0.31	31.48±0.52	56.17±0.29	22.72±0.34
ARPL+cs	45.31±0.59	58.58±0.33	32.87±0.47	32.08±0.66	56.24±0.23	23.16±0.55
Proser	46.63±0.90	62.79±0.68	36.37±0.70	31.01±0.83	58.44±0.54	23.59±0.60
SC-OSG	46.61±0.87	60.86±0.39	35.44±0.79	30.50±0.94	56.77±0.45	22.94±0.91
BN Stats	56.09±0.06	63.94±0.03	43.16±0.04	49.78±0.03	61.89±0.02	37.96±0.03
Tent	25.03±2.82	54.68±0.61	17.05±2.08	15.82±2.77	53.26±0.88	10.59±2.01
EATA	58.58±0.05	64.54±0.07	44.95±0.05	52.40±0.15	62.49±0.09	39.81±0.14
LAME	55.94±0.05	63.77±0.10	43.77±0.04	49.48±0.06	61.17±0.07	37.92±0.08
CoTTA	10.50±0.34	51.36±0.77	6.91±0.19	9.28±0.17	51.52±0.46	6.11±0.08
OTA	**59.45±0.06**	**65.33±0.03**	**47.06±0.04**	**53.89±0.05**	**63.13±0.04**	**41.93±0.03**

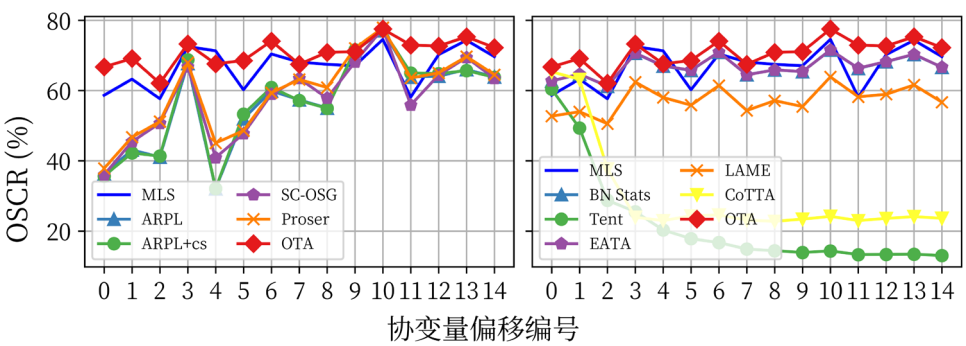

图 4 OTA 方法与开集识别方法、测试时适应方法的详细性能比较

RQ2：OTA 方法提出的自适应熵损失与开集熵损失是否分别有效地提升了已见类别的分类性能与未见类别的识别性能？

表 3 OTA 方法的消融实验

OTA 组件		分布偏移程度=3			分布偏移程度=5		
自适应熵损失	开集熵损失	Close-set Acc	AUROC	OSCR	Close-set Acc	AUROC	OSCR
		79.24	73.94	64.50	69.33	67.71	53.44
√		**83.03**	77.71	69.80	**78.61**	73.89	63.94
√	√	82.14	**80.15**	**70.88**	77.59	**75.70**	**64.39**

为了回答这个问题,我们对 OTA 方法进行消融实验,逐个验证 OTA 方法每个部件的有效性.如表 3 中所展

示的结果,当 OTA 方法加入自适应熵损失后,其分类已见类别的准确率显著提升.这说明利用自适应熵损失,OTA 方法能够稳健的选择置信的已见类别样本并利用所选样本更新模型.值得一提的是,在更新过程中,模型的特征表示也被更新了,因此,模型区分已见类别与未见类别的性能也相应得到了提升.当OTA方法加入开集熵损失后,模型的区分已见类别与未见类别的性能得到了进一步的提升,但是其分类已见类别的准确率略有下降.这不意味着开集熵损失对模型性能有害,因为模型容量一定时,分类已见类的准确率与识别未见类的性能之间天然存在权衡.OTA 方法加入开集熵损失后,能够综合衡量模型分类已见类与识别未见类能力的指标 OSCR 有所提升,这说明模型的综合能力在加入开集熵损失后有所提升.表 3 的实验结果证明了只有将 OTA 方法的几个部件组合在一起,才能够达到最优的性能.

4.5 其他讨论

超参数鲁棒性.OTA 方法在选择置信的已见类样本集合 D_t^{Kn} 与未见类样本集合 D_t^{OS} 分别使用超参数 α,δ.在本文的所有实验中,我们将 OTA 方法的超参数统一设置为 $\alpha=0.9,\delta=0.3$,证明了 OTA 方法的超参数值设置对于具体数据集不敏感.图 5 分别展示了超参数 α 在{0.85, 0.86, 0.87, 0.88, 0.89, 0.90, 0.91, 0.92, 0.93, 0.94}间取值、超参数 δ 在{0.25, 0.26, 0.27, 0.28, 0.29, 0.30, 0.31, 0.32, 0.33, 0.34}间取值的的性能表现.图 5 中的结果表明,OTA 方法对于超参数值的设置鲁棒,即使超参数 α,δ 在本文推荐值周围扰动,OTA 方法的性能也不会受到大幅影响.

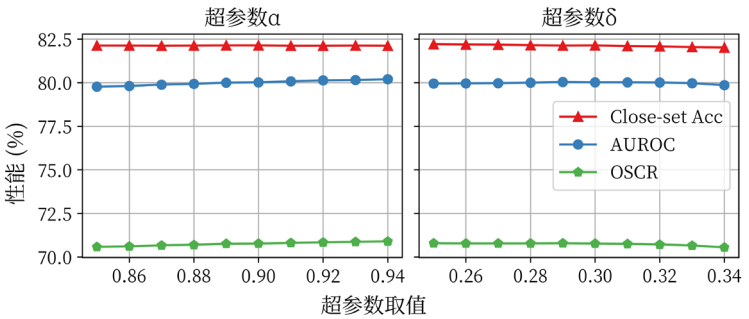

图 5 超参数鲁棒性分析

特征可视化.为了判断 OTA 方法是否有效地使模型适应于变化的协变量分布,我们将 MLS 方法、Proser 方法与本文所提的 OTA 方法的特征表示使用 T-SNE 算法在图 6 中进行可视化.由于 MLS 方法与 Proser 方法并未对模型及时更新.它们在面对协变量分布偏移时,未见类样本与已见类样本混合严重,同时,不同已见类样本间也混合严重.本文所提 OTA 方法的特征表示的判别能力明显更优,已见类别分布在表示空间的不同区域,同时,已见类与未见类样本间也有明显界限.这证明本文所提OTA方法能够有效地更新模型,使其适应于变化的协变量分布.

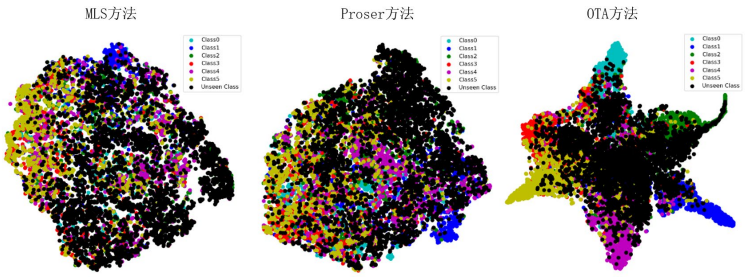

图 6 MLS 方法、Proser 方法与 OTA 方法的特征表示可视化

运行时间.我们进一步研究了 OTA 算法与对比方法的运行时间.在表 4 中,我们展示了基线开集识别算法

MLS,测试时适应算法 Tent、LAME 与 CoTTA 和本文所提算法 OTA 在预测一种协变量分布偏移情形下全部样本的运行时间.表 4 中的结果表明,OTA 算法的运行时间与基线算法、测试时适应算法的运行时间接近,并没有数量级上的差异,但相对于基线方法与对比方法能够有效地提升性能表现.因此,表 4 中的结果说明了 OTA 方法能够在测试阶段以较低的资源使模型适于协变量分布变化,避免了重新训练模型的资源开销与数据收集成本.

表 4 算法的运行时间

算法名称	MLS	Tent	LAME	CoTTA	OTA
运行时间	4.96s	11.77s	22.21s	329.07s	44.21s
OSCR	55.96%	30.29%	51.33%	28.00%	64.21%

5 总结

开集识别是机器学习的重要问题之一,其旨在准确分类已见类别的同时,识别并拒绝未见类别.然而现有开集识别方法在面对协变量分布偏移的问题时,面临严重的性能下降问题,其性能表现甚至不如基线方法.基于这一观察,本文提出了开放世界适应问题 AOW,旨在使开集识别模型稳健地分类已见类别并拒绝未见类别的同时,还不断更新模型使其适应于变化的协变量分布.针对此问题,我们设计了开放测试时适应方法 OTA.该方法利用自适应熵损失和开放熵损失在测试时自适应地更新模型.一方面,它消除了未见类样本在更新过程中对已见类判别能力的不利影响;另一方面,它利用未见类样本加强了模型对未见类别的识别能力.此外,OTA 方法还利用了参数正则化损失,以防止模型在更新过程中出现灾难性的遗忘问题.在不同偏移程度的基准数据集上的实验验证了 OTA 方法相比已有的开集识别方法和测试时适应方法具有更先进的性能.

References:

[1] Deng J, Dong W, Socher R, *et al*. Imagenet: a large-scale hierarchical image database. In: Proc. of the 2009 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[2] Reddy DR. Speech recognition by machine: a review. In: Proc. of the IEEE, 1976, 64(4): 501–531. [doi:10.1109/PROC.1976.10158]

[3] Stahlberg F. Neural machine translation: a review. Journal of Artificial Intelligence Research, 2020, 69: 343–418. [doi:10.1613/jair.1.12007]

[5] Neal L, Olson M, Fern X, *et al*. Open set learning with counterfactual images. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 613–628. [doi: 10.1007/978-3-030-01231-1_38]

[7] Zhou Z, Guo L-Z, Cheng Z-Z, *et al*. STEP: out-of-distribution detection in the presence of limited in-distribution labeled data. Advances in Neural Information Processing Systems. 2021. 29168–29180.

[8] Zhou D-W, Yang Y, Zhan D-C. Learning to classify with incremental new class. IEEE Trans. on Neural Networks and Learning Systems, 2021, 33(6): 2429–2443. [doi:10.1109/TNNLS.2021.3104882]

[9] Guo L-Z, Zhang Z-Y, Jiang Y, *et al*. Safe deep semi-supervised learning for unseen-class unlabeled data. In: Proc. of the 37th Int'l Conf. on Machine Learning. 2020. 3897–3906.

[10] Wong K, Wang S, Ren M, *et al*. Identifying unknown instances for autonomous driving. In: Proc. of the Conf. on Robot Learning. 2020. 384–393.

[11] Geng C, Huang S, Chen S. Recent advances in open set recognition: a survey. IEEE Trans on Pattern Analysis and Machine Intelligence, 2020, 43(10): 3614–3631. [doi:10.1109/TPAMI.2020.2981604]

[12] Zhu Y-N, Li Y-F. Semi-supervised streaming learning with emerging new labels. In: Proc. of the 34th AAAI Conf.on Artificial Intelligence. 2020. 7015–7022. [doi:10.1609/aaai.v34i04.6186]

[13] Vaze S, Han K, Vedaldi A, *et al*. Open-set recognition: a good closed-set classifier is all you need? In: Proc. of the 10th Int'l Conf. on Learning Representations. 2022.

[14] Bendale A, Boult TE. Towards open set deep networks. In: Proc. of the 2016 IEEE/CVF Conf. on Computer Vision and Pattern

Recognition (CVPR). 2016. 1563–1572. [doi: 10.1109/CVPR.2016.173]

[15] Zhou D-W, Ye H-J, Zhan D-C. Learning placeholders for open-set recognition. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2021. 4401–4410. [doi:10.1109/CVPR46437.2021.00438]

[16] Chen G, Peng P, Wang X, Tian Y. Adversarial reciprocal points learning for open set recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(11): 8065–8081. [doi:10.1109/TPAMI.2021.3106743]

[17] Sun X, Yang Z, Zhang C, *et al*. Conditional gaussian distribution learning for open set recognition. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2020. 13480–13489. [doi:10.1109/CVPR42600.2020.01349]

[18] Geng C, Chen S. Collective decision for open set recognition. IEEE Transactions on Knowledge and Data Engineering, IEEE, 2020, 34(1): 192–204. [doi:10.1109/TKDE.2020.2978199]

[19] Shao J-J, Guo L-Z, Yang X-W, *et al*. LOG: active model adaptation for label-efficient ood generalization. Advances in Neural Information Processing Systems. 2022.

[20] Guo L-Z, Zhou Z, Li Y-F. RECORD: resource constrained semi-supervised learning under distribution shift. In: Proc. of the 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2020. 1636–1644. [doi:10.1145/3394486.3403214]

[21] Wang D, Shelhamer E, Liu S, *et al*. Tent: fully test-time adaptation by entropy minimization. In: Proc. of the 8th Int'l Conf. on Learning Representations. 2020.

[22] Niu S, Wu J, Zhang Y, *et al*. Efficient test-time model adaptation without forgetting. In: Proc. of the 39th Int'l Conf. on Machine Learning. 2022. 16888–16905.

[23] Bartler A, Bühler A, Wiewel F, *et al*. MT3: meta test-time training for self-supervised test-time adaption. In: Proc. of the 25th Int'l Conf. on Artificial Intelligence and Statistics. 2022. 3080–3090.

[24] Wang Q, Fink O, Van Gool L, *et al*. Continual test-time domain adaptation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2022. 7191–7201. [doi:10.1109/CVPR52688.2022.00706]

[25] Boudiaf M, Mueller R, Ayed IB, *et al*. Parameter-free online test-time adaptation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2022. 8334–8343. [doi:10.1109/CVPR52688.2022.00816]

[26] Niu S, Wu J, Zhang Y, *et al*. Towards stable test-time adaptation in dynamic wild world. In: Proc. of the 11th Int'l Conf. on Learning Representations. 2023.

[27] Gong T, Jeong J, Kim T, *et al*. Robust continual test-time adaptation: instance-aware bn and prediction-balanced memory. Advances in Neural Information Processing Systems. 2022.

[28] Scheirer WJ, De Rezende Rocha A, Sapkota A, Boult TE. Toward open set recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012, 35(7): 1757–1772. [doi:10.1109/TPAMI.2012.256]

[29] Jain LP, Scheirer WJ, Boult TE. Multi-class open set recognition using probability of inclusion. In: Proc. of the European Conf. on Computer Vision (ECCV). 2014. 393–409. [doi:10.1007/978-3-319-10578-9_26]

[30] Ge Z, Demyanov S, Chen Z, *et al*. Generative openmax for multi-class open set classification. In: Proc. of the British Machine Vision Conf. 2017.

[31] Oza P, Patel VM. C2ae: class conditioned auto-encoder for open-set recognition. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). 2019. 2307–2316. [doi:10.1109/CVPR.2019.00241]

[32] Shao J-J, Yang X-W, Guo L-Z. Open-set learning under covariate shift. Machine Learning, 2022. [doi:10.1007/s10994-022-06237-1]

[33] Sun Y, Wang X, Liu Z, *et al*. Test-time training with self-supervision for generalization under distribution shifts. In: Proc. of the 37th Int'l Conf. on Machine Learning. 2020. 9229–9248.

[34] Liu Y, Kothari P, Van Delft B, *et al*. TTT++: when does self-supervised test-time training fail or thrive? Advances in Neural Information Processing Systems. 2021. 21808–21820.

[35] Schneider S, Rusak E, Eck L, *et al*. Improving robustness against common corruptions by covariate shift adaptation. Advances in Neural Information Processing Systems. 2020. 11539–11551.

[36] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proc. of the 32nd Int'l Conf. on Machine Learning. 2015. 448–456.

[37] Shanmugam D, Blalock D, Balakrishnan G, *et al*. Better aggregation in test-time augmentation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). 2021. 1214–1223. [doi:10.1109/ICCV48922.2021.00125]

[38] Phung VH, Rhee EJ. A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. Applied Sciences, 2019, 9(21): 4500. [doi:10.3390/app9214500]

[39] Zhou Z-H. Ensemble learning. Springer, 2021.

[40] Saito K, Saenko K. OVANet: one-vs-all network for universal domain adaptation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). 2021. 8980–8989. [doi:10.1109/ICCV48922.2021.00887]

[41] Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. In: Proc. of the 7th Int'l Conf. on Learning Representations. 2019.

[42] Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2017.

[43] Chen G, Qiao L, Shi Y, *et al*. Learning open set network with discriminative reciprocal points. In: Proc. of the European Conf. on Computer Vision (ECCV). 2020. 507–522. [doi:10.1007/978-3-030-58580-8_30]

[44] He K, Zhang X, Ren S, *et al*. Identity mappings in deep residual networks. In: Proc. of the European Conf. on Computer Vision (ECCV). 2020. 507–522. 2016. 630–645. [doi:10.1007/978-3-319-46493-0_38]

[45] Dhamija AR, Günther M, Boult T. Reducing network agnostophobia. Advances in Neural Information Processing Systems. 2018. 9175–9186.

附中文参考文献:

[4] 窦文,王怀民,贾焰,邹鹏.构造基于推荐的 Peer-to-Peer 环境下的 Trust 模型.软件学报,2004,15(4):571-583. http://www.jos.org.cn/1000-9825/15/571.htm

[6] 朱鹏飞,张琬迎,王煜,胡清华.考虑多粒度类相关性的对比式开放集识别方法.软件学报,2022,33(4):1156-1169. https://www.jos.org.cn/html/2022/4/6468.htm [doi:10.13328/j.cnki.jos.006468]