基于深度学习的类别增量学习算法综述

周大蔚 汪福运 叶翰嘉 詹德川
(计算机软件新技术国家重点实验室(南京大学) 南京 210023)

摘 要 近年来,深度学习模型在众多领域取得了广泛成功.现有的深度学习模型大多部署在静态环境下,依赖提前收集好的数据集进行离线训练,模型一经确定,便无法进一步更新.然而,现实中开放动态的环境往往存在以流形式不断到来的数据,包括随时间演进不断产生的新类别数据.因此,理想的机器学习模型应能够从流式数据中不断学习新类,从而增强自身的判别能力.这样的学习范式被称作"类别增量学习"(class-incremental learning),且近年来已成为机器学习领域的研究热点.面对流式数据,直接使用新类别样本训练模型会使其遗忘旧类别的数据,造成整体性能的下降.因此,设计增量学习模型时,需确保模型在学习新类的同时也能够抵抗灾难性遗忘.本文从机器学习的三个重要方面(数据层面、参数层面、算法层面)着眼,总结和归纳近几年基于深度学习的类别增量学习算法.此外,本文还在基准数据集上对 10 种典型算法进行了实验验证,并从中总结出适应类别增量学习的一般性规律.最后,本文对基于深度学习的类别增量学习算法目前存在的挑战加以分析,并展望未来的发展趋势.

关键词 类别增量学习;持续学习;开放动态环境;灾难性遗忘;模型复用
中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2023.01577

Deep Learning for Class-Incremental Learning: A Survey

ZHOU Da-Wei WANG Fu-Yun YE Han-Jia ZHAN De-Chuan
(*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023)

Abstract Recent years have witnessed the progress of deep learning in many fields, e.g., image classification and face recognition. Current deep models are deployed under the static environment, which requires collecting all the training data before the learning process. The deep model is unable to conduct further updating processes when the training process is terminated. However, data in the real world often come in stream format, which contains incoming instances from new classes. For example, in the opinion monitoring system, new topics will emerge as time goes by; in the electronic commerce platform, new types of products will arise day by day; in the robot learning scenario, the robot is required to learn new orders continually. As a result, an ideal model should learn from stream data and enhance its learning ability incrementally. Such a learning process, namely Class-Incremental Learning (CIL), is now drawing more and more attention from the machine learning community. Directly updating the incremental model with new class data will cause the forgetting of old ones and destroy the total performance, which is denoted as catastrophic forgetting in literature. As a result, the class-incremental learning model should incorporate new classes and meanwhile resist catastrophic forgetting over old ones. In this paper, we summarize and classify recent deep-leering-based class-incremental learning algorithms from three aspects, i.e., input, parameter, and algorithm. Typical class-incremental learning methods from the input

收稿日期:2022-04-08;在线发布日期:2022-10-08.本课题得到国家自然科学基金(61773198,61921006,62006112)、国家自然科学基金委员会与韩国国家研究基金会合作研究项目(61861146001)、江苏省自然科学基金(BK20200313)和计算机软件新技术协同创新中心资助.
周大蔚,博士研究生,中国计算机学会(CCF)会员,主要研究方向为增量学习、开放集识别.E-mail:zhoudw@lamda.nju.edu.cn.**汪福运**,本科生,主要研究方向为增量学习.**叶翰嘉**(通信作者),博士,副研究员,主要研究方向为度量学习、元学习.E-mail:yehj@lamda.nju.edu.cn.**詹德川**,博士,教授,主要研究领域为机器学习、数据挖掘.

计算机学报

aspect try to solve incremental learning tasks by regularizing and rehearsing the exemplar set, which can be divided into data replay-based and data restriction-based methods. Similar to the human learning process, data replay-based CIL methods aim to replay former instances when learning new ones, which obtain a trade-off between learning new knowledge and remembering old ones. Data restriction-based methods utilize the former examples as the regularization to restrict the direction of model updating. Class-incremental learning methods from the parameter aspect try to solve incremental learning tasks by regularizing model updating and adjusting network structure, which can be divided into parameter regularization-based and dynamic architecture-based methods. Parameter regularization-based methods weigh the importance of each parameter and restrict important parameters from being changed to overcome forgetting. On the other hand, dynamic architecture-based methods aim to dynamically adjust the network structure to meet the requirements of incoming new classes. Class-incremental learning methods from the algorithm aspect try to solve incremental learning tasks by model mapping and reducing inductive bias, which can be divided into knowledge distillation-based and post-tuning-based methods. Knowledge distillation-based CIL methods utilize the former model as the teacher to restrict the updating process of the current model. Post-tuning-based methods try to reduce the bias in the incremental model to get an unbiased prediction. In this paper, we conduct extensive experimental verification with ten typical algorithms on the benchmark datasets, i. e., CIFAR 100 and ImageNet ILSVRC2012. We analyze the behaviors of incremental models, including the accuracy trend, running time, memory budget, performance decay, and confusion matrix. We also summarize the common rules for class-incremental learning algorithms. Finally, we analyze the challenges and future trends and conclude this paper.

Keywords class-incremental learning; continual learning; dynamic environment; catastrophic forgetting; model reuse

1 引 言

近年来,机器学习方法在众多领域取得了显著的成效,并被广泛应用到图片分类[1]、聚类[2-3]、图像检索[4-5]、用户商品推荐[6-7]等场景中.传统的机器学习模型要求在训练前取得所有训练样本以进行离线训练,并且在训练结束后无法继续更新.然而,在开放动态环境中,训练样本往往以数据流的形式到来[8],或因存储、隐私等问题仅在一段时间内可以获得[9].理想情况下的机器学习模型应当能够仅利用数据流中的新样本更新模型,而无需耗费大量计算资源进行重新训练.因此,增量学习①(incremental learning)这一概念被提出,旨在设计具有持续学习能力的机器学习模型.面对不断到来的新数据,直接使用它们更新模型会引发灾难性遗忘[10-11](catastrophic forgetting)——模型在学习新数据的同时会遗忘以往学得旧数据的模式,失去对旧数据的判别能力,从而导致模型分类性能的急剧下降.因此,如何在模型持续学习新数据的同时抵抗对旧数据的灾难性遗忘便成为增量学习问题的研究重点.

真实世界中的场景往往会随时间演进产生新的类别[12-17].例如,在社交媒体中,新类型的新闻事件层出不穷[18-19];在电商平台上,新类型的商品会不断涌现[20].机器学习模型不断学习新增的类别无疑会遭受灾难性遗忘.这种新类别在数据流中不断到来的增量学习场景被称为类别增量学习(class-incremental learning).理想的学习模型应当能按顺序地学习一系列不断到来的新类,从而使自身的判别能力不断增强——这种学习过程和人类学习新事物的过程存在共性[21-22].图 1 展示了类别增量学习的训练和测试过程:模型首先在任务 1 上进行训练,学习分类鸟和水母.之后,需要基于当前模型分别在任务 2 中学习鹅和北极狐、在任务 3 中学习狗和螃蟹.按顺序完

① 又被称作持续学习(continual/continuous learning)、连续学习(sequential learning),或终身学习(lifelong learning),本文采用最普遍的命名方式,称其为增量学习.

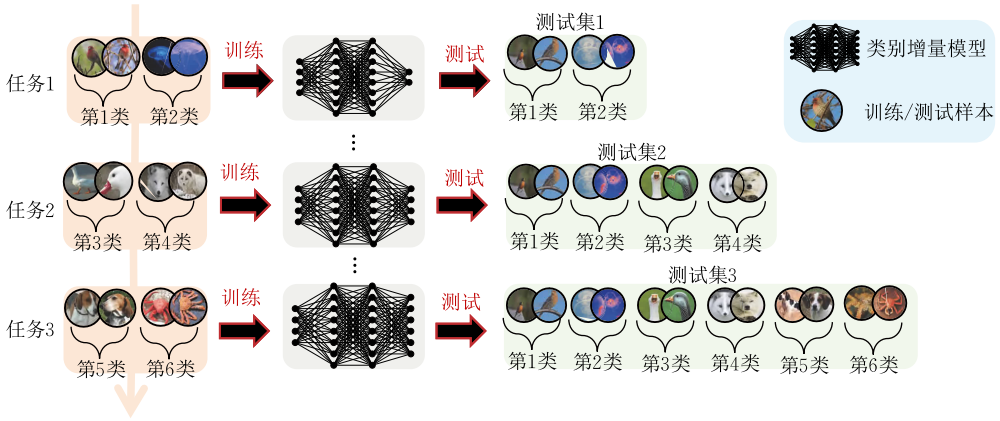


图 1　类别增量学习问题设定(模型需要按顺序学习不断到来的新类别,并在所有见过的类别上进行测试)

成训练后,模型需要在所有训练过的类别上进行评估,一个优秀的类别增量模型能既学得新类知识,又不遗忘旧类知识.在类别增量学习过程中,学习新类意味着模型要尽可能适配新类别的特征,抵抗灾难性遗忘则要求模型仍然反映旧类别的特征,因此二者存在学习过程中的权衡(trade-off).这种权衡最早在人类和鼠类的神经系统中被研究,又被称作稳定性-可塑性窘境[23](stability-plasticity dilemma),其中稳定性指模型保持已有知识的能力,而可塑性指模型学得新知识的能力.因此,理想的类别增量学习模型应当既能高效地学习新类别的知识,又不遗忘已有类别的旧知识.

除类别增量学习以外,依照数据流中新数据的特征和测试阶段的输入,增量学习任务还可以被分为任务增量学习[9,24-25](task-incremental learning)和域增量学习[26-28](domain-incremental learning)等子类问题.其中任务增量学习的设定和类别增量学习非常相似,二者区别在于,任务增量学习的测试阶段会为每一个样本提供额外的任务下标,模型只需要在给定任务的标记空间中进行预测,因而难度比类别增量学习更小,从而成为早期增量学习研究的主要设定.如图 2 所示,任务增量学习和类别增量学习的训练/测试集设定完全一致,但类别增量学习要求模型在测试阶段在所有已知类别中进行预测,而任务增量学习则只要求在给定任务的标记空间中进行预测.域增量学习则主要关注数据流发生概念漂移[29-31](concept drift)和分布变化[32-34]的场景:每个增量任务中都包含所有类别,但不同增量任务中同一类别的样本分布会发生变化.如图 2 所示,域增量学习要求模型首先学习真实拍摄的勺子和床两种类别,之后学习剪贴画风格的勺子和床,两个域之间存在分布变化.在测试阶段则要求模型能够对两种不同风格的图片给出准确的预测.由于在更新模型后要求模型同时区分新类和旧类,类别增量学习问题的研究相比任务增量学习和域增量学习都更具挑战性,对于构建真实世界的鲁棒分类器也更加具有现实意义,因此成为了近年来增量学习问题研究的重点和难点.本文主要着眼于类别增量学习算法,并对该领域最新研究成果进行分类和总结.

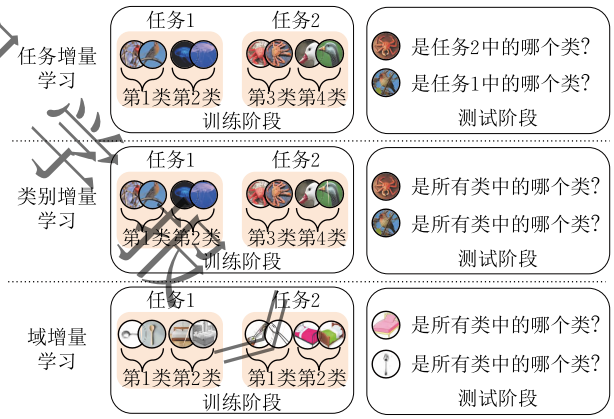


图 2　三种不同的增量学习问题设定

对于增量学习问题的研究可以追溯到对灾难性遗忘现象的观察[10-11,35],早期文献基于传统机器学习方法对增量学习问题进行研究[36-39],但其大多仅关注单阶段的增量学习过程,因此难以应对持续性、大规模的数据输入,无法满足现今开放动态环境场景下的应用需求.随着机器学习领域的发展和深度学习的成熟,基于深度学习的增量学习问题研究近年来成为机器学习、数据挖掘等领域主流会议上的热点.而类别增量学习因其应用面最广,难度最大,在所有增量学习问题的设定中受到最多关注.截至目前仍缺少有关基于深度学习的类别增量学习算法全面而深入的方法综述,已有的综述文献大多着眼

于早期的任务增量学习算法[9,40],并缺少在大规模数据集上的验证对比.其他综述或关注不同应用场景下的增量学习研究,文献[19]主要关注增量学习算法在自然语言处理领域的应用.文献[41]主要关注增量学习在机器人领域的应用.文献[42]主要从生物学角度对当前增量学习算法进行分析.与上述已有综述不同,本文主要关注基于深度学习的类别增量学习算法,更加全面深入地对当前类别增量学习算法进行划分和综述.考虑到机器学习的三个重要层面——数据层面、参数层面和算法层面,本文依此对已有的类别增量学习算法进行分类和总结.此外,本文也对10种典型的类别增量学习算法在多个基准公开数据集上进行了广泛的验证和比较.通过在基准数据集上的实验比对,本文观察到,从数据层面考虑,进行数据重放可以极大地提升增量学习模型的性能;从参数层面考虑,基于动态模型结构的类别增量学习算法能够取得当前最优的性能,同时也消耗了最多的存储开销;而从算法层面考虑的知识蒸馏和滞后调节方法能够在性能与存储方面折中.

本文的主要贡献如下:

(1)基于机器学习的三个重要层面——数据层面、参数层面和算法层面,对已有的基于深度学习的类别增量学习算法进行分类和总结,同时深入分析了不同方法之间的优势与不足.

(2)在图片分类、文本分类等多个基准数据集上对10种典型的类别增量学习算法进行了对比评估.通过多种评价准则,在多种数据设定下进行了比较,并从存储开销、运行时间、混淆矩阵、消融实验等多方面对已有方法进行深入对比分析.

(3)讨论了基于深度学习的类别增量学习的主要挑战和未来研究方向,从学习场景、数据形式、网络结构、优化方式、学习范式、知识迁移等角度对类别增量学习的未来研究方向进行展望.

本文第2节介绍相关工作;第3节给出类别增量学习问题的定义;第4节从三个层面对当前基于深度学习的类别增量学习算法进行分类和概述,分别对经典和最新的类别增量学习算法进行了介绍;第5节对当前类别增量学习的基准数据集进行总结,对于10种典型的类别增量学习算法进行系统性的复现,并在基准数据集上进行了测试和分析.除了模型分类性能以外,还从模型运行时间、存储开销等多方面分析了不同算法的优劣势;最后,第6节讨论未来增量学习的发展方向并总结全文.

2 相关工作

本文主要关注基于深度学习的类别增量学习算法,本节主要对相关领域,如域增量学习、任务增量学习,以及非深度类别增量学习进行讨论.同时,也对相关综述与增量学习在其他领域中的应用进行分析和讨论.

域增量学习.如图2所示,域增量学习主要关注数据流发生概念漂移[29-31](concept drift)和分布变化[32-34]的场景,这一设定也与领域自适应[43-44]、迁移学习[45]等任务存在相似性.文献[28]提出通过评估参数重要性构造参数正则化项约束模型更新,文献[46]提出通过维护缓存样本作为指示器,约束模型在新的分布上的数据的更新过程.文献[32]提出在模型约束的同时引入分类器集成,通过自适应模型加权进行动态模型结合.文献[47]将域增量学习任务抽象为双目标学习,并利用自组织映射网络进行模型持续更新.文献[48]进一步地将类别增量学习任务与领域自适应任务进行结合,在统一的框架内解决存在分布变化的类别增量学习问题,这一设定被文献[49-50]进一步扩展,并在期望最大化框架下进行端到端学习.域增量学习任务问题的核心是如何设计算法使得模型能够应对分布变化,而类别增量学习问题则假设所有任务为同一个类别,不存在分布变化,因此域增量学习算法设计的重点与类别增量学习有所不同.

任务增量学习.如图2所示,任务增量学习与类别增量学习的问题设定基本一致,区别仅在于测试阶段任务增量学习仅关注子任务内部的分类,因此其难度小于类别增量学习问题.早期的增量学习研究也主要关注于任务增量学习.由于测试阶段能够获得任务标记,大量基于动态模型结构的方法在任务增量学习场景中取得了优越的性能.模型通过分阶段对结构进行扩展,并在测试阶段选择性激活模型参数进行预测,可以有效地应对任务增量学习场景.文献[51]主张在面对新任务时,将模型结构进行复制,并建立旧模型和新模型之间的连接关系,促进知识向新模型迁移.相似地,文献[52]提出为每个新任务复制一个模型,并使用门结构学习样本到任务的映射,从而选择出最适合的模型用于测试.为了缓解由于序列化任务复制模型带来的爆炸性存储开销,文献[53]主张仅复制一次模型以节约存储.文献[54]进一步地在扩张剪枝的基础上引入了神经元复

《计算机学报》

制技术，从而帮助模型固定已有知识. 文献[27]则主张使用强化学习算法作为模型扩张的搜索指导. 因为任务增量学习场景与类别增量学习场景的区别仅在于测试阶段能否获得测试数据的任务标记，因此所有类别增量学习算法均可以适用于任务增量学习场景，而任务增量学习算法则不一定可以适用于类别增量学习场景. 因此，设计有效的算法应对类别增量学习问题对于在真实世界中设计学习系统更为重要.

非深度类别增量学习. 在深度学习兴起之前，已经存在一些基于传统机器学习算法的类别增量学习研究，Zhou 等人[36]分别从样本增量学习、类别增量学习和特征增量学习角度其进行了分类和总结. 基于传统机器学习的类别增量学习算法旨在发现和解决灾难性遗忘现象[35]，但其大多关注单阶段的增量学习过程，因此难以应对持续性、大规模的数据输入，无法满足现今开放动态环境场景下的应用需求. 文献[55-56]和[57-58]分别研究了使用梯度反向传播训练的神经网络和 Hopfield 神经网络中的灾难性遗忘现象. 文献[37]提出了一种在线学习场景下基于感知机的新类学习算法. 文献[38]提出一种集成学习方法，通过动态地调节分类器的权重实现类别增量学习. 文献[59]使用最小二乘支持向量机解决类别增量学习问题，但仅限于一次性学习一个新类. 相似地，Da 等人[39]提出 LACU-SVM，通过引入无标注数据训练支持向量机，优化了误分类风险. 然而 LACU-SVM 也仅能一次学习单个新类，无法适应复杂环境下同时学习多个新类的需求. 文献[14]在 LACU-SVM 的基础上进一步地通过构建无偏估计统计量以学习新类，并给出良好的理论保障，然而也仅限于一次学习一个新类.

相关综述讨论. 得益于深度学习的应用，近年来增量学习领域的研究蓬勃发展. 增量学习的研究对于当前构建鲁棒、可解释的机器学习系统具有重要意义，但由于兴起时间晚、新的增量学习算法不断涌现，已有的增量学习算法综述大多着眼于早期的任务增量学习算法. De Lange 等人[9]总结了早期任务增量学习的方法和关键问题，并在长尾分布设定下的任务增量学习场景进行了测试. van de Ven 等人[40]总结了三种典型增量学习的场景，并在简单数据集上进行了比较测试，受限于技术发展，文中仅涉及了早期的少量增量学习算法，并缺少在大规模数据集上的验证对比. Mai 等人[60]总结了在线学习场景[61-62]下的增量学习算法，并主要从是否支持模型在线更新的角度对已有方法进行了分类总结. 文献[63]从三个方面对类别增量学习问题进行了分类和验证，但缺乏对数据约束和动态模型结构的方法的分类和讨论. Biesialska 等人[19]则主要关注增量学习算法在自然语言处理领域的应用. 相似地，Lesort 等人[41]主要关注增量学习在机器人领域的应用. Parisi 等人[42]主要从生物学角度对当前增量学习算法进行分析，并主要分析了这些系统如何抵抗灾难性遗忘，然而并未对算法进行系统性的验证和比对.

其他领域的类别增量学习. 国内文献的相关工作[64-67]主要关注于如何在流式数据中动态更新模型. 此外，基于深度学习的类别增量学习也在生物信息学、自然语言处理、推荐系统等领域被广泛应用. 文献[68]在免疫系统中的外周血单个核细胞分类任务中引入类别增量学习. 文献[69]则提出适用于 miRNA 疾病分类的类别增量学习模型. 文献[70]则提出利用临床诊断记录进行类别增量学习以进行疾病诊断. 文献[71]在自然语言处理序列化生成任务中引入增量学习算法，文献[72-73]在文章情感分类任务中对模型进行增量训练. 类别增量学习算法也被文献[74-75]广泛应用到推荐系统等诸多领域中.

3　问题定义

本节首先给出类别增量学习问题的定义. 考虑到范例集是当前多数类别增量学习算法采用的训练技巧，为了后续章节的连贯性，本节也会给出范例集的定义.

定义 1.　类别增量学习. 类别增量学习旨在从数据流中不断学习新类[76]. 假设存在 B 个不存在类别重合的训练集 $\{\mathcal{D}^1,\mathcal{D}^2,\cdots,\mathcal{D}^B\}$，其中 $b=\{1,2,\cdots,B\}$ 表示训练集序号，$\mathcal{D}^b=\{(\boldsymbol{x}_i^b,y_i^b)\}_{i=1}^{n_b}$ 表示第 b 个增量学习阶段的训练集，又称作训练任务(task). 其中 n_b 表示第 b 阶段训练集样本总数. $\boldsymbol{x}_i^b\in\mathbb{R}^C$ 是来自于类别 $y_i\in Y_b$ 的一个训练样本，其中 Y_b 是第 b 个任务的标记空间，C 代表样本输入向量的维度. 不同训练任务间不存在类别重合，即对于 $b\neq b'$，有 $Y_b\bigcap Y_{b'}=\varnothing$. 在学习第 b 个任务的过程中，只能使用当前阶段的训练数据集 $\mathcal{D}^b$ 更新模型. 类别增量模型的目标不仅是学得当前数据集 $\mathcal{D}^b$ 中新类的知识，同时

计算机学报

也要保持之前所有学过类别的知识免遭遗忘.因此,类别增量学习任务考虑模型在所有已知类集合$\mathcal{Y}_b=Y_1\cup\cdots\cup Y_b$上的判别能力评估其学习能力.将增量学习模型对样本$\boldsymbol{x}$的输出记作$f(\boldsymbol{x})$,则模型要优化的期望风险描述为

$$\mathbb{E}_{(\boldsymbol{x}_j,y_j)\sim\mathcal{D}_t^1\cup\cdots\cup\mathcal{D}_t^b}[\ell(f(\boldsymbol{x}_j),y_j)] \tag{1}$$

其中$\mathcal{D}_t^b$表示第b个任务的样本分布.$\ell(\cdot,\cdot)$评估输入之间的差异,在分类任务中一般使用交叉熵损失函数.由于模型需要同时在学过的所有分布上最小化期望风险,能够最小化公式(1)的模型能够在学习新类的同时不遗忘旧类的知识.

进一步地,可以将深度神经网络按照特征提取层和线性分类器层进行解耦,则模型$f(\boldsymbol{x})$由特征提取模块$\phi(\cdot):\mathbb{R}^C\to\mathbb{R}^d$和线性分类器$\boldsymbol{W}\in\mathbb{R}^{d\times|\mathcal{Y}_b|}$组成,即$f(\boldsymbol{x})=\boldsymbol{W}^\top\phi(\boldsymbol{x})$.为了表述方便,可将线性分类器$\boldsymbol{W}$进一步表示成每个类别分类器$\boldsymbol{w}_i$的组合:$\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_{|Y_b|}]$

需要注意的是,典型的类别增量学习问题设定假设$Y_b\cap Y_{b'}=\varnothing$.然而在很多实际场景中,不同阶段的任务往往存在重合的类别.这种存在类别重合的增量学习设定被称为模糊类别增量学习[77-78](blurry class-incremental learning).新任务中存在的旧类别样本使得模型能够回顾以往的类别信息,因此难度小于典型类别增量学习设定,典型的类别增量学习算法也可直接应用于此种场景,因此本文主要关注典型类别增量学习问题设定.另一方面,当前的任务增量学习问题假设新类别在B个阶段内逐次出现,而数据流往往难以满足这一假设.当前工作大多基于已知的任务边界设计算法,而更广义的类别增量学习场景则应关注任务无关的类别增量学习场景[79](task-free class-incremental learning).

定义 2. 范例集.按照类别增量学习过程的定义,模型在学习第b个任务时仅能获取当前的训练集$\mathcal{D}^b$.仅使用$\mathcal{D}^b$更新模型很容易遭受灾难性遗忘.因此,当前类别增量学习的主流算法提出保存一个额外的范例集合(exemplar/example set),记作$\mathcal{E}=\{(\boldsymbol{x}_j,y_j)\}_{j=1}^M$,用于为每个见过的类别保存一定数目的代表性样本.保存的样本可以在学习新类的过程中辅助模型抵抗灾难性遗忘.目前有两种范例集储存方式,第一种考虑对每个类存储固定数目的范例样本,例如K个,这种情况下模型的存储开销会随增量学习任务的增多而线性增长——在学习完b个增量学习任务后,模型将存储$|\mathcal{Y}_b|K$个范例样本,其中$|\cdot|$表示集合大小.第二种储存方式考虑存储固定数目的范例样本,例如M个,模型在每次学习新类后,会删除部分旧类的范例样本,并加入部分新类的范例样本,保证每个类别储存$\left[\frac{M}{|\mathcal{Y}_b|}\right]$个样本,$[\cdot]$表示向下取整.由于后一种存储方式不会造成模型容量随增量学习任务增长带来的额外存储开销,目前的研究方法主要使用后一种思路存储范例集,本文也采用后一种存储方式.

对于范例集中样本$\boldsymbol{x}_j$的选择方式,目前主要有两种方法.第一种方法随机对所有样本进行采样,这样会导致采样到的范例样本方差较大.第二种则采用群聚[80](herding)思路,对每个类计算类别中心,并按照样本到类别中心的距离进行升序排序,优先选择那些离样本中心更近的样本作为范例样本.这样做能够保证采样到的范例样本更加贴近类中心,从而更具有代表性.

4 基于深度学习的类别增量学习算法分类

考虑到机器学习的三个重要层面——数据层面、参数层面和算法层面,本文依此对当前的类别增量学习算法进行分类和总结.数据层面的类别增量学习算法主要关注如何利用范例集增广训练集和如何利用范例集样本约束模型更新过程,并可以被细分为数据重放(data replay)和数据约束(data restriction)两种子类型.参数层面的类别增量学习算法主要关注如何从参数重要性对模型进行约束和如何动态调整模型的网络结构/参数数目,并可以被分为参数正则(parameter regularization)和动态结构(dynamic architecture)两种子类型.算法层面的类别增量学习算法则主要关注如何设计有效的学习范式以维护模型知识和如何发掘类别增量学习模型在训练过程中的偏好并进行调节,并可以被分为基于知识蒸馏(knowledge distillation)和滞后调节(post tuning)两种子类型.本文主要关注以上三个层面的六种子类型学习算法,并整理和总结相关的研究内容.对于类别增量学习的算法的分类如表 1 所示.接下来,将按照上述思路对每一种层面的类别增量学习算法进行总结和回顾.

计算机学报

表 1　类别增量学习方法分类与代表性方法

分类层面	子类型	代表性方法	特点
数据层面(4.1 节)	数据重放(4.1.1 节)	SR[81],GR[82]	易于操作,但容易过拟合
	数据约束(4.1.2 节)	GEM[46],A-GEM[83]	优化复杂度高,依赖假设
参数层面(4.2 节)	参数正则(4.2.1 节)	EWC[28],SI[84],IADM[32]	需存储约束矩阵,开销大
	动态结构(4.2.2 节)	DEN[54],DER[85]	模型参数量随数据线性增长
算法层面(4.3 节)	知识蒸馏(4.3.1 节)	LwF[86],iCaRL[76]	约束直观,易于实现
	滞后调节(4.3.2 节)	Rebalancing[87],BiC[88],WA[89]	可消除模型偏置

4.1　数据层面的类别增量学习算法

数据是机器学习任务的核心,因此,数据层面的类别增量学习算法主要关注如何有效地利用数据、存储数据、生成数据以抵抗灾难性遗忘.具体来说,基于数据重放的类别增量学习算法主要关注如何有效地利用旧类数据构成的范例集,通过在学习新类的同时复习旧类数据,保证模型在学习新类别的时候不遗忘旧类.基于数据约束的增量学习算法则主要关注如何利用旧类数据约束增量模型的更新过程,从而使更新后的模型仍能反映旧类数据的特征.

4.1.1　基于数据重放的类别增量学习算法

模型学习新类的过程和人脑学习新知识的过程类似,那么是否可以采取人类学习的技巧,在学习新知识的时候主动地复习之前学过的知识[90-92]?根据定义 2,模型在类别增量学习过程中对每个旧类均可以存储一定数量的范例集,记作$\mathcal{E}$,那么上述的复习过程可以被描述为

$$\mathcal{L}=\sum_{(\boldsymbol{x}_i,y_i)\in(\mathcal{D}^b\cup\mathcal{E})}\ell(f(\boldsymbol{x}_i),y_i) \tag{2}$$

式(2)表明模型在学习新任务 D^b 时,需要同时考虑其在以往的旧类别范例集上的损失,优化模型 $f(\cdot)$使得其能够同时拥有对旧类和新类的判别能力.易见,如果同时使用所有的训练集$\mathcal{D}^1\cup\mathcal{D}^2\cup\cdots\cup\mathcal{D}^B$离线训练模型,就可以全盘考虑所有类别信息,从而使模型获得涵盖所有类别的判别能力,这种学习特例被称作“先知”(Oracle),是所有类别增量学习任务的性能上界.

基于数据重放的类别增量学习算法的思想非常直观,因此引发了大量的相关研究.在范例集采样方面,文献[81,93]提出使用蓄水池采样以保证每个类保有固定数目的独立同分布范例样本.Aljundi 等人[94]在线增量学习场景中提出了一种贪心的范例集选择方法,以最大化范例样本多样性.文献[78]通过数据增广的方式衡量模型对样本的不确定性,并基于此采样对模型训练影响更大的样本以提升模型性能.De Lange 等人[95]将最近类中心分类器[96](nearest center mean)和数据重放思路相结合,解决了存在概念漂移情况下的类别增量学习任务.文献[97]在数据重放的基础上引入元学习,在更新模型的同时锚定旧类的知识.以上的数据重放方法考虑在输入空间保存样本,然而图片数据集的样本输入 $\boldsymbol{x}$ 可能非常大,为模型引入了额外的存储开销.因此,文献[98]主张可以直接存储模型对样本的特征表示,即 $\phi(\boldsymbol{x})$.一方面,存储提取后的特征能够极大程度地减少存储开销;另一方面,使用旧模型提取出的特征训练模型能够约束模型的分类器层,从而进一步防止灾难性遗忘.

除了存储样本以外,另外一些方法主张通过学习生成式模型对旧类别进行建模.在模型更新阶段,他们使用生成式模型生成出的数据构造范例集进行数据重放.Shin 等人[82]第一个提出使用生成式模型生成旧类样本,Hu 等人[99]在此基础上将模型参数分解为共享参数和动态参数,并利用生成式模型动态地生成数据帮助分类器模型调整动态参数,以适应网络演进.FearNet[100]是一种脑启发的方法,主张设计多个生成式模型并利用额外的子系统调整数据生成过程.文献[101-102]基于条件生成对抗网络生成特定类别的样本帮助模型训练,同时引入网络掩码约束模型重要参数防止遗忘.相似的想法在文献[103]中被引入到半监督增量学习问题中,用于解决无标注数据的生成和训练问题.除了生成式模型以外,还有其他对样本分布建模的方法,文献[104]基于变分自动编码器生成旧类样本,文献[105]提出可以将类别增量学习过程中的每个类建模为高斯分布,这样便可以通过维护均值和方差近似地采样出旧类样本.然而,基于生成式模型的方法的性能严格受制于生成图片的质量,尤其是在复杂的自然图像数据集中.同时,生成式模型和判别式模型一样,也会由于序列化训练遭受灾难性遗忘[106-108],这进一步加剧了数据重放训练的困难性.

基于数据重放的类别增量学习算法因其简单易于操作,也被大量应用到诸如语义分割[109-110]、概念漂移数据流学习[111]、推荐系统[75]、虚拟现实中的摄像头定位[112]等任务中.然而,由于模型的回顾过程

仅依赖于小规模范例集，基于数据重放的算法往往会遭受过拟合问题[46]. 文献[113]分析了这种过拟合对模型泛化性的影响. 另一方面，由于存储的范例集大小远小于每个增量学习阶段的训练集大小，这种回顾性训练也为模型引入了数据集不平衡[114]的潜在影响.

4.1.2 基于数据约束的类别增量学习算法

利用旧类数据可以进行数据重放，通过复习旧类知识抵抗灾难性遗忘. 那么，是否还存在其他思路利用旧类范例集？基于数据约束的算法旨在利用旧类数据约束模型的更新过程，并使用范例集指导模型的更新策略，以抵抗灾难性遗忘. 其中最具代表性的工作是 GEM[46]，GEM 希望找到最优的模型 f^* 满足：

$$f^*=\min_{f}\sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{D}^b}\ell(f(\boldsymbol{x}_i),y_i)$$

$$\text{s.t.}\sum_{(\boldsymbol{x}_j,y_j)\in\mathcal{E}}\ell(f(\boldsymbol{x}_j),y_j)\leq\sum_{(\boldsymbol{x}_j,y_j)\in\mathcal{E}}\ell(f^{b-1}(\boldsymbol{x}_j),y_j)\quad(3)$$

其中，f^{b-1} 代表上一阶段训练结束时的增量模型.

式(3)表明，GEM 的优化目标是找到能够在新任务上最小化分类损失的模型 f，同时要保证其在范例集上的分类损失不大于上一阶段模型在范例集上的损失. 由于范例集中的样本均来自旧类，这一约束一定程度上维持了模型在旧类上的性能. 为了考察模型在样本上的损失，GEM 计算模型在范例集上的梯度 $g_{\mathcal{E}}$ 和在当前任务样本集上的梯度 g. 如果二者夹角大于 90°，则将当前任务的梯度投影到距离其夹角最小的范例集的梯度方向上，以满足上述约束. 这个问题被进一步转化为二次规划问题进行求解. GEM 的想法非常直观：只要能够保证模型在范例集上的损失不增大，那么就认为模型在旧类上的性能就不会下降. 然而，由于式(3)的约束是定义在所有范例集上，模型每次更新都需要重新计算在范例集上的损失并求解二次规划问题，导致模型更新速度十分缓慢. 因此 A-GEM[83]被提出，通过将范例集上的约束松弛为在随机批次上的样本约束，同时无需求解二次规划问题，极大地加快了模型的训练速度. 这一思路也被文献[79]应用于无任务边界的在线增量学习场景中.

相似地，还有其他方法通过范例集样本对模型更新进行约束. Zeng 等人[115]在学习新类的过程中，仅允许参数在正交于以往任务张成的子空间的方向上进行更新，从而保证新类学习不会干扰旧类的性能. 进一步地，Tang 等人[116]将梯度解耦为共享梯度和特定梯度，模型更新过程的梯度需要贴近新任务上的梯度，与旧任务的共享梯度一致，并与旧任务的特定梯度张成的空间正交. Wang 等人[117]提出在更新阶段将梯度投影到之前任务的零空间上，以维持旧类上的分类性能.

然而，基于数据约束的类别增量学习算法依赖一系列假设，例如，GEM 认为只要保证模型在范例集上的损失不增大，模型在旧类上的性能便不会变差，这种假设往往是难以立足的. 这也导致基于数据约束的类别增量学习算法在实际增量学习过程中无法取得较好的性能.

4.1.3 数据层面的类别增量学习算法总结

本节主要探讨了两大类数据层面的类别增量学习算法，分别从数据的不同层面抵抗增量学习中的灾难性遗忘现象. 数据层面的增量学习算法因其简单易于操作，也被大量应用到其他机器学习任务中. 然而，这两类算法在模型更新过程中均依赖存储已有类别样本构造范例集. 存储旧类样本可能在某些场景下破坏用户隐私[118]，因此，在用户隐私较为重要或由于存储开销历史数据难以获得的情况下，应考虑不依赖范例集的类别增量学习算法.

此外，对于基于数据重放的类别增量学习算法，由于模型的回顾过程仅依赖于小规模范例集，这类算法往往会遭受过拟合和数据集不平横的潜在影响. 基于数据约束的类别增量学习算法将范例集样本视作指示器，认为只要保证模型在范例集上的损失不增大，模型在旧类上的性能便不会变差. 这种假设往往是难以立足的. 这也导致基于数据约束的类别增量学习算法在实际增量学习过程中无法取得较好的性能.

4.2 参数层面的类别增量学习算法

参数是构成机器学习模型的重要组成部分，因此，参数层面的类别增量学习算法主要关注如何控制模型参数以匹配不断到来的增量数据. 具体来说，基于参数正则的类别增量学习算法通过评估模型参数重要程度，并限制重要参数的偏移以巩固模型在以往任务上的知识. 基于动态模型结构的类别增量学习算法主要考虑设计神经网络的扩张和剪枝算法，以使模型的网络结构动态地匹配增量学习任务的需求.

4.2.1 基于参数正则的类别增量学习算法

在类别增量学习任务中，如果每个参数对模型分类的贡献不同，那么是否只需要保证对分类任务更重要的参数不改变，就可以维持模型的判别能力？基于参数正则的类别增量学习算法从贝叶斯框架[119]

出发，评估神经网络中的参数不确定性，并将其作为在学习新任务时的先验. 考虑到深度神经网络中的参数规模，以上评估过程往往假设参数之间相互独立，从而可以维持一个和模型大小同等规模的参数重要性矩阵. 在学习新的增量学习任务时，便可以基于参数重要性矩阵对更重要的参数施以更大的正则化约束项，从而维持模型在旧类别上的判别能力. 具体来说，将模型对每个参数的重要性评估的结果记作重要性矩阵 $\boldsymbol{\Omega}$，那么模型在学习新任务时的损失就可以被表述为

$$\mathcal{L}=\mathcal{L}_{\text{new}}+\lambda\mathcal{L}_{\text{reg}} \tag{4}$$

$$\text{where } \mathcal{L}_{\text{reg}}=\frac{1}{2}\sum_k\Omega_k(\theta_k^{b-1}-\theta_k)^2,$$

其中 $\mathcal{L}_{\text{new}}$ 代表模型在新任务上的学习损失，比如新类数据集上的交叉熵损失，$\mathcal{L}_{\text{reg}}$ 代表参数正则项，λ 是正则化项的权重. θ_k 是模型的第 k 个参数，θ_k^{b-1} 是模型在训练完上一阶段任务后第 k 个参数的值，$\Omega_k\geq 0$ 对应模型第 k 个参数的重要性.

式(4)表明，在学习新任务时，需要控制模型参数的偏移程度——若某个参数的重要程度 Ω_k 越大，则模型会施加更大的正则项，保证其不会偏移上一阶段的最终值太多. 通过控制参数 θ_k 不偏离模型上一阶段的参数 θ_k^{b-1}，可以维持模型在上一阶段任务的判别性能，从而在顺序化的学习任务中抵抗灾难性遗忘. 接下来将介绍不同评估参数重要性矩阵 $\boldsymbol{\Omega}$ 的方式，不同的评估方式构成了不同的类别增量学习算法.

在这一方面，最早提出评估参数重要程度的方法是 EWC[28](Elastic Weight Consolidation)，该方法认为在增量学习过程中，前序任务的后验构成了后续模型的先验，并可以依此帮助模型进行持续学习. EWC 使用费雪信息矩阵(fisher information matrix)来近似地估计参数 θ_k 在增量学习第 b 个任务的分布 $\mathcal{D}_t^b$ 上的重要程度 Ω_k：

$$\mathbb{E}_{\boldsymbol{x}\sim D_t^b,y\sim p_\theta(y|\boldsymbol{x})}\left[\left(\frac{\partial\log p_\theta(y|\boldsymbol{x})}{\partial\theta_k}\right)\left(\frac{\partial\log p_\theta(y|\boldsymbol{x})}{\partial\theta_k}\right)^{\top}\right] \tag{5}$$

注意到对数似然 $\log p_\theta(y|\boldsymbol{x})$ 在分类任务上等价于负信息熵损失，因此式(5)可以被视为导数协方差矩阵的期望损失. 然而，最初提出的 EWC 算法需要对每个增量学习任务维护一个参数重要程度矩阵，该矩阵和模型大小相当，因而导致了大量额外的存储开销. 因此，文献[120-121]从优化费雪信息矩阵估计的角度出发，对 EWC 算法进行了改进.

另一方面，由于 EWC 在训练完每一阶段增量任务之后才进行参数重要性评估，SI[84](Synaptic Intelligence)则主张在线地评估参数的重要程度，并通过考虑该参数对模型损失下降的贡献来对其重要程度进行估计. 文献[122]将 SI 与 EWC 对参数重要性的估计方式进行了结合. MAS[123](Memory Aware Synapses)提出使用额外的无标记数据在线地对参数重要性进行评估，并被文献[124]扩展到无任务边界的在线增量学习任务中. IMM[125](Incremental Moment Matching)使用在线更新的方式对参数重要性进行累计. 在测试阶段，IMM 会将多组不同任务的后验合并为一个高斯分布，并以此防止灾难性遗忘. Shi 等人[126]研究如何通过比特级参数更新维持信息增益，并将费雪信息矩阵和信息增益进一步结合，作为模型正则项. Yang 等人[32]基于在线评估参数重要性的思路，提出了增量自适应模型 IADM(Incremental Adaptive Deep Model). 该工作分析了神经网络的容量和持续性，认为神经网络中不同深度的特征具有演进特性，并在不同训练阶段各有优势——浅层网络收敛快，深层网络对数据的拟合性能强. 因此，可以为神经网络的每个隐含层附加线性分类器层，并对不同深度的神经网络预测结果进行集成. IADM 被文献[127]进一步扩展到半监督主动增量学习场景. 除了上述算法层面的改进以外，基于参数正则的类别增量学习算法也被广泛地应用到图片去雨[128]、自然语言生成[129]等任务中.

基于参数正则的类别增量学习算法考虑到了不同参数对于类别增量学习任务的重要程度，因此成为增量学习早期研究的重点. 然而在实际应用中，算法需要为每个增量学习阶段维持一个和模型等大小的参数重要性矩阵，使得算法存储开销随时间线性增长. 另一方面，由于不同任务对于不同参数的偏好不同，不同阶段的参数重要性矩阵也会出现矛盾[130]，这限制了该类方法的广泛应用.

4.2.2　基于动态模型结构的类别增量学习算法

在类别增量学习的场景下，由于新类不断到来，模型描绘的特征也应当随之变化. 假设模型一开始学习的类别是老虎，则模型倾向于抽取“胡须”、“花纹”等特征，如果后续到来的新类是鸟类，那么模型会倾向于抽取“鸟喙”、“爪子”等特征. 模型对新特征的学习往往伴随着对旧特征的覆盖和遗忘，因此在鸟类上更新后的模型往往不再适合再对老虎进行分类. 那么是否能够设计合适的学习算法，使得模型能够抽取的特征随任务增加不断增长，以动态地适配增量数据的学习过程呢？基于动态模型结构的类别

计算机学报

增量学习算法旨在设计有效的模型扩张和剪枝算法,通过对模型结构进行修改,匹配类别增量学习过程.在这方面,早期的研究通过复制模型结构进行动态扩张.文献[51]主张在面对新任务时,将模型结构进行复制,并建立旧模型和新模型之间的连接关系,促进知识向新模型迁移.相似地,文献[52]提出为每个新任务复制一个模型,并使用门结构学习样本到任务的映射,从而选择出最适合的模型用于测试.为了缓解由于序列化任务复制模型带来的爆炸性存储,文献[53]主张仅复制一次模型以节约存储,并基于 EWC 算法缓解灾难性遗忘.

复制模型会导致模型结构过于臃肿,且会引发存储开销的爆炸性增长,一些文献为了解决这一问题,主张动态地调节模型的宽度和深度,而非全部复制模型的所有部分.文献[131]认为模型学习到的特征可被分解为任务共享特征和任务特定特征,其中任务共享特征不容易遗忘.因此设计了一种扩充特征维度的方式用于学习任务的特定特征.文献[132]通过对模型剪枝得到网络的紧凑表示,之后在新任务到来时动态地扩张特征维度,以辅助模型适配增量任务的需求.DEN[54]则在扩张剪枝的基础上额外引入了神经元复制技术,从而帮助模型固定已有知识.以上算法启发性地对模型结构进行扩张,然而文献[27]主张使用强化学习算法作为模型扩张的搜索指导.进一步地,文献[133]引入了神经网络结构搜索(neural architecture search)过程,帮助模型找到最适应增量数据的网络结构.

除此之外,还有一些算法提出通过为部分神经元设计掩码(mask)以构建适用于当前任务的子网络(sub-network),从而将单个模型解耦成多个子模型解决网络的扩张问题.文献[134]通过注意力机制获取网络的掩码,文献[135]将网络结构建模为子模块,并通过学习模块间的随机连接构造子网络结构.文献[136]则将掩码应用到卷积神经网络的过滤器上,通过门结构获取过滤器的掩码,以获得任务特定的子网络结构.相似地,可以将掩码应用到网络权重上[137]以构造子网络.DER[85]是目前类别增量学习性能最好的算法,通过将模型复制和模型掩码相结合,在多个基准数据集上均实现了最优性能.

然而,基于动态模型结构的算法由于设计了多组网络结构,多数只能应用在测试阶段提供任务标记的任务增量学习场景.在类别增量学习场景中部署动态模型结构算法则需要设计额外的任务预测器[135-136],额外加重了模型的训练和测试开销.同时,对网络结构的复制会造成大量的存储开销[9],因此,基于动态模型结构的类别增量学习算法往往难以被实际部署在较长的数据流中.

4.2.3 参数层面的类别增量学习算法总结

本节主要探讨了两大类参数层面的类别增量学习算法,分别通过参数正则和扩充模型结构抵抗灾难性遗忘现象.基于参数正则的类别增量学习算法需要对模型存储等大小的参数重要性矩阵,显式地增加了流数据学习过程中的存储开销.类似地,基于动态模型结构的类别增量学习算法需要随模型更新不断地增加模型规模,这也会导致模型存储开销在学习过程中不断增加,因而导致参数层面的类别增量学习算法难以在资源受限的真实场景中进行应用.

此外,基于参数正则的类别增量学习算法考虑到了不同参数对于类别增量学习任务的重要程度,由于不同任务对于不同参数的偏好不同,不同阶段的参数重要性矩阵也会出现矛盾.另一方面,基于动态模型结构的算法由于设计了多组网络结构,多数只能应用在测试阶段提供任务标记的任务增量学习场景.在类别增量学习场景中部署动态模型结构算法则需要设计额外的任务预测器,额外加重了模型的训练和测试开销.

4.3 算法层面的类别增量学习算法

算法层面的类别增量学习算法主要通过设计行之有效的模型更新方法,或通过发掘模型在增量更新后存在的归纳偏好(inductive bias),实现对增量模型的进一步调节.具体来说,前者主要使用知识蒸馏方式,基于模型间预测结果的映射关系构造新的监督学习目标,从而将旧模型潜在的判别能力传递给新模型.后者观察到的归纳偏好则包括模型特征上的偏移、线性分类器层的不平衡和输出概率上的偏置,并通过归一化放缩等手段对模型进行滞后调节.

4.3.1 基于知识蒸馏的类别增量学习算法

知识蒸馏(knowledge distillation)由 Hinton 等人[138]提出,该学习范式能够将预训练好的教师模型的判别能力迁移到学生模型上,因而在模型压缩和加速[139-141]、迁移学习[142-143]、隐私保护[144]等领域被广泛应用[145].考虑到类别增量学习的目的是维持模型在旧类别上的判别能力,那么是否可以借助知识蒸馏这一手段,以旧模型作为老师模型,以新模型作为学生模型进行知识蒸馏?大量基于知识蒸馏的类别增量学习算法关注如何设计有效的知识迁移思路以辅助模型持续学习.

LwF(Learning without Forgetting)[86]是首个提出利用知识蒸馏进行任务增量学习的工作，这个想法后来被 iCaRL[76]进一步扩展到类别增量学习中，成为类别增量学习中最普遍应用的基准方法. iCaRL 利用旧模型为新模型建立监督关系，并把这种监督关系作为正则项，防止灾难性遗忘：

$$\mathcal{L}=(1-\lambda)\mathcal{L}_{\text{new}}+\lambda\mathcal{L}_{\text{KD}} \tag{6}$$

$$\text{where } \mathcal{L}_{\text{KD}}=\sum_{k=1}^{|\mathcal{Y}_{b-1}|}-S_k(f^{b-1}(\boldsymbol{x}))\log S_k(f(\boldsymbol{x})),$$

其中$\mathcal{Y}_{b-1}=Y_0\cup\cdots\cup Y_{b-1}$指所有旧类别的集合，$S_k(\cdot)$指模型输出经过 softmax 函数后输出在第 k 类上的预测概率，即：

$$S_k(f(\boldsymbol{x}))=S_k(\boldsymbol{W}^{\top}\phi(\boldsymbol{x}))=\frac{\exp^{\boldsymbol{w}_k^{\top}\phi(\boldsymbol{x})/\tau}}{\sum_{m=1}^{|\mathcal{Y}_b|}\exp^{\boldsymbol{w}_m^{\top}\phi(\boldsymbol{x})/\tau}} \tag{7}$$

$f^{b-1}(\boldsymbol{x})$指上一阶段训练完后的模型，该模型参数均被冻结不更新.

式(6)中的$\mathcal{L}_{\text{new}}$指模型在新任务上的损失，旨在帮助模型学习新类. $\mathcal{L}_{\text{KD}}$是知识蒸馏项，作为正则项损失. 由于旧模型 $f^{b-1}(\boldsymbol{x})$是在上一阶段训练得到的，能够较好地反映旧类别的特性，式(6)中的知识蒸馏项使得新模型 $f(\boldsymbol{x})$对于同一个样本输出和旧模型一致的预测结果. 因此，iCaRL 通过对齐新旧模型在旧类上的预测概率，使得新模型在旧类上保持了和旧模型一致的判别能力，从而抵抗灾难性遗忘. 式(6)中的超参数 λ 在学习新类和保持旧类知识间进行权衡，$\lambda=0$ 表明模型仅关注如何学习新类，会遭受灾难性遗忘，$\lambda=1$ 意味着模型仅考虑保持旧类知识，无法学习新类知识. 因此，文献[88]建议设定 $\lambda=\frac{|\mathcal{Y}_{b-1}|}{|\mathcal{Y}_b|}$，它代表了旧类和新类的数目比例. 需要注意的是，iCaRL 在分类阶段并不使用分类器 $\boldsymbol{W}^{\top}\phi(\boldsymbol{x})$进行分类，而采用最近类别中心[96]分类器.

iCaRL 方法的提出引发了大量关注，基于知识蒸馏的方法也成为类别增量学习中一个重要分支. EEIL[146]在 iCaRL 的基础上引入数据增广和分类器微调，进一步提升了 iCaRL 的性能. 文献[147]提出在每个增量学习阶段均训练一个新模型，再利用额外的无标记数据将多个分类器模型蒸馏成一个分类器模型. 类似地，文献[148]利用了额外的无标记样本进行三阶段知识蒸馏. Hou 等人[87]主张对新旧模型提取出的特征 $\phi(\boldsymbol{x})$进行知识蒸馏. 相似地，文献[149]提出对卷积神经网络进行不同的池化操作，并对池化层输出的不同产物进行知识蒸馏. 进一步地，文献[150]在知识蒸馏的基础上额外引入了对注意力机制的知识蒸馏项，通过约束前后模型注意力的结果防止灾难性遗忘. 文献[151]则认为式(6)中的$\mathcal{L}_{\text{new}}$也应当被替换为知识蒸馏损失. 一般实现$\mathcal{L}_{\text{new}}$的方式为计算当前模型预测输出与真实标记间的交叉熵，而该文献主张首先使用新数据训练新模型，再以新模型作为教师模型对增量模型进行知识蒸馏. 由于知识蒸馏作用在范例集样本上，文献[152]提出了一种可以优化范例集的元学习方式. 文献[153]则提出在不保存范例集的情况下，可以使用分类器模型逆生成样本[154]，并基于生成出来的样本进行知识蒸馏. 相似地，文献[155]将每个类别建模为高斯分布，并从高斯分布中采样旧类样本作为模型输入进行知识蒸馏. 文献[156]基于样本间的相似关系构造约束项，文献[157]利用弹性赫布图构造约束项，文献[158]利用神经气体网络，在模型更新过程中进行拓扑关系保持，均可以被视作知识蒸馏约束的变体. 文献[159]则通过保持样本距离排序，维持了模型在以往任务上的性能.

值得注意的是，目前基于知识蒸馏的方法仅关注如何利用旧模型帮助新模型的学习，而 Zhou 等人[130]则认为新模型和旧模型之间应该进行知识互迁移(co-transport)，主张加入新模型向旧模型的知识蒸馏项，从而进行双向知识迁移. 文献[160]提出在旧模型和新模型的特征空间分别进行知识蒸馏，以抵抗灾难性遗忘. 文献[161]提出在知识蒸馏的基础上，基于旧模型的分类器层向新模型的分类器层做迁移.

除了分类任务以外，基于知识蒸馏的类别增量学习方法也被广泛地应用到其他任务中. Douillard 等人[162]在语义分割任务中引入类别增量学习，并建立了旧模型和新模型之间的多尺度局部蒸馏损失. 文献[163]在行人重识别任务中对新旧任务的图结构进行知识蒸馏. 文献[164]在增量式动作识别任务中引入时间维度的知识蒸馏. 文献[165]从因果效应出发，建立模型因果效应之间的知识蒸馏，以抵抗灾难性遗忘. 文献[166]利用知识蒸馏手段解决了增量训练生成式模型时存在的灾难性遗忘问题，文献[167]将类别增量学习问题扩展到视频动作分类中，并提出对应的知识巩固方法以缓解灾难性遗忘.

基于知识蒸馏的类别增量学习算法因其算法直观，易于实现，因此被最为广泛地应用. 然而，仅考虑旧模型对新模型的约束限制了模型的特性，使得模型倾向于维持提取利于判别旧类的特征，不利于新

类的学习.理想的类别增量学习模型应能够对特征进行进一步提炼和巩固,在维持旧类判别能力的同时考虑新类的影响.

4.3.2 基于滞后调节的类别增量学习算法

由于类别增量学习过程中数据以流的形式到来,其分布不满足独立同分布(i.i.d.)假设,依顺序训练不断到来的新类样本会使得模型具有某种归纳偏好.基于滞后调节的类别增量学习算法主要通过观察增量模型的输出、权重方面的归纳偏好/偏置(bias),并采取相应的矫正策略.例如,若增量模型倾向于将样本预测为最新学得的新类,则应采取调节策略使模型对所有类别具有均等的预测概率.

最早观察到模型偏置的文献[87]发现随着模型进行增量训练,分类器的权重会产生偏移,具体体现为当前任务中的类别的分类器权重范数$\|\boldsymbol{w}\|$更大,而以往任务中旧类别的分类器权重范数较小.由于模型的预测基于$\boldsymbol{W}^{\top}\phi(\boldsymbol{x})$,模型会倾向于将样本分类到具有更大分类器权重的类别——新类中,同时不把样本分类到旧类中,导致模型具有明显的分类偏好.为此,作者提出使用余弦分类器替换原始分类器的输出:

$$f(\boldsymbol{x})=\frac{\boldsymbol{W}^{\top}}{\|\boldsymbol{W}\|}\frac{\phi(\boldsymbol{x})}{\|\phi(\boldsymbol{x})\|} \tag{8}$$

使用余弦分类器有效地缓解了权重不平衡带来的分类偏好.相似地,文献[146]通过采样均匀分布的范例集对线性分类器$\boldsymbol{W}$进行额外调整,从而解决了因为权重带来的分类偏好.文献[168]发现贪心地选择均匀分布的范例集并在增量学习过程中使用其重新训练模型,可以在在线增量学习任务中取得超越最佳性能的效果,并被文献[169]应用到增量目标检测任务中.进一步地,WA[89](Weight Aligning)提出可以直接在每个任务的训练阶段完成后对分类器权重$\boldsymbol{W}$进行归一化,从而保证所有分类器的模长均为1.由于模型的特征输出$\phi(\boldsymbol{x})$经过了线性整流函数ReLU[170]的映射,因此全部为正值,WA也主张对分类器权重进行裁剪,将所有小于0的权重置为0.于是,分类器权重和模型输出概率正相关,WA通过正则化和权重裁剪实现了模型滞后调节.

另一方面,一些文献关注到模型的另一种分类偏好,即增量模型倾向于对新学的类别给出更高的预测输出.因此,文献[171]提出使用双重存储的方法IL2M,以对模型的预测结果进行一次额外的调整.IL2M方法在存储范例集以外,还额外地存储了不同阶段模型对于不同类的预测输出的平均值,并以此对模型的预测输出进行校正:

$$\hat{f}(\boldsymbol{x})_k=\begin{cases}\boldsymbol{w}_k^{\top}\phi(\boldsymbol{x}), & \arg\max_m \boldsymbol{w}_m^{\top}\phi(\boldsymbol{x})\notin Y_b\\ \boldsymbol{w}_k^{\top}\phi(\boldsymbol{x})\times\dfrac{\mu^P(k)}{\mu^N(k)}\times\dfrac{\mu(N)}{\mu(P)}, & \text{其他}\end{cases} \tag{9}$$

其中$\hat{f}(\boldsymbol{x})_k$代表模型在第$k$类上的矫正输出.

式(9)表明,如果模型将一个样本预测为新类,则模型输出在每个类上的输出都需要进行校正,否则不需要进行校正.模型校正输出的方式是在每个类的原始输出$\boldsymbol{w}_k^{\top}\phi(\boldsymbol{x})$上乘以两个缩放因子.其中第一个缩放因子$\frac{\mu^P(k)}{\mu^N(k)}$的分子和分母分别指模型在学习完第$k$类时,在第$k$类样本上的平均输出,和当前阶段模型在第$k$类样本上的平均输出.第二个缩放因子$\frac{\mu(N)}{\mu(P)}$的分子和分母分别指模型在当前阶段和学习完第$k$类的阶段在所有样本上输出置信度的均值.由于模型偏置使其倾向于将样本预测为新类,若模型将样本$\boldsymbol{x}$预测为新类,式(9)将对所有类别上的输出进行校正,通过提升在旧类上的预测概率和降低在新类上的预测概率消除模型的归纳偏好,从而实现滞后调节.类似地,观察到模型在新类上的输出高于旧类,Wu等人[88]提出BiC(Bias Correction).BiC首先将范例集中的样本划分为训练集和验证集,并借助验证集中的样本训练一个额外的线性校正层.该层只有两个参数α,β,并依此对模型在新类上的输出进行调节:

$$\hat{f}(\boldsymbol{x})_k=\begin{cases}\alpha\boldsymbol{w}_k^{\top}\phi(\boldsymbol{x})+\beta, & k\in Y_b\\ \boldsymbol{w}_k^{\top}\phi(\boldsymbol{x}), & \text{其他}\end{cases} \tag{10}$$

式(10)表明模型会对在新类上的输出结果进行调整,模型学得的参数α往往处于$(0,1)$,β往往小于0,这表明模型自动地在验证集上学得了对新类的校准方式,并倾向于降低模型预测在新类上的概率.

此外,文献[172]利用范例集样本训练了用于聚合特征残差的权重函数,使模型能够自适应地对预测输出进行调整.SDC[173]提出对增量学习过程中类别中心的偏移进行滞后调节,通过训练过程中新类样本点的偏移估计旧类中心的偏移,增强了最近类中心分类器的性能.

增量模型产生偏置的原因主要来源于数据流中样本的非独立同分布采样,在类别增量学习设定中,样本分布具有显著的不平衡特性,因此一些工作也将解决长尾分布的算法[114,174]应用到类别增量学习中[175-176],并取得了不错的效果.基于滞后调节的类

别增量学习算法因其易于实现，性能良好，同时具有较强的可解释性，因此成为近期类别增量学习算法的研究热点.

基于滞后调节的类别增量学习算法是一种新兴的方法，它不关注模型本身的增量学习过程，转而关注模型训练结束阶段的再调节. 这种调节只是对模型部分组件的归纳偏置的缓解，从而达到消除模型偏置的目的. 由于后调节过程与模型训练过程解耦，这样的做法无法从模型训练阶段根除灾难性遗忘. 另一方面，模型的消除偏置过程往往引入了额外的需求，例如 BiC 算法中的验证集和 IL2M 算法中的统计量信息，这无疑增加了算法运行的复杂度.

4.3.3　算法层面的类别增量学习算法总结

本节主要探讨了两大类算法层面的类别增量学习算法，分别通过知识蒸馏和模型滞后调节抵抗灾难性遗忘现象. 其中滞后调节的过程也多数与知识蒸馏进行结合，被应用最为广泛. 然而，由于知识蒸馏仅考虑旧模型对新模型的约束，因此限制了模型的特性，使得模型倾向于维持提取利于判别旧类的特征，从而不利于新类的学习. 理想的类别增量学习模型应能够对特征进行进一步提炼和巩固，在维持旧类判别能力的同时考虑新类的影响.

此外，基于知识蒸馏的类别增量学习算法需要在模型更新过程中保持上一阶段的模型. 另一方面，对于同一样本需要前向计算两次以得到监督信息，显式地增加了模型训练开销和存储开销. 另外，基于滞后调节的类别增量学习算法在模型偏置消除过程往往引入了额外的统计量信息，这同样增加了算法运行的复杂度和存储开销.

4.4　关于类别增量学习算法划分的另一视角

基于深度学习的类别增量学习算法关注如何在增量更新模型的同时抵抗灾难性遗忘. 因其技术问题多样，单个算法常常同时涉及多个子领域，子领域之间的技术也经常互相借鉴. 例如，基于数据重放的技术目前已成为类别增量学习过程中的固定范式，并被广泛地应用到诸如动态模型结构的方法[85]和知识蒸馏[76]的方法中.

因此，本文对算法的分类方式也仅为诸多分类方式中的一种. 例如，还可以针对算法是否依赖范例集进行分类，将类别增量学习算法分为基于范例集和不基于范例集的算法，基于范例集的算法又可被进一步分类为基于数据重放和基于数据约束的方法，不基于范例集的算法则可被分为基于参数正则和动态模型结构的方法. 从另一方面，也可以将类别增量学习算法按照解决问题的角度进行归类，则可划分为数据重放、模型正则、模型扩张三个角度. 本文采用的分类方式涵盖了算法技术细节（数据、参数、算法层面），也考虑到了解决问题的角度（数据重放、数据约束、参数正则、动态结构、知识蒸馏、滞后调节）.

5　实验验证

本节首先介绍类别增量学习常用的基准数据集、数据划分与评测指标，之后在基准数据集上复现了 10 种典型的类别增量学习算法，并从多个角度分析了不同算法的优劣.

5.1　基准数据集与类别划分

类别增量学习问题的基准数据集与设定首次在文献[76]中被提出，并已受到广泛的仿效和对比. 目前被广泛使用的数据集和设定列举如下：

CIFAR100[177] 含有来自 100 个类别的 50000 张训练集样本和 10000 张测试集样本，每张图片的大小为 32×32.

CUB200-2011[178] 是一个细粒度鸟类分类数据集，包含 200 种鸟类的 11788 张图片，这个数据集被称作 CUB200. 此外，文献[173]从 200 个类中采样了 100 个子类①，构成 CUB100 数据集.

ImageNet ILSVRC2012[179] 是一个大规模图片分类数据集，具有来自 1000 个类别的 1281167 张训练集和 50000 张测试集. 这个数据集被称作 ImageNet1000. 类似于 CUB 数据集，可以从中采样 100 个子类，这构成了 ImageNet100 数据集.

本文对以上三个基准数据集进行了可视化，并展示在图 3 中，每行的样本来自于同一类. 此外，不同文章对于数据集每个增量学习任务中的类别的划分有所不同，并主要分成两种划分思路. 首先将数据集中的所有类别随机打乱，第一种划分将所有类别均分到每一个阶段中[76,89]. 第二种划分则将一半的类别作为第一个阶段的训练集，并将其余类别划分为等量的多个阶段[87,173]. 本文在小规模图像数据集 CIFAR100 和大规模图像数据集 ImageNet100 上进行实验，并同时验证以上两种不同的数据集划分对结果的影响，分别探究当前的增量学习算法在不同规模增量学习任务中的性能. 本文使用数据集原本的测试集以全面地衡量增量模型在新类和旧类上的

① 在类别增量学习基准设定中[76]，类别采样和重排均使用 NumPy 随机种子 1993 以保证可重复性. 若非特殊说明，本文中提到随机采样均指使用该随机种子进行.

计算机学报

(a) CIFAR100 (b) ImageNet100 (c) CUB200

图 3 类别增量学习基准数据集可视化

学习能力.

5.2 对比方法和评价指标

5.2.1 对比方法

基于表 1 对当前类别增量学习算法的分类,本文在基准数据集上对比了以下代表性方法的性能. Finetune:不考虑抵抗灾难性遗忘,在每个新数据集上直接使用交叉熵更新模型. Replay:基于式(2)实现的数据重放算法,利用范例集样本进行数据重放. 此外,本文还对比了基于数据约束的方法 GEM[46]、基于知识蒸馏的方法 LwF[86]、iCaRL[76]、PODNet[149]、基于后调节的方法 WA[89]、BiC[88]、基于参数正则的方法 EWC[28]和基于动态模型结构的算法 DER[85]. 本文也汇报了模型在离线情况下对所有数据进行多轮训练的结果,记作 Oracle,Oracle 是所有类别增量学习方法的性能上界. 本文对于所有方法的复现已经开源在 https://github.com/G-U-N/PyCIL,以供国内研究者使用.

5.2.2 实验设定

所有的实验均使用 Pytorch[180]实现,在 NVIDIA 3090 上运行. 所有对比方法使用 SGD 优化器[181]训练 170 轮,初始学习率为 0.1,并在第 80 和 120 轮衰减为 0.1 倍. 优化器的动量(momentum)参数设定为 0.9,权重衰减系数(weight decay)设定为 2e-4,模型训练阶段的 batchsize 设定为 128. 优化器学习率在每个新的增量任务到来时重置为 0.1. 对于 CIFAR100,使用 ResNet32[182]作为主干网络(backbone),对于 ImageNet100,使用 ResNet18 作为主干网络. 按照基准设定,存储 2000 个样本作为范例集样本,即为每个类存储 20 个范例集样本. 这些样本是基于群聚[80]方式采样,选择存储距离类别中心最近的样本. 所有方法的实现均基于原始论文中的默认参数,例如,对于 BiC,选择 10%的范例集样本作为验证集;对于 DER,使用 10 轮 warm-up 进行预训练,并设定温度系数 τ 为 5;对于 EWC,设置其正则项权重为 1000,对于 WA,使用 ℓ_2 范数对全连接层进行归一化.

5.2.3 评测指标

典型的类别增量学习方法主要考虑以下 3 个评测指标:

(1)一个直观的评测指标是考察每个任务训练结束后在所有已知类别上的分类准确率,将其记作 $\mathcal{A}_b$,其中 b 代表任务下标. 因此具有更高 $\mathcal{A}_b$ 的算法在类别增量学习学完第 b 个任务之后的分类性能更强. 为了更显著地衡量算法之间的差异性,一般采用 $\mathcal{A}_B$,即最后一阶段训练结束后的模型分类准确率进行对比.

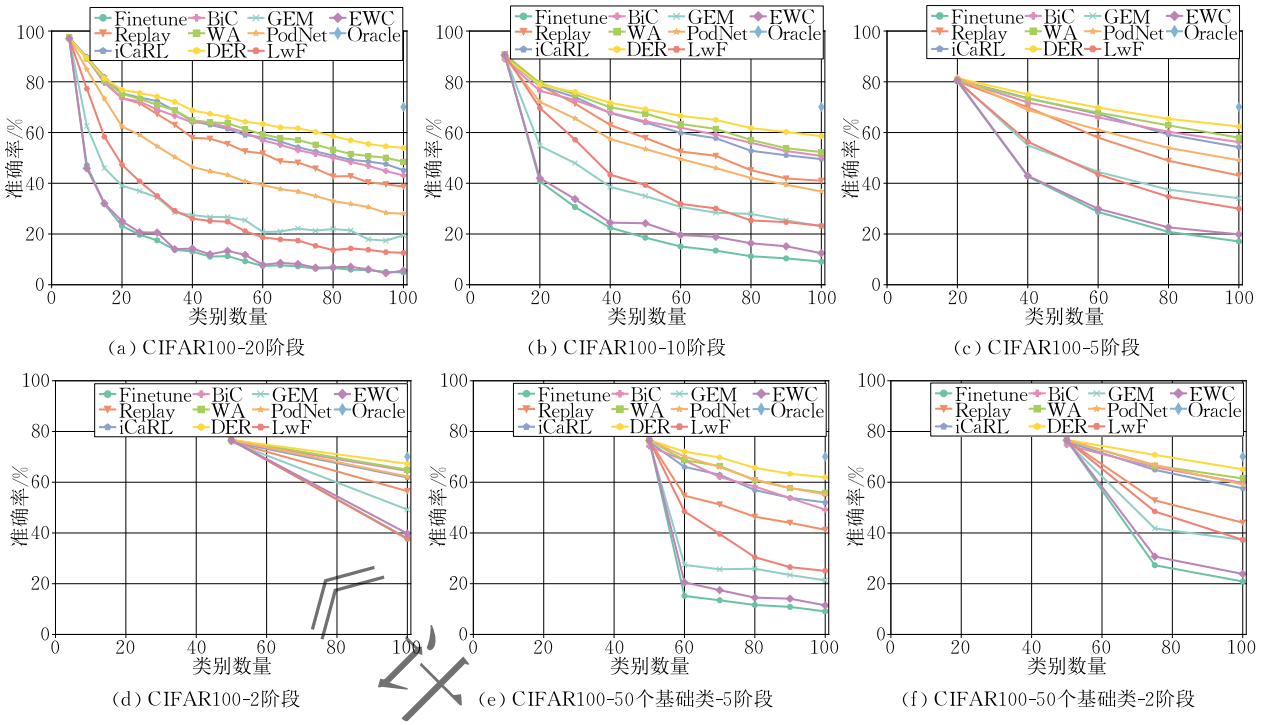
(2)为了协同考虑增量学习算法在数据流学习过程中的整体性能,可以将每个阶段的模型准确率进行平均,得到模型的平均准确率:$\bar{\mathcal{A}}=\frac{1}{B}\sum_{b=1}^{B}\mathcal{A}_b$,具有更高 $\bar{\mathcal{A}}$ 的模型在整个增量学习的过程中的性能更好.

(3)为了反映模型遭受灾难性遗忘的程度,本文同时沿用文献[158]中使用的性能下降率(Performance Dropping rate):$PD=\mathcal{A}_1-\mathcal{A}_B$,其中 $\mathcal{A}_B$ 代表模型训练完最后一个阶段时的分类准确率. PD 衡量了模型在开始和结束增量学习时的性能之差,性能下降率越高,表明模型初始的表现和最终的表现差距越大,因此遭受的灾难性遗忘程度也越大. 性能下降率越低,表明模型初始表现和最终表现差距不大,灾难性遗忘越不明显.

5.3 实验对比

5.3.1 小规模数据集 CIFAR100

本文首先在 CIFAR100 上对比所有方法的性能,实验结果如图 4 所示. 按照两种不同的数据集划分,分别在图 4(a)、(b)、(c)、(d)中报告了将 100 个



(a) CIFAR100-20阶段　(b) CIFAR100-10阶段　(c) CIFAR100-5阶段

(d) CIFAR100-2阶段　(e) CIFAR100-50个基础类-5阶段　(f) CIFAR100-50个基础类-2阶段

图 4　增量学习算法在 CIFAR100 数据集上的性能对比

类均分为 20 个、10 个、5 个和 2 个均等大小类别空间的增量任务的设定，并在图 4(e)、(f)中报告了以其中 50 个类作为第一个任务，并将其余类别均分为 5 个、2 个均等大小的类别空间的增量任务的设定.

从图 4 中可以发现，使用 Finetune 在类别增量学习任务上进行微调，会导致模型仅关注于新类的信息，而忽视了重要的旧类信息. 因此，在多轮更新后，增量模型只能体现当前任务中新类的特征，而无法对以往的旧类具有较好的判别能力，故 finetune 遭受到严重的灾难性遗忘，获得了所有方法中最差的性能. 对应地，当使用范例集对每个阶段的训练集样本进行增广，即按照式(2)的方式进行数据重放，便得到了 Replay 方法. Replay 方法相比 finetune，可以极大地提升模型性能(在 20 阶段的 CIFAR100 任务上提升最终准确率约 35%)，这证明了数据重放在增量学习过程中的有效性. 相比之下，EWC 算法由于仅考虑了参数层面的有效性，却没有利用范例集样本对模型进行约束，其相对于 finetune 方法的提升非常有限. LwF 在模型更新过程中引入了知识蒸馏损失，从而建立了旧模型对新模型的监督，以抵抗灾难性遗忘. iCaRL 在 LwF 的基础上进一步地使用了范例集数据重放和最近类中心分类器，因此更加适合类别增量学习的场景. 对比 Replay 和 GEM 的结果可以发现，仅基于数据重放已经足以实现对旧类数据较好的利用. 同样使用了范例集样本，GEM 主张的梯度投影在类别增量学习任务中并不能取得像 Replay 一样好的结果. 基于后调节的类别增量学习算法 WA 和 BiC 是基于 iCaRL 算法进行的，并改善了 iCaRL 模型中的归纳偏好，因此能够取得比 iCaRL 更好的结果. 得益于模型的复制和特征的拼接，基于动态模型结构的 DER 算法在所有的设定上都取得了最好的性能. 此外，可以观察到模型之间的性能差异随着类别增量学习阶段数目的增加而逐渐增大，这一结论非常直观——分类误差会在类别增量学习过程中不断累积. 同样地，对比两种不同的实验设定，即以一半类别作为第一个任务的基础类别，或将所有类别平均分配在所有任务中，可以发现：算法间的性能差异在后一种设定下更大. 这是由于模型在初始阶段的性能是一致的，因此在具有更少初始类的设定下，模型间的增量学习性能差异会随增量学习过程愈发明显.

5.3.2　大规模数据集 ImageNet100

除了小规模数据集 CIFAR100，本文也在大规模数据集 ImageNet100 上实证了各种方法的增量分类性能. 当前基准设定中对 ImageNet 的性能度量主要分为两种，文献[173]使用 Top-1 准确率，文献[76]使用 Top-5 准确率. 因此，本文在图 5 中汇报 ImageNet100 上的 Top-1 准确率的曲线，在图 6 中汇报 ImageNet100 上的 Top-5 准确率的曲线. 对于 ImageNet100 数据集，本文采用和 CIFAR100 中

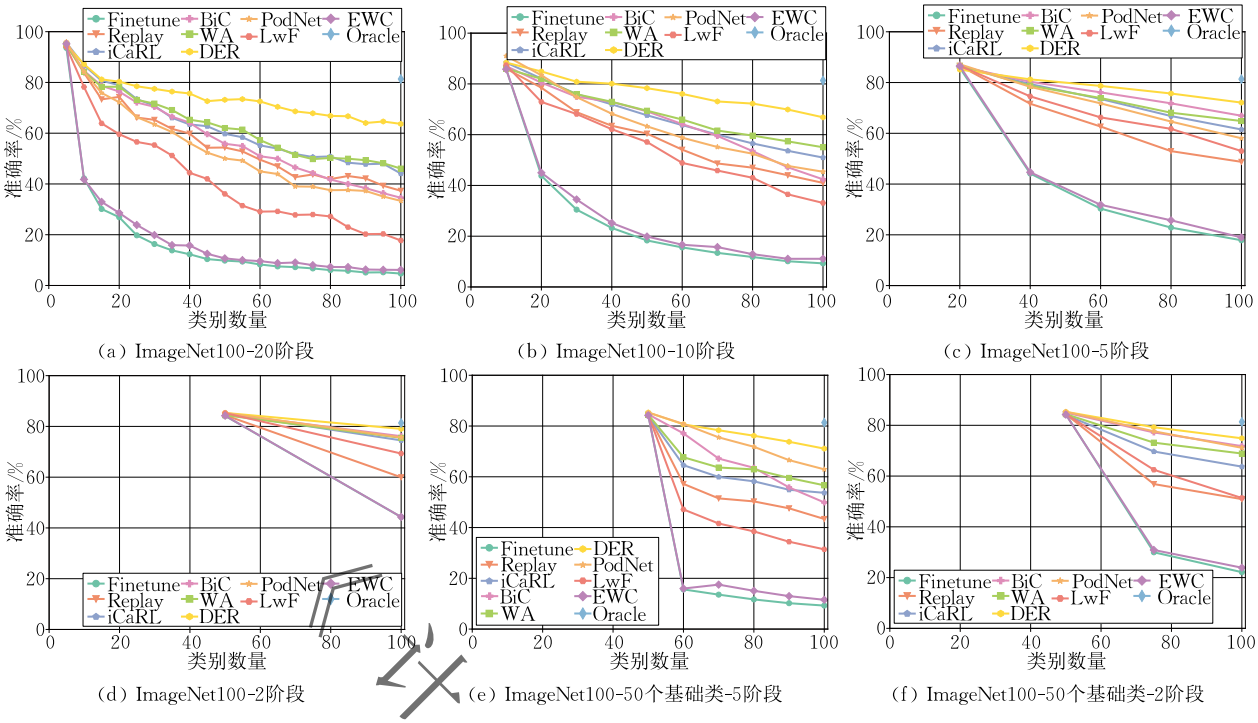


图 5 增量学习算法在 ImageNet100 数据集上的 Top-1 准确率

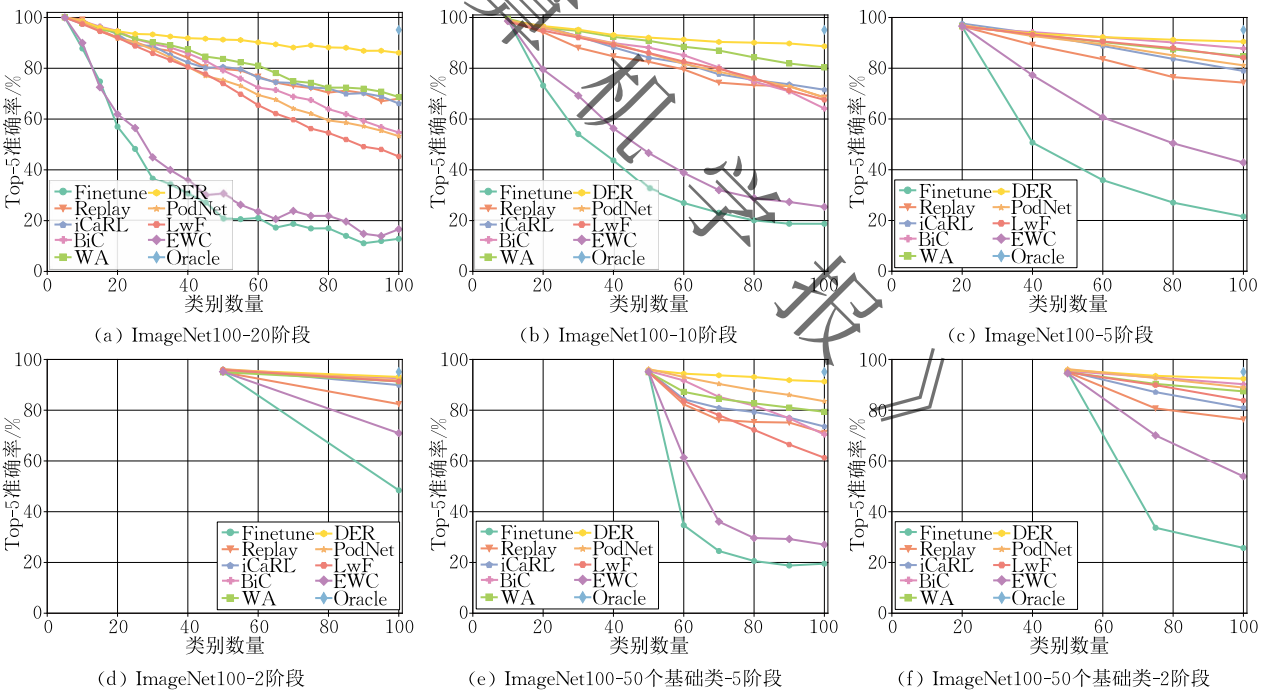


图 6 增量学习算法在 ImageNet100 数据集上的 Top-5 准确率

一致的类别划分,构成了两组不同设定下的 6 个增量学习类别划分. GEM 方法在求解二次规划过程中引入了参数规模乘以任务规模的存储开销,因此无法部署在大规模网络结构上进行求解. 通过观察图 5 和图 6 中的结果,本文得到了和小规模数据集 CIFAR100 上几乎一致的结论.

按照 5.2.3 小节中对模型性能度量的描述,除了已经在上述图中给出的增量模型准确率,本文也将两个数据集所有类别均分为 10 个增量任务进行类别增量学习实验. 表 2 中汇报了模型的平均准确率 $\bar{A}$,表 3 中汇报了模型的性能下降率 PD,并将性能最好的算法结果加粗,将性能第二名的算法结果加下划线. 其中 GEM 算法结果中的"—"代表由于额外存储开销无法在 ImageNet100 上运行.

表 2　不同算法的平均准确率 $\bar{A}$ 对比，具有更大 $\bar{A}$ 的算法在增量学习过程中性能更好（性能最好的方法被加粗表示，性能第二好的方法被加下划线表示）

实验设定	CIFAR100-10 阶段	ImageNet100-10 阶段 Top-1	ImageNet100-10 阶段 Top-5
Finetune	26.25	26.19	40.99
Replay	59.31	59.21	81.67
GEM	40.18	—	—
LwF	43.56	55.50	83.79
iCaRL	64.42	67.11	84.08
EWC	29.73	27.78	50.26
WA	<u>67.09</u>	<u>68.60</u>	<u>89.53</u>
PODNet	55.22	64.03	84.06
BiC	65.08	65.13	84.04
DER	**69.74**	**77.08**	**92.59**

表 3　不同算法的性能下降率 *PD* 对比，具有更小 *PD* 的算法在增量学习过程中遭受灾难性遗忘的程度更少，性能越稳定（性能最好的方法被加粗表示，性能第二好的方法被加下划线表示）

实验设定	CIFAR100-10 阶段	ImageNet100-10 阶段 Top-1	ImageNet100-10 阶段 Top-5
Finetune	81.71	76.50	80.26
Replay	49.79	45.00	30.16
GEM	67.17	—	—
LwF	67.55	54.30	30.84
iCaRL	39.78	37.42	27.48
EWC	77.96	74.70	73.20
WA	38.50	<u>30.96</u>	<u>19.18</u>
PODNet	52.92	45.60	30.22
BiC	<u>38.01</u>	44.60	34.46
DER	**31.01**	**21.56**	**10.16**

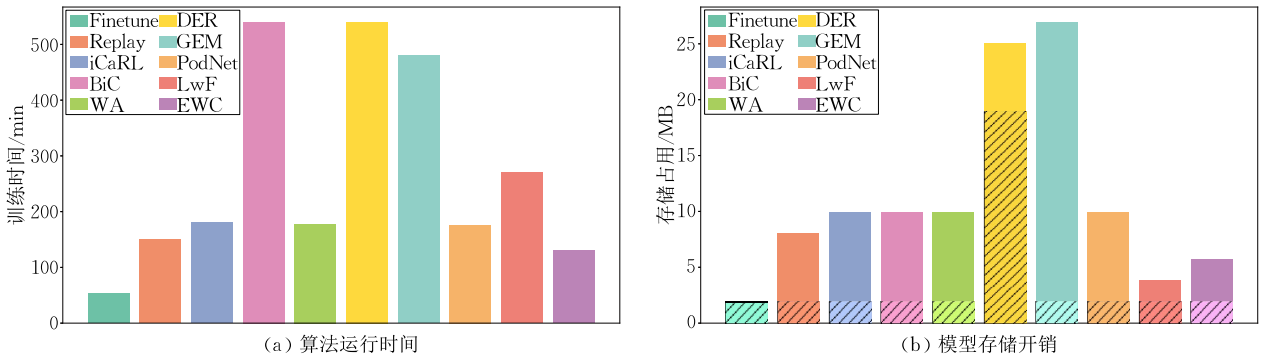
从两个表格中可以看出，本文得出的结论与图 5 和图 6 一致. DER 算法在两个数据集的多种指标上均取得了最好的性能，性能第二的算法则分别是基于滞后调节的算法 WA 和 BiC. 实验结论证明，基于滞后调节和动态模型结构的类别增量学习算法能够在当前类别增量学习的设定下取得较好的性能.

5.3.3　运行时间和模型大小分析

除了模型的增量学习性能，类别增量学习作为真实世界的应用场景，也要求模型运行速度快，占用空间少. 因此，本文记录了 10 阶段 CIFAR100 的训练过程，并在图 7(a)中对比了以上方法的运行时间，在图 7(b)中对比了以上算法的模型存储开销.

分析图 7(a)可以发现，仅基于新数据调整模型的算法 finetune 消耗了最少的运行时间，相应地性能也最差. DER、BiC、GEM 占用了最长的模型的运行时间（约为 finetune 的 10 倍），其中 DER 和 BiC 在类别增量学习任务上的表现更好，这意味着他们带来额外性能的同时消耗了更多的计算资源. 相比之下，其他方法如 iCaRL 虽然没有取得最优的模型性能，却仅比 finetune 增加了少量的模型运行时间，在要求实时响应的机器学习系统中则应当考虑这些方法.

在图 7(b)中，对于每一支柱状图，下部分的阴影代表模型的实际存储开销，上部分的非阴影部分代表模型的额外存储开销，包括范例集的存储开销、参数重要性矩阵的存储开销、基于知识蒸馏的旧模型的存储开销、基于数据约束的二次规划矩阵存储开销等. 因此，整个柱状图的长度代表了模型在训练过程中实际消耗的存储空间. 从图中可以发现，DER 算法由于对每一个任务复制了一个模型，其模型的存储开销是所有算法中最大的——模型在性能上的增益来自于更大的存储开销. GEM 算法由于需要维护一个参数规模乘以任务数的二次规划矩阵，其消耗的额外存储开销是所有算法中最多的. 总体来说，所有方法参数规模之间的大小关系和算法运行时间的长短关系基本一致.



(a) 算法运行时间　　(b) 模型存储开销

图 7　运行时间和模型存储开销对比

因此，若在资源受限的终端设备上执行算法，则应在算法性能外同时考虑模型训练效率和容量大小，而不应单纯地以算法性能为唯一指标.

5.3.4　混淆矩阵分析

为了可视化不同方法在预测阶段的偏好，本节对典型类别增量学习方法的预测混淆矩阵进行了可视化，如图 8 所示. 混淆矩阵对角线上的明暗程度反映了类别增量学习模型对该类别的判别性能，理想的模型应当对于所有类别均具有较好的预测性能.

可以观察到，直接对模型顺序化地微调(图 8(a))

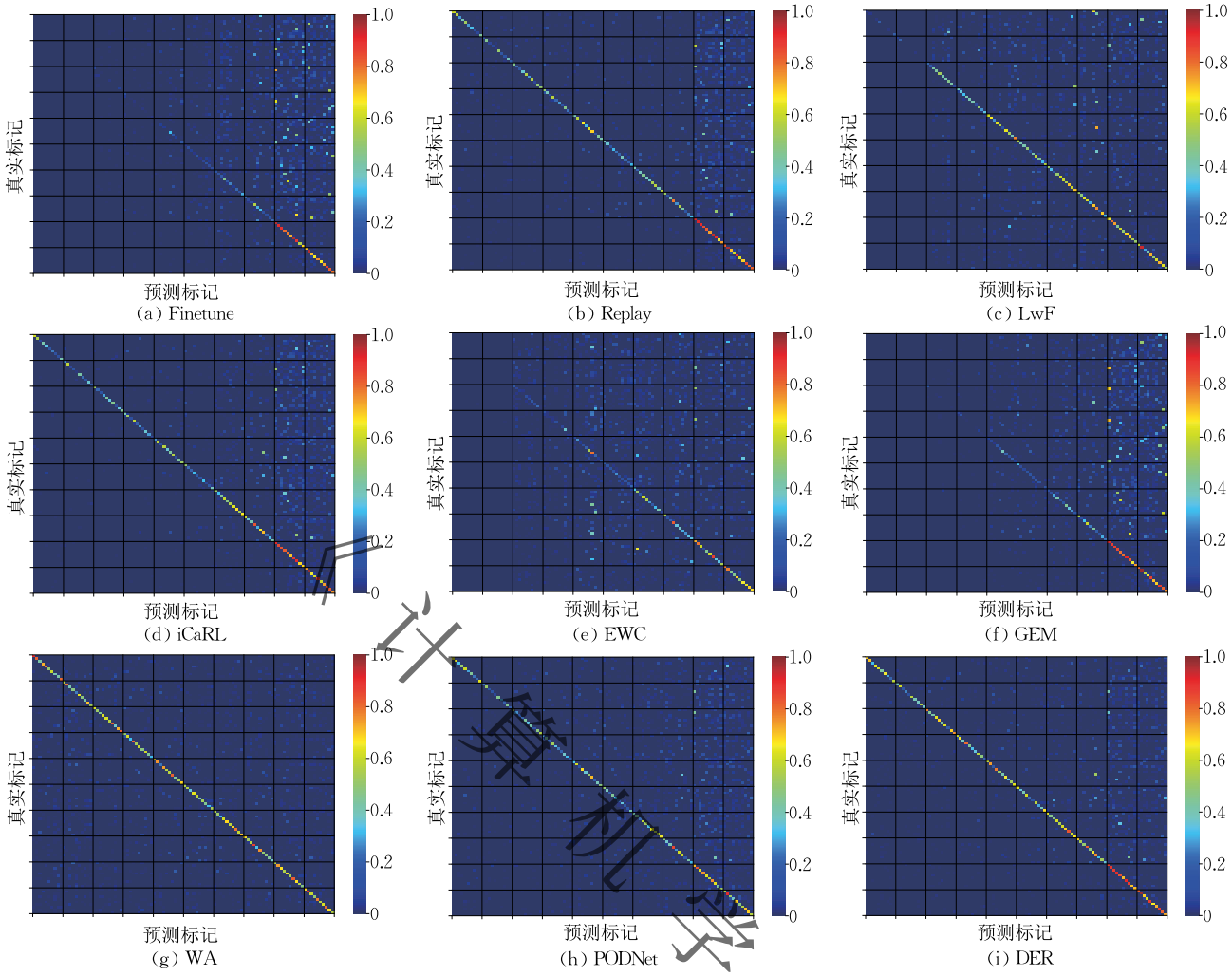


图 8 不同类别增量学习算法在 CIFAR100 数据集上预测结果的混淆矩阵

会导致模型的预测偏向于最后一个阶段,在混淆矩阵上显示出内部明显的不均匀.在加入范例集进行数据重放后,模型能够在一定程度上缓解这种预测上的偏好,并一定程度上地缓解灾难性遗忘(图 8(b)).对比 LwF(图 8(c))和 iCaRL(图 8(d)),可以发现在范例集的帮助下,iCaRL 进一步地缓解了灾难性遗忘,并展示出优秀的分类性能.观察 EWC(图 8(e))和 GEM(图 8(f))的混淆矩阵可以发现,基于参数正则和基于数据约束的类别增量学习方法并不能在当前设定下取得较好的性能,并遭受了严重的灾难性遗忘,这种现象与论文中的分类指标一致.最后,观察 WA(图 8(g))、PODNet(图 8(e))和 DER(图 8(i))可以发现,这三种方法在抵抗灾难性遗忘和学习新类的性能上均得了较好的结果,在所有类别上均具有较好的判别性能.

5.3.5 超参数选择分析

由于涉及算法较多,不同算法在超参数的选择上存在不同的设定.在本节中,首先查阅了复现算法的原文,并将几种普遍使用的算法参数进行列举.对于每个算法的特殊参数,例如 iCaRL 中损失项的加权系数和 softmax 使用的温度参数,本文均采用原始论文中的默认值.因此,影响方法性能最重要的参数在于每个阶段的训练轮数($Epoch$)与优化器权重衰减系数 Weight Decay 值(WD).通过查阅相关文献,本文选择以下 5 组代表性的参数组合,在 CIFAR100(5 阶段)数据集上进行实验:

设定 1:参考 WA 的设定,$Epoch=100$,$WD=1e\text{-}5$;

设定 2:参考 BIC 的设定,$Epoch=200$,$WD=1e\text{-}5$;

设定 3:$Epoch=200$,$WD=1e\text{-}4$;

设定 4:$Epoch=150$,$WD=1e\text{-}4$;

设定 5:参考 DER 的设定(也是本文在上述实验中使用的默认参数),$Epoch=170$,$WD=2e\text{-}4$;

除了以上 5 组代表性参数得到的最终结果以外,本文也在表 4、表 5 中汇报了部分论文的官方结

果. 由于类别增量学习设定多样，部分论文在基础类别数目、每阶段增量类别数目、数据集选用等方面有所差异，因此无法获得所有算法在当前任务上的官方结果. 表 4、表 5 表明，本文使用的参数在所有参数的对比中，在绝大多数方法上均能获得优秀的性能，在复现的结果上也与官方论文汇报的原始结果相当.

表 4　不同类别增量学习算法使用不同参数的最终准确率对比(最优的结果被加粗表示. “—”表明原文未在该设定下进行实验)

实验设定	设定 1	设定 2	设定 3	设定 4	设定 5	原文结果
Finetune	**21.91**	21.58	17.10	17.49	17.07	—
Replay	43.07	43.67	43.52	43.42	**43.78**	—
GEM	36.10	**36.37**	34.64	34.97	34.09	—
LwF	**41.39**	39.78	37.16	40.01	40.00	35.20
iCaRL	54.30	54.55	54.58	54.87	**54.93**	53.20
EWC	20.24	**24.15**	21.08	22.23	19.87	—
WA	56.65	57.89	58.20	57.67	**58.97**	59.20
PODNet	44.21	43.11	45.45	45.27	**49.08**	—
BiC	48.12	49.17	53.64	51.79	**56.75**	56.60
DER	61.34	61.37	62.78	63.12	**64.40**	65.40

表 5　不同类别增量学习算法使用不同参数的平均准确率对比(最优的结果被加粗表示. “—”表明原文未在该设定下进行实验)

实验设定	设定 1	设定 2	设定 3	设定 4	设定 5	原文结果
Finetune	**40.77**	40.23	37.85	37.98	37.90	—
Replay	59.24	59.84	59.96	59.76	**60.03**	—
GEM	51.56	**51.85**	51.20	51.07	50.32	—
LwF	**54.33**	52.02	51.89	53.51	48.96	53.20
iCaRL	66.36	66.07	66.18	66.01	**67.00**	66.72
EWC	38.02	**41.89**	39.86	40.76	39.18	—
WA	68.00	67.97	68.27	67.91	**68.50**	66.60
PODNet	59.08	58.51	60.58	60.11	**62.99**	—
BiC	61.34	61.87	64.87	64.53	**67.80**	68.18
DER	70.67	70.27	71.17	71.09	**71.82**	72.31

5.3.6　新闻分类场景下的类别增量学习

本文 5.1 节主要对典型的类别增量学习图片数据集进行了介绍与实证. 考虑到类别增量学习问题在除图像分类以外的文本分类表格数据等其他诸多领域也具有广泛应用，本节针对新闻文本内容分类进行探讨，验证多种典型类别增量学习算法在该任务上的性能. 新闻文本分类属于典型的自然语言处理问题，有效的类别增量学习算法能够使得分类器模型不断地掌握对新出现的新闻话题分类的能力，同时维持对以往话题分类的能力不遭受灾难性遗忘.

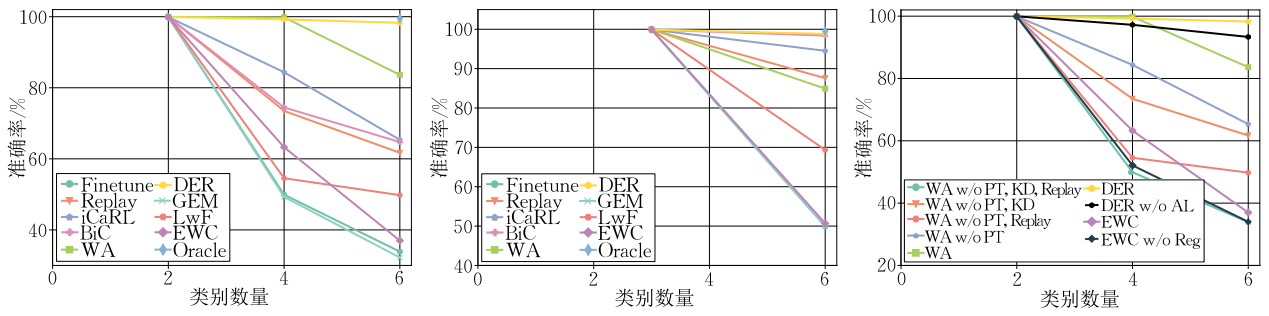
对于新闻文本分类数据集，本文参考文献[15-16]中公开的 NYTimes 数据集. NYTimes 数据集基于纽约时报 API 爬取了 2014 至 2017 年的 35 000 条纽约时报新闻数据. 每篇新闻内容依照纽约时报标签，被划分为“艺术”、“经济”、“运动”、“美国”、“科技”与“国际”，共 6 类. 每篇新闻内容均使用词向量(word2vec)[183]转化为 100 维的向量. 本文将以上 6 类数据分别划分为 2 阶段(每阶段 3 个类)和 3 阶段(每阶段 2 个类)进行增量学习训练，使用 5 层全连接神经网络(隐藏层维度为 100)作为主干网络，并参考 5.2.2 节的内容进行了超参数设置.

10 种不同类别增量学习算法在 NYTimes 数据集上的增量学习性能如图 9 所示. 可以从中总结出以下实验结论：

(1) 在新闻文本分类问题上，不同方法的整体性能趋势与图片分类问题中基本保持一致.

(2) 直接基于新类别的数据微调模型会使其遭受灾难性遗忘，基于数据约束(GEM)和参数正则(EWC)的方法无法显著克服灾难性遗忘. 相反，基于数据重放的方法(Replay)能够明显地提升模型性能，这表明通过数据重放回顾旧类知识无论在图片还是文本数据集上均能取得不错的效果.

(3) 基于知识蒸馏的类别增量学习算法(iCaRL)相比数据重放算法进一步提升了分类性能，基于滞后调节的算法(WA，BiC)进一步地缓解了分类器偏置. 基于动态模型结构的算法(DER)取得了最好的性能.



(a) NYTimes-3阶段　(b) NYTimes-2阶段　(c) 消融实验

图 9　不同类别增量学习算法在新闻分类数据集 NYTimes 上的实验分析

5.3.7 消融实验

本文主要对 10 种典型的类别增量学习算法进行了对比分析,考虑到不同算法引入不同的模块,对各个算法进行消融分析能够帮助理解不同算法不同模块的重要程度.因此,本节主要针对上述类别增量学习算法在 NYTimes 数据集上进行消融实验,采用 NYTimes-3 阶段设定,保持其他设置与 5.3.6 节中一致.不同算法的消融实验结果如图 9(c)所示.

需要注意的是,以上 10 种典型类别增量学习算法由于在发展过程中存在先后关系,因此存在互为消融的关系.例如,基于后调节的算法 WA,BiC 消去后调节过程后便变为了基于知识蒸馏的算法 iCaRL.iCaRL 消去数据重放过程则成为了 LwF 算法,若消去知识蒸馏损失则成为了 Replay 算法,Replay 算法再进一步消去数据重放过程则成为了基线算法 Finetune.因此,本节的消融实验主要考虑 WA、DER 以及 EWC.

图 9(c)中"w/o"表示删去某个模块,PT 代表 WA 算法中的后调节过程,KD 代表知识蒸馏损失,Replay 代表数据重放过程,AL 代表 DER 算法中的多样性损失,Reg 代表 EWC 算法中的参数正则项.从图 9(c)的消融实验中可以分析得出以下结论:

(1)在 WA 算法中,对性能影响较大的是后调节和数据重放过程,消融后分别会导致模型在当前任务最终性能上 18%和 15%.

(2)DER 算法中的多样性损失和 EWC 算法中的参数正则项对当前任务的性能影响不大.

5.4 实验结论

综合以上几节的实验内容,本文得出以下 3 个主要结论:

(1)相比于数据约束,利用范例集进行数据重放可以极大地缓解灾难性遗忘,是一种更高效且简单地利用范例集提升模型性能的手段.

(2)基于知识蒸馏的类别增量学习算法能够有效地利用范例集样本,基于后调节的类别增量学习算法可以进一步地缓解其中因类别不平衡导致的归纳偏好.

(3)基于动态模型结构的算法能够在多个数据集和多种设定下取得当前类别增量学习任务中最好的性能,但是性能的提升来源于额外的存储开销和运行时间.因此不适合部署在需要同时考虑算法运行效率和存储空间的场景.

6 类别增量学习的未来与展望

本节主要从类别增量学习的三个分类角度对其发展方向进行讨论,并结合实验章节中的一些结论,讨论开放动态环境下类别增量学习方法的发展方向.

6.1 数据层面

愈发一般的学习场景.当前的类别增量学习场景距离真实应用场景依然存在很多额外的限制.例如,要求模型存储一定数量的旧类样本作为范例集、要求给定增量学习任务的边界以评估参数重要程度或固定旧模型、在单个任务内要求进行多轮训练而非在线训练等.若将类别增量学习算法部署到开放动态环境中,则需要研究能够不借助范例集或额外生成式模型的类别增量学习算法[32]、能够不依赖任务边界进行增量学习的训练范式[79]、能够完全在线更新模型的训练技巧[184]等.对于不给定任务边界的增量学习场景,则可以考虑设计有效地任务边界检测器,用于判断数据流是否发生了概念漂移.若模型更新完全不能依赖范例集,则可以考虑基于无需数据的知识蒸馏方法[185]从模型中提炼范例样本.

愈发复杂的数据形式.开放动态环境中,数据往往呈现小样本[186]、多模态[187]、无标记[188]、弱监督[189]、有噪声[190]等复杂特性.适应开放动态环境的类别增量学习算法应当能够鲁棒地应对复杂环境带来的多样化模型输入.目前,已经有一些工作关注复杂环境下的增量学习问题.文献[156,158,191-194]关注如何改进知识蒸馏方法以将类别增量学习算法扩展到小样本类别增量学习任务中.文献[195]提出适用于视觉对话生成的多模态增量学习算法.文献[196]旨在设计算法,利用无标记样本进行增量表示学习.文献[197]关注数据存在噪声环境下的类别增量学习研究.对于小样本类别增量学习场景,文献[194,198]创造性地提出"向前兼容性"(forward compatibility)这一概念,通过在初始训练阶段为后续模型更新做好准备,使模型不依赖后续调整就可以获得有效的特征表示.对于多模态类别增量学习场景,则应当对应地考虑模态间的相互作用和关联,通过模态关联性抵抗灾难性遗忘.

6.2 参数层面

参数精简的网络结构.当前的类别增量学习算法往往部署在较短的数据流中,然而真实应用中的增量学习模型可能需要部署在移动终端上,并进行长期的增量学习和模型更新过程.使用动态模型结构的类别增量学习算法需要引入额外的模型存储,这种额外的存储开销往往会随着增量学习任务数的增多而线性增长.因此,开放动态环境下的应用需要设计适应长数据流的类别增量学习模型结构,并保证整个增量学习过程中模型参数具有平缓的增长速

《计算机学报》

度. 文献[98,199]研究了有效替代范例集的存储方式，然而目前尚无工作关注如何设计类别增量学习模型的紧凑表示. 对于模型大小受限的类别增量学习过程，文献[200]创造性地提出先扩张后压缩的学习框架，通过模型扩张增强判别能力，而后通过知识蒸馏进行模型压缩，减少存储开销. 同样地，也可以考虑模型剪枝[201-203]手段，在不伤害模型判别能力的情况下改善模型的存储开销.

任务维度的参数优化. 当前类别增量学习算法对于模型参数的优化方式是基于样本维度的，而元学习[204-205](meta-learning)是一种更高阶的参数优化方式，旨在通过学习大量采样出的任务学得模型优化的一般性方式，并抽取出适应任务的归纳偏好. 由于元学习算法能够利用旧任务的学习经验帮助新任务的学习，因此在类别增量学习领域具有广泛的应用前景. 文献[186]研究了如何借助元学习进行单阶段类别增量学习. 文献[206]从梯度优化方面论证了元学习对增量学习后续任务的帮助. 相似地，文献[207]从特征提取角度验证了元学习对增量学习的有效性. 文献[208]研究了如何利用元学习辅助任务增量学习过程. 可通过在增量学习过程中构造元学习任务[198,205,209]，将模型在元学习任务中学得的可泛化的学习能力应用到真实的类别增量学习任务中.

6.3　算法层面

开放世界的学习范式. 在开放动态环境下，模型应当不仅能学习新类，更应当拥有检测未知新类的能力. 在这种场景下，分类器能够自主地检测和学习未知新类，从而实现自动化的学习过程[20]. 检测新类要求模型具有开放集识别[18](open-set recognition)和新类发现[210](novel class discovery)的能力. 其中开放集识别指模型能够在区分已知类的同时检测数据集中的未知类. 新类发现指模型能够从多个未知类构成的集合中发掘出不同新类的子簇. 将以上二者和类别增量学习过程结合，便实现了开放世界学习[211](open-world learning). 目前机器学习领域正在关注开放世界学习的研究，包括语义分割[212]、人脸检测[213]、图像分类[20]等. 若要求类别增量学习模型应对开放动态环境的输入，则应将其与开放集检测模型进行结合，使模型能够检测未知类别的输入，并同时拥有顺序化学习新类的能力[16-17,20,214].

双向传递的知识迁移. 当前基于知识蒸馏的类别增量学习算法使用旧模型对新模型进行指导，从而缓解模型在旧类别上的灾难性遗忘. 然而，很少有方法关注如何利用旧模型帮助新类别的学习，这样的学习范式被称作双向知识迁移. 基于零样本学习[215](zero-shot learning)的方法关注如何使旧知识辅助新任务的学习，或可对知识的双向传递有所帮助. 在这方面，Zhou 等人[130]提出了利用增量学习过程中新类和旧类间的语义相似关系[216-217]指导分类器间最优输运[218-219]的增量学习算法，实现了从旧模型到新模型的知识迁移. 若要增强模型的前向知识迁移能力，可以考虑不同的预训练手段[194,220]和无监督学习[188]范式，在模型的初始训练阶段增强判别能力以辅助后续的学习过程. 另一方面，可通过存储旧模型[221]，并基于旧模型合成新增量模型的方式进行前向知识迁移.

7　结束语

设计行之有效的类别增量学习算法对于在开放动态环境下构建鲁棒、可拓展的学习模型具有重大意义，并引发了大量关注. 本文主要着眼于基于深度学习的类别增量学习算法，并从三方面对当前已有的研究成果进行了分类和总结. 此外，本文还将 10 种不同类型的类别增量学习的算法在基准数据集上进行了广泛的实验验证，希望对相关研究人员提供些许参考.

致　谢　感谢审稿专家提出的宝贵意见！

参　考　文　献

[1] Xu M, Guo L Z. Learning from group supervision: The impact of supervision deficiency on multi-label learning. Science China Information Sciences, 2021, 64(3): 1-13

[2] Yang C, Liu G, Yan C, et al. A clustering-based flexible weighting method in adaboost and its application to transaction fraud detection. Science China Information Sciences, 2021, 64(222101): 1-11

[3] Shang T, Zhao Z, Ren X, et al. Differential identifiability clustering algorithms for big data analysis. Science China Information Sciences, 2021, 64(5): 1-18

[4] Wang Z, Liu X, Lin J, et al. Multi-attention based cross-domain beauty product image retrieval. Science China Information Sciences, 2020, 63(2): 1-3

[5] Cheng W, Cai R, Zeng L, et al. IMCI: An efficient fingerprint retrieval approach based on 3D stacked memory. Science China Information Sciences, 2020, 63(7): 1-3

[6] Li G, Liu H, Li G, et al. LSTM-based argument recommendation for non-API methods. Science China Information Sciences, 2020, 63(9): 1-22

[7] Kong X, Han W, Liao L, et al. An analysis of correctness for API recommendation: Are the unmatched results useless? Science China Information Sciences, 2020, 63(9): 1-15

[8] Gomes H M, Barddal J P, Enembreck F, et al. A survey on ensemble learning for data stream classification. ACM Computing Surveys, 2017, 50(2): 1-36

[9] De Lange M, Aljundi R, Masana M, et al. A continual learning survey: Defying forgetting in classification tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear

[10] McCloskey M, Cohen N J. Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation, 1989, 24: 109-165

[11] McClelland J L, McNaughton B L, O'Reilly R C. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 1995, 102(3): 419-457

[12] Zhou Z H. Learnware: On the future of machine learning. Frontiers of Computer Science, 2016, 10(4): 589-590

[13] Zhang Y J, Yan Y H, Zhao P, et al. Towards enabling learnware to handle unseen jobs//Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 10964-10972

[14] Zhang Y J, Zhao P, Ma L, et al. An unbiased risk estimator for learning with augmented classes. Advances in Neural Information Processing Systems, 2020, 33: 10247-10258

[15] Mu X, Zhu F, Du J, et al. Streaming classification with emerging new class by class matrix sketching//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 2373-2379

[16] Mu X, Ting K M, Zhou Z H. Classification under streaming emerging new classes: A solution using completely-random trees. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(8): 1605-1618

[17] Wei X S, Ye H J, Mu X, et al. Multi-instance learning with emerging novel class. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(5): 2109-2120

[18] Zhou D W, Ye H J, Zhan D C. Learning placeholders for open-set recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4401-4410

[19] Biesialska M, Biesialska K, Costa-Jussà M R. Continual lifelong learning in natural language processing: A survey//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain, 2020: 6523-6541

[20] Zhou D W, Yang Y, Zhan D C. Learning to classify with incremental new class. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(6): 2429-2443

[21] Bremner A J, Lewkowicz D J, Spence C. Multisensory Development. Oxford, UK: Oxford University Press, 2012

[22] Tani J. Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena. New York, USA: Oxford University Press, 2016

[23] Grossberg S T. Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control: Volume 70. Berlin, Germany: Springer Science & Business Media, 2012

[24] Oren G, Wolf L. In defense of the learning without forgetting for task incremental learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2209-2218

[25] Masana M, Tuytelaars T, van de Weijer J. Ternary feature masks: Zero forgetting for task-incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3570-3579

[26] Bulat A, Kossaifi J, Tzimiropoulos G, et al. Incremental multi-domain learning with network latent tensor factorization //Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 10470-10477

[27] Xu J, Zhu Z. Reinforced continual learning//Proceedings of the Annual Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 899-908

[28] Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 2017, 114(13): 3521-3526

[29] Wang L, Zhang Y, Xu B, et al. Reliability concept drift online measurement for composite cloud systems. SCIENTIA SINICA Informationis, 2021, 51(9): 1438-1450

[30] Zhao P, Zhou Z H. Learning from distribution-changing data streams via decision tree model reuse. SCIENTIA SINICA Informationis, 2021, 51: 1-12

[31] Zhao P, Cai L W, Zhou Z H. Handling concept drift via model reuse. Machine Learning, 2020, 109(3): 533-568

[32] Yang Y, Zhou D W, Zhan D C, et al. Adaptive deep models for incremental learning: Considering capacity scalability and sustainability//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, USA, 2019: 74-82

[33] Zhao P, Wang G, Zhang L, et al. Bandit convex optimization in nonstationary environments. Journal of Machine Learning Research, 2021, 22(1): 5562-5606

[34] Zhao P, Zhang L, Jiang Y, et al. A simple approach for non-stationary linear bandits//Proceedings of the International Conference on Artificial Intelligence and Statistics. 2020: 746-755

[35] French R M. Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences, 1999, 3(4): 128-135

[36] Zhou Z H, Chen Z Q. Hybrid decision tree. Knowledge-Based Systems, 2002, 15(8): 515-528

[37] Fink M, Shalev-Shwartz S, Singer Y, et al. Online multiclass learning by interclass hypothesis sharing//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 313-320

[38] Muhlbaier M D, Topalis A, Polikar R. Learn^{++}.NC: Combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes. IEEE Transactions on Neural Networks and Learning Systems, 2008, 20(1): 152-168

[39] Da Q, Yu Y, Zhou Z H. Learning with augmented class by exploiting unlabeled data//Proceedings of the AAAI Conference on Artificial Intelligence. Québec City, Canada, 2014: 1760-1766

[40] van de Ven G M, Tolias A S. Three scenarios for continual learning. arXiv preprint arXiv:1904.07734, 2019

[41] Lesort T, Lomonaco V, Stoian A, et al. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. Information Fusion, 2020, 58: 52-68

[42] Parisi G I, Kemker R, Part J L, et al. Continual lifelong learning with neural networks: A review. Neural Networks, 2019, 113: 54-71

[43] You K, Long M, Cao Z, et al. Universal domain adaptation// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 2720-2729

[44] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks, 2010, 22(2): 199-210

[45] Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359

[46] Lopez-Paz D, Ranzato M. Gradient episodic memory for continual learning//Proceedings of the Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 6467-6476

[47] Tao X, Hong X, Chang X, et al. Bi-objective continual learning: Learning 'new' while consolidating 'known'// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 5989-5996

[48] Kundu J N, Venkatesh R M, Venkat N, et al. Class-incremental domain adaptation//Proceedings of the European Conference on Computer Vision. Cham, Swiss: Springer, 2020: 53-69

[49] Xie J, Yan S, He X. General incremental learning with domain-aware categorical representations. arXiv preprint arXiv: 2204.04078, 2022

[50] Simon C, Faraki M, Tsai Y H, et al. On generalizing beyond domains in cross-domain continual learning. arXiv preprint arXiv:2203.03970, 2022

[51] Rusu A A, Rabinowitz N C, Desjardins G, et al. Progressive neural networks. arXiv preprint arXiv:1606.04671, 2016

[52] Aljundi R, Chakravarty P, Tuytelaars T. Expert gate: Lifelong learning with a network of experts//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3366-3375

[53] Schwarz J, Czarnecki W, Luketina J, et al. Progress & compress: A scalable framework for continual learning//Proceedings of the International Conference on Machine Learning. Stockholmsmässan, Sweden, 2018: 4528-4537

[54] Yoon J, Yang E, Lee J, et al. Lifelong learning with dynamically expandable networks//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018

[55] Lewandowsky S, Li S C. Catastrophic interference in neural networks: Causes, solutions, and data//Frank D. Interference and Inhibition in Cognition. San Diego: Academic Press, 1995: 329-361

[56] Ratcliff R. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. Psychological Review, 1990, 97(2): 285

[57] Burgess N, Shapiro J, Moore M. Neural network models of list learning. Network: Computation in Neural Systems, 1991, 2(4): 399-422

[58] Nadal J, Toulouse G, Changeux J, et al. Networks of formal neurons and memory palimpsests. Europhysics Letters, 1986, 1(10): 535-542

[59] Kuzborskij I, Orabona F, Caputo B. From n to $n+1$: Multiclass transfer incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 3358-3365

[60] Mai Z, Li R, Jeong J, et al. Online continual learning in image classification: An empirical survey. Neurocomputing, 2022, 469: 28-51

[61] Zhang L. Online learning in changing environments//Proceedings of the 29th International Joint Conference on Artificial Intelligence. 2020: 5178-5182

[62] Wan Y, Zhang L. Efficient adaptive online learning via frequent directions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, to appear

[63] Masana M, Liu X, Twardowski B, et al. Class-incremental learning: Survey and performance evaluation on image classification. arXiv preprint arXiv:2010.15277, 2020

[64] Yu Han, Cai Hong-Ming, Zhang Yi-Fei, et al. An approach to constructing adaptive crowd groups based on incremental stream processing. Chinese Journal of Computers, 2020, 43(12): 2337-2351(in Chinese)
(于晗，蔡鸿明，张翼飞等. 基于增量式流处理的自适应群体划分方法. 计算机学报，2020，43(12)：2337-2351)

[65] Wang Zhuo, Chen Qun, Li Zhan-Huai, et al. An incremental partitioning strategy for data balance on MapReduce. Chinese Journal of Computers, 2016, 39(1): 19-35(in Chinese)
(王卓，陈群，李战怀等. 基于增量式分区策略的 MapReduce 数据均衡方法. 计算机学报，2016，39(1)：19-35)

[66] Fan Qiu-Shi, Zhou Min-Qi, Zhou Ao-Ying. A distributed join algorithm on separated data storage. Chinese Journal of Computers, 2016, 39(10): 2102-2113(in Chinese)
(樊秋实，周敏奇，周傲英. 基线与增量数据分离架构下的分布式连接算法. 计算机学报，2016，39(10)：2102-2113)

[67] Feng Jie-Ming, Li Zhan-Huai, Chen Qun, et al. Incremental locally weighted learning for adaptive cardinality estimation of query template. Chinese Journal of Computers, 2022, 45(1): 17-34(in Chinese)
(冯杰明，李战怀，陈群等. 基于增量局部加权学习的查询模板自适应基数估计. 计算机学报，2022，45(1)：17-34)

[68] Zhong J, Shaikh R A, Haoguo W, et al. Classification of PBMC cell types using scRNAseq, ANN, and incremental

计算机学报

learning//Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2020: 1351-1355

[69] Zheng K, You Z H, Wang L, et al. MISSIM: An incremental learning-based model with applications to the prediction of miRNA-disease association. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020, 18(5): 1733-1742

[70] Wang Z, Yang Y, Wen R, et al. Lifelong learning based disease diagnosis on clinical notes//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham, Swiss: Springer, 2021: 213-224

[71] Zhang Y, Wang X, Yang D. Continual sequence generation with adaptive compositional modules. arXiv preprint arXiv: 2203.10652, 2022

[72] Ke Z, Xu H, Liu B. Adapting BERT for continual learning of a sequence of aspect sentiment classification tasks. arXiv preprint arXiv:2112.03271, 2021

[73] Ke Z, Liu B, Xu H, et al. Classic: Continual and contrastive learning of aspect sentiment classification tasks. arXiv preprint arXiv:2112.02714, 2021

[74] Xu Y, Zhang Y, Guo W, et al. GraphSAIL: Graph structure aware incremental learning for recommender systems//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 2861-2868

[75] Mi F, Lin X, Faltings B. ADER: Adaptively distilled exemplar replay towards continual learning for session-based recommendation//Proceedings of the 14th ACM Conference on Recommender Systems. 2020: 408-413

[76] Rebuffi S A, Kolesnikov A, Sperl G, et al. iCaRL: Incremental classifier and representation learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2001-2010

[77] Koh H, Kim D, Ha J W, et al. Online continual learning on class incremental blurry task configuration with anytime inference. arXiv preprint arXiv:2110.10031, 2021

[78] Bang J, Kim H, Yoo Y, et al. Rainbow memory: Continual learning with a memory of diverse samples//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8218-8227

[79] Aljundi R, Lin M, Goujaud B, et al. Online continual learning with no task boundaries. arXiv preprint arXiv:1903.08671, 2019

[80] Welling M. Herding dynamical weights to learn//Proceedings of the International Conference on Machine Learning. Montreal, Canada, 2009: 1121-1128

[81] Isele D, Cosgun A. Selective experience replay for lifelong learning//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 3302-3309

[82] Shin H, Lee J K, Kim J, et al. Continual learning with deep generative replay//Proceedings of the Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 2990-2999

[83] Chaudhry A, Ranzato M, Rohrbach M, et al. Efficient lifelong learning with A-GEM//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2018

[84] Zenke F, Poole B, Ganguli S. Continual learning through synaptic intelligence//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 3987-3995

[85] Yan S, Xie J, He X. DER: Dynamically expandable representation for class incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3014-3023

[86] Li Z, Hoiem D. Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 2935-2947

[87] Hou S, Pan X, Loy C C, et al. Learning a unified classifier incrementally via rebalancing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 831-839

[88] Wu Y, Chen Y, Wang L, et al. Large scale incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 374-382

[89] Zhao B, Xiao X, Gan G, et al. Maintaining discrimination and fairness in class incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13208-13217

[90] Robins A. Catastrophic forgetting in neural networks: The role of rehearsal mechanisms//Proceedings 1993 the First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems. Dunedin, New Zealand, 1993: 65-68

[91] Robins A. Catastrophic forgetting, rehearsal and pseudorehearsal. Connection Science, 1995, 7(2): 123-146

[92] Chaudhry A, Rohrbach M, Elhoseiny M, et al. On tiny episodic memories in continual learning. arXiv preprint arXiv:1902.10486, 2019

[93] Rolnick D, Ahuja A, Schwarz J, et al. Experience replay for continual learning//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 350-360

[94] Aljundi R, Lin M, Goujaud B, et al. Gradient based sample selection for online continual learning//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 11816-11825

[95] De Lange M, Tuytelaars T. Continual prototype evolution: Learning online from non-stationary data streams//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 8250-8259

[96] Mensink T, Verbeek J, Perronnin F, et al. Distance-based image classification: Generalizing to new classes at near-zero cost. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(11): 2624-2637

[97] Chaudhry A, Gordo A, Dokania P K, et al. Using hindsight to anchor past knowledge in continual learning//Proceedings

of the AAAI Conference on Artificial Intelligence. 2020: 6993-7001

[98] Iscen A, Zhang J, Lazebnik S, et al. Memory-efficient incremental learning through feature adaptation//Proceedings of the European Conference on Computer Vision. Springer, 2020: 699-715

[99] Hu W, Lin Z, Liu B, et al. Overcoming catastrophic forgetting for continual learning via model adaptation//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019

[100] Kemker R, Kanan C. FearNet: Brain-inspired model for incremental learning//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018

[101] Ostapenko O, Puscas M, Klein T, et al. Learning to remember: A synaptic plasticity driven framework for continual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 11321-11329

[102] Xiang Y, Fu Y, Ji P, et al. Incremental learning using conditional adversarial networks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 6619-6628

[103] Wang L, Yang K, Li C, et al. ORDisCo: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 5383-5392

[104] Jiang J, Cetin E, Celiktutan O. IB-DRR-incremental learning with information-back discrete representation replay//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3533-3542

[105] Zhu K, Cao Y, Zhai W, et al. Self-promoted prototype refinement for few-shot class-incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 6801-6810

[106] Wu C, Herranz L, Liu X, et al. Memory replay GANs: Learning to generate images from new categories without forgetting//Proceedings of the Annual Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 5966-5976

[107] Zhai M, Chen L, Mori G. Hyper-lifelongGAN: Scalable lifelong learning for image conditioned generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2246-2255

[108] Verma V K, Liang K J, Mehta N, et al. Efficient feature transformations for discriminative and generative continual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13865-13875

[109] Maracani A, Michieli U, Toldo M, et al. RECALL: Replay-based continual learning in semantic segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 7026-7035

[110] Kim S C, Yoo Y, Moon T, et al. SSUL: Semantic segmentation with unknown label for exemplar-based class-incremental learning. arXiv preprint arXiv:2106.11562, 2021

[111] Korycki L, Krawczyk B. Class-incremental experience replay for continual learning under concept drift//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3649-3658

[112] Wang S, Laskar Z, Melekhov I, et al. Continual learning for image-based camera localization//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 3252-3262

[113] Verwimp E, De Lange M, Tuytelaars T. Rehearsal revealed: The limits and merits of revisiting samples in continual learning. arXiv preprint arXiv:2104.07446, 2021

[114] Ye H J, Zhan D C, Chao W L. Procrustean training for imbalanced deep learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 92-102

[115] Zeng G, Chen Y, Cui B, et al. Continual learning of context-dependent processing in neural networks. Nature Machine Intelligence, 2019, 1(8): 364-372

[116] Tang S, Chen D, Zhu J, et al. Layerwise optimization by gradient decomposition for continual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9634-9643

[117] Wang S, Li X, Sun J, et al. Training networks in null space of feature covariance for continual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 184-193

[118] Dong J, Wang L, Fang Z, et al. Federated class-incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10164-10173

[119] MacKay D J. A practical Bayesian framework for backpropagation networks. Neural Computation, 1992, 4(3): 448-472

[120] Liu X, Masana M, Herranz L, et al. Rotate your networks: Better weight consolidation and less catastrophic forgetting//Proceedings of the 24th International Conference on Pattern Recognition (ICPR). Beijing, China, 2018: 2262-2268

[121] Lee J, Hong H G, Joo D, et al. Continual learning with extended Kronecker-factored approximate curvature//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9001-9010

[122] Chaudhry A, Dokania P K, Ajanthan T, et al. Riemannian walk for incremental learning: Understanding forgetting and intransigence//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 532-547

[123] Aljundi R, Babiloni F, Elhoseiny M, et al. Memory aware synapses: Learning what (not) to forget//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 139-154

[124] Aljundi R, Kelchtermans K, Tuytelaars T. Task-free continual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 11254-11263

计算机学报

[125] Lee S W, Kim J H, Jun J, et al. Overcoming catastrophic forgetting by incremental moment matching//Proceedings of the Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 4652-4662

[126] Shi Y, Yuan L, Chen Y, et al. Continual learning via bit-level information preserving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16674-16683

[127] Yang Y, Zhou D W, Zhan D C, et al. Cost-effective incremental deep model: Matching model capacity with the least sampling. IEEE Transactions on Knowledge and Data Engineering, 2021, to appear

[128] Zhou M, Xiao J, Chang Y, et al. Image de-raining via continual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4907-4916

[129] Mi F, Chen L, Zhao M, et al. Continual learning for natural language generation in task-oriented dialog systems//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings: 2020: 3461-3474

[130] Zhou D W, Ye H J, Zhan D C. Co-transport for class-incremental learning//Proceedings of the 29th ACM International Conference on Multimedia, 2021: 1645-1654

[131] Ebrahimi S, Meier F, Calandra R, et al. Adversarial continual learning//Proceedings of the European Conference on Computer Vision. Springer, 2020: 386-402

[132] Hung C Y, Tu C H, Wu C E, et al. Compacting, picking and growing for unforgetting continual learning//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019, 32: 13669-13679

[133] Li X, Zhou Y, Wu T, et al. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 3925-3934

[134] Serra J, Suris D, Miron M, et al. Overcoming catastrophic forgetting with hard attention to the task//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 4548-4557

[135] Rajasegaran J, Hayat M, Khan S, et al. Random path selection for incremental learning. Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 12669-12679

[136] Abati D, Tomczak J, Blankevoort T, et al. Conditional channel gated networks for task-aware continual learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3931-3940

[137] Mallya A, Davis D, Lazebnik S. Piggyback: Adapting a single network to multiple tasks by learning to mask weights//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 67-82

[138] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015

[139] Tang J, Wang K. Ranking distillation: Learning compact ranking models with high performance for recommender system//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 2289-2298

[140] Pan Y, He F, Yu H. A novel enhanced collaborative autoencoder with knowledge distillation for top-*n* recommender systems. Neurocomputing, 2019, 332: 137-148

[141] Zhou Z H, Jiang Y. NeC4.5: Neural ensemble based C4.5. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(6): 770-773

[142] Yim J, Joo D, Bae J, et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4133-4141

[143] Ahn S, Hu S X, Damianou A, et al. Variational information distillation for knowledge transfer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9163-9171

[144] Papernot N, Abadi M, Erlingsson U, et al. Semi-supervised knowledge transfer for deep learning from private training data//Proceedings of the International Conference on Learning Representations. Toulon, France, 2017

[145] Gou J, Yu B, Maybank S J, et al. Knowledge distillation: A survey. International Journal of Computer Vision, 2021, 129(6): 1789-1819

[146] Castro F M, Marín-Jiménez M J, Guil N, et al. End-to-end incremental learning//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 233-248

[147] Zhang J, Zhang J, Ghosh S, et al. Class-incremental learning via deep model consolidation//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 1131-1140

[148] Lee K, Lee K, Shin J, et al. Overcoming catastrophic forgetting with unlabeled data in the wild//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 312-321

[149] Douillard A, Cord M, Ollion C, et al. PODNet: Pooled outputs distillation for small-tasks incremental learning//Proceedings of the European Conference on Computer Vision. Springer, 2020: 86-102

[150] Dhar P, Singh R V, Peng K C, et al. Learning without memorizing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5138-5146

[151] Hou S, Pan X, Loy C C, et al. Lifelong learning via progressive distillation and retrospection//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 437-452

[152] Liu Y, Su Y, Liu A A, et al. Mnemonics training: Multi-class incremental learning without forgetting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12245-12254

[153] Smith J, Hsu Y C, Balloch J, et al. Always be dreaming: A new approach for data-free class-incremental learning. arXiv preprint arXiv:2106.09701, 2021

[154] Yin H, Molchanov P, Alvarez J M, et al. Dreaming to distill: Data-free knowledge transfer via deepinversion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8715-8724

[155] Zhu F, Zhang X Y, Wang C, et al. Prototype augmentation and self-supervision for incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 5871-5880

[156] Dong S, Hong X, Tao X, et al. Few-shot class-incremental learning via relation knowledge distillation//Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 1255-1263

[157] Tao X, Chang X, Hong X, et al. Topology-preserving class-incremental learning//Proceedings of the European Conference on Computer Vision. Springer, 2020: 254-270

[158] Tao X, Hong X, Chang X, et al. Few-shot class-incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12183-12192

[159] Liu Y, Hong X, Tao X, et al. Model behavior preserving for class-incremental learning. IEEE Transactions on Neural Networks and Learning Systems, 2022

[160] Simon C, Koniusz P, Harandi M. On learning the geodesic path for incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1591-1600

[161] Hu W, Qin Q, Wang M, et al. Continual learning by using information of each class holistically//Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 7797-7805

[162] Douillard A, Chen Y, Dapogny A, et al. PLOP: Learning without forgetting for continual semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4040-4050

[163] Pu N, Chen W, Liu Y, et al. Lifelong person re-identification via adaptive knowledge accumulation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 7901-7910

[164] Park J, Kang M, Han B. Class-incremental learning for action recognition in videos//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13698-13707

[165] Hu X, Tang K, Miao C, et al. Distilling causal effect of data in class-incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3957-3966

[166] Zhai M, Chen L, Tung F, et al. Lifelong GAN: Continual learning for conditional image generation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 2759-2768

[167] Ma J, Tao X, Ma J, et al. Class incremental learning for video action classification//Proceedings of the 2021 IEEE International Conference on Image Processing. 2021: 504-508

[168] Prabhu A, Torr P H, Dokania P K. GDumb: A simple approach that questions our progress in continual learning//Proceedings of the European Conference on Computer Vision. Springer, 2020: 524-540

[169] Joseph K, Khan S, Khan F S, et al. Towards open world object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 5830-5840

[170] Li Y, Yuan Y. Convergence analysis of two-layer neural networks with ReLU activation. arXiv preprint arXiv: 1705.09886, 2017

[171] Belouadah E, Popescu A. IL2M: Class incremental learning with dual memory//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 583-592

[172] Liu Y, Schiele B, Sun Q. Adaptive aggregation networks for class-incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2544-2553

[173] Yu L, Twardowski B, Liu X, et al. Semantic drift compensation for class-incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6982-6991

[174] Ye H J, Chen H Y, Zhan D C, et al. Identifying and compensating for feature deviation in imbalanced deep learning. arXiv preprint arXiv:2001.01385, 2020

[175] He C, Wang R, Chen X. A tale of two CILs: The connections between class incremental learning and class imbalanced learning, and beyond//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3559-3569

[176] Ahn H, Kwak J, Lim S, et al. SS-IL: Separated softmax for incremental learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 844-853

[177] Krizhevsky A, Hinton G, et al. Learning multiple layers of features from tiny images. Toronto, Canada: University of Toronto, Technical report, 1, 2009

[178] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 Dataset. Pasadena, Canada: California Institute of Technology: CNS-TR-2011-001, 2011

[179] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248-255

[180] Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 8026-8037

[181] Bottou L. Stochastic gradient descent tricks//Klaus-Robert M. Neural networks: Tricks of the trade. Berlin, Germany: Springer, 2012: 421-436

[182] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2015: 770-778

[183] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space//Proceedings of the International Conference on Learning Representations (ICLR). Scottsdale, USA, 2013

[184] He J, Mao R, Shao Z, et al. Incremental learning in online scenario//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13926-13935

[185] Lopes R G, Fenu S, Starner T. Data-free knowledge distillation for deep neural networks. arXiv preprint arXiv: 1710.07535, 2017

[186] Ye H J, Hu H, Zhan D C. Learning adaptive classifiers synthesis for generalized few-shot learning. International Journal of Computer Vision, 2021, 129(6): 1930-1953

[187] Ramachandram D, Taylor G W. Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine, 2017, 34(6): 96-108

[188] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9729-9738

[189] Berthelot D, Carlini N, Goodfellow I, et al. MixMatch: A holistic approach to semi-supervised learning//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 5050-5060

[190] Xia X, Liu T, Wang N, et al. Are anchor points really indispensable in label-noise learning?//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019, 32: 6838-6849

[191] Zhang C, Song N, Lin G, et al. Few-shot incremental learning with continually evolved classifiers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12455-12464

[192] Shi G, Chen J, Zhang W, et al. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima//Proceedings of the Annual Conference on Neural Information Processing Systems. 2021: 6747-6761

[193] Cheraghian A, Rahman S, Ramasinghe S, et al. Synthesized feature based few-shot class-incremental learning on a mixture of subspaces//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 8661-8670

[194] Zhou D W, Wang F Y, Ye H J, et al. Forward compatible few-shot class-incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 9046-9056

[195] Chen F, Meng F, Chen X, et al. Multimodal incremental transformer with visual grounding for visual dialogue generation. arXiv preprint arXiv:2109.08478, 2021

[196] Cha H, Lee J, Shin J. Co2l: Contrastive continual learning //Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 9516-9525

[197] Kim C D, Jeong J, Moon S, et al. Continual learning on noisy data streams via self-purified replay//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 537-547

[198] Zhou D W, Ye H J, Ma L, et al. Few-shot class-incremental learning by sampling multi-phase tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, in Press

[199] Hayes T L, Kafle K, Shrestha R, et al. Remind your neural network to prevent catastrophic forgetting//Proceedings of the European Conference on Computer Vision. Springer, 2020: 466-483

[200] Wang F Y, Zhou D W, Ye H J, et al. FOSTER: Feature boosting and compression for class-incremental learning. arXiv preprint arXiv:2204.04662, 2022

[201] Gao S, Huang F, Cai W, et al. Network pruning via performance maximization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9270-9280

[202] Qian C, Yu Y, Zhou Z H. Pareto ensemble pruning//Proceedings of the AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2935-2941

[203] Li N, Yu Y, Zhou Z H. Diversity regularized ensemble pruning//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg, 2012: 330-345

[204] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 1126-1135

[205] Ye H J, Sheng X R, Zhan D C. Few-shot learning with adaptively initialized task optimizer: A practical meta-learning approach. Machine Learning, 2020, 109(3): 643-664

[206] Riemer M, Cases I, Ajemian R, et al. Learning to learn without forgetting by maximizing transfer and minimizing interference//Proceedings of the International Conference on Learning Representations (ICLR). New Orleans, USA, 2018

[207] Javed K, White M. Meta-learning representations for continual learning//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 1820-1830

[208] Rajasegaran J, Khan S, Hayat M, et al. iTAML: An incremental task agnostic meta-learning approach//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13588-13597

[209] Chao W L, Ye H J, Zhan D C, et al. Revisiting meta-learning as supervised learning. arXiv preprint arXiv:2002.00573, 2020

[210] Fini E, Sangineto E, Lathuilière S, et al. A unified objective for novel class discovery//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 9284-9292

[211] Bendale A, Boult T. Towards open world recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1893-1902

[212] Cen J, Yun P, Cai J, et al. Deep metric learning for open world semantic segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 15333-15342

[213] Rostami M, Spinoulas L, Hussein M, et al. Detection and continual learning of novel face presentation attacks//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 14851-14860

[214] Zhou D W, Yang Y, Zhan D C. Detecting sequentially novel classes with stable generalization ability//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham, Swiss: Springer, 2021: 371-382

[215] Romera-Paredes B, Torr P. An embarrassingly simple approach to zero-shot learning//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 2152-2161

[216] Ye H J, Zhan D C, Jiang Y, et al. Heterogeneous few-shot model rectification with semantic mapping. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(11): 3878-3891

[217] Ye H J, Zhou D W, Hong L, et al. Contextualizing multiple tasks via learning to decompose. arXiv preprint arXiv: 2106.08112, 2021

[218] Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport//Proceedings of the Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 2292-2300

[219] Villani C. Optimal Transport: Old and New: Volume 338. Berlin, Germany: Springer Science & Business Media, 2008

[220] Shi Y, Zhou K, Liang J, et al. Mimicking the oracle: An initial phase decorrelation approach for class incremental learning. arXiv preprint arXiv: 2112.04731, 2021

[221] Zhou D W, Wang Q W, Ye H J, et al. A model or 603 exemplars: Towards memory-efficient class-incremental learning. arXiv preprint arXiv: 2205.13218, 2022

ZHOU Da-Wei, Ph. D. candidate. His research interests include incremental learning and open-set recognition.

WANG Fu-Yun, Undergraduate. His research interests include incremental learning.

YE Han-Jia, Ph. D., associate researcher. His research interests include metric learning and meta-learning.

ZHAN De-Chuan, Ph. D., professor. His research interests include machine learning and data mining.

Background

Deep models have achieved or even surpassed human-level performance in many tasks in recent years. Current deep models are deployed under the static environment, which requires the entire data before the learning process. However, the model cannot conduct further updating after the training process. By contrast, data in the real world often come with the stream format, containing incoming new classes. As a result, an ideal model should learn from streaming data and enhance its learning ability. Such a learning process, namely Class-Incremental learning, draws more attention from the machine learning community. Directly updating the model with new class data will cause the forgetting of old ones and destroy the total performance. As a result, the incremental model should learn new classes and meanwhile resist catastrophic forgetting.

This paper deeply summarizes and classifies some traditional and state-of-the-art algorithms for class-incremental learning from three aspects, i.e., input, parameters, and algorithm. The research of class-incremental learning is divided into several main perspectives, e.g., data rehearsal, data restriction, parameter regularization, dynamic architecture, knowledge distillation, and post-tuning. Solving class-incremental learning helps to understand the behavior of learning and forgetting in the learning systems. It will facilitate the design of robust and explainable models in the open world.

Besides, this paper conducts extensive experimental verification with ten typical algorithms under various settings and summarizes the common rules for class-incremental learning.

This research was supported by the National Natural Science Foundation of China (Nos. 61773198, 61921006, 62006112), the NSFC-NRF Joint Research Project under Grant No. 61861146001, the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the NSF of Jiangsu Province (No. BK20200313).