ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 5,Sep. 2025,pp. 1122－1138
DOI:10. 16337/j. 1004-9037. 2025. 05. 002

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

多模态持续学习方法研究进展

张　伟，钱龙玥，张　林，李　腾

（山东大学控制科学与工程学院，济南 250061）

摘　要：多模态持续学习（Multimodal continual learning，MMCL）作为机器学习和人工智能领域的一个重要研究方向，旨在通过融合多种模态数据（如图像、文本或语音等）来实现持续的知识积累与任务适应。相较于传统单模态学习方法，MMCL不仅能够并行处理多源异构数据，还能在有效保持已有知识的同时适应新任务需求，展现出在智能系统中的巨大应用潜力。本文系统性地对多模态持续学习进行综述。首先，从基本概念、评估体系和经典单模态持续学习方法3个维度阐述了MMCL的基础理论框架。其次，深入剖析了MMCL在实际应用中的优势与挑战：尽管其在多模态信息融合方面具有显著优势，但仍面临模态不平衡、异构性融合等关键挑战，这些挑战既制约了当前方法的性能表现，也为未来研究指明了方向。基于此，本文随后从基于回放、正则化、参数隔离和大模型4个主要方面，全面梳理了MMCL方法的研究现状与最新进展。最后，对MMCL的未来发展趋势进行了前瞻性展望。

关键词：多模态持续学习；模态对齐；灾难性遗忘；预训练模型；任务适应性

中图分类号：TP391.41　　**文献标志码：**A

Research Progress on Multimodal Continual Learning Methods

ZHANG Wei, QIAN Longyue, ZHANG Lin, LI Teng

(School of Control Science and Engineering, Shandong University, Jinan 250061, China)

Abstract: Multimodal continual learning (MMCL), as a significant research direction in the fields of machine learning and artificial intelligence, aims to achieve continuous knowledge accumulation and task adaptation through the integration of multiple modal data (such as images, text, audio, etc.). Compared with traditional single-modal learning methods, MMCL not only enables parallel processing of multi-source heterogeneous data, but also effectively retains existing knowledge while adapting to new task requirements, demonstrating immense application potential in intelligent systems. This paper provides a systematic review of multimodal continual learning. Firstly, the fundamental theoretical framework of MMCL is elaborated from three dimensions: Basic concepts, evaluation systems, and classical single-modal continual learning methods. Secondly, the advantages and challenges of MMCL in practical applications are thoroughly analyzed: Despite its significant advantages in multimodal information fusion, it still faces critical challenges such as modal imbalance and heterogeneous fusion, which not only constrain the performance of current methods but also indicate future research directions. Based on this, the paper then comprehensively reviews the research status and latest advancements in MMCL methods from four main aspects: Replay-based, regularization-based, parameter isolation-based, and large model-based

基金项目：新一代人工智能国家科技重大专项（2021ZD0112002）。
收稿日期：2025-03-07；**修订日期：**2025-06-04

approaches. Finally, a forward-looking perspective on the future development trends of MMCL is presented.

Key words: multimodal continual learning (MMCL); modality alignment; catastrophic forgetting; pre-trained models; task adaptation

引 言

传统的单模态持续学习(Unimodal continual learning, UMCL)方法在单一数据类型(如图像、文本或语音)上取得了显著进展,然而这些方法难以真实模拟人类处理信息的复杂过程。人类在感知和理解世界时,通常会依赖多种感知通道(如视觉、听觉、语言等)的协同作用[1],而单模态学习方法由于仅依赖于单一类型的数据,难以全面捕捉场景的丰富信息,容易受到数据质量或模态固有局限性的影响。这种局限性使得单模态学习在动态环境下的持续学习(Continual learning, CL)问题中表现不佳,特别是当模型需要不断学习新任务时,往往会因为新知识的引入而遗忘先前学到的知识,即所谓的灾难性遗忘问题[2],这严重制约了模型在实际应用中的表现。

相比之下,多模态学习通过融合来自不同模态的信息,能够弥补单一模态的不足,从而获得更加全面和多元的知识表示。利用多模态数据的互补信息,模型可以学习到更准确、更稳健的表示,显著减轻对单一模态规律的依赖,并有效缓解灾难性遗忘问题[2]。尤其是在复杂的真实场景中,多模态学习往往能够带来意料之外的良好效果,在提升模型的适应性、稳定性和泛化能力方面展现出巨大的潜力。因此,多模态持续学习(Multimodal continual learning, MMCL)逐渐成为当前人工智能领域的研究热点之一。然而,尽管MMCL具有显著的优势,其在增量学习过程中仍然面临诸多挑战。与UMCL相比,MMCL不仅需要平衡新任务的学习与旧任务知识的保留,还需要解决不同模态之间的知识对齐问题[3]。特别是在增量学习过程中,如何有效融合多模态信息并避免模态间的冲突,成为了一个亟待解决的关键问题。此外,多模态数据的异构性[4]和模态间的不平衡性[5]进一步增加了问题的复杂性。预训练模型(Pre-trained model, PTM)激发了广泛的研究兴趣,特别是如何利用PTM强大的表征能力来推动CL的发展[6]。PTM基于大规模数据集和先进技术进行训练,具有较强的泛化能力[7],在跨模态特征对齐和泛化能力方面表现出色,为MMCL提供了新的思路[8-10]和坚实的研究基础[11]。

本文将从理论基础、优势挑战和研究现状3个维度对MMCL进行系统性综述。通过全面回顾该领域的最新研究进展,本文不仅总结了当前的发展现状和重要成果,还提供了清晰的研究脉络和实用的参考文献,并且通过深入分析该领域面临的挑战和机遇,为未来研究指明了潜在方向。

1 预备知识

1.1 基本概念

多模态持续学习是一种结合多模态学习和持续学习的前沿研究方向,旨在使模型能够从多种数据模态(如文本、图像或音频等)中持续学习新任务,同时避免遗忘已学知识。多模态学习[12]通过整合不同模态的互补信息提升模型性能,而持续学习[13]则要求模型在不断接收新数据时动态适应,同时克服“灾难性遗忘”问题。多模态持续学习的核心挑战在于如何有效融合多模态信息,处理数据分布变化,并在持续学习过程中保持对旧任务的记忆。

在持续学习的背景下,单模态的持续学习通常根据研究场景分为类别增量、任务增量和域增量任务[14]。然而,多模态持续学习进一步增加了任务划分的复杂性。由于涉及多种模态,任务划分的维度

不仅限于上述3种,还可能包括模态增量任务[15],即模型需要逐步学习处理新的模态数据。此外,数据流的动态性和任务划分的多样性也对多模态持续学习提出了更高的要求。

数据集在多模态持续学习中扮演着关键角色。根据各个模态领域的任务要求和目标不同,有各自专属的任务数据集。根据任务的不同(如任务增量、类别增量或域增量),数据集需要进行不同的划分和处理。例如,在类别增量任务中,数据集可能被划分为多个阶段,每个阶段都引入新的类别;在域增量任务中,可能用不同的数据集表示不同的域。这种灵活的划分方式使得多模态持续学习能够更好地模拟真实世界中的动态环境,从而提升模型的适应性和实用性。

1.2 评估体系

在MMCL框架中,评估模型的性能需要综合考虑其在多个任务上的表现。由于持续学习框架可以应用于不同的下游任务(如分类、检索等),因此评估指标也会因任务类型而异。本节主要介绍UMCL与MMCL框架的通用评估指标。设共有 T 个序列任务,$R_{i,t}$ 为模型完成第 i 个任务训练之后在第 j 个任务测试集上的表现。

(1)平均表现(Average performance, AP)衡量模型在所有任务上的整体性能,计算公式为

$$\mathrm{AP}=\frac{1}{T}\sum_{i=1}^{T}R_{T,i} \tag{1}$$

式中 $R_{T,i}$ 表示模型在完成所有任务训练后,在第 i 个任务上的测试准确率。AP值越高,表明模型在所有任务上的整体性能越好。

(2)遗忘度量(Forgetting measure, FM)[16]评估持续学习方法的抗遗忘能力。第 k 个任务的遗忘度量定义为

$$\mathrm{FM}_k=\frac{1}{k-1}\sum_{j=1}^{k-1}\max_{l\in 1,2,\cdots,k-1}R_{l,j}-R_{k,j}\quad \forall j<k \tag{2}$$

式中:$R_{l,j}$ 表示模型在完成第 l 个任务训练后,在第 j 个任务上的测试性能;$R_{k,j}$ 表示模型在完成第 k 个任务训练后,在第 j 个任务上的测试性能。FM值越小,表明模型的抗遗忘能力越强。

(3)前向迁移(Forward transfer, FWT)[17]衡量模型对未来任务学习的帮助,计算公式为

$$\mathrm{FWT}=\frac{1}{T-1}\sum_{i=2}^{T}(R_{i-1,i}-\overline{b_i}) \tag{3}$$

式中 $\overline{b_i}$ 表示模型在随机初始化时在第 i 个任务上的测试准确率。FWT值为正则表示模型在学习当前任务时对未学习任务有正向迁移。对于任务序列中的最后一个任务,FWT不存在。

(4)反向迁移(Backward transfer, BWT)[17]衡量模型在学习新任务时对已学习任务性能的影响,计算公式为

$$\mathrm{BWT}=\frac{1}{T-1}\sum_{i=1}^{T-1}(R_{T,i}-R_{i,i}) \tag{4}$$

式中 $R_{i,i}$ 表示模型在完成第 i 个任务训练后,在同一任务上的测试准确率。BWT值为正则表示模型在学习新任务时对已学习任务有正向迁移,为负则表示存在灾难性遗忘。第1个任务不存在BWT。

(5)零样本迁移(Zero-shot transfer, ZT)[18]衡量模型对后续任务的零样本迁移能力,在基于预训练模型的持续学习方法中衡量零样本学习场景中的知识迁移能力。其计算公式为

$$\mathrm{ZT}_t=\frac{1}{t-1}\sum_{i=1}^{t-1}R_{i,t} \tag{5}$$

式中$R_{i,t}$表示模型在完成第i个任务训练后,在第t个任务上的零样本准确率。ZT值越高,表明模型的零样本迁移能力越强。

(6)宏表现(Macro performance, MP)[19]衡量模型在整个任务周期中的平均表现,其计算公式为

$$\mathrm{MP}=\frac{1}{K_t}\sum_{i=1}^{K_t}r_{t,i} \tag{6}$$

式中:K_t表示第t个任务中可见类别的数量,$r_{t,i}$表示任务t的成功率。MP相对于AP值在某些人物场景下更能体现模型的整体性能,其值越高,表明模型在各类别上的泛化能力越强。

此外,Díaz-Rodríguez等[20]还提出了模型大小(Model size, MS)、样本储存效率(Samples storage size, SSS)和计算效率(Computational efficiency, CE)等用于评估持续学习方法存储成本和计算成本的指标,为全面衡量持续学习方法的实际应用性能提供了重要依据。特别是在多模态持续学习场景中,这些指标的优化对于实现高效、轻量化的多模态融合模型具有重要意义。

上述评估指标针对持续学习通用框架,可以全面衡量模型在持续学习过程中的性能表现。对于MMCL而言,任务性能的提升可以间接反映多模态数据之间的模态知识对齐效果,这些指标同样适用。在实际应用中,可以根据具体任务需求选择合适的指标进行评估,或结合多个指标进行综合分析。

1.3 经典的单模态持续学习方法

本文将单模态的持续学习方法分为基于回放的方法[17,21-24]、基于正则化的方法[25-27]、基于参数隔离的方法[28-31]和基于预训练模型的方法[32-34]。

基于回放的方法直接保留旧任务样本在新任务训练时进行回放来缓解旧知识的遗忘。Rolnick等[21]提出通过动态调整旧数据的保留数量,避免计算成本随任务不断到来而过分增长。Rebuffi等[22]选择与类别特征均值距离更近的样本作为代表性旧数据,在新任务到来时参与训练,但受限于固定的存储空间,算法在多次增量学习任务之后不得不逐步减少旧任务中的样本数量,导致模型在学习新任务时面临新旧任务样本不均衡的问题,这一问题会导致模型在训练过程中更加偏向于新任务。Lopez-Paz等[17]使用一个记忆模块来存储旧任务的部分实例样本,并利用这些样本的梯度信息来保护旧任务表现不被新任务影响。Shin等[23]利用生成模型学习历史阶段的数据分布,并在当前阶段重放生成的数据,缓解数据存储和隐私问题。Jodelet等[24]通过扩散模型生成属于旧类别的伪样本,创新性地结合了知识蒸馏与伪样本重放策略。

基于正则化的方法通过在损失函数中引入正则项,防止模型在学习新知识时覆盖旧任务的知识。Li等[25]第一个将知识蒸馏[26]融入到持续学习领域,通过蒸馏损失正则化项约束让新数据在学生模型上的输出接近在教师模型上的输出实现模型优化,免于对旧数据的依赖。Kirkpatrick等[27]提出的弹性权重巩固(Elastic weight consolidation, EWC)是一种代表性的方法,它引入了一个和参数有关的通用正则损失,鼓励新任务训练参数尽量保护旧任务表现,在参数更新时做到了新旧任务衰减程度的权衡。

基于参数隔离的方法核心在于新旧任务参数的隔离,模型容量随着模型结构进行扩展,因此模型的性能不会受到初始容量的限制[4]。Serra等[28]提出了一种硬注意力机制(Hard attention to the task, HAT),通过学习任务特定的几乎二进制注意力向量来保留旧任务的信息,并有效减少灾难性遗忘。Mallya等[29-30]借鉴网络量化和剪枝的思想,设置包含基础网络和掩码的双层结构,并训练任务特定掩码以获得特定网络参数以适应当前任务。Zhou等[31]引入了表示合并技术来应对多任务学习后出现的特征漂移问题,并通过新旧分类器权重合并进行分类器巩固,实现了历史知识的整合与更新,旨在解决领域增量学习中的灾难性遗忘问题。

单模态的预训练模型无疑为单模态的持续学习方法提供了灵感。Janson等[32]通过实验认证发现，预训练模型具备足够强的表征能力，即使在不进行额外训练或复杂的持续学习机制的情况下，仍然在下游任务中保持良好表现。基于此，Zheng等[33]通过多种文本单模态实验，验证了预训练大模型在仅编码器（如BERT系列）和仅解码器（如GPT2、Pythia）架构下的强大表征能力，再次验证了这些预训练模型在文本分类和意图分类等下游任务中展现出跨任务泛化能力。Jia等[34]将预训练的ViT[35]和Swin Transformer[36]进行冻结，仅对少量参数进行提示微调来让大模型的先验知识适配到下游任务。Wang等[37]以预训练的ViT-B/16[38]模型为基础，创新性地采用可学习的提示参数来实现知识的紧凑编码，使用提示池替代传统的重放缓冲区，摆脱了对重放机制的依赖，并通过设计基于关键-查询的匹配策略，让模型从提示池中检索与实例相对应的特定提示，引导模型复用已学习的表征而非重新学习新表征。这一方法成功地将新任务的学习问题转化为在冻结的预训练模型上优化小规模提示参数的过程。他们发现这种方法使用单一提示池在多任务之间传递知识，难以区分任务通用特征与每个任务的独特特征，在此基础上又提出DualPrompt[39]通过直接在高层提示空间进行解耦，分别学习任务不变和任务特定的知识。Smith等[40]在ViT-B/16[38]的基础上提出了一种基于注意力的提示加权方法，以增强提示检索的有效性。Zhou等[41]将视觉原型分类器作为可靠基准，构建了一个可以与任何参数高效调优方法正交结合的通用框架。基于此，McDonnell等[42]以预训练的ViT-B/16[38]为骨干网络，提出了冻结随机投影（Random projection, RP）和去相关机制扩展该框架。Zhou等[43]提出了可扩展子空间集合（Expandable subspace ensemble, EASE），通过多子空间融合决策，将多个任务特定骨干网络的特征表示拼接在一起，并引入轻量级适配器以在冻结的预训练模型之上构建低成本的任务特定子空间，缓解新旧任务之间的特征冲突。

2 多模态持续学习的优势和挑战

2.1 优势

多模态持续学习通过整合多模态学习与持续学习的双重优势，在处理复杂动态环境时展现出卓越性能。本节将从多模态学习、持续学习以及二者的协同效应3个维度系统阐述其优势。

在多模态学习中，不同模态的数据往往提供互补的信息。例如图像可以提供丰富的视觉信息，而文本可以提供语义信息或上下文描述。通过融合多模态信息，模型能够获得更全面、更准确的理解，在医疗诊断[44-45]和自动驾驶[46]等领域中提供更准确的判断和感知。当某一模态的数据存在噪声或缺失时，其他模态的数据可以提供补充信息[12]，这种多模态的冗余性能够提高模型的鲁棒性，使得模型在复杂场景中表现更加稳定[47]。另外，多模态学习可以帮助模型学习到更通用的特征表示，从而提高其在未见数据上的泛化能力。例如，通过联合学习图像和文本，模型可以更好地理解视觉概念与语义之间的关系[38]，从而在跨模态任务（如图文检索、图像生成文本）中表现更优。这种泛化能力在多模态持续学习中尤为重要。

对于持续学习，现实世界中的环境是不断变化的，持续学习使模型能够在新任务或新数据到来时不断更新和优化，同时避免遗忘之前学到的知识[13]。多模态持续学习可以同时利用多种模态的信息来应对复杂的变化，能够进一步增强这种适应性。并且通过存储模型参数[29]或部分关键数据[21]来保留知识。这种方法显著减少了数据存储的开销，尤其是在多模态场景数据量通常非常庞大的情况下，持续学习能力尤为重要。另外，持续学习可以利用之前学到的知识来加速新任务的学习[17]，也就是将从一个任务中学到的特征迁移到另一个相关任务中，从而减少训练时间和资源消耗。在多模态持续学习

中,模型可以同时利用多种模态的知识更有效地加速新任务的学习。

多模态持续学习将多模态学习和持续学习的优势结合起来,使其在复杂动态环境中捕捉更丰富的特征,表现出更强的环境适应性和更高的资源效率,展现出独特的优势。

2.2 挑战

与UMCL相比,MMCL除了要面临灾难性遗忘挑战,还因其多模态数据的复杂性引入了更多可能加剧遗忘的挑战。

(1)模态不平衡与异构性问题

多模态数据通常有不同的来源(如图像、文本、语音等),这些模态在数据分布、特征表示和语义信息上存在显著差异。例如,图像通常以二维矩阵的形式表示,而文本则是一维序列,语音则是时间序列信号。这种模态之间的不平衡性[5]和异构性[4]使得模型难以有效地融合多模态信息,可能导致某些模态在训练过程中被忽视,从而影响模型的整体性能[48-49]。例如,在视频分析任务中,视觉模态可能占据主导地位,而音频模态的信息可能被忽略,导致模型无法充分利用多模态数据的互补性。如何有效地表示和融合这些异构的多模态信息,是多模态持续学习中的一个重要挑战。

(2)模态对齐与交互优化

在多模态学习中,一个核心挑战是如何让不同模态(如图像、文本等)的特征表示在语义上保持一致,这一过程称为模态对齐[12]。特别是在增量学习场景下,不同模态的特征表示可能会随着任务的增加而发生变化[2],导致不同模态间的语义对齐变得更加复杂[50]。例如,在图文匹配任务中,图像中的物体需要与文本描述中的词语进行对齐,但随着新任务的引入,图像和文本的语义关系可能会发生变化,从而增加对齐的难度。此外,如何在3种或更多模态,例如视觉、音频和文本模态等的情况下实现精细的模态交互,仍然是一个未完全解决的问题。

(3)预训练知识的维护

许多多模态持续学习方法依赖于预训练模型,这些模型在大规模数据集上学习到了通用的特征表示。然而,在MMCL的持续微调过程中,一些来自预训练先验知识可能会退化,导致未来任务的性能严重下降[18],产生负向迁移[51]。在跨模态任务中,预训练模型可能已经学习到了图像和文本之间的语义关系,但在持续学习过程中,这些关系可能会被新任务的数据所覆盖,从而导致模型在旧任务上的性能下降[52]。如何有效地维护和利用预训练知识,是多模态持续学习中的一个关键问题。

(4)计算成本与效率问题

多模态持续学习不仅需要承担多任务序列训练的计算成本,还因为添加了数据模态不可避免地增加了参数量,并且通常涉及大规模数据和复杂的模型结构[53],导致训练和推理过程中的计算成本显著增加。训练一个多模态模型可能需要同时处理图像、文本和音频数据,这不仅增加了数据存储和传输的开销,还使得模型的训练和推理时间大幅延长。如何在保证模型性能的同时提高计算效率,是一个亟待解决的问题。

3 MMCL方法研究现状

本文在单模态经典持续学习方法的基础上,系统梳理并总结了MMCL方法的研究现状,详细阐述了各类方法的核心特点及其创新之处。为了更直观地呈现这些方法的框架与机制,本文在图1中展示了相应的框架概览,并在表1中对每个类别具体的方法使用的预训练网络和涉及的模态进行了统计。

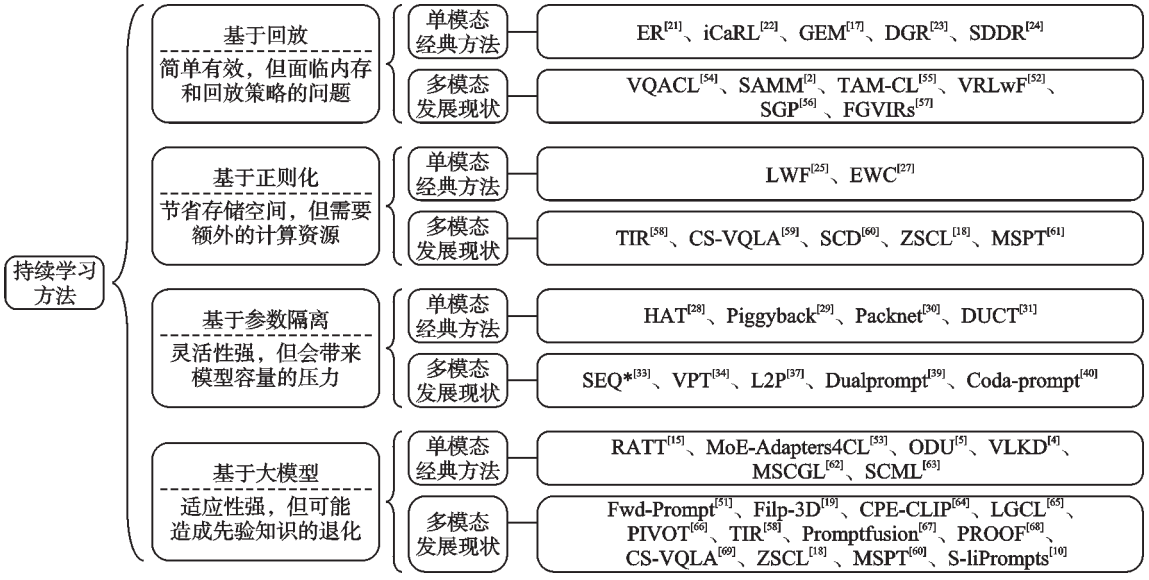

图1 单模态与多模态持续学习方法框架概览

Fig.1 Overview of unimodal and multimodal continual learning methods

表1 多模态持续学习方法的骨干网络与相关模态

Table 1 Backbone networks and involved modalities in multimodal continual learning methods

类别	方法	骨干网络	视觉	文本	音频	加速度	陀螺仪	图谱	触觉
基于回放的方法	VQACL[54]	—	√	√					
	SAMM[2]	—	√		√				
	TAM-CL[55]	—	√	√					
	SGP[56]	—	√	√					
	FGVIRs[57]	—	√			√	√		
基于正则化的方法	TIR[58]	BLIP2[70], InstructBLIP[71]	√	√					
	CS-VQLA[59]	VisualBERT[72]	√	√					
	SCD[60]	ViLT	√	√					
	ZSCL[18]	CLIP[38]	√	√					
	MSPT[61]	CLIP[38]	√	√					
基于参数隔离的方法	RATT[15]	—	√	√					
	MoE-Adapters4CL[53]	CLIP[38]	√	√					
	ODU[5]	GoogleNet networks	√		√				√
	VLKD[4]	—	√	√					
	MSCGL[62]	—	√	√				√	
	SCML[63]	—	√	√					
基于大模型的方法	Fwd-Prompt[51]	BLIP2[70], InstructBLIP[71]	√	√					
	Filp-3D[19]	CLIP[38]	√	√					
	CPE-CLIP[64]	CLIP[38]	√	√					
	LGCL[65]	CLIP[38]	√	√					
	PIVOT[66]	CLIP[38]	√	√					
	TIR[58]	BLIP2[70], InstructBLIP[71]	√	√					
	CS-VQLA[59]	VisualBERT[72]	√	√					
	ZSCL[18]	CLIP[38]	√	√					
	MSPT[61]	CLIP[38]	√	√					

3.1　基于回放的方法

作为一种直白的持续学习策略,回放的方法将旧任务知识在学习新任务时加入训练,通过“复习”来避免遗忘,本文在图2中展示了基于回放方法的代表性架构。

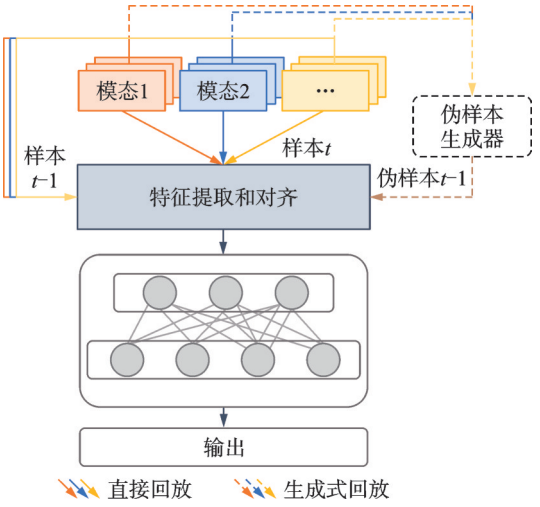

图2　基于回放的多模态持续学习方法代表性架构

Fig.2　Representative architectures of replay-based multimodal continual learning methods

回放真实的历史样本需要将部分旧任务训练实例存储在情节记忆中,由于内存空间有限,这些方法的关键在于如何选择具有代表性的数据样本。对于多模态数据,一个直观的回放思路是直接随机选择和回放各模态旧任务样本。Zhang等[54]沿用了ER[21]的思路对旧任务训练样本进行随机采样和回放,实现视觉和文本模态的视觉问答任务持续学习,并在VQA v2[73]和NExT-QA[74]数据集的标准测试和新组合测试中均取得了最优表现,相较ER[21]等主流基于回放的持续学习方法,在平均精度上提升高达6.1%,在遗忘率上最高减少4.83%,显示出其在维持旧知识与适应新任务之间优越的平衡能力。Sarfraz等[2]利用视觉和音频模态数据点之间的关系结构相似性,融合和对齐信息,并随机采样进行回放,实现分类任务持续学习,其在多种多模态持续学习设置中均优于基线方法ER[21],甚至在音频模态上相较于单模态模型取得数量级的性能提升。Cai等[55]采用知识蒸馏的方法,让学生模型的输出接近教师模型,由此约束历史样本选择和存储,在4个视觉-语言基准数据集序列任务中的遗忘率显著优于直接微调、EWC[13,27]和ER[21]三个基准方法。Ding等[52]探究了微调大模型导致的泛化能力退化,并针对基于CLIP[38]的持续学习问题提出了基于重放词汇的学习方法(Learning without forgetting via replayed vocabulary, VRLwF),通过在文本编码器上将重放的词汇视为伪类别,并在它们的logits上执行蒸馏,模拟了零-shot学习和先前类别,在多阶段训练设置下,该方法实现了高达60.97%的平均准确率,相较于微调及LwF[25]等基准方法最高提升达10.21%,展现出良好的泛化能力。

单模态回放方法所面临的存储压力和隐私保护等问题,在多模态场景中同样存在。为此,生成式回放在多模态持续学习中受到了广泛关注。然而,与单模态生成不同,多模态回放需要生成包含多种模态且高度相关的数据元组,同时要求这些元组具有详细的标注和较高的准确性,这使得多模态数据生成任务变得异常复杂和具有挑战性。为了应对这些困难,一些研究将重点转向生成替代数据或部分数据。Lei等[56]使用场景图作为视觉信息的简洁符号表示,代替传统图像,帮助模型回放场景图-问题-答案三元组进行知识回顾,在不同任务顺序和场景下,此方法在性能上均优于基线方法和SOTA(State of the art)方法,甚至在两种任务设置下,其平均准确率分别比额外保存了真实图像、答案和相应的任务标签的VQG[75]方法高出3.86%和10.65%。另外,Peng等[76]研究发现,模态间优化不平衡导致模型在多模态任务中的表现未能充分发挥其潜力。基于此,Kim等[57]设计激励训练方案结合表示分布估计和虚拟表示生成增强惯性测量单元(Inertial measuring unit, IMU)模态的表达能力,解决视觉-IMU多模态网络持续学习中的模态不平衡问题,并在多个数据集上验证了其在分类任务和检索任务中的有效性。基于回放的多模态持续学习方法能够根据具体的模态任务和模型特征设计回放策略,针对性地选择回放数据,具有较大的灵活性和研究前景。

3.2 基于正则化的方法

基于PTM在多模态融合和任务适应性方面表现出强大的优势，对多模态PTM进行微调成为一种极具潜力的研究方向[77]。基于正则化的方法在模型微调过程中引入约束机制，防止参数过度偏离旧任务的最优状态，从而缓解灾难性遗忘问题，其架构如图3所示。He等[58]分别在BLIP2[70]和InstructBLIP[71]上，基于EWC[27]提出了任务相似性信息的正则化和模型扩展方法，进行持续指令微调过程中进行参数正则化，并在两种序列任务持续学习基准设置下验证了其相对于EWC[27]、ER[21]和A-GEM[69]等持续学习方法均有优势。

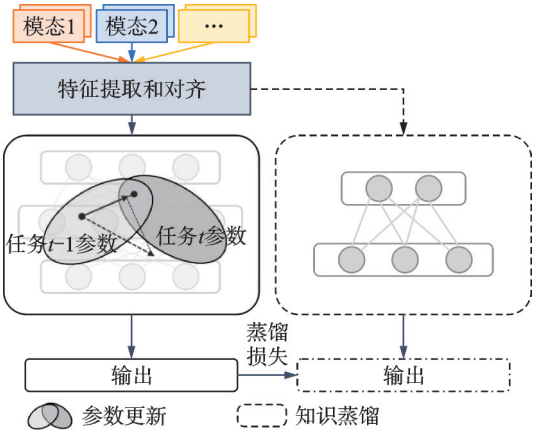

图3 基于正则化的多模态持续学习方法代表性架构

Fig.3 Representative architectures of regularization-based multimodal continual learning methods

在处理多模态复杂数据时，若对所有可训练参数施加统一的约束，不仅未能充分利用多模态数据相较于单模态数据的独特优势，反而抑制了多模态持续学习的潜力。鉴于LWF[25]利用知识蒸馏的策略，只有在参数变化最终导致模型输出变化时，才对其施加惩罚，允许参数更自由地变化。Bai等[59]在非示例的持续外科视觉问答定位回答框架下，提出了刚性-可塑性意识蒸馏和自校准异质蒸馏策略，分别针对最终输出和中间特征图进行蒸馏，帮助保留旧知识，在非重叠类别及新旧重叠类别上均表现出更优的灾难遗忘缓解能力，在分类任务中，其在$t=1$与$t=2$时的整体平均准确率分别相比次优方法提升了2.60%和2.68%，而在定位任务中，其平均交并比分别提升0.44与0.94，展现出在稳定性-可塑性平衡方面的显著优势。Lao等[60]提出了无数据存储的自我批判蒸馏(Self-critical distillation, SCD)方法，在logits层和特征层上传递领域知识，在不同顺序任务下的遗忘度明显低于主流蒸馏方法，对旧知识保持更好。与传统的蒸馏方法不同，Zheng等[18]将CLIP[38]作为基础模型，通过在特征空间进行蒸馏和在参数空间进行权重集成，避免了模型在持续学习中的性能显著下降，在无数据存储设置下显现出强大的优势。Chen等[61]在同样的基础模型上提出了多模态稳定性-可塑性变换器(Multimodal stability-plasticity transformer, MSPT)，旨在解决多模态知识图谱构建在持续学习中的问题，通过梯度调节技术，解决了不同模态之间的学习动态不平衡，使用多模态交互与注意力蒸馏，精细调节多模态交互的遗忘率，避免次要模态的遗忘，并在两个增量多模态基准任务中显著优于现有多模态和单模态增量学习方法，有效减缓旧任务的遗忘问题，展现出优越的稳定性与可塑性平衡能力。

3.3 基于参数隔离的方法

基于参数隔离的方法通过启用不同的模型参数来应对不同的任务，相比于基于正则化的方法和基于回放的方法，避免了任务间干扰[15]。根据模型设计，同样可分为静态结构和动态结构，图4中展示了其代表性架构。

静态结构主要依赖对不同的任务对应激活不同的参数实现保护旧知识。Del等[15]针对图像描述任务中的序列生成问题，引入了基于注意力机制的对短暂任务的循环注意力，通过注意力掩码来动态调整模型对不同任务的关注程度，在MS-COCO[78]和Flickr30K[79]两个数据集的实验上，相比传统的EWC[27]与LwF[25]在防止遗忘方面表现突出，在所有任务上均实现了近乎零遗忘，且在多个任务上还实现了负遗忘，体现了其对旧任务知识的稳固保持和跨任务迁移能力。

动态结构方法能够随着新任务的到来添加新的模块,最直接的思路就是在新任务到来时,直接在网络中添加对应的新模块,实现任务和特定模块之间的直接对应。Yu等[53]提出了一种基于专家混合适配器的参数高效的持续学习框架MoE-Adapters4CL,使用冻结的CLIP[38]作为基础模型,MoE(Mixture of experts)适配器作为稀疏的专家,在每个新任务到来时逐步添加到基础模型中,根据任务特定路由器动态选择加权不同的适配器,通过设计分布判别自选器(Distribution discriminative auto-selector, DDAS),将分布内的测试数据自动分配给合适的专家适配器或将分布外的数据分配给预训练的CLIP[38]模型,利用其零样本迁移能力实现新任务上的有效预测,在域增量任务和类增量任务上,该方法在“Transfer”“Average”和“Last”指标上均优于大多数现有方法,尤其在域增量任务中,3种指标较第2名方法分别提高了0.8%、1.3%和1.4%,并且在少量样本设置下也优于其他主流方法。Sun等[5]通过字典学习和共享知识库策略有效地处理多模态数据和任务的内在关系,面对任务序列中间可能出现的模态缺失问题,通过在不同任务间共享知识表示和其他模态作为辅助训练分类器,此方法在实验中通过3D加速度传感器收集触觉信息,与视觉和声音共同作为信息源识别不同材料。相比于传统的固定字典和知识库方法,此方法能够更好地适应新任务,显著提高了多模态融合的识别效果,尤其在视觉和触觉模态用于声音识别以及视觉和声音模态用于触觉识别的情况下,性能表现尤为突出。

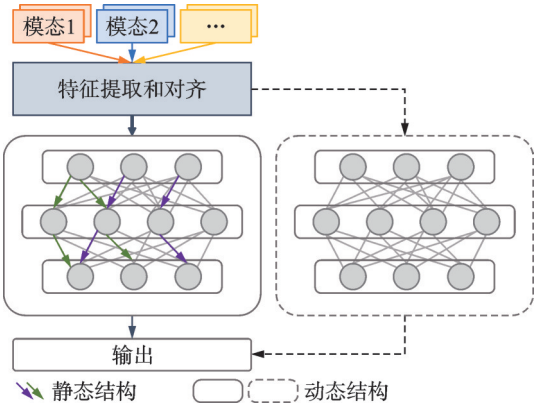

图4　基于参数隔离的多模态持续学习方法代表性架构

Fig.4　Representative architectures of parameter isolation-based multimodal continual learning methods

在这些基础上,Peng等[4]引入了自适应网络扩展,实现网络自适应调整在学习新知识的同时保持旧任务表现,识别并保持跨领域知识的关键,避免了冗余计算压力。Cai等[62]提出了多模态结构演化持续图学习模型(Multimodal structure-evolving continual graph learning, MSCGL),自适应探索模型架构的同时保持历史信息,设计了自适应多模态图神经网络模型,结合共享策略,避免了对类似任务进行不必要的架构扩展。多个多模态图分类任务上的实验结果表明,MSCGL在准确率和遗忘率方面均优于现有方法,平均准确率达到89.44%,相比GEM[17]提升了11.51%,在任务数量增加时,展现出更强的稳定性与扩展性。

鉴于多模态数据对比单模态在持续学习任务中增加的复杂性,Song等[63]模拟人类跨模态学习的认知机制,提出了一种按照模态顺序学习的持续学习机制,针对不同模态设计专用编码器,并将不同模态的特征映射到共同的特征空间,免去了繁琐的模态之间特征的显式对齐操作,在现实场景不确定的模态分布变化中具有更强的灵活性和适应能力,并利用大量实验证明了该方法在不同模态顺序、参数设置及持续任务长度下均表现出良好的稳定性和适应性。基于参数隔离的方法有无尽的模型架构变化潜力,具有较好的自由度和灵活性,发展潜力巨大。

3.4　基于大模型的方法

从零训练的传统持续学习方法,相当于从婴儿开始培养,使其逐步学习知识。基于大模型的方法则类似于让一名已有丰富知识的成年人继续学习新任务[7],在多种下游任务上展现出强大的泛化能力。受预训练的大模型[35,80-82]在单一领域中取得成功的启发,多模态预训练大模型近年来也越来越受到关注,特别是在多模态持续学习方向带来了新的希望与突破。早期代表模型VisualBERT[72]将图像区域

特征与文本嵌入拼接输入BERT[80]，通过统一的Transformer[83]架构实现语义融合，为多模态预训练奠定了基础。随后，CLIP[38]模型的提出彻底打破了对固定类别标签的依赖，它利用大规模图文对学习跨模态表示，通过并行的图像编码器和文本编码器对图像与文本进行嵌入对齐，从而显著提升了模型的通用性与开放词汇能力。BLIP2[70]在此基础上引入了两阶段的感知-语言桥接框架，通过视觉编码器和Q-former模块提取跨模态信息，再结合语言大模型实现灵活的图文问答与生成，进一步拓宽了模型适应复杂任务的能力。InstructBLIP[71]则通过指令微调机制强化模型的任务指令感知能力，使其在开放式多模态对话中更加贴近人类语言习惯与任务意图。这些大模型本身是基于多模态数据进行训练的，具备强大的上下文建模和跨模态理解能力[12]，能够在处理文本、图像甚至音频数据保留和强化多模态数据的内在关联性的同时，学习通用的且高层次的表征，在新任务到来时实现知识的协同利用[47]，使得大模型在面对多模态持续学习任务时展现出独特的适应性。同时，大模型的少样本学习能力进一步降低了持续学习对数据规模和标注资源的依赖，使其能够在动态扩展的多模态环境中保持稳定的性能。这些特性使得大模型在跨模态特征对齐和多任务适应能力方面具有天然优势，构成了大模型支持多模态持续学习的内在机制。

预训练模型在多模态持续学习中的作用主要体现在两个方面：首先，通过在大规模数据集上进行训练，预训练模型能够学习到通用的特征表示，为多种任务提供强有力的初始知识基础；其次，通过微调机制，模型可以高效地适应新任务，同时保留对旧任务的记忆，从而有效缓解灾难性遗忘问题。图5展示了这类方法的代表性架构。

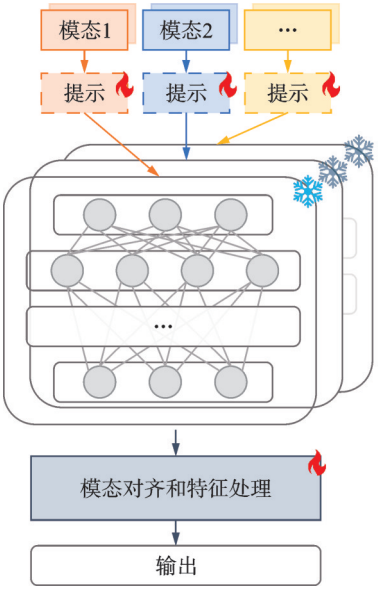

图5 基于大模型的多模态持续学习方法代表性架构

Fig.5 Representative architectures of large model-based multimodal continual learning methods

除了前面提到的将大模型与正则化方法结合的方法[18,58-59,61]之外，Zheng等[51]在BLIP2[70]和InstructBLIP[71]等视觉语言预训练大模型的基础上，提出了正向迁移提示调优方法，通过将提示梯度投影到残差空间和预训练空间，分别实现抗遗忘和正向迁移，显著提升了模型的持续学习能力，相比于序列指令微调及EWC[27]等基准方法，此方法在处理以BLIP2[70]和InstructBLIP[71]作为主干模型进行不同顺序的图像问答持续学习任务时均能表现最优。类似地，Xu等[19]和D'Alessandro等[64]利用CLIP[38]作为骨干网络，分别在视觉和文本模态上设计了特征增强机制和轻量级可学习提示，成功实现了少样本类增量学习。Khan等[65]提出语言引导的持续学习(Language guidance for prompt-based continual learning, LGCL)，分别利用预训练的ViT[35]和CLIP[38]的文本编码器来提取图像特征和生成任务的语言表示，强调语言作为统一的语义空间可以为不同任务提供共享表示，通过任务级别和特征级别两个层次的对齐机制提升模型的泛化能力，实验结果表明，LGCL显著优于传统的持续学习方法，与单模态的DualPrompt[39]和L2P[37]方法相比，在Split ImageNet-R数据集[39]上的平均准确率分别提升了1.33%和0.94%，并在不同骨干网络的对比中保持优势，验证了基于大模型的语言引导能够有效地提升模型的性能。Villa等[66]提出了PIVOT(Prompting for video continual learning)视频持续学习策略，利用CLIP[38]视觉编码器，并将其映射到适用于视频理解的特征空间，通过时空提示机制和多模态分类器实现无需视频预训练的知识迁移。此方法通过在预训练的大模型上引入任务特定的时空Prompt机制，在3个数据集的10任务和

20任务设定下都全面领先于现有方法,甚至距离上限(全监督训练)只差1%~2%,几乎达到了视频持续学习任务的极限性能。

这些方法普遍倾向于基于同一预训练模型训练多个分类器,该策略忽视了模型多样性所带来的知识多样性优势[67]。基于这个发现,Chen等[67]结合了预训练ViT[35]和CLIP[38],并在推理时动态融合它们的logits,提出稳定器和增强器来解耦稳定性和可塑性,在类增量和域增量任务中均优于现有方法,尤其是在类增量任务数据集Split-Imagenet-R[39]上,此方法准确率表现出巨大优势,分别以11.4%和17%的优势领先于L2P[37]和DualPrompt[39]方法。Zhou等[68]基于CLIP[38]提出了结合图像和文本特征的投影融合方法(Projection fusion, PROOF),利用CLIP[38]的跨模态视觉-文本匹配能力,设计了一个三级集成方法,分别考虑图像-文本、图像-图像原型以及图像-调整后文本的跨模态融合,将图像和文本特征投影到共享的投影空间。在每个新任务的训练过程中,模型冻结旧投影层,仅优化当前任务的投影层,并通过对单任务扩展的投影层进行聚合来构建统一的表示。在实验中,不论是相比于iCaRL[22]等经典的持续学习方法还是L2P[37]等基于提示的单模态大模型方法,均表现出显著优势,验证了图文信息的协同适配与预训练大模型的先验知识对于多模态持续学习的重要性。Wang等[10]基于CLIP[38]的S-liP-rompts在不同领域之间独立学习提示,逐步构建提示池,通过冻结预训练模型,仅调整当前领域相关的提示和分类器,实现跨领域无遗忘。此方法在无需样本存储的前提下,超越了当前所有主流方法,其平均准确率甚至以3.15%的优势超过了上限(联合训练),展示出较好的应用和发展潜力。

随着新型预训练模型与方法的不断涌现,持续学习的实现机制已突破单一方法的局限,逐渐向多种机制的深度融合发展。这一趋势不仅显著提升了模型的整体性能,也为解决多模态持续学习中复杂问题提供了新的研究思路。然而,基于大模型的方法依然面临诸多关键挑战。例如,在增量任务与预训练知识发生冲突时,模型的泛化能力往往会出现退化[18,51],如何在保持预训练能力与适应新任务之间实现有效权衡仍是一个具有挑战性的问题。此外,扩展模型的持续学习能力通常以牺牲计算效率为代价,既增加了模型的结构复杂性,也带来了较高的计算成本。从表1可以看出,当前广泛应用的大模型主要集中在视觉、文本和音频等常见模态,对于如何将其扩展至如深度传感器、生物传感器[84]等非典型模态,并实现跨模态特征对齐,现有研究仍显不足[11]。这些挑战亟需在算法设计、模型结构及训练范式等方面实现突破性进展。

4 总结与展望

随着多模态模型的快速发展,多模态持续学习已成为机器学习和人工智能领域一个活跃且极具前景的研究方向。本文从基本概念、评估体系和经典单模态方法3个方面介绍了MMCL的预备知识,并深入分析了其在实际应用中的优势与挑战,从基于回放、正则化、参数隔离和大模型4个类别全面梳理了MMCL方法的研究现状与最新进展。MMCL不仅继承了UMCL的基本框架,还引入了多模态数据的复杂性和多样性,这使得其在模态融合、知识迁移和任务适应性等方面展现出巨大的潜力。

多模态持续学习的研究可从多个维度深入拓展。首先,随着多模态数据的不断丰富,如何维持各模态之间的平衡与协同仍是亟待解决的核心问题[48,76]。在引入3种及以上模态的场景中,模态对齐的复杂度显著提升[85],亟需设计高效的对齐机制,以提升模型性能并适应更复杂的任务需求。其次,当前研究主要聚焦于视觉、文本和音频等常见模态,而对如生物传感器[84]等非典型模态关注较少,限制了MMCL在多学科、多场景中的进一步拓展。未来应构建具备异构模态兼容性的通用框架,以应对由模态差异带来的协同建模与对齐挑战。此外,在大模型基础上设计MMCL方法时,应高度重视对预训练知识的保留,防止在序列任务中破坏已有能力[18,51]。当前的参数隔离和提示调优等策略可为实现稳定迁移与持续学习提供有益启发。最后,现有评估体系多沿用单模态持续学习的评估标准,难以全面反

映多模态持续学习的特性与能力。未来应构建更具针对性的多模态评估框架,从模态知识平衡、迁移效率到任务表现等多个维度进行系统评估,为MMCL在更广泛领域的应用奠定坚实基础。未来研究将在模态扩展、模态对齐、知识维护、任务适应和模型稳定性等方面继续深入探索,推动该领域的理论创新和技术突破。通过解决多模态数据融合、模态间交互优化以及知识迁移等关键问题,MMCL有望在智能感知、人机交互及自动驾驶等领域发挥更大的作用,为人工智能的发展注入新的活力。

参考文献:

[1] MROCZKO-WĄSOWICZ A. Editorial: Perception-cognition interface and cross-modal experiences: Insights into unified consciousness[J]. Frontiers in Psychology, 2016, 7: 1593.

[2] SARFRAZ F, ZONOOZ B, ARANI E. Beyond unimodal learning: The importance of integrating multiple modalities for lifelong learning[EB/OL]. (2024-05-04). https://arxiv.org/abs/2405.02766v1.

[3] WANG H, ZHOU S, WU Q, et al. Confusion mixup regularized multimodal fusion network for continual egocentric activity recognition[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Paris, France: IEEE, 2023: 3552-3561.

[4] PENG Y, QI J, YE Z, et al. Hierarchical visual-textual knowledge distillation for life-long correlation learning[J]. International Journal of Computer Vision, 2021, 129(4): 921-941.

[5] SUN F, LIU H, YANG C, et al. Multimodal continual learning using online dictionary updating[J]. IEEE Transactions on Cognitive and Developmental Systems, 2021, 13(1): 171-178.

[6] ZHOU D W, SUN H L, NING J, et al. Continual learning with pre-trained models: A survey[EB/OL]. (2024-01-29). https://arxiv.org/abs/2401.16386v1.

[7] STEINER A, KOLESNIKOV A, ZHAI X, et al. How to train your vit? data, augmentation, and regularization in vision transformers[EB/OL]. (2021-01-28). https://arxiv.org/abs/2106.10270v1.

[8] LEE K Y, ZHONG Y, WANG Y X. Do pre-trained models benefit equally in continual learning?[C]//Preceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE, 2023: 6474-6482.

[9] MEHTA S V, PATIL D, CHANDAR S, et al. An empirical investigation of the role of pre-training in lifelong learning[J]. Journal of Machine Learning Research, 2023, 24(214): 1-50.

[10] WANG Y, HUANG Z, HONG X. S-prompts learning with pre-trained transformers: An Occam's razor for domain incremental learning[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2022: 5682-5695.

[11] YU D, ZHANG X, CHEN Y, et al. Recent advances of multimodal continual learning: A comprehensive survey[EB/OL]. (2024-10-07). https://arxiv.org/abs/2410.05352.

[12] BALTRUŠAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423-443.

[13] WANG L, ZHANG X, SU H, et al. A comprehensive survey of continual learning: Theory, method and application[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5362-5383.

[14] VAN DE VEN G M, TOLIAS A S. Three scenarios for continual learning[EB/OL]. (2019-04-15). https://arxiv.org/abs/1904.07734.

[15] DEL CHIARO R, TWARDOWSKI B, BAGDANOV A, et al. RATT: Recurrent attention to transient tasks for continual image captioning[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2020, 16736-16748.

[16] CHAUDHRY A, DOKANIA P K, AJANTHAN T, et al. Riemannian walk for incremental learning: Understanding forgetting and intransigence[C]//Proceedings of Computer Vision—ECCV 2018. Cham: Springer International Publishing, 2018: 556-572.

[17] LOPEZ-PAZ D, RANZATO M, LOPEZ-PAZ D, et al. Gradient episodic memory for continual learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: ACM, 2017:

6470-6479.

[18] ZHENG Z, MA M, WANG K, et al. Preventing zero-shot transfer degradation in continual learning of vision-language models [C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 19068-19079.

[19] XU W, HUANG T, QU T, et al. Filp-3D: Enhancing 3D few-shot class-incremental learning with pre-trained vision-language models[EB/OL]. (2023-12-28). https://arxiv.org/abs/2312.17051v2.

[20] DÍAZ-RODRÍGUEZ N, LOMONACO V, FILLIAT D, et al. Don't forget, there is more than forgetting: New metrics for continual learning[EB/OL].(2018-10-31). https://arxiv.org/abs/1810.13166.

[21] ROLNICK D, AHUJA A, SCHWARZ J, et al. Experience replay for continual learning[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2019: 350-360.

[22] REBUFFI S A, KOLESNIKOV A, SPERL G, et al. iCaRL: Incremental classifier and representation learning[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 5533-5542.

[23] SHIN H, LEE J K, KIM J, et al. Continual learning with deep generative replay[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: ACM, 2017: 2994-3003.

[24] JODELET Q, LIU X, PHUA Y J, et al. Class-incremental learning using diffusion model for distillation and replay[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Paris, France: IEEE, 2023: 3417-3425.

[25] LI Z, HOIEM D. Learning without forgetting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40 (12): 2935-2947.

[26] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL].(2017-11-17). https://arxiv.org/abs/1711.09784.

[27] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(13): 3521-3526.

[28] SERRA J, SURIS D, MIRON M, et al. Overcoming catastrophic forgetting with hard attention to the task[C]//Proceedings of the 35th International Conference on Machine Learning. Massachusetts, USA: PMLR, 2018: 4548-4557.

[29] MALLYA A, DAVIS D, LAZEBNIK S, et al. Piggyback: Adapting a single network to multiple tasks by learning to mask weights[C]//Proceedings of Computer Vision—ECCV 2018. [S.l.]: ACM, 2018: 72-88.

[30] MALLYA A, LAZEBNIK S. PackNet: Adding multiple tasks to a single network by iterative pruning[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7765-7773.

[31] ZHOU D W, CAI Z W, YE H J, et al. Dual consolidation for pre-trained model-based domain-incremental learning[C]//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2025.

[32] JANSON P, ZHANG W, ALJUNDI R, et al. A simple baseline that questions the use of pretrained-models in continual learning[EB/OL]. (2022-10-10). https://arxiv.org/abs/2210.04428.

[33] ZHENG J, QIU S, MA Q. Learn or recall? revisiting incremental learning with pre-trained language models[EB/OL]. (2023-12-13). https://arxiv.org/abs/2312.07887.

[34] JIA M, TANG L, CHEN B C, et al. Visual prompt tuning[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2022: 709-727.

[35] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2022-10-22). https://arxiv.org/abs/2010.11929.

[36] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 9992-10002.

[37] WANG Z, ZHANG Z, LEE C Y, et al. Learning to prompt for continual learning[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 139-149.

[38] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning. Massachusetts, USA: PMLR, 2021: 8748-8763.

[39] WANG Z, ZHANG Z, EBRAHIMI S, et al. DualPrompt: Complementary prompting for rehearsal-free continual learning [C]//Proceedings of Computer Vision—ECCV 2022. Cham: Springer Nature Switzerland, 2022: 631-648.

[40] SMITH J S, KARLINSKY L, GUTTA V, et al. CODA-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 11909-11919.

[41] ZHOU D W, CAI Z W, YE H J, et al. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need[J]. International Journal of Computer Vision, 2025, 133(3): 1012-1032.

[42] MCDONNELL M D, GONG D, PARVANEH A, et al. RanPAC: Random projections and pre-trained models for continual learning[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2023: 12022-12053.

[43] ZHOU D W, SUN H L, YE H J, et al. Expandable subspace ensemble for pre-trained model-based class-incremental learning [C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 23554-23564.

[44] TAN W, TIWARI P, PANDEY H M, et al. Multimodal medical image fusion algorithm in the era of big data[J]. Neural Computing and Applications, 2020. DOI:10.1007/s00521-020-05173-2.

[45] 杨印凯, 万鹏, 石航, 等. 基于多模态超声对比学习的肝癌诊断方法[J]. 数据采集与处理, 2024, 39(4): 874-885.
YANG Yinkai, WAN Peng, SHI Hang, et al. Liver cancer diagnosis method based on multi-modal ultrasound contrast learning [J]. Journal of Data Acquisition and Processing, 2024, 39(4): 874-885.

[46] CUI C, MA Y, CAO X, et al. A survey on multimodal large language models for autonomous driving[C]//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Waikoloa, HI, USA: IEEE, 2024: 958-979.

[47] MROUEH Y, MARCHERET E, GOEL V. Deep multimodal learning for audio-visual speech recognition[C]//Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, QLD, Australia: IEEE, 2015: 2130-2134.

[48] HE C, CHENG S, QIU Z, et al. Continual egocentric activity recognition with foreseeable-generalized visual—IMU representations[J]. IEEE Sensors Journal, 2024, 24(8): 12934-12945.

[49] CHENG S, HE C, CHEN K, et al. Vision-sensor attention based continual multimodal egocentric activity recognition[C]//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea: IEEE, 2024: 6300-6304.

[50] NI Z, WEI L, TANG S, et al. Continual vision-language representation learning with off-diagonal information[C]//Proceedings of the 40th International Conference on Machine Learning. Massachusetts, USA: PMLR, 2023: 26129-26149.

[51] ZHENG J, MA Q, LIU Z, et al. Beyond anti-forgetting: Multimodal continual instruction tuning with positive forward transfer [EB/OL]. (2024-01-17). https://arxiv.org/abs/2401.09181.

[52] DING Y, LIU L, TIAN C, et al. Don't stop learning: Towards continual learning for the clip model[EB/OL]. (2022-07-19). https://arxiv.org/abs/2207.09248.

[53] YU J, ZHUGE Y, ZHANG L, et al. Boosting continual learning of vision-language models via mixture-of-experts adapters [C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 23219-23230.

[54] ZHANG X, ZHANG F, XU C. VQACL: A novel visual question answering continual learning setting[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 19102-19112.

[55] CAI Y, THOMASON J, ROSTAMI M. Task-attentive transformer architecture for continual learning of vision-and-language tasks using knowledge distillation[EB/OL]. (2024-01-27). https://arxiv.org/abs/2401.15275.

[56] LEI S W, GAO D, WU J Z, et al. Symbolic replay: Scene graph as prompt for continual learning on VQA task[J]. Proceedings

of the AAAI Conference on Artificial Intelligence, 2023, 37(1): 1250-1259.

[57] KIM W, SON B, KIM I. VILT: Vision-and-language transformer without convolution or region supervision[C]//Proceedings of the 38th International Conference on Machine Learning. Massachusetts, USA: PMLR, 2021: 5583-5594.

[58] HE J, GUO H, TANG M, et al. Continual instruction tuning for large multimodal models[EB/OL]. (2023-11-27). https://arxiv.org/abs/2311.16206.

[59] BAI L, ISLAM M, REN H. Revisiting distillation for continual learning on visual question localized-answering in robotic surgery[C]//Proceedings of Medical Image Computing and Computer Assisted Intervention—MICCAI 2023. Cham: Springer Nature Switzerland, 2023: 68-78.

[60] LAO M, PU N, LIU Y, et al. Multi-domain lifelong visual question answering *via* self-critical distillation[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM, 2023: 4747-4758.

[61] CHEN X, ZHANG J, WANG X, et al. Continual multimodal knowledge graph construction[C]//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. [S.l.]: ACM, 2024.

[62] CAI J, WANG X, GUAN C, et al. Multimodal continual graph learning with neural architecture search[C]//Proceedings of the ACM Web Conference 2022. Virtual Event, Lyon, France: ACM, 2022: 1292-1300.

[63] SONG G, TAN X. Real-world cross-modal retrieval via sequential learning[J]. IEEE Transactions on Multimedia, 2020, 23: 1708-1721.

[64] D'ALESSANDRO M, ALONSO A, CALABRÉS E, et al. Multimodal parameter-efficient few-shot class incremental learning[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Paris, France: IEEE, 2023: 3385-3395.

[65] KHAN M G Z A, NAEEM M F, VAN GOOL L, et al. Introducing language guidance in prompt-based continual learning [C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 11429-11439.

[66] VILLA A, ALCÁZAR J L, ALFARRA M, et al. PIVOT: Prompting for video continual learning[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 24214-24223.

[67] CHEN H, WU Z, HAN X, et al. Prompt fusion: Decoupling stability and plasticity for continual learning[C]//Proceedings of Computer Vision—ECCV 2024. Cham: Springer Nature Switzerland, 2025: 196-212.

[68] ZHOU D W, ZHANG Y, WANG Y, et al. Learning without forgetting for vision-language models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(6): 4489-4504.

[69] CHAUDHRY A, RANZATO M A, ROHRBACH M, et al. Efficient lifelong learning with A-GEM[EB/OL]. (2018-12-02). https://arxiv.org/abs/1812.00420.

[70] LI J, LI D, SAVARESE S, et al. BLIP-2[C]//Proceedings of the 40th International Conference on Machine Learning. Honolulu, Hawaii, USA: ACM, 2023: 19730-19742.

[71] DAI W, LI J, LI D, et al. InstructBLIP: Towards general-purpose vision-language models with instruction tuning[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2023: 49250-49267.

[72] LI L H, YATSKAR M, YIN D, et al. VisualBERT: A simple and performant baseline for vision and language[EB/OL]. (2019-08-03). https://arxiv.org/abs/1908.03557.

[73] GOYAL Y, KHOT T, SUMMERS-STAY D, et al. Making the V in VQA matter: Elevating the role of image understanding in visual question answering[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 6325-6334.

[74] XIAO J, SHANG X, YAO A, et al. NExT-QA: Next phase of question-answering to explaining temporal actions[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 9772-9781.

[75] KRISHNA R, BERNSTEIN M, LI F F. Information maximizing visual question generation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 2008-2018.

[76] PENG X, WEI Y, DENG A, et al. Balanced multimodal learning via on-the-fly gradient modulation[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 8228-8237.

[77] HUANG L, CAO X, LU H, et al. Class-incremental learning with CLIP: Adaptive representation adjustment and parameter fusion[C]//Proceedings of Computer Vision—ECCV 2024. Cham: Springer Nature Switzerland, 2025: 214-231.

[78] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Proceedings of Computer Vision— ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[79] PLUMMER B A, WANG L, CERVANTES C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 2641-2649.

[80] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019: 4171-4186.

[81] MANN B, RYDER N, SUBBIAH M, et al. Language models are few-shot learners for prognostic prediction[EB/OL]. (2023-02-24). https://arxiv.org/abs/2302.12692.

[82] 高志强, 沈佳楠, 姬纬通, 等. 大模型技术的军事应用综述[J]. 南京航空航天大学学报, 2024, 56(5): 801-814.
GAO Zhiqiang, SHEN Jianan, JI Weitong, et al. Review of military applications of foundation model technology[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2024, 56(5): 801-814.

[83] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc., 2017: 6000-6010.

[84] CUI F, YUE Y, ZHANG Y, et al. Advancing biosensors with machine learning[J]. ACS Sensors, 2020, 5(11): 3346-3364.

[85] XU P, ZHU X, CLIFTON D A. Multimodal learning with transformers: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12113-12132.

作者简介:

张伟(1981-),**通信作者**,男,博士,教授,博士生导师,研究方向:计算机视觉、机器人,E-mail:davidzhang@sdu.edu.cn。

钱龙玥(2000-),女,硕士研究生,研究方向:机器学习、计算机视觉。E-mail:202314856@mail.sdu.edu.cn。

张林(1993-),男,博士后,研究方向:图表示学习、机器学习、计算机视觉。

李腾(1989-),男,博士,研究员,硕士生导师,研究方向:深度学习、机器人感知与规划、智能系统。

(编辑:张黄群)