



## 面向大模型时代的持续学习方法论演变

王全子昂 王仁振 孟德宇 徐宗本

### The Evolution of Continual Learning Methodologies in the Era of Large Models

WANG Quan-Zi-Ang, WANG Ren-Zhen, MENG De-Yu, XU Zong-Ben

在线阅读 View online: <https://doi.org/10.16383/j.aas.c240805>

---

## 您可能感兴趣的其他文章

### 类别增量学习研究进展和性能评价

Class Incremental Learning: A Review and Performance Evaluation

自动化学报. 2023, 49(3): 635–660 <https://doi.org/10.16383/j.aas.c220588>

### 随机梯度下降算法研究进展

Research Advances on Stochastic Gradient Descent Algorithms

自动化学报. 2021, 47(9): 2103–2119 <https://doi.org/10.16383/j.aas.c190260>

### 一种基于随机权神经网络的类增量学习与记忆融合方法

A Class Incremental Learning and Memory Fusion Method Using Random Weight Neural Networks

自动化学报. 2023, 49(12): 2467–2480 <https://doi.org/10.16383/j.aas.c220312>

### 深度域适应综述: 一般情况与复杂情况

A Review of Deep Domain Adaptation: General Situation and Complex Situation

自动化学报. 2021, 47(3): 515–548 <https://doi.org/10.16383/j.aas.c200238>

### 深度对比学习综述

Deep Contrastive Learning: A Survey

自动化学报. 2023, 49(1): 15–39 <https://doi.org/10.16383/j.aas.c220421>

### 基于深度学习的多目标跟踪关联模型设计

Designing Affinity Model for Multiple Object Tracking Based on Deep Learning

自动化学报. 2020, 46(12): 2690–2700 <https://doi.org/10.16383/j.aas.c180528>

# 面向大模型时代的持续学习方法论演变

王全子昂<sup>1</sup> 王仁振<sup>1</sup> 孟德宇<sup>1,2,3</sup> 徐宗本<sup>1,2,3</sup>

**摘要** 以深度学习为代表的机器学习方法已经在多个领域取得显著进展,然而大多方法局限于静态场景,难以像人类一样在开放世界的动态场景中不断学习新知识,同时保持已经学过的知识.为解决该挑战,持续学习受到越来越多的关注.现有的持续学习方法大致可以分为两类,即传统的非预训练模型持续学习方法以及大模型时代下逐步演进的预训练模型持续学习方法.本文旨在对这两类方法的研究进展进行详细的综述,主要从四个层面对比非预训练模型和预训练模型方法的异同点,即数据层面、模型层面、损失/优化层面以及理论层面.着重分析从应用非预训练模型的方法发展到应用预训练模型的方法的技术变化,并分析出现此类差异的内在本质.最后,总结并展望未来持续学习发展的趋势.

**关键词** 持续学习,灾难性遗忘,预训练模型,机器学习,深度学习

**引用格式** 王全子昂,王仁振,孟德宇,徐宗本.面向大模型时代的持续学习方法论演变.自动化学报,2025,51(8):1-27

**DOI** 10.16383/j.aas.c240805 **CSTR** 32138.14.j.aas.c240805

## The Evolution of Continual Learning Methodologies in the Era of Large Models

WANG Quan-Zi-Ang<sup>1</sup> WANG Ren-Zhen<sup>1</sup> MENG De-Yu<sup>1,2,3</sup> XU Zong-Ben<sup>1,2,3</sup>

**Abstract** Machine learning methods, especially deep learning, have achieved remarkable progress across various fields. However, most approaches are limited to static scenarios and struggle to continually learn new knowledge in dynamic, open-world scenarios while retaining previously acquired knowledge, unlike humans. To address this challenge, continual learning (CL) has attracted increasing attention. Existing CL methods can be broadly categorized into two types: Traditional CL methods based on non-pretrained models, and CL methods based on pretrained models that have emerged with the advent of large models. This paper aims to provide a detailed review of the research progress in these two categories of methods, mainly comparing the similarities and differences between non-pretrained and pretrained model approaches from four perspectives: Data level, model level, loss/optimization level and theoretical level. We focus on analyzing the technical changes from methods employing non-pretrained models to those employing pretrained models, and analyze the underlying reasons for these differences. Finally, we summarize and envision the future trends in continual learning development.

**Key words** Continual learning, catastrophic forgetting, pretrained model, machine learning, deep learning

**Citation** Wang Quan-Zi-Ang, Wang Ren-Zhen, Meng De-Yu, Xu Zong-Ben. The evolution of continual learning methodologies in the era of large models. *Acta Automatica Sinica*, 2025, 51(8): 1-27

随着机器学习技术的发展,人工智能模型在很

多场景得到广泛的应用.以计算机视觉为例,在图像分类<sup>[1-4]</sup>、图像分割<sup>[5-7]</sup>、图像生成<sup>[8-10]</sup>等领域已经取得很好的效果.然而,当前的机器学习模型大多仅面向封闭的静态场景设计和开发,仍然依赖于数据的独立同分布 (Independent and identically distributed, i.i.d) 假设.这些方法通常需要人工收集并处理大量样本用于模型训练,训练后的模型在同分布的测试数据上能够取得较好的预测结果,然而面对非同分布或动态变化的数据分布往往表现不佳.如图1所示,在真实世界中,数据分布往往随时间和场景在持续地发生变化,例如电商领域的推荐系统,随着环境和时间变化,人们对不同商品的喜好程度是时刻变化的,并且会出现大量新商品需要模型拟合并预测,传统的机器学习模型难以处理这样复杂的动态场景.这就需要机器学习模型能够模

收稿日期 2024-12-24 录用日期 2025-04-10

Manuscript received December 24, 2024; accepted April 10, 2025

国家重点研发计划 (2020YFA0713900), 鹏城实验室重大项目 (PCL2024A06), 国家自然科学基金天元基金天元数学西北中心强化项目 (12426105), 国家自然科学基金 (62272375, 62306233) 资助

Supported by National Key R&D Program of China (2020 YFA0713900), Major Key Project of PCL (PCL2024A06), Tianyuan Fund for Mathematics of the National Natural Science Foundation of China (12426105), and National Natural Science Foundation of China (62272375, 62306233)

本文责任编辑 林宙辰

Recommended by Associate Editor LIN Zhou-Chen

1. 西安交通大学数学与统计学院 西安 710049 2. 西安交通大学智能网络与网络安全教育部重点实验室 西安 710049 3. 鹏城实验室 深圳 518052

1. School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049 2. Ministry of Education Key Laboratory of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049 3. Pengcheng Laboratory, Shenzhen 518052

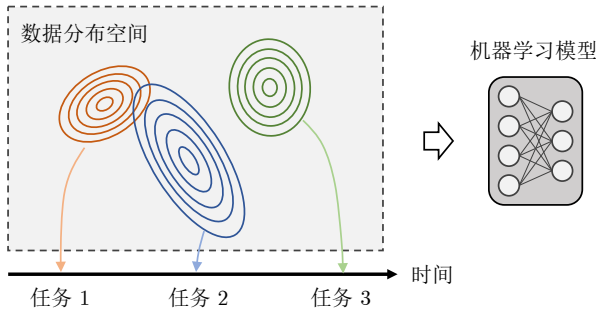


图 1 动态场景下的机器学习图示

Fig.1 Illustration of machine learning under the dynamic scenarios

拟人类应对动态场景的能力, 不断学习新知识同时保留已经学过的历史任务的知识. 这即是持续学习 (Continual learning, CL) 的问题基本设定<sup>[11-13]</sup>.

持续学习旨在模拟人类的终身学习知识的能力, 故而也称为终身学习 (Lifelong learning)、增量学习 (Incremental learning). 如图 2 所示, 持续学习希望机器学习模型按时序学习一系列任务的知识, 并且能够在所有见过的任务上表现良好. 由于学习新任务知识时难以接触到历史任务的数据, 所以模型很容易出现灾难性遗忘 (Catastrophic forgetting)<sup>[14]</sup> 问题, 即新任务到来后模型很快会遗忘已经学过的关于历史任务的知识, 导致模型在新任务上表现良好, 而过去见过的旧任务性能出现大幅下降. 所以持续学习方法希望尽可能保留处理旧任务的能力, 在此基础上融合新旧任务知识, 使得机器学习模型可以持续处理更加复杂多样的任务.

值得说明的是, 持续学习的目标和设定与多任务学习 (Multi-task learning)<sup>[15-16]</sup> 以及在线学习

(Online learning)<sup>[17]</sup> 不同. 多任务学习的目标是同时学习多个给定的不同任务, 希望模型在这些有限的任务上表现良好. 与持续学习的主要区别在于: 多任务学习的任务集合是固定的, 不会出现遗忘问题; 而持续学习的任务集合是增量式的, 未来任务的数量、类型和数据分布都是未知的, 处理新任务时极易出现灾难性遗忘. 另一方面, 在线学习的目标是在流式数据上训练机器学习模型, 其数据依然服从独立同分布假设, 类别集合始终保持固定; 而持续学习仅假设每个任务的数据服从独立同分布假设, 随着新任务的到来, 持续学习模型需要处理的类别集合会不断扩张. 当前的一些持续学习方法也考虑更加实际且具有挑战性的在线持续学习 (Online continual learning) 设置, 即假设每个任务都是流式数据, 导致整个持续学习过程中数据分布变化更加剧烈且不稳定, 其具体问题设置在后续介绍.

近年来, 很多优秀工作不断推动持续学习算法研究, 在图像分类、文本分类等多种问题上都取得很好的结果, 本文则主要关注图像分类问题的持续学习方法. 早期的持续学习方法主要针对规模相对较小的神经网络进行研究, 例如使用多层感知机 (Multi-layer perception, MLP) 或残差网络 (ResNet)<sup>[2]</sup> 作为模型的主干网络. 这些方法的网络参数通常是随机初始化 (Xavier 初始化<sup>[18]</sup>、He 初始化<sup>[19]</sup> 等). 在整个持续学习过程中需要不断优化网络的全部参数, 从而增强应对各个任务的能力. 然而, 神经网络通常是在一个庞大的搜索空间中基于梯度下降进行优化, 这就导致网络参数很容易随数据分布的变化而快速变化, 对于历史任务的表征不稳定. 尤其当模型遇到新任务时, 网络参数对于从未见过的

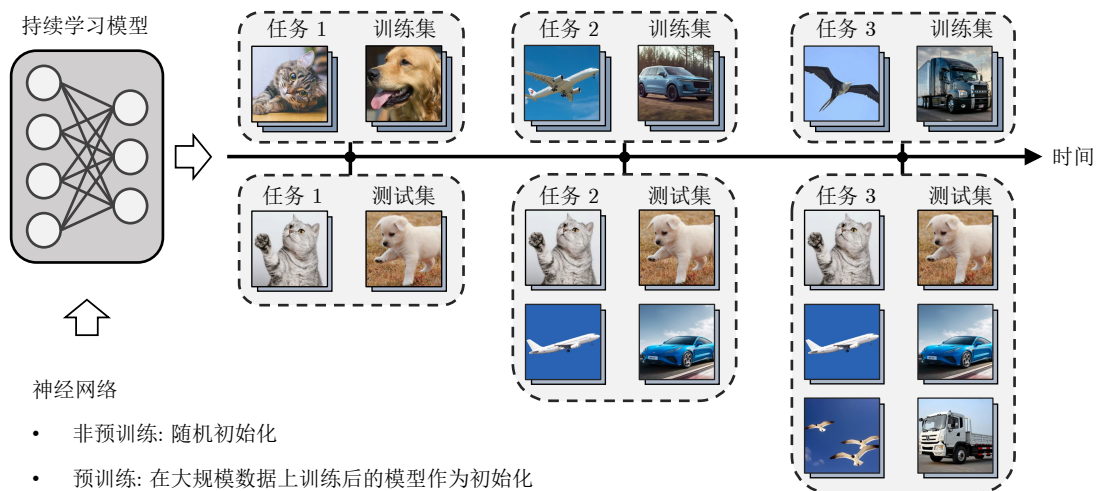


图 2 持续学习设置图示

Fig.2 Illustration of the continual learning setting

新任务数据往往表现得较为敏感, 其梯度量级要远大于对于历史任务数据的梯度, 使得模型很容易过度偏向沿着新任务梯度进行优化, 导致对于历史任务的表征受到破坏, 并出现灾难性遗忘. 所以早期持续学习方法尽管从不同角度提出很多有效且实用的方法来减轻遗忘问题, 但是受到主干网络优化的稳定性的限制, 整体的预测精度依然不高.

随着机器学习技术和开源社区的发展, 特别是随着大模型时代的到来, 研究者开始尝试使用一些大规模预训练神经网络作为机器学习模型的主干网络. 这些预训练模型通常是在庞大的数据集上进行训练, 其参数量远超过先前工作中使用的多层感知机或残差网络, 可以适用于多种下游任务并达到惊人性能<sup>[3-4, 10]</sup>. 因此, 这些预训练模型也被称为基础模型 (Foundation model)<sup>[20]</sup>, 如预训练 ViT<sup>[3]</sup>、CLIP<sup>[21]</sup> 以及各种大语言模型<sup>[22]</sup> 等. 预训练基础模型通常具有强大的特征提取能力和泛化能力, 例如在零样本泛化 (Zero-shot generalization) 问题上就能取得较为不错的表现. 自然地, 人们希望这些大规模的预训练模型也可以像人类一样不断学习新知识, 从而应对更为复杂的开放世界的动态场景问题. 由于预训练模型参数量非常庞大, 直接调整其参数会耗费大量计算资源成本, 所以一些工作提出参数高效微调 (Parameter efficient fine-tuning, PEFT)<sup>[23-24]</sup> 策略, 如 Prompt tuning<sup>[25-26]</sup>、Adaptor<sup>[27]</sup> 等方法. 然而在一系列下游任务上, 无论是直接调整预训练模型参数还是应用微调策略, 预训练基础模型仍然会出现严重的遗忘问题, 即下游任务表现良好, 但是预训练模型原有的特征表达和泛化能力会快速下降, 这就使得机器学习模型难以应对所有见过的任务, 无法处理动态场景问题. 为此, 近期出现大量基于预训练模型的持续学习方法, 希望利用预训练基础模型强大的表征能力来避免灾难性遗忘问题<sup>[28-30]</sup>.

为方便起见, 在持续学习领域中, 本文将先前使用参数量较少的神经网络作为主干网络且没有预训练的方法称为非预训练持续学习方法, 将应用大规模预训练基础模型作为主干网络的方法称为预训练持续学习方法. 现有的持续学习领域的综述文章并没有深入探讨传统的非预训练持续学习方法和预训练持续学习方法之间的本质联系和区别. 具体而言, 综述 [11] 是较早的关于持续学习的完整综述, 仅整理当时已有的非预训练持续学习方法; 综述 [12-13] 简单介绍预训练持续学习方法, 并将其看作是一类独立的方法, 并未深入考虑这些方法和非预训练持续学习方法的差异和联系; 而综述 [28] 则主要关注预训练持续学习方法的发展, 对其进行更细

致的划分和整理. 另外, 还有一些综述文章仅针对特定的设置或领域进行总结, 如综述 [31-32] 仅关注小样本持续学习的设置; 综述 [29-30] 对基于预训练大模型的持续学习方法进行整理; 综述 [33] 则重点关注多模态持续学习领域的发展并对相关方法进行总结. 本文重点关注视觉相关的分类问题, 重新梳理持续学习整体发展路线, 整理并分析从非预训练持续学习方法到预训练持续学习方法的思路和技术上的变化, 着重探讨目前更加前沿的基于预训练大模型的持续学习依然存在的问题, 深入探讨持续学习未来可能的发展趋势.

本文的整体结构如下: 第 1 节主要介绍持续学习的基本框架、常见设置以及评价指标; 第 2 节分别从数据、模型、损失/优化以及理论四个层面介绍非预训练模型和预训练模型的方法, 对比两类方法的技术变化并分析其差异化的原因, 最后探讨未来持续学习技术的发展趋势; 第 3 节简要介绍持续学习技术在其他相关领域的应用; 第 4 节对全文进行总结与展望.

## 1 持续学习基本设置

### 1.1 基本框架

持续学习通常考虑在一系列不同任务的集合上训练机器学习模型, 并且在训练某个特定任务时, 先前见过的任务数据通常难以重新获得. 以有监督分类问题为例, 对于第  $t$  个任务训练数据  $\mathcal{D}_{trn}^t = \{\mathcal{X}_{trn}^t, \mathcal{Y}_{trn}^t\}$ , 其中  $\mathcal{X}_{trn}^t$  和  $\mathcal{Y}_{trn}^t$  分别表示输入的训练数据和对应标签, 标签  $\mathcal{Y}_{trn}^t$  所包含的类别集合记为  $\mathcal{C}^t$ , 持续学习要求模型能够学习从输入数据到对应标签的映射. 在第  $t$  个任务训练结束后, 经过训练的持续学习模型需要能够在所有见过的任务的测试数据  $\mathcal{D}_{tst}^{1:T} = \cup_{s=1}^t \mathcal{D}_{tst}^s$  上表现良好, 其中  $\mathcal{D}_{tst}^s = \{\mathcal{X}_{tst}^s, \mathcal{Y}_{tst}^s\}$  表示第  $s$  个任务的测试集,  $\mathcal{X}_{tst}^s$  和  $\mathcal{Y}_{tst}^s$  分别表示测试数据及其标签.

如图 2 所示, 持续学习的目标是在一系列任务  $\mathcal{D}_{trn}^{1:T} = \{\mathcal{D}_{trn}^1, \mathcal{D}_{trn}^2, \dots, \mathcal{D}_{trn}^T\}$  上优化模型参数  $\theta$ . 假设模型参数服从先验概率  $p(\theta)$ , 从贝叶斯角度来看, 在观测到一系列任务后的后验概率可以表示为

$$p(\theta|\mathcal{D}_{trn}^{1:T}) \propto p(\theta|\mathcal{D}_{trn}^{1:T-1})p(\mathcal{D}_{trn}^T|\theta) \propto p(\theta) \prod_{t=1}^T p(\mathcal{D}_{trn}^t|\theta) \quad (1)$$

其中, 后验概率  $p(\theta|\mathcal{D}_{trn}^{1:T})$  正比于模型参数的先验概率  $p(\theta)$  和各个任务上的似然概率  $p(\mathcal{D}_{trn}^t|\theta)$  的乘积. 可以看到, 模型参数的先验概率分布对整个持续学



习模型都有着至关重要的作用,而非预训练持续学习模型和预训练持续学习模型的主要区别就在于主干模型参数服从的先验概率分布不同。

具体而言,非预训练持续学习模型通常采用常规的初始化方法获得初始参数  $\theta$ ,例如常见的 Xavier 初始化<sup>[18]</sup>或 He 初始化<sup>[19]</sup>等。这些先验分布主要是为保证模型较为稳定地训练和收敛,避免梯度爆炸或梯度消失等问题,而无法提供额外信息以增强模型能力,所以非预训练模型只能从每个任务中逐渐学习知识并增强判别能力。

然而,对于预训练持续学习模型,其主干网络通常是在大规模数据集  $\mathcal{D}_{pre}$  上训练过的基础模型,其参数  $\theta$  服从的先验概率分布就等于预训练基础模型参数的后验概率,即

$$p(\theta) = p(\theta|\mathcal{D}_{pre}) \quad (2)$$

其中,预训练阶段的后验概率  $p(\theta|\mathcal{D}_{pre})$  建模丰富的数据信息,提供强大的特征提取能力和泛化能力,使得预训练持续学习方法出现一些技术上的变化。

## 1.2 常见设置

至今持续学习已经发展出很多不同的设置,这些设置看起来非常繁杂,但大多可以通过以下几种方式对现有设置进行划分:

1) 任务不相交/任务混淆. 根据持续学习各个任务的数据空间设置的不同,可以将持续学习设置大致分为以下两类:

- 任务不相交 (Task disjoint). 持续学习中的各个任务不相交,即对于任意的第  $i$  个和第  $j$  个任务,其数据集满足  $\mathcal{D}^i \cap \mathcal{D}^j = \emptyset$ 。

- 任务混淆 (Task blurry). 各个任务的数据并非完全不相交,即满足  $\mathcal{D}^i \cap \mathcal{D}^j \neq \emptyset$ 。具体而言,假设第  $k$  个类别的数据主要出现在第  $i$  个任务,然而有少量 ( $K\%$ ) 的第  $k$  个类别的数据会出现在其他任务中,即所谓的 Blurry- $K$  设置。

2) 类/任务/域增量学习. 根据任务数据的标签空间的假设,一种主流的分类方法将持续学习设置分为以下几类<sup>[34]</sup>:

- 类增量学习 (Class incremental learning, CIL). 对于给定的测试样本  $x \in \mathcal{D}^t$ , 其任务编号  $t$  未知,需要模型自行判断该样本属于哪个任务并进行分类。

- 任务增量学习 (Task incremental learning, TIL). 对于给定的测试样本  $x \in \mathcal{D}^t$ , 其任务编号  $t$  未知,模型可以将任务编号作为先验知识来指导分类预测。

- 域增量学习 (Domain incremental learning,

DIL). 对于给定的测试样本  $x \in \mathcal{D}^t$ , 其任务编号  $t$  未知;同时,假设各个任务仅数据空间发生域迁移,其标签空间保持一致。

- 无任务持续学习 (Task-free continual learning)<sup>[35-36]</sup>. 对于给定的训练样本或测试样本,其任务编号都未知,需要模型在训练阶段自动判别任务切换并适应非稳态数据分布。

这些设置默认假设任务不相交,但也可以与任务混淆设置相结合,例如任务混淆类增量设置<sup>[37-38]</sup>。

3) 在线/离线持续学习. 根据每个任务的训练是在线还是离线学习可以分为以下两类:

- 在线持续学习 (Online CL). 在计算资源受限的情况下,模型需要持续地在线处理流式数据,即对于每个新任务的训练样本,模型仅能训练一次。这就要求模型快速适应数据分布变化,在尽可能避免遗忘的同时快速学习新知识。

- 离线持续学习 (Offline CL). 在训练新任务时,该任务的数据可以暂时储存,并反复用于模型迭代优化以获得更好的表征能力。相比在线持续学习,离线持续学习的数据分布更加稳定,通常可以获得更稳定的模型和更好的性能表现。

4) 单模态/多模态持续学习. 根据持续学习过程中的各个任务模态将设置分为以下两类:

- 单模态持续学习 (Unimodal CL). 在整个持续学习过程中的所有任务保持单个模态。例如,所有任务仅包含视觉图像一个模态。

- 多模态持续学习 (Multimodal CL). 所有任务都是多模态任务,并且在持续学习过程中,不同任务的模态有可能不同。

本文主要关注视觉模态相关的持续学习方法,包括仅包含视觉模态的单模态持续学习方法以及利用文本、音频或其他模态促进视觉模态任务的持续学习方法。其他例如文本单模态的持续学习方法仅在最后做简单总结。

以上介绍的一些主流分类方法还可以互相组合以构成更加复杂的设置,并且很多工作还尝试结合噪声标签<sup>[39-40]</sup>、不均衡<sup>[41-42]</sup>、少样本<sup>[31-32]</sup>等设置以解决更加实际的持续学习问题<sup>[43]</sup>。还有一些其他前沿的研究方向也涉及持续学习相关技术,本文将在最后进行简要介绍。

## 1.3 评价指标

根据持续学习的目标,在分类问题中的主要评价指标有三个<sup>[11-13]</sup>,即平均精度、遗忘程度以及前向传播程度,分别用于衡量持续学习模型在所有见过的任务上的性能、关于历史任务的遗忘程度以及模

型学习新任务的能力. 具体而言, 三种评价指标的计算如下:

1) 平均精度 (Average accuracy, ACC). 为衡量持续学习模型的性能, 可以直接平均最终所有任务的预测精度. 记  $a_{(s, t)}$  表示持续学习模型在第  $t$  个任务训练结束后, 在第  $s$  个任务的测试数据上计算的预测精度, 其中  $s \leq t$ . 则当模型在所有的  $T$  个任务训练结束后的平均精度表达式为:

$$\text{ACC} = \frac{1}{T} \sum_{t=1}^T a_{(t, T)} \quad (3)$$

2) 遗忘程度 (Forgetting measure, FM). 为衡量在持续学习训练过程结束后的先前每个历史任务的遗忘程度, 可以用该任务的最高预测精度和最终该任务精度的差异作为遗忘程度的度量. 记  $a_t^*$  表示持续学习模型在整个持续学习过程中第  $t$  个任务的最高测试精度, 则  $a_t^* - a_{(t, T)}$  表示第  $t$  个任务到训练结束时的遗忘程度. 由此, 当模型在所有的  $T$  个任务训练结束后的遗忘程度表达式为:

$$\text{FM} = \frac{1}{T-1} \sum_{t=1}^{T-1} (a_t^* - a_{(t, T)}) \quad (4)$$

3) 前向传播程度 (Forward transfer, FWT). 为衡量模型学习新任务的难易程度, 考虑计算在新任务上训练后的模型相较于随机模型的性能提升. 假设当前训练第  $t$  个新任务, 则构造一个随机初始化的模型在该任务上的预测精度  $\tilde{a}_t$ , 以及持续学习模型在新任务上训练后的预测精度  $a_{(t, t)}$ . 由此, 当模型在所有的  $T$  个任务训练结束后的前向传播程度表达式为:

$$\text{FWT} = \frac{1}{T-1} \sum_{t=2}^T (a_{(t, t)} - \tilde{a}_t) \quad (5)$$

该指标越高, 则表明持续学习模型可以更好地学习新知识, 说明模型学习过的历史任务知识有可能会更好地促进新任务知识的学习.

值得我们思考的是, 在预训练持续学习相关方法的评价体系中, 有可能存在“虚假繁荣”的问题. 这是因为模型的主干网络在预训练阶段已经在很多数据类别上表现良好, 即使后续不再训练并直接用于某些数据上进行测试, 依然可以依靠预训练模型自身的零样本泛化能力获得不错的预测结果, 然而这并不能说明持续学习方法的有效性. 例如, 在 ImageNet-1K 上预训练的基础模型直接在 CIFAR-10 这样的小规模数据上测试, 其效果甚至比非预训练持续学习模型的效果还要好, 因此在这样强大的主干网络基础上设计得到的持续学习方法需要更加

细致和谨慎的对比以验证方法的有效性.

## 2 持续学习方法研究脉络

根据机器学习的基本流程, 以下将持续学习方法大致分为三个层面进行介绍: 数据层面、模型层面以及损失/优化层面. 此外, 本节也将介绍持续学习理论层面的研究进展. 对于每个层面的持续学习方法, 主要侧重于分析非预训练持续学习模型到预训练持续学习模型的研究思路和方法的演化趋势, 着重分析出现这些演进变化的原因并探讨未来持续学习可能的发展方向与趋势.

### 2.1 数据层面

对于机器学习算法而言, 数据输入是最为基础的步骤. 根据静态场景下传统机器学习的经验, 如果模型输入的是高质量的训练数据, 往往也能得到较好的结果. 这样的经验也可以迁移到动态场景下, 并且已经有很多工作尝试从数据层面出发对持续学习方法进行改进, 已经取得很好的效果. 本节将分别介绍非预训练持续学习模型和预训练持续学习模型是如何处理输入数据, 从而在避免遗忘的同时达到更好的预测精度.

#### 2.1.1 数据层面的非预训练持续学习模型

在数据层面上, 核心思想即数据重放 (Data replay), 是指在学习新任务知识时, 将少量历史任务数据和新任务数据一同进行训练, 从而在学习新知识的同时有效避免遗忘. 如图 3 所示, 这一脉络的持续学习方法论主要的思路有两种: 保存少量真实历史任务数据以及生成大量历史任务数据.

1) 基于重放的方法 (Rehearsal). 重放方法<sup>[44-48]</sup>假设在训练历史任务的过程中可以将少量样本保存在记忆缓存 (Memory buffer) 中, 在后续新任务的训练过程中不断重放. 记记忆缓存为  $\mathcal{M} = \{\mathcal{X}_{buf}^{1:t-1}, \mathcal{Y}_{buf}^{1:t-1}\}$ , 其中  $\mathcal{X}_{buf}^{1:t-1}$  和  $\mathcal{Y}_{buf}^{1:t-1}$  包含所有见过的历史任务的少量数据和对应标签, 通常记忆缓存大小  $|\mathcal{M}|$  会受到限制, 无法保存所有的历史数据. 由此, 重复回放方法的训练损失基本可以表达为如下形式:

$$\min_{\theta} \mathcal{L}(f_{\theta}(\mathcal{X}_{trn}^t), \mathcal{Y}_{trn}^t) + \mathcal{L}(f_{\theta}(\mathcal{X}_{buf}^{1:t-1}), \mathcal{Y}_{buf}^{1:t-1}) \quad (6)$$

其中,  $f_{\theta}(\cdot)$  表示持续学习模型的主干网络; 第一项损失表示新任务数据的损失; 第二项表示记忆缓存中历史任务数据的重放损失, 可以看作是正则项以避免模型训练过度倾向于新任务, 从而减轻遗忘问题.

然而, 由于记忆缓存通常只能保存极为少量的数据, 这使得在长期的持续学习过程中, 模型很容易在记忆缓存数据上发生过拟合, 降低模型对历史

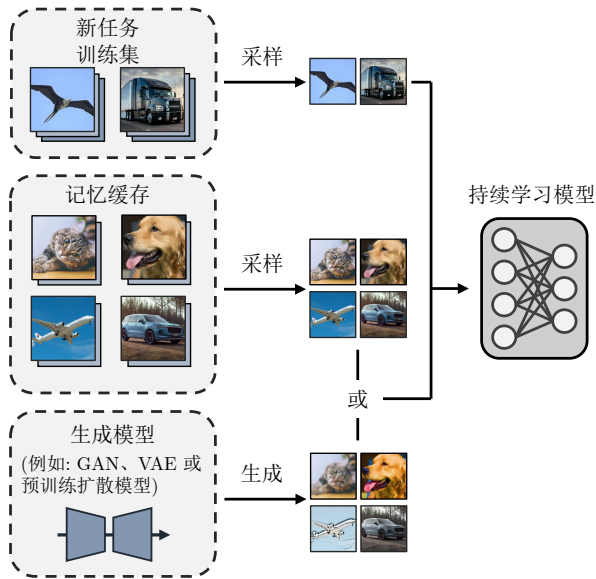


图3 数据层面的持续学习方法图示

Fig.3 Illustration of CL methods on the perspective of data

任务的泛化性,从而出现更严重的灾难性遗忘现象.为缓解这一问题,很多工作从不同角度提出改进方案以进一步提高记忆缓存数据的利用效率,主要可以划分为如下三类方法:

- 数据增广. 数据增广是一种常用的避免数据过拟合的方法,主要思想是对原始数据图像进行随机裁剪、随机翻转等操作以增强数据的多样性<sup>[49-51]</sup>.在持续学习过程中,记忆缓存中的少量数据很容易被神经网络过度拟合,而对记忆缓存数据应用不同的数据增广<sup>[47, 52]</sup>可以有效改善过拟合问题,进而改善遗忘程度.例如 RAR<sup>[53]</sup>对缓存数据进行随机增广 (RandAug) 以减轻过拟合问题的发生.

- 数据表征. 除数据增广这种简单常用的方式,一些工作考虑通过其他技术来增强缓存数据的表征能力.例如 DER++<sup>[47]</sup>额外保存历史模型对记忆缓存数据的预测概率,用于进一步约束当前网络的预测概率分布; MRDC<sup>[54]</sup>通过 JPEG<sup>[55]</sup>压缩历史数据从而在有限的记忆缓存内保存更多的历史任务数据,可以有效增加缓存数据的多样性和表征能力; SER<sup>[48]</sup>则通过显著性预测图作为额外的约束使得模型更加关注分类目标,从而增强数据表征.

- 数据选择. 记忆缓存的目的实际上是通过缓存数据近似过去历史任务的数据分布,然而不同数据所包含的信息量不同,很多工作关注于如何选择有限的样本以更好地近似历史任务数据分布.一些工作<sup>[37-38, 46, 56-59]</sup>通过手工设计数据选择标准进行选择.如 iCaRL<sup>[46]</sup>选择和类中心距离最近的样本存入记忆缓存<sup>[60]</sup>; RM<sup>[38]</sup>设计样本代表性指标,根据代表

性等间隔采样并存入记忆缓存;还有一些工作通过梯度<sup>[37]</sup>或信息论<sup>[59]</sup>等方法选择记忆缓存数据.除此之外,一些工作<sup>[61-64]</sup>则通过双层优化 (Bi-level optimization) 方法从每个任务数据集中自动选择样本,其核心思想是希望在内层优化中选择出来的样本可以在外层优化中近似使用全部数据训练网络的结果.

这些方法在很多持续学习设置下都能取得很好的结果并且能极大降低模型遗忘程度,然而在实际情况中的很多场景下无法获得历史任务数据,尤其是保存的缓存数据可能会存在隐私问题,如医疗数据<sup>[65]</sup>.为此,一些工作考虑根据历史模型生成过去任务的数据进行重放.

2) 基于伪重放的方法. 随着生成模型的发展,一些持续学习方法除主干网络外还构造一个生成模型,如 GAN<sup>[8]</sup>、VAE<sup>[9]</sup>等,并在持续学习过程中同时训练主干网络和生成模型,这类方法也称为伪重复回放 (Pseudo rehearsal) 方法<sup>[66-73]</sup>.具体而言,生成模型用于生成所有见过任务的数据,并将生成数据和当前任务数据一同训练主干网络.除此之外,还有很多方法考虑使用不同的生成方法合成历史任务数据.例如数据集蒸馏 (Dataset distillation)<sup>[74-75]</sup>,即通过优化方法将历史任务数据集的信息压缩至少量合成样本上<sup>[76-78]</sup>;或是通过模型反演 (Model inversion)<sup>[79-80]</sup>,在新任务训练前从上一个任务训练结束后的模型恢复出历史任务数据<sup>[81]</sup>.

另一方面,由于图像空间极其复杂,生成模型需要对每个像素点进行合理预测才能构成一个可以使用的合成样本,该过程训练难度较大,对于生成模型的要求也较高.为降低生成图像样本的难度,一些工作<sup>[82-83]</sup>考虑在主干网络的特征空间中生成特征向量用于重放.因为特征空间通常相对简单,一般可以认为特征向量分布在一个低维流形上,由此可以极大程度上降低生成模型的训练和推理难度.生成的历史任务的特征向量可以用于训练分类器,增强模型跨任务的判别能力.

总的来说,生成方法虽然可以有效解决数据隐私问题,同时可以轻松获得大量历史任务数据,然而依然存在三个关键问题: a) 生成方法的计算量通常较大,优化较为困难; b) 很多生成方法存在模式塌缩问题,即生成的样本较为相似,导致生成模型生成的大量合成数据多样性仍然不足; c) 生成数据和真实数据的分布差异较大,导致主干网络在数据重放时训练出现偏差,其遗忘问题并未得到较好的改善.

### 2.1.2 数据层面的预训练持续学习模型

基于预训练模型,研究者们关注更加真实且具



有挑战的设置, 即无重放 (Rehearsal-free 或 Exemplar-free) 设置<sup>[84]</sup>, 大多此类方法更关注使用预训练生成模型来合成历史任务数据. 在图像生成领域, 扩散模型 (Diffusion model)<sup>[10]</sup> 已经取得惊人的效果, 可以生成视觉质量很高的图像. 一个直观的思路就是应用预训练的扩散模型持续生成旧任务数据以避免遗忘<sup>[85-90]</sup>. 在持续学习过程中, 仅需对预训练的扩散模型进行少量微调即可适应当前任务的数据分布, 从而生成较高视觉质量的合成样本用于重放. 相比先前利用 GAN 等传统生成方法的模型, 这些方法不再需要从头训练生成模型, 其计算资源和训练难度显著下降, 同时这些方法的生成性能也显著提升. 另一方面, 由于预训练扩散模型具有较强的外推能力, 有可能生成不真实的分布外数据, 从而污染合成历史任务数据集. 如果在网络训练中使用过多的被污染的合成数据 (Data contamination), 反而有可能损害模型性能<sup>[91-92]</sup>. 为此, 一些研究者尝试降低在持续学习问题中合成数据带来的负面影响<sup>[93]</sup>.

类似非预训练持续学习模型的思路, 还有一些预训练方法也尝试采用特征重放以及模型反演的思路进行持续学习. 例如, H-Prompts<sup>[94]</sup> 假设预训练模型提取的特征服从高斯分布, 保存各个类别的特征向量的均值和协方差矩阵, 在后续任务中从高斯分布采样即可获得生成特征向量, 用于特征重放以减轻遗忘; 也有一些工作<sup>[95]</sup> 发现利用预训练模型进行模型反演得到的数据质量更高, 能获得更好的持续学习性能.

另外, 在处理一些实际问题中的遗忘现象时, 如果可以不考虑数据隐私问题, 保存少量历史任务数据并进行重放仍然是一种非常高效的减轻遗忘的做法. 例如, 在多模态持续学习问题中, TAM-CL<sup>[96]</sup>、VQACL<sup>[97]</sup>、KDR<sup>[98]</sup> 等方法直接保存少量数据进行重放, IncCLIP<sup>[99]</sup>、SGP<sup>[100]</sup>、AID<sup>[101]</sup> 等方法则采用生成历史数据的方法进行伪重复回放.

### 2.1.3 发展趋势

随着近年来预训练模型的快速发展, 更多研究者开始关注如何利用预训练大模型的能力来提升持续学习任务的性能. 预训练模型本身具有很强的特征提取和泛化能力, 遗忘程度相比非预训练模型有很大程度下降, 同时也不需要大量数据来维持网络的判别能力. 因此, 研究者可以更加关注保护数据隐私, 着重考虑更加困难的无重放持续学习方法的开发.

在生成历史任务数据方面, 尽管预训练扩散模型生成数据的视觉质量很高, 然而对于机器学习模型而言, 生成数据和真实数据之间仍然存在分布差

异. 合成数据在持续学习过程中仍然不能精确近似历史任务数据分布, 导致模型可能出现偏差, 无法很好地保留对于历史任务的判别能力. 所以如何利用预训练生成模型更好地近似历史任务的真实数据分布并促进主干网络训练, 仍然是一个未来需要继续深入探索的重要问题.

另外值得注意的是, 尽管目前有大量工作尝试无重放或伪重放问题, 然而在实际问题中, 如果数据隐私允许, 保存少量历史任务数据仍然是一种非常简单且高效的方法, 通过重放可以有效帮助持续学习模型回忆学过的知识, 从而极大程度上减轻遗忘. 所以即使是在大模型时代, 高效的重放策略依然是实际问题中的常用解决方案. 因此, 如何选择更具有代表性的历史数据, 或通过数据集蒸馏等方法将历史任务信息整合到少量合成数据中也是未来持续学习的一个重点研究方向.

## 2.2 模型层面

在机器学习中, 选择合适的模型对于性能表现也是至关重要的. 在持续学习方法中, 有诸多研究专注于设计合理的网络模型以适配持续学习问题, 从模型自身结构出发来减轻遗忘问题. 本节将展示非预训练持续学习模型和预训练持续学习模型在模型层面的改进, 总结它们之间的内在联系并分析在此层面持续学习方法论可能的未来发展趋势.

### 2.2.1 模型层面的非预训练持续学习模型

对于传统的非预训练持续学习模型, 整个主干网络的参数  $\theta$  都需要不断优化, 不同任务之间很容易互相干扰. 为增强模型的表征能力并减轻任务之间相互干扰的问题, 对模型本身的结构进行改进是非常直接有效的策略, 相关的改进大致可以分为三个方面:

1) 模型表征. 由于历史任务数据难以获得, 主干网络的特征提取能力有限, 并且网络在训练过程中容易走“捷径” (Shortcut)<sup>[102]</sup>, 即网络倾向于利用数据中的简单规律或偏差来完成任务, 而非具有普遍意义的特征, 这就进一步导致模型极易出现遗忘问题<sup>[103-104]</sup>. 例如图 2 中展示的第 2 个和第 3 个任务, 假设第 2 个分类任务学习判别飞机和小汽车两个类别, 网络很容易捕捉到飞机的“蓝天”背景以及小汽车中“轮子”的结构, 依靠这两个显著特征就可以很好地区分飞机和小汽车的图像. 然而当第 3 个任务学习鸟和货车两个类别时, 由于鸟的背景往往也是“蓝天”, 并且货车也有“轮子”结构, 那么网络就会出现混淆, 难以判别飞机和鸟这两类图像, 也很难区分小汽车和货车, 从而使训练模型出现严重



遗忘的问题。

为此, 增强持续学习过程中模型的表征能力至关重要。很多方法采用自监督学习<sup>[105-108]</sup>、因果学习<sup>[109-111]</sup>等方法来增强主干网络的表征能力; 也有利用大量无标记数据结合当前任务数据进行半监督学习<sup>[112-114]</sup>, 希望模型通过处理大量多样化的数据以避免学习到一些有偏的特征。这些方法可以有效提高模型的特征提取能力, 进而在一定程度上避免任务之间的干扰并减轻遗忘问题。

除此之外, 还有一些工作从模型集成的角度增强特征表达。受生物学领域互补学习系统 (Complementary learning systems, CLS)<sup>[115]</sup> 启发, 一些工作<sup>[116-118]</sup> 尝试构造两个主干网络分别用于模拟大脑中快速学习的海马体 (Hippocampus) 以及慢速学习的新皮层 (Neocortex)。其中快速学习网络通过梯度下降等优化算法尽快适应新任务知识, 而慢速学习网络则可以通过滑动平均 (Moving average) 等方式缓慢整合快速学习网络知识, 最终测试时通过集成快速学习网络和慢速学习网络的预测概率来获得最终预测概率。这样的预测概率既考虑到新任务知识, 同时也保证尽可能稳定地巩固所有旧任务的知识, 有效增强模型表征并提升了预测精度。

2) 模型偏差。随着持续学习的训练, 主干网络中的线性分类器容易出现任务近因性偏差 (Task-recency bias)<sup>[119-121]</sup>, 该偏差是指模型很容易将历史任务的测试样本预测为新任务的类别, 导致新任务预测精度较高而旧任务性能出现急剧下降。一些方法<sup>[46, 119]</sup> 直接在测试阶段将线性分类器替换为最近类别均值 (Nearest class mean, NCM) 分类器, 即在训练阶段计算各个类别特征向量的均值作为类别中心, 对于测试样本仅需计算特征向量和各个类别中心距离, 选择最近的类别作为预测结果。这种方法虽然规避了线性分类器的任务近因性偏差, 但是高度依赖于特征提取器的表征能力, 尤其是在线持续学习设置下, 特征提取器经常处于欠拟合状态, 导致这种做法的性能表现不佳<sup>[121]</sup>。另外, 这类方法在训练阶段仍然使用的是线性分类器, 然而任务近因性偏差也会导致分类器回传梯度不准确, 进而影响特征提取器的学习, 影响了模型的测试性能。

为探究这一偏差出现的原因, 很多研究做出深入细致的分析。研究者最早简单地认为这是由于缺少旧任务数据造成的, 即使在记忆缓存中保存少量的历史任务数据, 其数据量也要远小于新任务的数据量, 这就导致新旧任务之间存在不平衡 (Imbalance) 问题。由此, 一些工作<sup>[122-125]</sup> 通过分别调整线性分类器关于新旧任务类别的分类向量的范数来解决

该偏差。随着研究的深入, 一些工作发现出现该偏差的原因较为复杂, 正如前面的分析, 特征提取器和分类器参数是耦合在一起的, 训练过程中互相影响, 单独对分类器进行处理可能是不充分的, 特征提取器也可能存在偏差<sup>[83, 126-127]</sup>。另外, 一些工作<sup>[128-130]</sup> 发现特征提取器的 BN (Batch normalization) 层的统计量也存在倾向于新任务的偏差, 进而导致模型出现任务近因性偏差。更进一步, 近期工作<sup>[121]</sup> 提出持续偏差适配器, 直接矫正主干网络输出的后验概率分布, 并通过双层优化联合更新偏差适配器和主干网络参数, 显著矫正了模型的任务近因性偏差问题<sup>[131]</sup>。

3) 模型结构。由于每个任务之间可能会互相干扰, 一个直观的思想是针对每个任务的数据自动选择不同的参数进行预测, 由此避免新任务对旧任务的扰动从而造成遗忘。如图 4(a)、图 4(b) 所示, 这类方法根据模型容量是否发生变化可以大致分为扩展模型和路径模型两类:

- 扩展模型。如图 4(a) 所示, 这类方法的核心思想是在持续学习过程中根据需求不断扩展网络结构。早期的扩展模型方法, 如 PNN<sup>[132]</sup>、DER<sup>[133]</sup> 等, 直接为每个任务训练一个特征提取器, 通过选择机制或是直接扩展特征向量来增强模型表征能力。尽管这些方法具有良好的表现, 但是不断增加的模型容量导致存储和推理资源消耗巨大, 尤其难以应对具有较多任务的持续学习问题。为减轻计算和存储代价, 一些工作<sup>[134-139]</sup> 考虑在原始主干网络的基础上构造多个专家模型, 其中每个专家模型的体量都不大, 不同任务使用不同的专家模型进行预测, 由此尽可能避免任务间干扰。为避免模型容量随任务数量无限地线性增长, 很多工作考虑根据任务相似性判断是否增加专家模型: 如果新任务较为复杂且和先前任务差异较大, 则构建新的专家模型来帮助训练和推理; 反之, 如果新任务和某些历史任务较为相似, 则利用现有的专家模型进行训练和推理, 不会扩展模型容量。这种动态扩展的专家模型思路无论是在表征能力还是存储代价上都取得了比较好的平衡。

- 路径模型。如图 4(b) 所示, 与扩展模型思路不同, 这一类方法为避免模型容量的不断增大, 考虑基于固定容量的主干网络, 对于不同任务的数据通过门控机制<sup>[140]</sup> 或掩码<sup>[141-145]</sup> 等方式选择不同的参数构成子网络进行推理。只需要在持续学习过程中针对每个任务训练不同的选择方式, 就可以实现在固定容量的主干网络中动态选择路径进行推理, 从而产生特定性的预测结果, 以此减弱任务间的互相干扰问题。

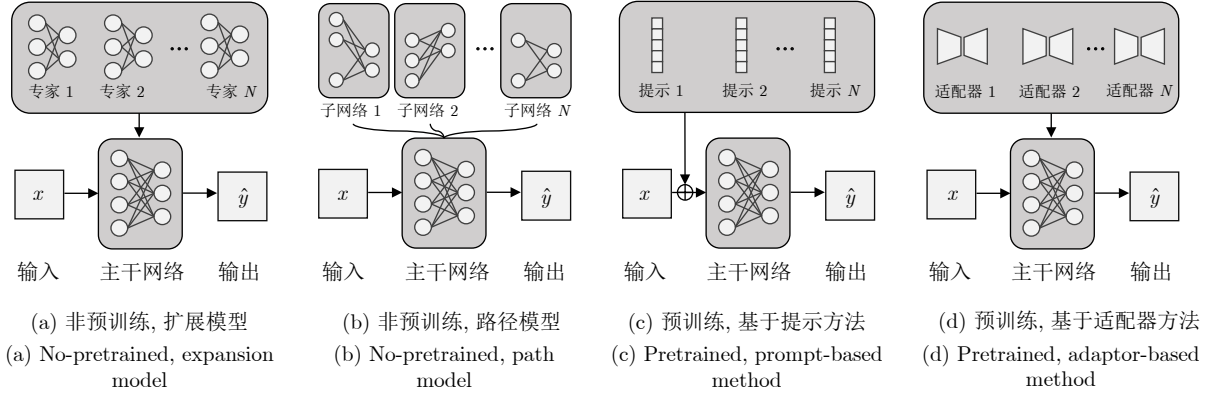


图 4 模型结构相关持续学习方法图示

Fig. 4 Illustration of CL methods on model architecture

由于这些方法通过对模型网络进行改进, 极大程度上避免不同任务之间的互相干扰, 通常都能达到较好的预测精度. 然而无论是扩展模型还是路径模型的训练难度通常较大, 而且专家模型容易出现退化现象, 即不同任务在选择专家模型时会趋于同样的选择, 而其他未被选择的专家模型很少被训练且无法提供额外的信息, 进而导致严重的遗忘, 所以在训练专家模型时往往需要添加额外的正则项迫使专家模型多样化、差异化.

### 2.2.2 模型层面的预训练持续学习模型

当使用预训练模型作为持续学习方法的主干网络时, 其中大量参数无需调整, 仅需要微调一小部分参数即可实现对新任务的适应, 基于此思想, 发展出一大批和先前工作看似完全不同的策略来进行持续学习. 然而, 这些方法的本质思想并未发生变化, 也可以类似地大致划分为模型表征、模型偏差以及模型结构三类:

1) 模型表征. 类似传统的非预训练持续学习模型, 一些工作<sup>[146-147]</sup>为进一步增强预训练模型的特征提取能力, 通过融合自监督等方法进行训练, 也取得不错的效果. 另外, 最近一些工作<sup>[148]</sup>发现使用随机映射 (Random projection, RP) 将预训练模型提取的数据特征映射至高维空间即可提高分类判别性, 其中随机映射是随机初始化的, 并且在训练过程中保持固定, 在几乎不增加训练代价的同时可获得更好的分类能力. 除此之外, 还有一些工作<sup>[149-150]</sup>尝试使用较小的学习率微调预训练模型的特征提取器并用较大的学习率调整分类器, 由此可以更好地适应下游任务, 同时不会过分遗忘.

2) 模型偏差. 一些工作发现即使应用预训练模型作为主干网络, 其分类器也可能存在偏差. 然而和非预训练持续学习模型的任务近因性偏差不同,

预训练模型中分类器的偏差主要表现为缺乏跨任务判别性, 即每个任务内部的分类通常能达到较好的性能, 然而对于不同任务的样本有可能出现混淆. 其主要原因是在训练分类器时, 为避免任务间干扰, 通常会固定先前训练过的关于历史任务的分类向量, 仅更新当前任务对应类别的分类向量, 由于缺乏历史任务数据, 导致分类器没有协同训练各个任务的分类向量, 进而使得分类器缺乏跨任务判别性. 为此, 类似一些先前的非预训练持续学习方法<sup>[46]</sup>, 一些工作<sup>[148, 151]</sup>尝试直接将线性分类器替换为最近类别均值 (NCM) 分类器, 在推理阶段首先计算样本特征和 NCM 分类器中各个类别的特征中心的距离, 然后选择距离最近的类别作为预测结果. 因为预训练模型可以提供较强的特征提取能力, NCM 分类器可以获得具有代表性的类别中心用于分类, 进而取得较好的结果.

另外, 也有方法<sup>[152-153]</sup>直接通过计算线性分类器解析解来获得分类器参数, 从而解决分类器偏差. 具体而言, 给定第  $t$  个任务的样本特征和对应标签, 求解分类器参数可以看作是一个最小二乘问题. 为更好地阐述该类方法的基本思想, 假设考虑仅包含两个任务的持续学习问题. 记第  $t$  ( $t = 1, 2$ ) 个任务的样本特征  $Z^t \in \mathbf{R}^{N^t \times D}$  和对应 0-1 标签  $Y^t \in \mathbf{R}^{N^t \times C^t}$ , 模型最后的线性分类器参数记为  $W^t \in \mathbf{R}^{D \times C^t}$ , 其中,  $N^t$  表示第  $t$  个任务的样本数量,  $D$  表示特征维度,  $C^t$  表示第  $t$  个任务的类别数量, 则对于第  $t$  个任务的最优分类器的优化问题可以表示为:

$$\min_{W^t} \|Y^t - Z^t W^t\|_F^2 \quad (7)$$

其最优解为:

$$W^{t*} = ((Z^t)^T Z^t)^{-1} (Z^t)^T Y^t \quad (8)$$

此时考虑联合两个任务的数据求解最优分类

器, 记联合样本特征和联合标签分别为:

$$\begin{cases} Z = \begin{bmatrix} Z^1 \\ Z^2 \end{bmatrix} \in \mathbf{R}^{(N^1+N^2) \times D} \\ Y = \begin{bmatrix} Y^1 & \mathbf{0}_{N^1 \times C^2} \\ \mathbf{0}_{N^2 \times C^1} & Y^2 \end{bmatrix} \in \mathbf{R}^{(N^1+N^2) \times (C^1+C^2)} \end{cases} \quad (9)$$

其中,  $\mathbf{0}_{m \times n}$  表示所有元素全为 0 的  $m \times n$  维矩阵. 由此, 联合分类器最优解可以表达为如下形式:

$$W^* = (Z^T Z)^{-1} (Z^T Y) = ((Z^1)^T Z^1 + (Z^2)^T Z^2)^{-1} [(Z^1)^T Y^1, (Z^2)^T Y^2] \quad (10)$$

由此可以看出, 仅需保存历史任务 (本例中即第 1 个任务) 的特征-特征内积矩阵  $(Z^1)^T Z^1 \in \mathbf{R}^{D \times D}$  和特征-标签内积矩阵  $(Z^1)^T Y^1 \in \mathbf{R}^{D \times C^1}$ , 在新任务到来后直接计算新的特征-特征内积矩阵  $(Z^2)^T Z^2 \in \mathbf{R}^{D \times D}$  和特征-标签内积矩阵  $(Z^2)^T Y^2 \in \mathbf{R}^{D \times C^2}$ , 再通过式 (10) 的方式计算即可得到所有任务上的最优分类器参数, 从而解决分类器偏差问题.

然而该方法要求预训练模型对于先前任务数据表征不发生变化, 即要求预训练模型参数始终保持固定, 这一做法使得该方法难以通过微调等手段保证预训练模型很好地适应新任务数据分布, 从而无法获得更好的数据表征, 导致性能受限.

3) 模型结构. 除模型表征和模型偏差方法外, 应用预训练模型的持续学习方法主要集中在模型结构这一方面. 如图 4(c)、图 4(d) 所示, 这类方法本质上借鉴传统非预训练模型中扩展模型相关的方法 (如图 4(a) 所示), 其核心思想是通过添加提示 (Prompt) 或是添加额外适配器 (Adaptor) 等方式对固定的预训练模型进行参数高效微调, 并对不同的任务使用不同的微调参数以获得具有特异性的输出. 这种思路的优势在于, 可以以很小的训练代价来更新少量参数, 即使增加新的提示或适配器也不会对计算资源和存储空间有较大要求. 下面分别从提示微调和适配器这两个主要的微调策略介绍相关持续学习方法.

a) 提示微调的方法最早出现在大语言模型中, 由于预训练语言模型的参数量通常较为庞大, 为减少计算量, 研究者尝试对预训练模型的输入添加一组可学习参数, 从而影响模型对于输入的预测, 其中的可学习参数就称为提示. 添加提示的具体做法有很多种, 例如提示微调 (Prompt tuning)、前缀微调 (Prefix tuning) 等.

如图 4(c) 所示, 应用提示进行持续学习的一个

经典方法是 L2P<sup>[154]</sup>. 该方法构造一个提示池 (Prompt pool) 记为  $P = \{P_1, \dots, P_M\}$ , 其中  $M$  表示提示池包含提示的数量, 每个提示记为  $P_i$  ( $i = 1, \dots, M$ ). 对于每个样本, 从提示池中自动选择  $N$  ( $1 \leq N \leq M$ ) 个提示  $\{P_{s_1}, \dots, P_{s_N}\}$  与该样本的输入令牌 (Token)  $x_e$  合并作为预训练模型特征提取器的输入, 可以表示为:

$$x_p = [P_{s_1}; \dots; P_{s_N}; x_e] \quad (11)$$

为实现自动选择提示, 该方法采用键-查询 (Key-query) 匹配的思路. 具体而言, 为每个提示  $P_i$  构造一个键 (Key)  $K_i$ , 也是一组可学习参数; 同时将样本的输入令牌  $x_e$  直接送入预训练模型得到该样本特征并将其作为查询 (Query)  $Q(x_e)$ , 然后即可根据该样本查询与所有提示的键之间的距离选择最近的  $N$  个提示. 该方法在持续学习过程中仅需要优化提示池中的提示及其对应的键  $\{P_i, K_i\}_{i=1}^M$  和网络最后的线性分类器  $\phi$ , 其余参数均保持固定, 其待优化参数量要远小于传统的非预训练持续学习模型, 同时, 借助预训练模型的特征提取能力, 该方法可以达到远超非预训练持续学习模型的效果.

基于该提示微调持续学习方法, 后续有一大批工作<sup>[155-161]</sup> 尝试对提示的设计和选择进行改进. 例如, DualPrompt<sup>[155]</sup>、HiDePrompt<sup>[157]</sup> 等方法考虑为每个任务构造一组任务特定提示 (Task-specific prompt), 并为所有任务构造一组任务无关提示 (Task-agnostic prompt), 其中任务特定提示仅在对应任务训练期间进行优化, 并在训练结束后保持固定, 在推理阶段对测试样本通过键-查询方式选择任务特定提示, 而任务无关提示则通常在整个持续学习过程中保持不断优化. 这些方法希望任务特定提示可以学习对应任务的信息, 同时希望任务无关提示可以学习持续学习所有任务的通用信息, 结合两种提示信息进一步增强模型表现. 然而任务无关提示是持续优化的, 也可能存在遗忘问题, 并且无法保证更新后的任务无关提示不会对先前训练好的任务特定提示造成负面影响, 所以如何在未来任务未知且缺少过去任务数据的情况下学习任务无关提示仍然是值得思考的问题.

另外还有一些工作发现键-查询匹配机制在跨任务时有可能出现匹配不准确的问题, 这是因为所有提示的键都在持续更新, 有可能导致历史任务的数据选择的提示发生变化, 进而选择不准确的提示作为输入. 为此, CODAprompt<sup>[156]</sup> 采用学习加权的方式集成提示池中所有提示的信息, 以软加权的方式替代键-查询这种硬选择的方式. OnePrompt<sup>[162]</sup> 则在少样本持续学习设置下发现仅设置一个提示即



可, 无需构造提示池并从中选择提示. 另外, 还有一些方法<sup>[163-165]</sup>考虑设计生成器来针对不同任务样本生成特定提示, 从而避免提示选择的问题, 并且对于任意样本都可以获得不同的具有针对性的提示.

在多模态持续学习中, 也有很多工作尝试借鉴添加提示的思路, 在减轻遗忘的同时降低计算量. 例如, Fwd-Prompt<sup>[166]</sup>添加共享的提示; CPE-CLIP<sup>[167]</sup>和 TRIPLET<sup>[168]</sup>则为每层网络设计提示组, 每个提示组包含任务特定提示和任务共享提示. 除此之外, 多模态持续学习不仅需要考虑提示对于持续学习的作用, 还需要考虑不同模态之间提示的相互作用. 具体而言, CPE-CLIP 将视觉提示看作语言提示的函数, 在优化语言提示的同时协同调整视觉提示; TRIPLET 提出提示解耦和提示交互策略, 旨在解耦不同模态的提示以提取各个模态特征, 然后再融合不同模态信息以整体优化多模态学习.

b) 如图 4(d) 所示, 另一类方法是在固定预训练模型的部分参数上添加适配器<sup>[169-177]</sup>. 对于主干模型中的第  $l$  层网络参数  $W_l$  及其输入特征嵌入 (Embedding)  $X_l$ , 其输出可以简单表示为  $X_{l+1} = \sigma(W_l X_l)$ , 其中  $\sigma(\cdot)$  表示网络其余操作, 如激活函数等. 适配器则是在  $W_l$  的基础上额外添加一组参数  $\Delta W_l$ , 此时网络输出变为  $X_{l+1} = \sigma((W_l + \Delta W_l) X_l)$ , 通过调整适配器参数即可影响预训练模型的输出, 使其更加适应新任务的数据分布. 为进一步降低适配器参数量, 可以使用低秩分解的思路将适配器参数分解为多个参数量较小的矩阵组合. 例如 LoRA<sup>[27]</sup>使用两个低秩矩阵乘积获得适配器参数, 假设  $\Delta W_l$  是  $k \times d$  维矩阵, 则 LoRA 将其分解为  $\Delta W_l = A \cdot B$ , 其中  $A \in \mathbf{R}^{k \times r}$ ,  $B \in \mathbf{R}^{r \times d}$ ,  $r \ll \min(d, k)$  表示低秩维度, “ $\cdot$ ”表示矩阵乘法. 在微调过程中仅需更新  $A$  和  $B$  的参数, 其参数量远小于原始适配器的参数量. 针对持续学习问题, 借鉴先前非预训练持续学习模型中的专家模型思想, 可以针对不同任务设置不同的适配器参数, 并尽可能减少各个适配器之间的互相干扰, 由此实现避免遗忘的目的. 由于主干网络具有较强的特征提取能力且始终保持固定, 增加的适配器参数量通常较小, 这极大降低了训练难度并且有效提升了性能.

然而, 对每个任务都设置独立的适配器仍然会造成参数量的快速增长, 一些方法<sup>[178-184]</sup>考虑构造专家模型. 类似先前非预训练的持续学习专家模型方法, 只有在现有专家模型不足以适应当前数据分布的情况下才会引入新的专家模型, 并通过加权等方式集成多个适配器的结果以增强模型对任务数据分布的适应. 该策略可以有效降低参数量的增长,

并且相比基于传统非预训练模型的专家模型方法, 基于预训练模型的专家模型更加容易训练且能够实现更好的性能. 除此之外, 添加适配器以及专家模型的思想在多模态持续学习<sup>[185-190]</sup>中也有广泛应用并取得了较好的性能.

### 2.2.3 发展趋势

回顾模型层面预训练和非预训练持续学习模型的方法, 表面上看起来两类模型发展出不同的技术路线: 在非预训练模型阶段, 研究者们更加关注整个主干网络各个部分在持续学习训练过程中出现的各种问题并提出改进方案, 比如增强模型表征、矫正模型训练时出现的偏差以及尽可能减弱不同任务之间的干扰等; 而在预训练模型阶段, 研究者们尝试对预训练的基础模型结构进行高效微调, 进一步挖掘并增强预训练模型针对动态数据分布的特征提取和泛化能力, 例如使用微调策略使预训练模型更加适应新任务数据分布、使用高维随机映射增强判别能力以及基于预训练模型特征改善分类器偏差等. 然而实际上预训练持续学习模型继承传统非预训练模型的研究思路, 如为不同任务构造不同的微调参数以获得特定输出, 从而减轻任务之间干扰; 利用自监督等方法进一步增强模型表征; 检验分类器的偏差, 并通过 NCM 等类似方法改善模型预测结果. 这表明先前非预训练模型的研究虽然效果相对较弱, 但是关于持续学习本身的研究思路仍然值得借鉴和参考.

另外, 基于预训练的基础模型的持续学习方法依然存在一些自身的问题需要解决, 例如在基于提示方法中, 键-查询匹配不准确; 并且现有很多工作添加的任务无关提示无法保证在不影响其他任务特定提示的同时提取全局信息. 另外, 在基于适配器的方法中, 专家模型可能出现退化现象; 以及在多模态持续学习中, 如何高效利用多个模态的信息协同训练并互相促进. 除了这些问题的研究, 还有一些更具有挑战性的问题包括: 在保证泛化性的同时进一步挖掘预训练基础模型自身的特征提取的潜力; 尽可能降低专家模型等算法复杂度以促进实际应用; 如何自动适应未知数据分布并降低新知识对预训练基础模型的冲击以避免虚假遗忘<sup>[191]</sup>等问题.

## 2.3 损失/优化层面

除了数据和模型层面, 机器学习还必须考虑设计合理有效的损失和优化方法. 对于传统的面对静态场景的机器学习模型, 研究者针对不同目标开发出多种多样的损失函数和优化算法. 对于持续学习而言, 传统的损失和优化方法可能会面临严重的遗

忘问题,不足以完成持续学习的目标.为此,一大批工作尝试从损失/优化角度对持续学习模型进行改进,并且提出很多新颖且高效的算法.

### 2.3.1 损失/优化层面的非预训练持续学习模型

应用非预训练模型作为主干网络的方法需要不断优化整个主干网络的参数,该过程需要在庞大的参数空间中搜索最优解,并且很容易陷入局部极优,进而导致模型出现更加严重的遗忘,所以从损失以及优化角度进行改进是很有必要的.此类工作大致可以分为正则化、梯度对齐、损失平滑以及元持续学习等四类:

1) 正则化 (Regularization-based). 这类方法的基本思想是通过在损失函数添加正则项,从而实现机器学习模型优化进行约束,避免模型过分关注新任务知识而出现严重遗忘.一种简单且直观的方式是借鉴知识蒸馏 (Knowledge distillation)<sup>[192-193]</sup> 的思想,考虑保存上一个任务训练结束时的模型  $\{\theta^{t-1}, \phi^{t-1}\}$  来约束当前任务的模型  $\{\theta^t, \phi^t\}$  训练.以 LwF<sup>[194]</sup> 为例,将新任务数据输入当前模型和上一个任务的模型,并添加蒸馏 (Distillation) 损失,即要求二者输出的差异尽可能小,从而迫使当前模型和先前任务的模型具有近似的分布以避免遗忘,这种通过当前数据进行约束的方式也被称作数据正则.然而上一个任务模型并未见过当前任务数据,其输出未必包含准确且充分的信息,难以合理近似先前任务的表征空间.如果强行要求当前模型的输出近似旧任务模型的不准确输出,有可能损害模型对于新任务的学习,并难以保证对于旧任务预测能力的保留.

另一方面,一些工作尝试直接对模型参数变化进行约束.如果模型参数变化很小,则对旧任务的影响也会减小,从而减轻遗忘问题.然而限制模型的所有参数变化也会导致模型学习新任务的能力受到影响.受到模型剪枝思想的启发,网络中的大多数参数对于当前任务的性能是相对不重要的,仅有少量参数发挥较大的作用.所以在持续学习过程中尽量保持对于旧任务而言相对重要的参数不发生变化以减轻遗忘,同时在新任务训练时更新其他对于旧任务而言不重要的参数以保证模型学习新知识的能力.由此,一些工作考虑在持续学习过程中不断估计重要参数并约束其变化,以 EWC<sup>[195]</sup> 为例,该方法尝试使用 Hessian 阵来衡量参数重要性.为降低 Hessian 阵的计算复杂度和计算资源, EWC 忽略参数之间的相互依赖关系并采用对角 Fisher 信息阵估计网络参数的重要程度.

具体而言,考虑仅包含两个任务的持续学习问

题,在训练第 2 个任务时的 EWC 方法的损失函数可以表示为:

$$L(\theta) = L_2(\theta) + \lambda \sum_{\theta_i \in \theta} \frac{1}{2} F_i (\theta_i - \theta_{1,i}^*)^2 \quad (12)$$

其中,  $L_2(\cdot)$  表示模型在第 2 个任务数据上的训练损失;  $\lambda$  表示正则项的权重系数;  $F_i(\cdot)$  表示 Fisher 信息阵中关于参数  $\theta_i$  的元素;  $\theta_{1,i}^*$  表示在第 1 个任务上训练得到的最优参数集合. Fisher 信息阵的对角元素可以通过式 (13) 估计:

$$F_i = E_{(x,y) \sim \mathcal{D}_{train}^1} \left[ \left( \frac{\partial \log P(y|x, \theta)}{\partial \theta_i} \right)^2 \right] \quad (13)$$

其中,  $P(y|x, \theta)$  是模型对数据  $x$  和标签  $y$  的预测概率;  $E$  表示期望. 对角元素  $F_i$  越大则表明参数  $\theta_i$  越重要,在优化过程中需要施加更强的约束以减小其更新幅度;相反,不重要的参数则在新任务更新时有更大的自由度.基于 EWC,后续的一些工作<sup>[196-198]</sup> 设计更简单或更精确的方式来估计参数重要性,如 Online EWC<sup>[196]</sup> 等.这些方法可以应用于任何网络结构,且具有一定的数学解释.

然而这类方法需要通过 Fisher 信息阵等较为复杂的方法来估计参数重要性,并且在后续训练中需要保存所有历史任务的 Fisher 信息阵,其存储代价会随任务数量增加而线性增长,不利于实际应用.另外,从效果上来说,这类方法在面对具有较多任务的持续学习问题时通常表现不佳,这主要是因为持续学习过程中,旧任务的参数重要性有可能随着模型更新而发生变化,从而导致利用保存的历史任务 Fisher 信息阵估计得到的参数重要性越来越不准确,进而出现越来越严重的遗忘问题<sup>[199]</sup>.

2) 梯度对齐 (Gradient alignment). 通常神经网络的优化采用基于梯度的方法,如随机梯度下降 (Stochastic gradient descent, SGD) 或 Adam 等方法,其基本思路是希望模型参数朝着梯度下降的方向进行优化.然而在持续学习过程中,在优化新任务时的梯度有可能和先前任务的梯度方向不一致,任务之间因而会出现冲突,即新任务梯度下降而旧任务梯度却在上升,导致模型在旧任务上的性能急剧下降.基于梯度对齐的方法思路是在训练新任务时,希望新任务和旧任务的梯度尽可能对齐,迫使网络参数在新任务更新时旧任务的梯度不会上升,从而减弱此类冲突问题<sup>[200-202]</sup>.该方法思路如图 5(a)、图 5(b) 所示,  $g_1$  和  $g_2$  分别表示新旧任务梯度,若方向不一致会干扰旧任务的性能,反之则可以促进旧任务收敛.

此类方法的一个典型的实现思路是通过将新任

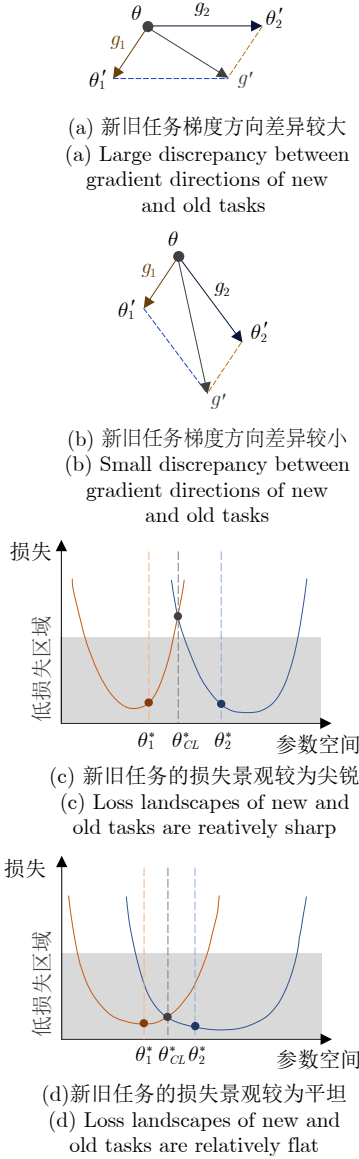


图5 损失/优化层面梯度的对齐和损失平滑图示

Fig.5 Illustration of gradient alignment and loss flatness on the perspective of loss and optimization

任务的梯度往旧任务梯度方向投影来实现梯度对齐. 以 GEM<sup>[200]</sup> 为例, 该方法在训练第  $t$  个任务时, 通过在记忆缓存中保存少量历史任务数据来估计旧任务梯度  $g_k$  ( $k < t$ ), 其目标函数可以表示为:

$$\min_g \frac{1}{2} \|g - g_t\|^2, \quad \text{s.t. } g^T g_k \geq 0, \quad \forall k < t \quad (14)$$

其中, 优化参数  $g$  表示当前更新的梯度. 该目标函数旨在令当前任务的更新方向不会增加旧任务的梯度, 并且尽可能与当前任务样本的真实梯度  $g_t$  接近. 尽管 GEM 通过二次规划求解该优化问题需要较大的计算量, 但是该思想对后续工作产生深远影响. 例如, A-GEM<sup>[201]</sup> 通过放松约束条件, 利用所有

任务的平均梯度替代式 (14) 中每个任务单独计算的历史任务梯度  $g_k$ , 由此大幅降低了计算复杂度. 然而这类方法通常对梯度的约束较为严格, 容易损害模型在新任务上的性能, 并且反复利用记忆缓存中的少量数据计算梯度作为投影方向, 导致记忆缓存数据极易过拟合, 反而损害历史任务性能.

另外, 还有一些工作<sup>[203-204]</sup> 从零空间 (Null space) 角度考虑对参数梯度优化方向进行约束, 即希望网络参数在旧任务的零空间内进行更新, 从而避免对旧任务的影响. 具体而言, 第  $l$  层的网络参数记为  $W_l$ , 以 SGD 优化算法为例, 记学习率为  $\alpha$ , 则在新任务上的一次迭代更新后的参数为  $W_l^{t+1} = W_l^t - \alpha g_{W_l}$ , 则参数变化为  $\Delta W_l = W_l^{t+1} - W_l^t = -\alpha g_{W_l}$ . 若对于旧任务数据的输入满足条件  $\Delta W_l X^{1:t-1} = 0$ , 则称梯度  $g_{W_l}$  落在旧任务的零空间内. 由此可得, 更新后的参数对于旧任务数据输入有  $W_l^{t+1} X^{1:t-1} = (W_l^t + \Delta W_l) X^{1:t-1} = W_l^t X^{1:t-1}$ , 即参数更新不会影响对于旧任务数据的预测结果, 从而避免遗忘问题. 同样地, 这类方法虽然数学上严格推导不会存在遗忘, 但由于无法获得所有旧数据, 通常还需要额外增加记忆缓存等手段来近似计算零空间. 另外, 对于新任务梯度的严格约束很容易破坏模型学习新任务的能力.

3) 损失平滑 (Loss flatness). 大多数的损失平滑方法<sup>[205-207]</sup> 核心思想是在训练期间扰动模型参数, 迫使模型对抗扰动并收敛到相对稳定的极优点, 在该极优点附近的损失景观 (Loss landscape) 较为平滑, 不会出现损失的剧烈变动, 由此增加模型的鲁棒性. 在神经网络的优化过程中, 平坦的损失景观意味着在当前模型参数处于低损失平滑区域, 对参数变化相对不敏感. 即使新任务到来导致模型朝着不同方向进行更新, 在一定程度内的参数变化依然使得模型处于关于旧任务的低损失区域, 不会过分影响旧任务的性能. 该方法思路如图 5(c)、图 5(d) 所示,  $\theta_1^*$  和  $\theta_2^*$  分别表示两个任务的最优参数,  $\theta_{CL}^*$  表示两个任务持续学习后的最优参数. 若两个任务的损失更加平滑, 则低损失区域范围更大且变化更平稳, 可以帮助持续学习模型收敛到具有更低损失的极小值. 一些工作<sup>[208-211]</sup> 尝试将现有的损失平滑方法应用于持续学习场景, 影响模型参数更新并使其收敛到相对平滑的极小值. 这些方法通常可以应用于多种持续学习基准方法, 作为一种即插即用的模块来有效提升模型性能并减轻遗忘问题.

4) 元持续学习 (Meta-continual learning). 元学习的核心是学习一个适应能力强的初始化或优化规则, 在持续学习中合理地应用元学习方法<sup>[212]</sup> 可以使得模型在面对新任务或环境时能够快速适应, 同



时尽量保留旧任务的知识,可以更加高效地应对任务切换和灾难性遗忘问题<sup>[213-218]</sup>。应用元学习的方式多种多样,通常采用的基本思路是: a) 先利用元学习隐式梯度对齐的特点,实现新旧任务的对齐以避免遗忘。这是因为元学习通常采用双层优化的方式求解,在求解中存在一个隐式的内外层损失梯度对齐的优化过程。利用这一点,内层数据使用记忆缓存数据(也可以是新任务数据和记忆缓存数据的合集),外层数据则使用新任务数据,从而隐式地实现新旧任务梯度对齐。b) 再通过设计元目标影响模型优化,如增强模型泛化性和迁移性,从而直接优化模型参数或是控制一些关键参数来间接影响模型训练。

以 MER<sup>[214]</sup> 为例,该方法通过新旧任务梯度内积的正负分别形式化迁移 (Transfer) 和干扰 (Interference), 在目标函数中显式地最小化新旧任务梯度之间的干扰,并以一阶元学习算法 Reptile 进行优化,建立 GEM 梯度对齐的目标和元学习之间的联系。再比如 La-MAML<sup>[215]</sup>,其通过元学习控制各个参数的学习率,进而隐式地控制新旧任务梯度对齐的程度:如果梯度对齐程度较高,则学习率较大;反之则说明新旧任务之间存在冲突,可以自动降低学习率以减缓模型在该批次数据上的训练,从而达到较好的收敛结果。

更进一步,最近的一项研究 VR-MCL<sup>[199]</sup> 从理论上建立元持续学习和先前基于正则化方法之间的联系,直观反映元持续学习方法对于改善遗忘问题的有效性。该方法通过元学习方法改善先前正则化方法存在参数重要性估计不准确的问题,获得更加稳定的 Hessian 阵估计,显著改善了元持续学习和基于正则化方法的性能。

另外,本团队近期研发的元持续学习框架<sup>[121, 131]</sup> 尝试学习模型在处理持续学习问题时的内在规律,并将其应用于模型泛化,在不利用预训练模型的情况下显著增强了模型的泛化性。具体而言, Dual-CBA<sup>[131]</sup> 发现每次学习新任务时,新任务的概率预测值和所有旧任务的概率预测值之间存在一定内在联系并且保持相对稳定的规律。为此设计双分支持续偏差适配器,其中的一个分支称为类别特定适配器,基于 CBA<sup>[121]</sup> 用于拟合各个类别之间在持续学习过程中出现的偏差问题;另一个分支称为类别无关适配器,用于拟合新任务和旧任务之间的偏差。借鉴元学习思想,通过元目标优化该双分支适配器,并隐式建模新旧任务的梯度对齐以避免遗忘。该框架优化得到的类别无关适配器可以有效改善新旧任务之间的偏差,并且可以直接用于多种不同的持续学习任务设置并改善其性能,展现出较强的迁移和泛化性。

## 2.3.2 损失/优化层面的预训练持续学习模型

回顾先前的非预训练模型在损失/优化层面的改进,很多方法对于计算代价有较高要求。例如基于正则方法需要估计 Hessian 阵;梯度对齐的方法需要约束所有参数的梯度方向;元持续学习方法则大多需要求解双层优化问题,其优化过程也需要计算网络参数的二阶信息等。然而预训练模型的参数量通常远远大于传统的非预训练模型,所以先前在损失/优化层面进行改进的方法难以直接应用于预训练持续学习模型。

对于预训练模型,一类可行的思路是对微调参数的更新添加约束:首先引入少量微调参数以帮助预训练基础模型适应当前数据分布,然后再对不同任务的微调参数添加正则化约束以避免互相干扰。例如,类似 EWC, EProj<sup>[219]</sup> 考虑大型多模态模型的持续学习问题,为每个任务设置一个投影层并冻结其他参数,再通过正则化方法根据不同任务的相似性作为投影层可学参数的重要性得分。InfLoRA<sup>[170]</sup> 则为每个任务设置一个 LoRA 微调模块,并通过正交性约束限制其更新方向,以保证每个任务的 LoRA 模块互不影响。这种正交性的正则策略可以将每个任务的微调参数看作是微调参数空间中的一组基底,通过对基底的组合即可获得任意表达,各个微调参数可以互相弥补信息的不足,共同作用于预训练基础模型,有效避免不同的微调参数之间互相干扰。还有一些工作采用其他正则化方法,如文献<sup>[220]</sup> 考虑对梯度施加正则,文献<sup>[221]</sup> 采用零空间的思想限制微调参数的更新。

同样,目前已有一些工作尝试将预训练基础模型和元持续学习结合,其基本思路也是对规模较小的专家网络<sup>[222]</sup> 或令牌<sup>[223]</sup> 使用元学习的策略,从而降低元学习的优化难度以满足实际需求。

## 2.3.3 发展趋势

在损失/优化这一层面的持续学习方法中,有很多基于正则的方法具有较强的数学推理作为支撑并在理论上实现较为严谨的避免遗忘,如 EWC、GEM 等。然而这类方法在实际应用中往往效果不佳,这是因为在实际应用中对网络训练施加过强的正则化约束会导致网络难以适应新任务的学习,同时也有可能过拟合到历史任务。所以如何弥补持续学习中数学推导和实际网络训练之间的差异成为这一类方法的主流问题。为此,研究者们尝试放松数学上的严格约束,例如元持续学习以隐式的梯度对齐方式避免遗忘,从而更契合网络的训练和优化,一定程度上改善了理论和模型性能的不统一。然而这一方向还有很大的进步空间,无论是理论的严谨度、

性能的稳定性以及理论和性能的平衡, 依然是未来的一个主要研究方向。

如前所述, 无论是基于正则还是元持续学习方法, 由于其计算复杂度较高, 都无法直接应用于大规模的预训练基础模型。目前的主要做法是结合模型结构方法中的添加提示或适配器的方法, 仅对少量的微调参数的持续学习更新添加正则化约束或进行元学习。然而如何从本质上降低损失/优化层面的持续学习方法的计算量和优化难度仍然是一个值得思考且极具挑战性的问题。

另外, 预训练持续学习模型的泛化性主要来源于基础模型自身的能力, 而非类似非预训练模型中元持续学习的部分方法, 通过学习持续学习内在规律获得较为本质的迁移和泛化能力。后续工作中的一个可能的发展方向就是结合预训练模型的强大表征能力与元持续学习迁移泛化的特点, 进一步捕捉机器学习方法在处理持续学习问题时表现出的一些规律, 从而进一步增强其泛化能力。

## 2.4 理论层面

一个完备的机器学习框架的主要因素不仅包括数据、模型以及损失/优化, 还需要严谨的理论框架和深入的理论分析, 这可以保障模型收敛以及模型性能, 并且对于算法理解至关重要。很多工作尝试从不同角度对持续学习模型和方法的理论发展做出贡献, 下面将进行详细的介绍和分析。

### 2.4.1 关于持续学习的理论研究进展

1) PAC-Bayesian 理论。在传统机器学习理论中, PAC-Bayesian 理论结合先验知识和数据对模型进行推断, 并通过概率方式推导出模型在测试数据上的泛化误差与其在训练数据上的经验误差之间的关系。基于此, 很多工作<sup>[138, 224-226]</sup>尝试在持续学习设置下推导模型的 PAC 泛化误差界, 并不断放松关于任务分布的假设, 以进一步考虑更加动态和复杂的场景。例如, Pentina 等<sup>[224]</sup>考虑未来任务的期望损失和已观测任务的平均损失的关系, 提出一个 PAC-Bayesian 泛化界, 并根据该泛化界提出参数迁移和表示迁移算法。该理论假设所有任务从一个固定的任务分布中采样得到, 而文献<sup>[225]</sup>进一步放松这一假设, 允许任务之间存在依赖关系, 并说明这种依赖性会导致收敛速度减慢; 另外, 还允许任务分布随时间变化, 提出使用迁移的方法从历史任务中提取知识, 并提供迁移算法的泛化界。在文献<sup>[138]</sup>中, 研究者分析任务分布之间的差异、损失景观的平坦度以及参数空间的覆盖程度, 为模型的稳定性和可塑性提供泛化误差界, 该理论指出, 找到更平坦的最优解或减小任务之间的差异有助于降

低泛化误差; 并根据理论分析提出 CoSCL 算法, 即构造多个子网络集成协作以提高稳定性和可塑性。

2) 概率模型。包括先前介绍的基于正则的方法 EWC<sup>[195]</sup> 在内, 很多工作通过建立持续学习的概率模型进而推导出一系列高效的算法<sup>[197-198, 227-228]</sup>。例如, VCL<sup>[227]</sup> 提出一个通用的变分持续学习框架, 结合在线变分推断和蒙特卡洛变分推断方法, 可以适用于判别和生成持续学习模型。Andle 等<sup>[228]</sup>分析持续学习任务序列中的网络信息流 (Information flow), 在学习新任务的同时优化网络层与层之间的信息保存, 以管理在整个层中传递的任务特定的知识, 从而实现模型对先前任务性能的保持。另外, 综述<sup>[12]</sup>也从贝叶斯角度梳理持续学习相关的理论工作, 同样对损失景观的平坦度、任务分布的差异以及参数空间的结构进行分析, 并得到和文献<sup>[138]</sup>类似的结论, 即平坦的损失有利于降低泛化误差界; 协同任务会增加泛化误差, 而竞争任务会降低泛化误差<sup>[226]</sup>; 同时也解释了固定结构的模型难以不断学习新任务的原因, 即参数空间会不断缩小, 可塑性会越来越低。

3) 线性模型。由于神经网络的高度非线性特质, 难以直接对其进行透彻的理论分析, 所以有很多工作<sup>[229-233]</sup>尝试从相对简单的线性模型出发, 考虑并分析线性分类/回归问题在持续学习场景下的参数变化及其性质。例如, Ding 等<sup>[232]</sup>定义遗忘和迁移的期望, 从而分析过参数化、任务相似性、任务顺序对持续学习的影响。Li 等<sup>[233]</sup>通过构建线性回归任务, 对混合专家模型进行理论分析, 并给出期望遗忘和泛化误差的显式表达式来表征专家模型对持续学习性能的作用。

4) 其他。还有一些其他工作借鉴其他领域的优秀理论来分析持续学习框架。例如一些方法<sup>[234-235]</sup>将持续学习中不断到来的流式任务和在线学习中的流式数据结合起来, 参考在线学习中的遗憾边界 (Regret bound) 理论进行分析, 从而开发出具有较好迁移性的方法。另外还有方法从神经正切核 (Neural tangent kernel, NTK)<sup>[236-237]</sup>、动态规划<sup>[238]</sup>、任务分解<sup>[239]</sup>、信息瓶颈理论<sup>[240]</sup>等不同角度对持续学习问题进行理论分析。

### 2.4.2 发展趋势

由此可见, 目前的大多数理论中, PAC-Bayesian 理论以及概率模型相关的工作发展较为完善, 但是仍然有较大的进步空间, 一个关键问题在于如何利用理论分析指导模型和算法的设计。例如, 很多理论工作指出, 具有较高相似度或依赖性的任务容易在持续学习过程中造成混淆, 从而对模型泛化性有负面影响<sup>[138, 226, 230, 232]</sup>, 然而这些分析大多是定



性的,任务相似性还没有较为明确的系统定义,相似任务带来的负面影响也无法定量分析.所以相关的理论研究还需要进一步讨论和探索,并且迫切需要更加严格且具有实际指导意义的理论工作.

除此之外,预训练持续学习的理论研究尚未形成完整的理论体系.尽管近期研究<sup>[241]</sup>尝试在预训练模型基础上建立理论和实际算法之间的联系,但是并未充分揭示预训练机制与持续学习动态交互的理论本质.值得注意的是,这些初步探索证实了理论分析的可行性,为后续研究提供重要启示.未来研究可以从以下两个维度突破:1)建立预训练表征的泛化性理论,明确预训练先验对持续学习任务序列的适应性边界;2)发展动态知识整合的数学框架,量化预训练参数空间与持续学习更新过程的互作用机制.通过融合深度学习理论与持续学习方法论,构建兼具解释性与实用性的理论基础.

### 3 持续学习应用及相关研究方向

在很多实际应用领域中,尽管它们看起来和持续学习毫无关系,但是实际上持续学习技术已经被广泛应用且成为这些领域中的关键组成部分.本文简要介绍以下几个现阶段较为前沿的研究方向,并简要阐述持续学习技术在其中的关键应用.

1)持续测试阶段自适应(Continual test-time adaptation, CTTA).测试阶段自适应(Test-time adaptation, TTA)<sup>[242-244]</sup>旨在解决测试数据与训练数据分布不一致的问题.与传统测试阶段固定模型不同,测试阶段自适应允许模型在训练结束后根据测试数据自动调整模型参数,无需重新训练或访问原始训练数据.常见的策略是利用无监督损失根据测试数据自动调整模型参数,从而适应测试数据的分布变化.

在此基础上结合持续学习,很多工作开始研究持续测试阶段自适应<sup>[242, 245-250]</sup>,希望模型在测试阶段会遇到随环境或时间持续动态变化的测试数据,并要求模型能够适应后续遇到的一系列测试数据,其核心挑战在于平衡新知识的适应与旧知识的保留,避免灾难性遗忘.因此这类方法也会使用持续学习相关技术,如基于重放<sup>[249-250]</sup>或约束重要参数更新<sup>[242, 246, 248]</sup>的方法.

2)模型忘却(Model unlearning).传统持续学习通常是避免在学习新知识时遗忘已经学过的旧知识,而模型忘却<sup>[243, 251-254]</sup>则是利用模型容易出现遗忘的特点,希望模型遗忘先前学过的某些特定的训练数据以保护用户的数据隐私,从而符合某些隐私规则、社会规范等.一些工作尝试实现精确忘却

(Exact unlearning)<sup>[252, 255-256]</sup>,该过程通常需要在剔除指定的训练数据后重新训练整个或部分模型.然而重新训练会耗费巨大的计算资源,所以很多研究开始探索近似忘却(Approximate unlearning)<sup>[257-259]</sup>,即希望最小化近似忘却模型与精确忘却模型之间的分布差异,再通过变分等方式求解忘却后的模型.

然而,模型在忘却指定训练数据时,很有可能同时遗忘其他训练数据,从而引发灾难性遗忘并严重损害模型泛化性.这就要求模型忘却算法需要在忘却指定样本和维持泛化性之间达到较好的平衡.为此,很多工作结合持续学习技术进行探索,比如 UNLEARN<sup>[260]</sup>、Forsaken<sup>[261]</sup>对参数更新添加约束, Srubbing procedure<sup>[258]</sup>在最大化忘却的同时还添加在其他数据上的损失以避免灾难性遗忘,  $O^3$ <sup>[262]</sup>则引入 LoRA 适配器用于持续地忘却指定数据,并设计一个分布外检测器来衡量输入和忘却数据的相似性.

3)强化学习(Reinforcement learning, RL).强化学习是指智能体通过与环境交互进行学习.具体而言,智能体在环境中采取行动,并从环境中获得观察和奖励.目标是学习一个动作(Action),即根据到目前为止所看到的观察(Observation)和奖励(Reward)来选择下一个动作,使得奖励最大化.

在长期的强化学习过程中,动作的变化会导致智能体与环境的交互发生变化,从而出现非稳态的情况,这会降低强化学习的效率.而基于重放的方法也是强化学习中的一个重要解决方案<sup>[263-264]</sup>.即在进行新的观测时,重放先前遇到的状态和动作,使得训练更加稳定.与传统持续学习不同,在强化学习中的重放策略并不关心存储资源.在重放的过程中,也会遇到和传统持续学习类似的问题,例如选择合适的重放数据<sup>[265]</sup>,过度重放会损害模型学习新任务的能力<sup>[266-267]</sup>等.

目前已有研究者尝试将持续学习在强化学习中的概念形式化,称为持续强化学习(Continual RL)<sup>[268-269]</sup>.该设置进一步强调在强化学习中的动态性,包括状态、动作、奖赏、观察等,都会随着时间的推移而发生变化.现有的工作已经应用很多持续学习技术,例如基于重放<sup>[270]</sup>、基于正则<sup>[271]</sup>、基于模型架构<sup>[272]</sup>等方法.

4)大语言模型中的持续学习.在自然语言处理(Natural language processing, NLP)领域,大语言模型(Large language model, LLM)在文本理解、对话、生成、翻译等领域都取得较强的性能<sup>[273-276]</sup>.然而大语言模型的训练极其复杂且耗费巨大,当需要整合新知识时,重新训练是不现实的.因此也会采用持续学习相关技术,希望在学习新知识的同时



不损害大语言模型原有的能力. 具体而言, 大语言模型中的持续学习可以粗略分为以下三类<sup>[29]</sup>:

- 持续预训练 (Continual pre-training, CPT). 在预训练阶段, 往往需要扩展模型对语言的基础理解, 例如学习新的事实性知识<sup>[277-278]</sup>、学习特定领域知识<sup>[279-280]</sup> 或扩展大语言模型支持的自然语言种类或代码语言种类<sup>[281]</sup> 等, 都需要在学习新知识的同时避免遗忘.

- 持续指令微调 (Continual instruction tuning, CIT). 预训练结束后, 大语言模型还需进行指令微调以适应用户的命令, 提升响应能力. 这就需要模型持续适应多种指令任务<sup>[282-283]</sup>, 适应特定领域的指令<sup>[284-285]</sup>, 或是持续扩展模型使用新工具的能力<sup>[286]</sup>.

- 持续对齐 (Continual alignment, CA). 最后还需要确保大语言模型的输出符合价值观、道德标准和社会规范, 比如持续对齐人类的价值观和偏好以及新的法律和标准规范<sup>[287-289]</sup>.

以上的设置大多从实际出发, 目的是解决大语言模型在训练和改进过程中出现的遗忘问题, 所以具体的问题设置和先前持续学习常用设置略有不同, 但是采用的技术基本思想一致, 包括基于重放和正则的方法, 或是扩展模型结构<sup>[290-292]</sup> 的策略.

## 4 结束语

持续学习作为人工智能领域的前沿方向, 旨在

赋予机器系统类人化的持续进化能力, 使其能够在动态开放环境中实现知识积累与能力延展, 该技术已经广泛应用于机器人、自动化系统、医疗保健等领域. 随着大规模预训练基础模型技术的快速发展, 持续学习研究正经历从传统非预训练范式向预训练范式的跃迁. 本文从数据、模型、损失/优化和理论四个层面系统性地梳理持续学习领域的研究脉络与发展趋势, 着重分析非预训练持续学习模型和预训练持续学习模型之间的区别与联系, 并在表 1 中整理这四个方面的非预训练和预训练持续学习方法的参考文献. 本文分析表明, 即使预训练持续学习方法逐渐展现出与传统非预训练持续学习方法不同的技术路线和发展趋势, 然而本质上预训练方法仍然极大程度地继承传统非预训练方法的设计思路和解决方案, 例如伪重放、增强模型表征、扩展模型结构等. 另一方面, 预训练持续学习方法也要面对许多严峻的挑战. 未来研究需重点突破三大方向: 1) 研发面向预训练大模型的轻量化持续学习机制, 通过参数高效微调等技术创新性地降低计算复杂度, 从而结合传统持续学习技术实现更强的性能; 2) 构建跨模态、跨任务的通用迁移框架, 考虑结合预训练基础模型和元学习理论, 尝试赋予持续学习模型自动应对持续变化的数据分布的能力, 释放持续学习模型的泛化潜能; 3) 继续探索持续学习的理论体系构建, 特别是在稳定性-可塑性平衡、知识表征动态

表 1 持续学习方法总结  
Table 1 Summary of continual learning methods

方法分类	非预训练持续学习方法	预训练持续学习方法
数据层面	数据增广: [47, 52-53]	
	基于重放	数据表征: [47-48, 54] [96-98]
	数据选择: [37-38, 46, 56-59, 61-64]	
	生成模型: [66-73]	生成预训练模型: [85-90, 93, 99-101]
模型层面	基于伪重放	合成数据集: [76-80] [95]
	特征重放: [82-83]	特征重放: [94]
	模型表征	[105-108, 110-114, 116-118] [146-150]
	模型偏差	[83, 119-131] [148, 151-153]
损失/优化层面	扩展模型: [132-139]	提示微调: [154-168]
	路径模型: [140-145]	适配器及专家模型: [169-190]
	正则化	[194-199] [170, 219]
	梯度对齐	[200-204] [220-221]
理论层面	损失平滑	[205-211]
	元持续学习	[121, 131, 199, 214-218] [222-223]
	PAC-Bayesian 理论	[138, 224-226]
	概率模型	[195, 197-198, 227-228]
	线性模型	[229-233]
	其他	[234-235, 237-240]

演化等基础理论层面亟待突破。这些突破不仅将推动持续学习技术向实用化迈进,更为构建具备终身学习能力的通用人工智能系统开辟崭新路径。

## References

- Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR, 2015. 1–14
- He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 770–778
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth  $16 \times 16$  words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929, 2020.
- Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 9992–10002
- Zhao H S, Shi J P, Qi X J, Wang X G, Jia J Y. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 6230–6239
- Zheng S X, Lu J C, Zhao H S, Zhu X T, Luo Z K, Wang Y B, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 6877–6886
- He K M, Chen X L, Xie S N, Li Y H, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 15979–15988
- Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Communications of the ACM*, 2020, **63**(11): 139–144
- Kingma D P, Welling M. Auto-encoding variational Bayes. In: Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR, 2014. 1–14
- Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 574
- de Lange M, Aljundi R, Masana M, Parisot S, Jia X, Leonardis G, et al. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(7): 3366–3385
- Wang L Y, Zhang X X, Su H, Zhu J. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(8): 5362–5383
- Zhou D W, Wang Q W, Qi Z H, Ye H J, Zhan D C, Liu Z W. Classincremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 9851–9873
- McCloskey M, Cohen N J. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 1989, **24**: 109–165
- Zhang Y, Yang Q. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022, **34**(12): 5586–5609
- Sener O, Koltun V. Multi-task learning as multi-objective optimization. In: Proceedings of the 32nd Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc. 2018. 525–536
- Hoi S C H, Sahoo D, Lu J, Zhao P L. Online learning: A comprehensive survey. *Neurocomputing*, 2021, **459**: 249–289
- Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: PMLR, 2010. 249–256
- He K M, Zhang X Y, Ren S Q, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1026–1034
- Bommasani R, Hudson D A, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv: 2108.07258, 2021.
- Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021. 8748–8763
- Zhao W X, Zhou K, Li J Y, Tang T Y, Wang X L, Hou Y P, et al. A survey of large language models. arXiv preprint arXiv: 2303.18223, 2023.
- Han Z Y, Gao C, Liu J Y, Zhang J, Zhang S Q. Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv: 2403.14608, 2024.
- Xin Y, Yang J J, Luo S Q, Zhou H D, Du J L, Liu X H, et al. Parameter-efficient fine-tuning for pre-trained vision models: A survey. arXiv preprint arXiv: 2402.02242, 2024.
- Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. 3045–3059
- Jia M L, Tang L M, Chen B C, Cardie C, Belongie S, Hariharan B, et al. Visual prompt tuning. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 709–727
- Hu E J, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, et al. LoRA: Low-rank adaptation of large language models. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022. 1–13
- Zhou D W, Sun H L, Ning J Y, Ye H J, Zhan D C. Continual learning with pre-trained models: A survey. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence. Jeju, South Korea: IJCAI, 2024. 8363–8371
- Wu T T, Luo L H, Li Y F, Pan S R, Vu T T, Haffari G. Continual learning for large language models: A survey. arXiv preprint arXiv: 2402.01364, 2024.
- Shi H Z, Xu Z H, Wang H Y, Qin W Y, Wang W Y, Wang Y B, et al. Continual learning of large language models: A comprehensive survey. arXiv preprint arXiv: 2404.16789, 2024.
- Zhang J H, Liu L, Silvén O, Pietikäinen M, Hu D W. Few-shot class-incremental learning for classification and object detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(4): 2924–2945
- Tian S S, Li L S, Li W J, Ran H, Ning X, Tiwari P. A survey on few-shot class-incremental learning. *Neural Networks*, 2024, **169**: 307–324
- Yu D Z, Zhang X N, Chen Y K, Liu A W, Zhang Y F, Yu P S, et al. Recent advances of multimodal continual learning: A comprehensive survey. arXiv preprint arXiv: 2410.0535, 2024.
- van de Ven G M, Tolias A S. Three scenarios for continual learning. arXiv preprint arXiv: 1904.07734, 2019.

- 35 Aljundi R, Kelchtermans K, Tuytelaars T. Task-free continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 11246–11255
- 36 Lee S, Ha J, Zhang D S, Kim G. A neural dirichlet process mixture model for task-free continual learning. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR, 2020. 1–22
- 37 Aljundi R, Lin M, Goujaud B, Bengio Y. Gradient based sample selection for online continual learning. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019. Article No. 1058
- 38 Bang J, Kim H, Yoo Y J, Ha J W, Choi J. Rainbow memory: Continual learning with a memory of diverse samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 8214–8223
- 39 Kim C D, Jeong J, Moon S, Kim G. Continual learning on noisy data streams via self-purified replay. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 517–527
- 40 Karim N, Khalid U, Esmaeili A, Rahnavard N. CNLL: A semi-supervised approach for continual noisy label learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: 2022. 3877–3887
- 41 Chrysakis A, Moens M F. Online continual learning from imbalanced data. In: Proceedings of the 37th International Conference on Machine Learning. Virtual Event: PMLR, 2020. 1952–1961
- 42 Kim C D, Jeong J, Kim G. Imbalanced continual learning with partitioning reservoir sampling. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 411–428
- 43 Koh H, Kim D, Ha J W, Choi J. Online continual learning on class incremental blurry task configuration with anytime inference. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022. 1–21
- 44 Ratcliff R. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 1990, **97**(2): 285–308
- 45 Robins A. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 1995, **7**(2): 123–146
- 46 Rebuffi S A, Kolesnikov A, Sperl G, Lampert C H. iCaRL: Incremental classifier and representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 5533–5542
- 47 Buzzega P, Boschini M, Porrello A, Abati D, Calderara S. Dark experience for general continual learning: A strong, simple baseline. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1335
- 48 Bellitto G, Salanitri F P, Pennisi M, Bonicelli L, Porrello A, Calderara S, et al. Saliency-driven experience replay for continual learning. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024. 1–28
- 49 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278–2324
- 50 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc., 2012. 1097–1105
- 51 Zhang H Y, Cisse M, Dauphin Y N, Lopez-Paz D. Mixup: Beyond empirical risk minimization. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018. 1–13
- 52 Buzzega P, Boschini M, Porrello A, Calderara S. Rethinking experience replay: A bag of tricks for continual learning. In: Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE, 2021. 2180–2187
- 53 Zhang Y Q, Pfahringer B, Frank E, Bifet A, Lim N J S, Jia Y Z. A simple but strong baseline for online continual learning: Repeated augmented rehearsal. In: Proceedings of the 36th Neural Information Processing Systems. New Orleans, USA: NeurIPS, 2022. 1–13
- 54 Wang L Y, Zhang X X, Yang K, Yu L H, Li C X, Hong L Q, et al. Memory replay with data compression for continual learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022. 1–25
- 55 Wallace G K. The JPEG still picture compression standard. *Communications of the ACM*, 1991, **34**(4): 30–44
- 56 Isele D, Cosgun A. Selective experience replay for lifelong learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI Press, 2018. 3302–3309
- 57 Killamsetty K, Sivasubramanian D, Ramakrishnan G, Iyer R. GLISTER: Generalization based data subset selection for efficient and robust learning. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press, 2021. 8110–8118
- 58 Yoon J, Madaan D, Yang E, Hwang S J. Online coreset selection for rehearsal-based continual learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022. 1–16
- 59 Sun S Y, Calandriello D, Hu H Y, Li A, Titsias M K. Information-theoretic online memory selection for continual learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022. 1–25
- 60 Welling M. Herding dynamical weights to learn. In: Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada: ACM, 2009. 1121–1128
- 61 Borsos Z, Mutný M, Krause A. Coresets via bilevel optimization for continual learning and streaming. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1247
- 62 Zhou X, Pi R J, Zhang W Z, Lin Y, Chen Z H, Zhang T. Probabilistic bilevel coreset selection. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 27287–27302
- 63 Hao J, Ji K Y, Liu M R. Bilevel coreset selection in continual learning: A new formulation and algorithm. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2220
- 64 Tong R L, Liu Y H, Shi J Q, Gong D. Coreset selection via reducible loss in continual learning. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: ICLR, 2025. 1–36
- 65 Verma T, Jin L Y, Zhou J, Huang J, Tan M R, Choong B C M, et al. Privacy-preserving continual learning methods for medical image classification: A comparative analysis. *Frontiers in Medicine*, 2023, **10**: Article No. 1227515
- 66 Mellado D, Saavedra C, Chabert S, Salas R. Pseudorehearsal approach for incremental learning of deep convolutional neural networks. In: Proceedings of Computational Neuroscience: First Latin American Workshop, LAWCON. Porto Alegre, Brazil: Springer International Publishing, 2017. 118–126
- 67 Shin H, Lee J K, Kim J, Kim J. Continual learning with deep



- generative replay. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 2994–3003
- 68 Wu C S, Herranz L, Liu X L, Wang W X, van de Weijer J, Raducanu B. Memory replay GANs: Learning to generate new categories without forgetting. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc., 2018. 5966–5976
- 69 Rios A, Itti L. Closed-loop memory GAN for continual learning. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: AAAI Press, 2019. 3332–3338
- 70 Xiang Y, Fu Y, Ji P, Huang H. Incremental learning using conditional adversarial networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019. 6618–6627
- 71 Wang Z, Liu L, Duan Y Q, Tao D C. Continual learning through retrieval and imagination. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press, 2022. 8594–8602
- 72 Ayub A, Wagner A R. EEC: Learning to encode and regenerate images for continual learning. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: ICLR, 2021. 1–16
- 73 Chen P H, Wei W, Hsieh C J, Dai B. Overcoming catastrophic forgetting by Bayesian generative regularization. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021. 1760–1770
- 74 Wang T Z, Zhu J Y, Torralba A, Efros A A. Dataset distillation. arXiv preprint arXiv: 1811.10959, 2018.
- 75 Lei S Y, Tao D C. A comprehensive survey of dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(1): 17–32
- 76 Wiewel F, Yang B. Condensed composite memory continual learning. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). Shenzhen, China: IEEE, 2021. 1–8
- 77 Sangermano M, Carta A, Cossu A, Bacciu D. Sample condensation in online continual learning. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). Padua, Italy: IEEE, 2022. 1–8
- 78 Gu J Y, Wang K, Jiang W, You Y. Summarizing stream data for memory-constrained online continual learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 12217–12225
- 79 Yin H X, Molchanov P, Alvarez J M, Li Z Z, Mallya A, Hoiem D, et al. Dreaming to distill: Data-free knowledge transfer via deepinversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 8712–8721
- 80 Yin H X, Mallya A, Vahdat A, Alvarez J M, Kautz J, Molchanov P. See through gradients: Image batch recovery via gradinversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE, 2021. 16332–16341
- 81 Smith J, Hsu Y C, Balloch J, Shen Y L, Jin H X, Kira Z. Always be dreaming: A new approach for data-free class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 9354–9364
- 82 Liu X L, Wu C S, Menta M, Herranz L, Raducanu B, Bagdanov A D, et al. Generative feature replay for class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE, 2020. 915–924
- 83 Iscen A, Zhang J, Lazebnik S, Schmid C. Memory-efficient incremental learning through feature adaptation. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 699–715
- 84 Smith J S, Tian J J, Halbe S, Hsu Y C, Kira Z. A closer look at rehearsal-free continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023. 2410–2420
- 85 Gao R, Liu W W. DDGR: Continual learning with deep diffusion-based generative replay. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR, 2023. 10744–10763
- 86 Smith J S, Hsu Y C, Zhang L Y, Hua T, Kira Z, Shen Y L, et al. Continual diffusion: Continual customization of text-to-image diffusion with C-LoRA. *Transactions on Machine Learning Research*, 2024, **2024**: 1–13
- 87 Jodelet Q, Liu X, Phua Y J, Murata T. Class-incremental learning using diffusion models for distillation and replay. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023. 3417–3425
- 88 Zajac M, Deja K, Kuzina A, Tomczak J M, Trzcinski T, Shkurti F, et al. Exploring continual learning of diffusion models. arXiv preprint arXiv: 2303.15342, 2023.
- 89 Masip S, Rodriguez P, Tuytelaars T, van de Ven G M. Continual learning of diffusion models with generative distillation. In: Proceedings of the 3rd Conference on Lifelong Learning Agents. Pisa, Italy: PMLR, 2025. 431–456
- 90 Cywinski B, Deja K, Trzcinski T, Twardowski B, Kucinski L. GUIDE: Guidance-based incremental learning with diffusion models. arXiv preprint arXiv: 2403.03938, 2024.
- 91 Hataya R, Bao H, Arai H. Will large-scale generative models corrupt future datasets? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023. 20498–20508
- 92 Martínez G, Watson L, Reviriego P, Hernández J A, Juárez M, Sarkar R. Towards understanding the interplay of generative artificial intelligence and the Internet. In: Proceedings of the 1st International Workshop on Epistemic Uncertainty in Artificial Intelligence. Pittsburgh, USA: Springer, 2023. 59–73
- 93 Wang M R, Michel N, Mao J F, Yamasaki T. Dealing with synthetic data contamination in online continual learning. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024. 1–28
- 94 Zuo Y K, Yao H T, Yu L, Zhuang L S, Xu C S. Hierarchical prompts for rehearsal-free continual learning. arXiv preprint arXiv: 2401.11544, 2024.
- 95 Hatamizadeh A, Yin H X, Roth H, Li W Q, Kautz J, Xu D G, et al. GradViT: Gradient inversion of vision Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 10011–10020
- 96 Cai Y L, Thomason J, Rostami M. Task-attentive Transformer architecture for continual learning of vision-and-language tasks using knowledge distillation. In: Proceedings of the Findings of the Association for Computational Linguistics (EMNLP). Singapore, Singapore: Association for Computational Linguistics, 2023. 6986–7000
- 97 Zhang X, Zhang F F, Xu C S. VQACL: A novel visual question answering continual learning setting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023. 19102–19112
- 98 Yang R, Wang S, Zhang H, Xu S Y, Guo Y H, Ye X T, et al. Knowledge decomposition and replay: A novel cross-modal image-text retrieval continual learning method. In: Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: Association for Computing Machinery, 2023. 6510–6519

- 99 Yan S P, Hong L Q, Xu H, Han J H, Tuytelaars T, Li Z F, et al. Generative negative text replay for continual vision-language pretraining. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 22–38
- 100 Lei S W, Gao D F, Wu J Z, Wang Y X, Liu W, Zhang M M, et al. Symbolic replay: Scene graph as prompt for continual learning on VQA task. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 1250–1259
- 101 Cheng S X, He C Y, Chen K L, Xu L F, Li H L, Meng F M, et al. Vision-sensor attention based continual multimodal ego-centric activity recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, South Korea: IEEE, 2024. 6300–6304
- 102 Geirhos R, Jacobsen J H, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020, **2**(11): 665–673
- 103 Wei Y J, Ye J X, Huang Z Z, Zhang J P, Shan H M. Online prototype learning for online continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023. 18718–18728
- 104 Kim D, Park D, Shin Y, Bang J, Song H, Lee J G. Adaptive shortcut debiasing for online continual learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 13122–13131
- 105 Jing L L, Tian Y L. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(11): 4037–4058
- 106 Cha H, Lee J, Shin J. Co2L: Contrastive continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 9496–9505
- 107 Gomez-Villa A, Twardowski B, Yu L, Bagdanov A D, van de Weijer J. Continually learning self-supervised representations with projected functional regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 3866–3876
- 108 Purushwalkam S, Morgado P, Gupta A. The challenges of continuous self-supervised learning. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 702–721
- 109 Yao L Y, Chu Z X, Li S Y, Li Y L, Gao J, Zhang A D. A survey on causal inference. *ACM Transactions on Knowledge Discovery From Data*, 2021, **15**(5): Article No. 74
- 110 Hu X T, Tang K H, Miao C Y, Hua X S, Zhang H W. Distilling causal effect of data in class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 3956–3965
- 111 Chu Z X, Li R P, Rathbun S, Li S. Continual causal inference with incremental observational data. In: Proceedings of the 39th International Conference on Data Engineering (ICDE). Anaheim, USA: IEEE, 2023. 3430–3439
- 112 Wang L Y, Yang K, Li C X, Hong L Q, Li Z G, Zhu J. OR-DisCo: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 5379–5388
- 113 Smith J, Balloch J, Hsu Y C, Kira Z. Memory-efficient semi-supervised continual learning: The world is its own replay buffer. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). Shenzhen, China: IEEE, 2021. 1–8
- 114 Luo Y, Wong Y K, Kankanhalli M, Zhao Q. Learning to predict gradients for semi-supervised continual learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025, **36**(2): 2593–2607
- 115 O'Reilly R C, Bhattacharyya R, Howard M D, Ketz N. Complementary learning systems. *Cognitive Science*, 2014, **38**(6): 1229–1248
- 116 Pham Q, Liu C H, Hoi S C H. DualNet: Continual learning, fast and slow. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 1234
- 117 Arani E, Sarfraz F, Zonooz B. Learning fast, learning slow: A general continual learning method based on complementary learning system. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022. 1–22
- 118 Ren X Y, Qin Y, Wang B, Cheng X Q, Jia L M. A complementary continual learning framework using incremental samples for remaining useful life prediction of machinery. *IEEE Transactions on Industrial Informatics*, 2024, **20**(12): 14330–14340
- 119 Mai Z D, Li R W, Kim H, Sanner S. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR). Nashville, USA: IEEE, 2021. 3584–3594
- 120 Rypeś G, Cygert S, Trzcinski T, Twardowski B. Task-recency bias strikes back: Adapting covariances in exemplar-free class incremental learning. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024. 63268–63289
- 121 Wang Q Z A, Wang R Z, Wu Y C, Jia X X, Meng D Y. CBA: Improving online continual learning via continual bias adaptor. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023. 19036–19046
- 122 Hou S H, Pan X Y, Loy C C, Wang Z L, Lin D H. Learning a unified classifier incrementally via rebalancing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 831–839
- 123 Ahn H, Kwak J, Lim S, Bang H, Kim H, Moon T. SS-IL: Separated softmax for incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 824–833
- 124 Wu Y, Chen Y P, Wang L J, Ye Y C, Liu Z X, Guo Y D, et al. Large scale incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 374–382
- 125 Caccia L, Aljundi R, Asadi N, Tuytelaars T, Pineau J, Belilovsky E. New insights on reducing abrupt representation change in online continual learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022. 1–27
- 126 Yu L, Twardowski B, Liu X L, Herranz L, Wang K, Cheng Y M, et al. Semantic drift compensation for class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 6980–6989
- 127 Zhu K, Zhai W, Cao Y, Luo J B, Zha Z J. Self-sustaining representation expansion for non-exemplar class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 9286–9295
- 128 Pham Q, Liu C H, Hoi S C H. Continual normalization: Rethinking batch normalization for online continual learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022. 1–20
- 129 Cha S M, Cho S, Hwang D, Hong S, Lee M, Moon T. Rebalancing batch normalization for exemplar-based class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC,

- Canada: IEEE, 2023. 20127–20136
- 130 Lyu Y L, Wang L Y, Zhang X X, Sun Z C, Su H, Zhu J, et al. Overcoming recency bias of normalization statistics in continual learning: Balance and adaptation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1108
  - 131 Wang Q Z, Wang R Z, Wu Y C, Jia X X, Zhou M H, Meng D Y. Dual-CBA: Improving online continual learning via dual continual bias adaptors from a bi-level optimization perspective. arXiv preprint arXiv: 2408.13991, 2024.
  - 132 Rusu A A, Rabinowitz N C, Desjardins G, Soyer H, Kirkpatrick J, Kavukcuoglu K, et al. Progressive neural networks. arXiv preprint arXiv: 1606.04671, 2016.
  - 133 Yan S P, Xie J W, He X M. DER: Dynamically expandable representation for class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 3013–3022
  - 134 Mallya A, Lazebnik S. PackNet: Adding multiple tasks to a single network by iterative pruning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7765–7773
  - 135 Golkar S, Kagan M, Cho K. Continual learning via neural pruning. arXiv preprint arXiv: 1903.04476, 2019.
  - 136 Yoon J, Kim S, Yang E, Hwang S J. Scalable and order-robust continual learning with additive parameter decomposition. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR, 2020. 1–15
  - 137 Hihn H, Braun D A. Mixture-of-variational-experts for continual learning. arXiv preprint arXiv: 2110.12667, 2021.
  - 138 Wang L Y, Zhang X X, Li Q, Zhu J, Zhong Y. CoSCL: Co-operation of small continual learners is stronger than a big one. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 254–271
  - 139 Zhou Y Q, Lei T, Liu H X, Du N, Huang Y P, Zhao V T, et al. Mixture-of-experts with expert choice routing. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 515
  - 140 Abati D, Tomczak J, Blankevoort T, Calderara S, Cucchiara R, Bejnordi B E. Conditional channel gated networks for task-aware continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 3930–3939
  - 141 Mallya A, Davis D, Lazebnik S. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 67–82
  - 142 Wortsman M, Ramanujan V, Liu R, Kembhavi A, Rastegari M, Yosinski J, et al. Supermasks in superposition. In: Proceedings of the 34th Annual Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2020. 15173–15184
  - 143 Kang H, Mina R J L, Madjid S R H, Yoon J, Hasegawa-Johnson M, Hwang S J, et al. Forget-free continual learning with winning subnetworks. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 10734–10750
  - 144 Kang H, Yoon J, Madjid S R H, Hwang S J, Yoo C D. On the soft-subnetwork for few-shot class incremental learning. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2023. 1–23
  - 145 Gao Q, Shan X J, Zhang Y C, Zhou F. Enhancing knowledge transfer for task incremental learning with data-free subnetwork. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2994
  - 146 Fini E, da Costa V G T, Alameda-Pineda X, Ricci E, Alahari K, Mairal J. Self-supervised models are continual learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 9611–9620
  - 147 Ye Y W, Xie Y T, Zhang J P, Chen Z Y, Wu Q, Xia Y. Continual self-supervised learning: Towards universal multi-modal medical data representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2024. 11114–11124
  - 148 McDonnell M D, Gong D, Parvaneh A, Abbasnejad E, van den Hengel A. RanPAC: Random projections and pre-trained models for continual learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2024. Article No. 526
  - 149 Zhang G W, Wang L Y, Kang G L, Cheng L, Wei Y C. SLCA: Slow learner with classifier alignment for continual learning on a pre-trained model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023. 19091–19101
  - 150 Zhang G W, Wang L Y, Kang G L, Cheng L, Wei Y C. SLCA++: Unleash the power of sequential fine-tuning for continual learning with pre-training. arXiv preprint arXiv: 2408.08295, 2024.
  - 151 He J P, Zhu F Q. Exemplar-free online continual learning. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). Bordeaux, France: IEEE, 2022. 541–545
  - 152 Zhuang H P, Weng Z Y, Wei H Z, Xie R C Z, Toh K A, Lin Z P. ACIL: Analytic class-incremental learning with absolute memorization and privacy protection. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 843
  - 153 Zhuang H P, He R, Tong K, Zeng Z Q, Chen C, Lin Z P. DS-AL: A dual-stream analytic learning for exemplar-free class-incremental learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 17237–17244
  - 154 Wang Z F, Zhang Z Z, Lee C Y, Zhang H, Sun R X, Ren X Q, et al. Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 139–149
  - 155 Wang Z F, Zhang Z Z, Ebrahimi S, Sun R X, Zhang H, Lee C Y, et al. DualPrompt: Complementary prompting for rehearsal-free continual learning. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 631–648
  - 156 Smith J S, Karlinsky L, Gutta V, Cascante-Bonilla P, Kim D, Arbelle A, et al. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023. 11909–11919
  - 157 Wang L Y, Xie J Y, Zhang X X, Huang M Y, Zhu J. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2024. Article No. 3022
  - 158 Chen H R, Wu Z X, Han X T, Jia M L, Jiang Y G. PromptFusion: Decoupling stability and plasticity for continual learning. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2023. 196–212
  - 159 Wang Y B, Huang Z W, Hong X P. S-prompts learning with pre-trained transformers: An Occam’s razor for domain incre-



- mental learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 411
- 160 Kang Z Q, Wang L Y, Zhang X X, Alahari K. Advancing prompt-based methods for replay-independent general continual learning. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: ICLR, 2025. 1–19
- 161 Gao Z, Cen J, Chang X. Consistent prompting for rehearsal-free continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: 2024. 28463–28473
- 162 Huang W C, Chen C F, Hsu H. OVOR: Oneprompt with virtual outlier regularization for rehearsal-free class-incremental learning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024. 1–20
- 163 Jung D, Han D, Bang J, Song H. Generating instance-level prompts for rehearsal-free continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023. 11813–11823
- 164 Tang Y M, Peng Y X, Zheng W S. When prompt-based incremental learning does not meet strong pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. 1706–1716
- 165 Yang C Y, Liu W T, Chen S S, Qi J Y, Zhou A M. Generating prompts in latent space for rehearsal-free continual learning. In: Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: Association for Computing Machinery, 2024. 8913–8922
- 166 Zheng J H, Ma Q L, Liu Z, Wu B Q, Feng H W. Beyond anti-forgetting: Multimodal continual instruction tuning with positive forward transfer. arXiv preprint arXiv: 2401.09181, 2024.
- 167 D'Alessandro M, Alonso A, Calabrés E, Galar M. Multimodal parameter-efficient few-shot class incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023. 3385–3395
- 168 Qian Z, Wang X, Duan X G, Qin P D, Li Y H, Zhu W W. Decouple before interact: Multi-modal prompt learning for continual visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023. 2941–2950
- 169 Li J S, Wang S K, Qian B, He Y H, Wei X, Gong Y H. Dynamic integration of task-specific adapters for class incremental learning. arXiv preprint arXiv: 2409.14983, 2024.
- 170 Liang Y S, Li W J. InfLoRA: Interference-free low-rank adaptation for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2024. 23638–23647
- 171 Zhao L L, Zhang X R, Yan K, Ding S H, Huang W R. SAFE: Slow and fast parameter-efficient tuning for continual learning with pre-trained models. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024. 1–25
- 172 Zhang X, Bai L, Yang X, Liang J Y. C-LoRA: Continual low-rank adaptation for pre-trained models. arXiv preprint arXiv: 2502.17920, 2025.
- 173 Wu Y C, Piao H M, Huang L K, Wang R Z, Li W H, Pfister H, et al. SD-LoRA: Scalable decoupled low-rank adaptation for class incremental learning. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: ICLR, 2025. 1–17
- 174 Wei X W, Li G H, Marculescu R. Online-LoRA: Task-free online continual learning via low rank adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Tucson, USA: IEEE, 2025. 6634–6645
- 175 He J P, Duan Z H, Zhu F Q. CL-LoRA: Continual low-rank adaptation for rehearsal-free class-incremental learning. arXiv preprint arXiv: 2505.24816, 2025.
- 176 Jin Y, Liu J, Chen S. Multi-LoRA continual learning based instruction tuning framework for universal information extraction. *Knowledge-Based Systems*, 2025, **308**: Article No. 112750
- 177 Liu X, Chang X B. LoRA subtraction for drift-resistant space in exemplar-free continual learning. arXiv preprint arXiv: 2503.18985, 2025.
- 178 Lu Y, Zhang S, Cheng D, Liang G, Xing Y, Wang N, et al. Training consistent mixture-of-experts-based prompt generator for continual learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 19152–19160
- 179 Le M, Nguyen A, Nguyen H, Nguyen T, Pham T, Van Ngo L, et al. Mixture of experts meets prompt-based continual learning. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024. 1–38
- 180 Jung M J, Kim J. PMoE: Progressive mixture of experts with asymmetric Transformer for continual learning. arXiv preprint arXiv: 2407.21571, 2024.
- 181 Yang S, Ali M A, Wang C L, Hu L J, Wang D. MoRAL: MoE augmented LoRA for LLMs' lifelong learning. arXiv preprint arXiv: 2402.11260, 2024.
- 182 Marouf I E, Roy S, Tartaglione E, Lathuilière S. Weighted ensemble models are strong continual learners. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2024. 306–324
- 183 Wang H Y, Lu H D, Yao L N, Gong D. Self-expansion of pre-trained models with mixture of adapters for continual learning. In: Proceedings of the 38th Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024. 1–20
- 184 Zhao H, Wang Z, Sun Q, Song K, Li Y, Hu X, et al. LLaVA-CMoE: Towards continual mixture of experts for large vision-language models. arXiv preprint arXiv: 2503.21227, 2025.
- 185 Song G, Tan X Y. Real-world cross-modal retrieval via sequential learning. *IEEE Transactions on Multimedia*, 2021, **23**: 1708–1721
- 186 Sun F C, Liu H P, Yang C, Fang B. Multimodal continual learning using online dictionary updating. *IEEE Transactions on Cognitive and Developmental Systems*, 2021, **13**(1): 171–178
- 187 Peng Y X, Qi J W, Ye Z D, Zhuo Y K. Hierarchical visual-textual knowledge distillation for life-long correlation learning. *International Journal of Computer Vision*, 2021, **129**(4): 921–941
- 188 Yu J Z, Zhuge Y Z, Zhang L, Hu P, Wang D, Lu H C, et al. Boosting continual learning of vision-language models via mixture-of-experts adapters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2024. 23219–23230
- 189 Jha S, Gong D, Yao L N. CLAP4CLIP: Continual learning with probabilistic finetuning for vision-language models. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024. 1–41
- 190 Gao Z J, Zhang X X, Xu K L, Mao X J, Wang H M. Stabilizing zero-shot prediction: A novel antidote to forgetting in continual vision-language tasks. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024. 1–27
- 191 Zheng J H, Cai X D, Qiu S J, Ma Q L. Spurious forgetting in continual learning of language models. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: ICLR, 2025. 1–66
- 192 Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv: 1503.02531, 2015.

- 193 Gou J P, Yu B S, Maybank S J, Tao D C. Knowledge distillation: A survey. *International Journal of Computer Vision*, 2021, **129**(6): 1789–1819
- 194 Li Z Z, Hoiem D. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(12): 2935–2947
- 195 Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu A A, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2017, **114**(13): 3521–3526
- 196 Huszár F. On quadratic penalties in elastic weight consolidation. arXiv preprint arXiv: 1712.03847, 2017.
- 197 Ritter H, Botev A, Barber D. Online structured Laplace approximations for overcoming catastrophic forgetting. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc., 2018. 3742–3752
- 198 Zenke F, Poole B, Ganguli S. Continual learning through synaptic intelligence. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017. 3987–3995
- 199 Wu Y C, Huang L K, Wang R Z, Meng D Y, Wei Y. Meta continual learning revisited: Implicitly enhancing online hessian approximation via variance reduction. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024. 1–30
- 200 Lopez-Paz D, Ranzato M A. Gradient episodic memory for continual learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6470–6479
- 201 Chaudhry A, Ranzato M A, Rohrbach M, Elhoseiny M. Efficient lifelong learning with A-GEM. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: ICLR, 2018. 1–20
- 202 Tang S X, Chen D P, Zhu J G, Yu S J, Ouyang W L. Layer-wise optimization by gradient decomposition for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 9629–9638
- 203 Wang S P, Li X R, Sun J, Xu Z B. Training networks in null space of feature covariance for continual learning. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 184–193
- 204 Kong Y J, Liu L, Wang Z, Tao D C. Balancing stability and plasticity through advanced null space in continual learning. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 219–236
- 205 Dinh L, Pascanu R, Bengio S, Bengio Y. Sharp minima can generalize for deep nets. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017. 1019–1028
- 206 Foret P, Kleiner A, Mobahi H, Neyshabur B. Sharpness-aware minimization for efficiently improving generalization. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: ICLR, 2021. 1–19
- 207 Liu Y, Mai S Q, Chen X N, Hsieh C H, You Y. Towards efficient and scalable sharpness-aware minimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 12350–12360
- 208 Yang E N, Shen L, Wang Z Y, Liu S W, Guo G B, Wang X W. Data augmented flatness-aware gradient projection for continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023. 5607–5616
- 209 Chen R H, Jing X Y, Wu F, Chen H W. Sharpness-aware gradient guidance for few-shot class-incremental learning. *Knowledge-Based Systems*, 2024, **299**: Article No. 112030
- 210 Yang E N, Shen L, Wang Z Y, Liu S W, Guo G B, Wang X W, et al. Revisiting flatness-aware optimization in continual learning with orthogonal gradient projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(5): 3895–3907
- 211 Bian A, Li W, Yuan H J, Yu C R, Wang M, Zhao Z X, et al. Make continual learning stronger via C-flat. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024. 7608–7630
- 212 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017. 1126–1135
- 213 Hospedales T, Antoniou A, Micaelli P, Storkey A. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(9): 5149–5169
- 214 Riemer M, Cases I, Ajemian R, Liu M, Rish I, Tu Y, et al. Learning to learn without forgetting by maximizing transfer and minimizing interference. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: ICLR, 2019. 1–31
- 215 Gupta G, Yadav K, Paull L. Look-ahead meta learning for continual learning. In: Proceedings of the 34th Annual Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2020. 11588–11598
- 216 Javed K, White M. Meta-learning representations for continual learning. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019. Article No. 163
- 217 He X, Sygnowski J, Galashov A, Rusu A A, Teh Y W, Pascanu R. Task agnostic continual learning via meta learning. In: Proceedings of the 37th International Conference on Machine Learning. Virtual Event: PMLR, 2020.
- 218 Beaulieu S, Frati L, Miconi T, Lehman J, Stanley K O, Clune J, et al. Learning to continually learn. In: Proceedings of the 24th European Conference on Artificial Intelligence. Santiago de Compostela, Spain: IOS Press, 2020. 992–1001
- 219 He J H, Guo H Y, Tang M, Wang J Q. Continual instruction tuning for large multimodal models. arXiv preprint arXiv: 2311.16206, 2023.
- 220 Qiao J Y, Zhang Z Z, Tan X, Chen C W, Qu Y Y, Peng Y, et al. Prompt gradient projection for continual learning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2023. 1–22
- 221 Lu Y, Zhang S Z, Cheng D, Xing Y H, Wang N N, Wang P, et al. Visual prompt tuning in null space for continual learning. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024. 1–24
- 222 Liu R H, Zhang J Y, Song Y Q, Zhang Y, Yang B L. Discarding the crutches: Adaptive parameter-efficient expert meta-learning for continual semantic parsing. In: Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: Association for Computational Linguistics, 2025. 3560–3578
- 223 Seo Y, Lee D, Yeo J. Train-attention: Meta-learning where to focus in continual knowledge learning. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2025. 58284–58308
- 224 Pentina A, Lampert C H. A PAC-Bayesian bound for lifelong learning. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China: PMLR, 2014. 991–999
- 225 Pentina A, Lampert C H. Lifelong learning with non-i.i.d.

- tasks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2015. 1540–1548
- 226 Ramesh R, Chaudhari P. Model zoo: A growing brain that learns continually. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022. 1–27
- 227 Nguyen C V, Li Y Z, Bui T D, Turner R E. Variational continual learning. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018. 1–18
- 228 Andle J, Yasaei S S. Theoretical understanding of the information flow on continual learning performance. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 86–101
- 229 Peng B H, Risteski A. Continual learning: A feature extraction formalization, an efficient algorithm, and barriers. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 2060
- 230 Lin S, Ju P Z, Liang Y B, Shroff N. Theory on forgetting and generalization of continual learning. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR, 2023. 21078–21100
- 231 Goldfarb D, Hand P. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In: Proceedings of the 26th International Conference on Artificial Intelligence and Statistics. Valencia, Spain: PMLR, 2023. 2975–2993
- 232 Ding M, Ji K Y, Wang D, Xu J H. Understanding forgetting in continual learning with linear regression. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. 10978–11001
- 233 Li H B, Lin S, Duan L J, Liang Y B, Shroff N B. Theory on mixture-of-experts in continual learning. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: ICLR, 2025. 1–38
- 234 Alquier P, Mai T T, Pontil M. Regret bounds for lifelong learning. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA: PMLR, 2017. 261–269
- 235 Wu Y S, Wang P A, Lu C J. Lifelong optimization with low regret. In: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. Naha, Japan: PMLR, 2019. 448–456
- 236 Jacot A, Gabriel F, Hongler C. Neural tangent kernel: Convergence and generalization in neural networks. In: Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing. Virtual Event: Association for Computing Machinery, 2018. Article No. 6
- 237 Doan T, Bannani M A, Mazouze B, Rabusseau G, Alquier P. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics. Virtual Event: PMLR, 2021. 1072–1080
- 238 Krishnan R, Balaprakash P. Formalizing the generalization-forgetting trade-off in continual learning. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 1322
- 239 Kim G, Xiao C N, Konishi T, Ke Z X, Liu B. A theoretical study on solving continual learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 366
- 240 Sun S Y, Calandriello D, Hu H Y, Li A, Titsias M. Information-theoretic online memory selection for continual learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022. 1–25
- 241 Peng L Z, Elenter J, Agterberg J, Ribeiro A, Vidal R. TSVD: Bridging theory and practice in continual learning with pre-trained models. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: ICLR, 2025. 1–47
- 242 Wang D Q, Shelhamer E, Liu S T, Olshausen B A, Darrell T. Tent: Fully test-time adaptation by entropy minimization. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: ICLR, 2021. 1–15
- 243 Wang Z Y, Yang E N, Shen L, Huang H. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(3): 1464–1483
- 244 Liang J, He R, Tan T N. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 2025, **133**(1): 31–64
- 245 Gong T, Jeong J, Kim T, Shin J, Lee S J. NOTE: Robust continual test-time adaptation against temporal correlation. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1976
- 246 Wang Q, Fink O, van Gool L, Dai D X. Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 7191–7201
- 247 Chen H R, Goldblum M, Wu Z X, Jiang Y G. Adaptive retention & correction: Test-time training for continual learning. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: ICLR, 2025. 1–13
- 248 Niu S C, Wu J X, Zhang Y F, Chen Y F, Zheng S J, Zhao P L, et al. Efficient test-time model adaptation without forgetting. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 16888–16905
- 249 Döbler M, Marsden R A, Yang B. Robust mean teacher for continual and gradual test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023. 7704–7714
- 250 Yang P N, Liang J, Cao J, He R. Auto: Adaptive outlier optimization for online test-time OOD detection. arXiv preprint arXiv: 2303.12267, 2023.
- 251 Cao Y Z, Yang J F. Towards making systems forget with machine unlearning. In: Proceedings of the IEEE Symposium on Security and Privacy. San Jose, USA: IEEE, 2015. 463–480
- 252 Bourtonle L, Chandrasekaran V, Choquette-Choo C A, Jia H R, Travers A, Zhang B W, et al. Machine unlearning. In: Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA: IEEE, 2021. 141–159
- 253 Nguyen T T, Huynh T T, Ren Z, Nguyen P L, Liew A W C, Yin H Z, et al. A survey of machine unlearning. arXiv preprint arXiv: 2209.02299, 2022.
- 254 Wang W Q, Tian Z Y, Zhang C H, Yu S. Machine unlearning: A comprehensive survey. arXiv preprint arXiv: 2405.07406, 2024.
- 255 Wu Y J, Dobriban E, Davidson S. DeltaGrad: Rapid retraining of machine learning models. In: Proceedings of the 37th International Conference on Machine Learning. Virtual Event: PMLR, 2020. 10355–10366
- 256 Sekhari A, Acharya J, Kamath G, Suresh A T. Remember what you want to forget: Algorithms for machine unlearning. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 1383
- 257 Guo C, Goldstein T, Hannun A, van der Maaten L. Certified data removal from machine learning models. In: Proceedings of



- the 37th International Conference on Machine Learning. Virtual Event: PMLR, 2020. 3832–3842
- 258 Golatkar A, Achille A, Soatto S. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 9301–9309
- 259 Nguyen Q P, Kian B, Low H, Jaillet P. Variational Bayesian unlearning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1344
- 260 Du M, Chen Z, Liu C, Oak R, Song D. Lifelong anomaly detection through unlearning. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. London, UK: Association for Computing Machinery, 2019. 1283–1297
- 261 Ma Z, Liu Y, Liu X M, Liu J, Ma J F, Ren K. Learn to forget: Machine unlearning via neuron masking. *IEEE Transactions on Dependable and Secure Computing*, 2023, **20**(4): 3194–3207
- 262 Gao C Y, Wang L X, Ding K Z, Weng C K, Wang X, Zhu Q. On large language model continual unlearning. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: ICLR, 2025. 1–30
- 263 Lin L J. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 1992, **8**(3): 293–321
- 264 Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529–533
- 265 Schaul T, Quan J, Antonoglou I, Silver D. Prioritized experience replay. In: Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico: ICLR, 2016. 1–21
- 266 Lyle C, Rowland M, Dabney W, Kwiatkowska M, Gal Y. Learning dynamics and generalization in deep reinforcement learning. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 14560–14581
- 267 Dohare S, Hernandez-Garcia J F, Lan Q F, Rahman P, Mahmood A R, Sutton R S. Loss of plasticity in deep continual learning. *Nature*, 2024, **632**(8026): 768–774
- 268 Kumar S, Marklund H, Rao A, Zhu Y F, Jeon H J, Liu Y Y, et al. Continual learning as computationally constrained reinforcement learning. arXiv preprint arXiv: 2307.04345, 2023.
- 269 Abel D, Barreto A, van Roy B, Precup D, van Hasselt H, Singh S. A definition of continual reinforcement learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2192
- 270 Daniels Z A, Raghavan A, Hostetler J, Rahman A, Sur I, Piacentino M, et al. Model-free generative replay for lifelong reinforcement learning: Application to StarCraft-2. In: Proceedings of the 1st Conference on Lifelong Learning Agents. Montréal, Canada: PMLR, 2022. 1120–1145
- 271 Igl M, Farquhar G, Luketina J, Boehmer W, Whiteson S. Transient non-stationarity and generalisation in deep reinforcement learning. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: ICLR, 2021. 1–16
- 272 Gaya J B, Doan T, Caccia L, Soulier L, Denoyer L, Raileanu R. Building a subspace of policies for scalable continual learning. In: Proceedings of the 11th International Conference of Learning Representations. Kigali, Rwanda: ICLR, 2023. 1–28
- 273 Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, et al. GPT-4 technical report. arXiv preprint arXiv: 2303.08774, 2023.
- 274 Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv: 2302.13971, 2023.
- 275 Bai J Z, Bai S, Chu Y F, Cui Z Y, Dang K, Deng X D, et al. Qwen technical report. arXiv preprint arXiv: 2309.16609, 2023.
- 276 Liu A X, Feng B, Xue B, Wang B X, Wu B C, Lu C D, et al. Deepseek-V3 technical report. arXiv preprint arXiv: 2412.19437, 2024.
- 277 Sun Y, Wang S H, Li Y K, Feng S K, Tian H, Wu H F, et al. ERNIE 2.0: A continual pre-training framework for language understanding. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020. 8968–8975
- 278 Jang J, Ye S, Yang S, Shin J, Han J, Kim G, et al. Towards continual knowledge learning of language models. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: ICLR, 2022. 1–27
- 279 Ke Z X, Shao Y J, Lin H W, Konishi T, Kim G, Liu B. Continual pre-training of language models. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2023. 1–16
- 280 Yang X J, Gao J F, Xue W X, Alexandersson E. PLLaMa: An open-source large language model for plant science. arXiv preprint arXiv: 2401.01600, 2024.
- 281 Gogoulou E, Lesort T, Boman M, Nivre J. A study of continual learning under language shift. In: Proceedings of the 27th International Conference on Text, Speech, and Dialogue. Brno, Czech Republic: Springer, 2023. 71–84
- 282 Razdaibiedina A, Mao Y B, Hou R, Khabsa M, Lewis M, Almahairi A. Progressive prompts: Continual learning for language models. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2023. 1–23
- 283 Peng B H, Tian Z T, Liu S, Yang M C, Jia J Y. Scalable language model with generalized continual learning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024. 1–23
- 284 Wang X, Zhang Y S, Chen T Z, Gao S Y, Jin S J, Yang X J, et al. TRACE: A comprehensive benchmark for continual learning in large language models. arXiv preprint arXiv: 2310.06762, 2023.
- 285 Song C Y, Han X, Zeng Z N, Li K, Chen C, Liu Z Y, et al. ConPET: Continual parameter-efficient tuning for large language models. arXiv preprint arXiv: 2309.14763, 2023.
- 286 Hao S B, Liu T Y, Wang Z, Hu Z T. ToolkenGPT: Augmenting frozen language models with massive tools via tool embeddings. In: Proceedings of the 37th Neural Information Processing Systems. New Orleans, USA: NeurIPS, 2024. 1–25
- 287 Zhang H, Gui L, Zhai Y Z, Wang H, Lei Y, Xu R F. COPF: Continual learning human preference through optimal policy fitting. arXiv preprint arXiv: 2310.15694, 2023.
- 288 Zhang H, Lei Y, Gui L, Yang M, He Y L, Wang H, et al. CPPO: Continual learning for reinforcement learning with human feedback. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: ICLR, 2024. 1–24
- 289 Suhr A, Artzi Y. Continual learning for instruction following from realtime feedback. In: Proceedings of the 37th Neural Information Processing Systems. New Orleans, USA: NeurIPS, 2024. 1–20
- 290 Wang X, Chen T Z, Ge Q M, Xia H, Bao R, Zheng R, et al. Orthogonal subspace learning for language model continual learning. In: Proceedings of the Findings of the Association for Computational Linguistics (EMNLP). Singapore, Singapore: Association for Computational Linguistics, 2023. 10658–10671
- 291 Jang J, Kim S, Ye S, Kim D, Logeswaran L, Lee M, et al. Exploring the benefits of training expert language models over in-

struction tuning. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR, 2023. 14702–14729

- 292 Qiao F L, Mahdavi M. Learn more, but bother less: Parameter efficient continual learning. In: Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2024. 1–23



**王全子昂** 西安交通大学数学与统计学院博士研究生. 2019 年获得西安交通大学学士学位. 主要研究方向为机器学习, 持续学习和半监督学习.

E-mail: [sniperwqza@stu.xjtu.edu.cn](mailto:sniperwqza@stu.xjtu.edu.cn)

(**WANG Quan-Zi-Ang** Ph.D. candidate at the School of Mathematics and Statistics, Xi'an Jiaotong University. He received his bachelor degree from Xi'an Jiaotong University in 2019. His research interest covers machine learning, continual learning and semi-supervised learning.)

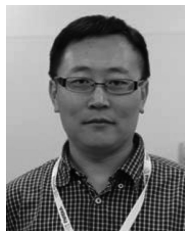


**王仁振** 西安交通大学数学与统计学院助理教授. 2022 年获得西安交通大学博士学位. 主要研究方向为半监督学习, 持续学习和医学图像分析.

E-mail: [rwang@xjtu.edu.cn](mailto:rwang@xjtu.edu.cn)

(**WANG Ren-Zhen** Assistant professor at the School of Mathematics and Statistics, Xi'an Jiaotong University. He received his Ph.D. degree from Xi'an Jiaotong University in 2022. His research interest covers semi-supervised

learning, continual learning and medical image analysis.)



**孟德宇** 西安交通大学数学与统计学院教授, 澳门科技大学特聘教授. 2008 年获得西安交通大学博士学位. 主要研究方向为机器学习, 人工智能和计算机视觉. 本文通信作者.

E-mail: [dymeng@mail.xjtu.edu.cn](mailto:dymeng@mail.xjtu.edu.cn)

(**MENG De-Yu** Professor at the School of Mathematics and Statistics, Xi'an Jiaotong University, and distinguished professor at Macau University of Science and Technology. He received his Ph.D. degree from Xi'an Jiaotong University in 2008. His research interest covers machine learning, artificial intelligence and computer vision. Corresponding author of this paper.)



**徐宗本** 中国科学院院士, 西安交通大学数学与统计学院教授. 1987 年获得西安交通大学博士学位. 主要研究方向为大数据与人工智能的数学基础与核心算法.

E-mail: [zbxu@mail.xjtu.edu.cn](mailto:zbxu@mail.xjtu.edu.cn)

(**XU Zong-Ben** Academician of the Chinese Academy of Sciences, and professor at the School of Mathematics and Statistics, Xi'an Jiaotong University. He received his Ph.D. degree from Xi'an Jiaotong University in 1987. His main research interest is mathematical foundation and core algorithms of big data and artificial intelligence.)