

深度模型的持续学习综述：理论、方法和应用

张东阳^① 陆子轩^① 刘军民^{*①} 李澜宇^{②③}

^①(西安交通大学 西安 710049)

^②(南京电子技术研究所 南京 210013)

^③(雷达探测感知全国重点实验室 南京 210039)

摘要：自然界中的生物需要在其一生中不断地学习并适应环境，这种持续学习的能力是生物学习系统的基础。尽管深度学习方法在计算机视觉和自然语言处理领域取得了重要进展，但它们在连续学习任务时面临严重的灾难性遗忘问题，即模型在学习新知识时会遗忘旧知识，这在很大程度上限制了深度学习方法的应用。持续学习研究对人工智能系统的改进和应用具有重要意义。该文对深度模型的持续学习进行了全面回顾。首先介绍了持续学习的定义和典型设定，阐述了问题的关键。其次，将现有持续学习方法划分为基于正则化、基于回放、基于梯度和基于网络结构4类，分析了各类方法的优点和局限性。同时，该文强调并总结了持续学习领域的理论分析进展，建立了理论与方法之间的联系。此外，提供了常用的数据集和评价指标，以公正评判不同方法。最后，从多个领域的应用价值出发，讨论了深度持续方法面临的问题、挑战和未来研究方向。

关键词：深度学习；持续学习；灾难性遗忘

中图分类号：TN911.7; TP181; TP183

文献标识码：A

文章编号：1009-5896(2024)10-3849-30

DOI: [10.11999/JEIT240095](https://doi.org/10.11999/JEIT240095)

A Survey of Continual Learning with Deep Networks: Theory, Method and Application

ZHANG Dongyang^① LU Zixuan^① LIU Junmin^① LI Lanyu^{②③}

^①(Xi'an Jiaotong University, Xi'an 710049, China)

^②(Nanjing Research Institute of Electronics Technology, Nanjing 210039, China)

^③(National Key Laboratory of Radar Detection and Sensing, Nanjing 210039, China)

Abstract: Biological organisms in nature are required to continuously learn from and adapt to the environment throughout their lifetime. This ongoing learning capacity serves as the fundamental basis for the biological learning systems. Despite the significant advancements in deep learning methods for computer vision and natural language processing, these models often encounter a serious issue, known as catastrophic forgetting, when learning tasks sequentially. This refers to the model's tendency to discard previously acquired knowledge when acquiring new information, which greatly hampers the practical application of deep learning models. Thus, the exploration of continual learning is paramount for enhancing and implementing artificial intelligence systems. This paper provides a comprehensive survey of continual learning with deep models. Firstly, the definition and typical settings of continual learning are introduced, followed by the key aspects of the problem. Secondly, existing methods are categorized into four main groups: regularization-based, replay-based, gradient-based and structure-based approaches, with an outline of the strengths and weaknesses of each group. Meanwhile, the paper highlights and summarizes the theoretical progress in continual learning, establishing a crucial nexus between theory and methodology. Additionally, commonly used datasets and evaluation metrics are provided to facilitate fair comparisons among these methods. Finally, the paper addresses current issues, challenges and outlines future research directions in deep continual learning, taking into account its potential applications across diverse fields.

Key words: Deep learning; Continual learning; Catastrophic forgetting

收稿日期：2024-02-22；改回日期：2024-07-18；网络出版：2024-08-28

*通信作者：刘军民 junminliu@mail.xjtu.edu.cn

基金项目：国家自然科学基金(62276208, 12326607, 11991023), 陕西省杰出青年科学基金(2024JC-JCQN-02)

Foundation Items: The National Natural Science Foundation of China (62276208, 12326607, 11991023), The Natural Science Basic Research Program of Shaanxi Province (2024JC-JCQN-02)

1 引言

学习是人类最基本的能力。近年来,深度学习方法在计算机视觉^[1-3]、自然语言处理^[4,5]、语音识别^[6]等领域取得了革命性突破,使神经网络在单一任务方面的学习能力几乎达到甚至超越了人类水平。然而,这些先进方法通常基于封闭世界假设,依赖静态数据进行反复训练,耗费时间和资源;并且训练和测试数据也往往被假定为独立同分布的^[7],一经部署,模型只能够识别在训练阶段所见过的样本,难以泛化到新的类别或新的数据分布。

现实世界是开放动态的,封闭世界假设和数据独立同分布假设在这样的背景下很难成立。模型需要适应环境的变化,从连续的信息流^[8,9]中学习知识。不仅要能够在新任务上学习知识,同时也要能够保留已有知识,这种持续获取新知识而不遗忘旧知识的能力被称为持续学习,也称为增量学习或终身学习。长期以来,如何使机器学习模型具备这种持续学习的能力一直备受关注,同时也面临着严峻的挑战。一方面,现实环境中的信息和数据呈爆炸式增长,存储所有数据来训练模型是低效耗时的;另一方面,仅使用新数据训练模型会导致灾难性遗忘^[10,11],即模型在学习新知识的过程中会遗忘已有知识,在最糟糕的情况下已有知识会被新知识完全覆盖。这些问题严重地制约了深度学习模型在实际应用中的发展。

与神经网络不同,人类和其他哺乳动物生来就擅长以持续学习的方式获取知识,这种能力是由复杂的神经认知系统和决策功能指导的。为了应对环境变化,哺乳动物的神经认知功能不断发展和进化。一些生物学工作^[12-14]发现,大脑依赖神经突触可塑性机制进化出复杂的神经认知功能,这种神经突触可塑性是大脑的一个基本特征,使得我们能够学习、记忆和适应动态环境。互补学习的理论^[15,16]指出,哺乳动物的海马体和新皮质系统相互协作,进行记忆的概括和巩固,海马体具有短期记忆功能能够快速学习知识,而新皮质系统主要负责长期记忆。海马体和新皮质系统的相互作用对于同时学习规律性和特异性是至关重要的。

这些生物学中的发现深刻影响了人工持续学习系统的研究和发展。神经突触可塑性机制启发了一系列基于正则化的方法,它们通过对网络权重施加不同程度的可塑性来保护知识,如EWC(Elastic Weight Consolidation)^[17]等。互补学习理论则影响了基于回放的方法,它们使用记忆模块来对历史知识进行存储。如Hinton等人^[18]在早期提出了一种具有双权重的计算网络,用来存储长期知识和短期知识;Kamra等人^[19]则使用深度生成模型进行记忆。

随着深度学习的快速发展,持续学习问题变得日益紧迫。神经网络通过随机梯度下降^[20]来拟合目标数据分布,这种训练方式具有数据依赖性,当数据分布发生变化时灾难性遗忘就会发生。针对这一问题,Lopez-Paz等人^[21]使用历史数据梯度对当前梯度进行约束,Zeng等人^[22]将梯度投影到历史数据的特征正交空间上,从而避免当前训练对历史任务的影响,这些作用在参数梯度上的方法被称为基于梯度的方法。此外,灾难性遗忘还可以通过为任务分配特定的资源来缓解,在训练过程中只有任务特定的参数是可学习的,而任务共享的参数被冻结,从而避免不同任务之间的相互干扰,例如Mallya等人^[23]通过剪枝技术将网络参数分配给不同任务,Yan等人^[24]为每个任务添加不同的特征提取网络,这些方法被称为基于网络结构的方法。近年来,视觉Transformer模型(Vision Transformer, ViT)^[8]和预训练模型的兴起在持续学习中引起了广泛的关注和讨论。Douillard等人^[25]利用ViT结构特性构建可扩展持续学习网络,Wang等人^[26]基于视觉预训练模型和参数高效微调(Parameter Efficient Finetuning, PEFT)^[27-32]技术,在持续学习任务上取得了优越的性能。

先前也有一些综述性工作对持续学习的研究进展进行了总结。例如,文献[33]结合动物的大脑机理整理和总结了一些受大脑启发的持续学习方法;文献[34]综合评估了持续学习中的任务增量学习问题;文献[35-37]侧重于对类别增量学习进行综述和性能评估。然而,随着持续学习领域的快速发展,许多工作日益涌现,极大地提升了任务性能。之前的研究要么聚焦于特定的方向,要么缺乏前沿领域的发展。

相比之下,本文是对深度持续学习的一次全面回顾,包括持续学习的方法、理论和应用多个方面。在方法上,本文调研了持续学习领域的最新工作进展,覆盖面广、时效性强。例如,在基于回放的方法中,本文调研了一些使用最新的图像生成模型(扩散模型)进行持续学习的方法;对于基于梯度的方法,本文将其单独列为一类并进行更细致的探讨,包含子空间投影、梯度情景记忆和平坦极小点3个子类;在基于网络结构的方法中,本文涵盖了近期发展的参数高效微调方法,该类方法借助预训练模型强大的泛化能力在持续学习任务上表现优越;这些前沿工作是文献[37]所缺乏的。在理论方面,本文综述了近年来持续学习领域的理论分析进展,旨在建立理论与方法之间的联系,促进二者的有机结合;而先前的综述^[33-37]普遍缺乏对理论工作

的总结和整理。在应用上, 本文介绍了持续学习在图像分类、检测、分割、自然语言处理、预训练模型和生成模型等任务上的应用, 包含了深度学习领域中的多个前沿问题。最后, 结合持续学习的研究现状, 本文分析了其面临的问题和挑战, 展望了未来研究方向。

本文的整体结构如下: 第1节简介持续学习的研究背景。第2节阐述持续学习的问题定义、问题分析、典型设定以及相关研究领域。第3节对持续学习的方法进行划分, 从基于正则化、基于回放、基于梯度和基于网络结构4个方面介绍持续学习的方法, 分析不同方法的区别和联系, 阐述优势和局限性。第4节总结了持续学习的理论分析进展, 建立了理论分析与方法之间的联系。第5节给出持续学习常用的实验数据集和评价指标。第6节介绍持续学习的应用。第7节分析持续学习面临的挑战, 展望未来研究方向。最后, 对全文进行总结。

2 持续学习介绍

2.1 问题定义

持续学习旨在依次地从不同的任务中学习知识, 在学习新知识的同时能够保留已有的知识, 减缓或克服灾难性遗忘问题。在持续学习的场景下, 任务数据并非是静态不变的, 而是以信息流的形式依次出现。本节假定每个任务均是分类任务, 且任务之间有明确的边界, 在此条件下给出持续学习的问题定义。

形式上, 假设有 T 个训练任务, 第 t 个任务的训练集表示为 $\mathcal{D}_t = \mathcal{X}_t \times \mathcal{Y}_t = \{(\mathbf{x}_t^i, y_t^i)\}_{i=1}^{n_t}$ 。其中, $\mathbf{x}_t^i \in \mathcal{X}_t$ 是任务 t 输入空间 \mathcal{X}_t 中的第 i 个样本, $y_t^i \in \mathcal{Y}_t$ 是其标签, n_t 是第 t 个任务的样本数量。 $\mathcal{X} = \bigcup_{t=1}^T \mathcal{X}_t$ 与 $\mathcal{Y} = \bigcup_{t=1}^T \mathcal{Y}_t$ 分别是所有任务的输入空间和标签空间。持续学习旨在让模型学习从所有任务的输入空间 \mathcal{X} 到标签空间 \mathcal{Y} 的映射 $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ 。在训练任务 t 时, 只有当前任务数据 \mathcal{D}_t 是可见的; 而在测试阶段, 模型需要在所有已见任务 $\mathcal{D}_{1:t} = \bigcup_{k=1}^t \mathcal{D}_k$ 上进行测试。

2.2 问题分析

根据问题定义, 持续学习的目标可以概括为两点。一方面, 模型需要对历史任务的知识进行保持和巩固, 称为模型的稳固性(stability); 另一方面, 模型需要从当前任务学习新知识, 称为模型的可塑性(plasticity)。持续学习的一个核心挑战是在历史任务与当前任务的性能之间进行权衡, 称为稳固性和可塑性权衡(stability-plasticity trade-off)^[38]。实验表明, 稳固性和可塑性总是此消彼长的, 在学习

任务初期, 可塑性往往比稳固性重要; 而随着所学知识的增多, 稳固性则比可塑性更加关键。此外, 由于先前任务数据在当前阶段是未知的, 持续学习的另一个挑战是对先前任务的信息进行有效的保留与估计, 根据不同的技术手段可以划分为不同的方法, 具体将在第3节中介绍。

2.3 典型设定

根据任务划分方式的不同, 持续学习可分为不同的设定。典型设定包括3种^[39]: 域增量学习、任务增量学习和类别增量学习, 见表1。3种设定下的不同任务均具有明显的分布差异, 其中, 在域增量学习设定下, 由于每个任务的类别相同, 因此模型只需在任务内的类别上进行判断; 在任务增量学习中, 尽管任务类别不同, 但由于任务标识在测试时已知, 因此模型同样只需要在任务内的类别上进行判断; 而在类别增量学习中, 由于各个任务的类别互不相同且任务标识在测试时未知, 因此模型需要同时判断样本所属的任务和具体类别, 相较前两者更加困难。

这3种设定均属于离线持续学习, 即任务边界是明确的, 当前训练阶段结束之后才会执行下一个阶段的训练; 在一些特殊场景下, 任务边界模糊不清, 称为模糊边界持续学习^[40]; 在真实场景中, 任务数据可能以连续数据流的形式呈现, 称为在线学习^[41,42]。此外, 上述3种设定中的学习任务均是标注数据的分类任务, 还有一些其他的持续学习任务, 例如, 持续目标检测^[43–46]、持续语义分割^[47–51]、持续生成模型^[52–55]、持续预训练模型^[56–58]等。

尽管主流的持续学习方法通常只在3种典型设定下进行实验评估, 但它们对于其他特殊的设定或者任务也同样适用。本文将在第6节介绍持续学习在其他任务中的应用, 在第7节分析其面临的问题和挑战。

2.4 相关研究领域

本节介绍持续学习和相关研究领域的区别, 如表2所示。

传统的监督学习通常假定训练和测试数据是独立同分布的, 模型在训练数据上学习后直接在测试集上测试。

多任务学习^[59]的目标是学习多个任务, 并实现

表1 持续学习的不同任务设定

任务设定	数据分布	任务标识
域增量学习	$p(\mathcal{X}_i) \neq p(\mathcal{X}_j), \mathcal{Y}_i = \mathcal{Y}_j, \forall i \neq j$	×
任务增量学习	$p(\mathcal{X}_i) \neq p(\mathcal{X}_j), \mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset, \forall i \neq j$	√
类别增量学习	$p(\mathcal{X}_i) \neq p(\mathcal{X}_j), \mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset, \forall i \neq j$	×

不同任务之间的相互促进。其训练数据是多个任务组成的数据集 $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_t$, 在训练阶段, 所有任务的训练数据均是可以访问的。而持续学习在训练时只能访问当前阶段的训练数据, 先前任务数据无法获取。

元学习^[60]旨在通过多个任务学习到泛化的表征, 进而适应到新的任务上。其训练数据是多个数据分布 $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_t$, 测试数据是新的数据分布 \mathcal{D}_t 。元学习在测试时只关注模型在测试集上的性能, 而持续学习关注模型在所有任务上的性能。

迁移学习^[61]旨在将知识从源域 \mathcal{D}_{src} , 迁移到目标域 \mathcal{D}_{tgt} 。模型先在源域上进行训练, 随后在目标域上进行微调适应。迁移学习和持续学习的主要区别包括: (1)迁移学习不是连续性的, 而持续学习是连续性的; (2)迁移学习旨在通过源域促进目标域的学习, 训练结束后源域性能可能下降; 而持续学习不仅仅期望提升模型在当前任务上的性能, 同时也要维持模型在历史任务上的性能。

域适应和域泛化^[62, 63]是迁移学习的两个子领域, 在域适应中, 目标域通常缺少标注信息; 域泛化则没有目标域的任何信息。而持续学习则可以获取当前任务信息。

3 持续学习方法

持续学习可以通过不同的技术思路实现。在损失层面, 添加适当的正则损失能够约束网络的训练, 从而巩固知识, 这类方法称为基于正则化的方法; 在数据层面, 回放历史任务数据可以帮助网络回忆知识, 这类方法称为基于回放的方法; 在训练算法上, 对反向传播过程中的梯度进行约束和修正能够保证当前训练不干扰历史任务, 这类方法称为基于梯度的方法; 在网络结构上, 可以为每个任务设置任务特定的网络参数或模块, 这类方法称为基于网络结构的方法。本文对现有工作按照时间和类别进行了整理和划分, 如图1所示。本节将对不同类型的方法进行细致的介绍与讨论。

表 2 持续学习与相关研究领域的区别

研究领域	训练数据	测试数据	额外限制
监督学习	$\mathcal{D}_{\text{train}}$	$\mathcal{D}_{\text{test}}$	$p(\mathcal{D}_{\text{train}}) = p(\mathcal{D}_{\text{test}})$
多任务学习	$\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_t$	$\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_t$	$p(\mathcal{D}_i) \neq p(\mathcal{D}_j), i \neq j$
元学习	$\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_{t-1}$	\mathcal{D}_t	$p(\mathcal{D}_i) \neq p(\mathcal{D}_t), i < t$
迁移学习	$\mathcal{D}_{\text{src}}, \mathcal{D}_{\text{tgt}}$	\mathcal{D}_{tgt}	$p(\mathcal{D}_{\text{src}}) \neq p(\mathcal{D}_{\text{tgt}})$
域适应	$\mathcal{D}_{\text{src}}, \mathcal{D}_{\text{tgt}}$	\mathcal{D}_{tgt}	\mathcal{D}_{tgt} 无标注信息
域泛化	\mathcal{D}_{src}	\mathcal{D}_{tgt}	\mathcal{D}_{tgt} 无法访问
持续学习	$\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_t$	$\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_t$	$\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_{t-1}$ 无法访问



图 1 持续学习方法时间线路图

需要注意的是, 一些工作结合了多种学习策略, 如iCaRL^[64]回放历史任务样本并配合知识蒸馏策略; GEM^[21]利用回放样本的梯度对当前训练梯度进行约束等。本文在对这些混合型方法进行分类时, 按照其工作重点进行划分。

3.1 基于正则化的方法

基于正则化的方法通过显式地添加正则损失, 在模型更新时强加约束, 来保留模型学习过的知识。根据不同正则化的目标, 可以分为参数正则化和数据正则化。此外, 由于训练过程的新旧数据不平衡会直接导致模型具有偏向性, 即模型偏向将数据预测为最近任务的类别, 本小节也将介绍一些任务偏向修正的方法, 作为参数正则化和数据正则化的补充。

3.1.1 参数正则化

大脑基于神经突触可塑性机制进化出复杂的神经认知功能, 基于这一思想, 参数正则化方法构建权重巩固机制, 对网络中重要的神经元连接施加保护。具体来讲, 在训练第 t 个任务时, 参数正则化的方法具有如下形式正则损失

$$\mathcal{L}_{\text{reg}} = \sum_i \Omega_{t-1}^i (\theta_t^i - \theta_{t-1}^i)^2$$

其中, θ_t^i 和 θ_{t-1}^i 分别是新模型和旧模型的第 i 个参数, Ω_{t-1}^i 为该参数对第 $t-1$ 个任务的重要程度。参数正则化的方法关键在于如何进行参数的重要性估计。Kirkpatrick等人^[17]提出弹性权重巩固(Elastic Weight Consolidation, EWC), 使用费希尔信息矩阵(Fisher Information Matrix, FIM)估计参数重要性

$$\Omega_t^i = \mathbb{E}_{(x,y) \sim \mathcal{D}_t} \left[\left(\frac{\partial \mathcal{L}}{\partial \theta_t^i} \right)^2 \right]$$

尽管可以为每个任务单独保存参数重要性矩阵, 但所需存储空间随着任务增多而线性增长。实际上, 也可以通过对所有任务进行累加惩罚^[65], 即 $\Omega_{1:t}^i = \sum_{k=1}^t \Omega_k^i$, 从而降低存储开销。此外, 对角的费希尔信息矩阵假设了网络参数之间相互独立, 这在许多情况下并不满足。R-EWC^[66]对参数空间进行旋转, 将参数梯度与坐标系对齐, 从而使得费希尔信息矩阵的对角化假设更加合理; Ritter等人^[67]使用了一种基于Kronecker分解的黑塞矩阵估计, 该方法只需假设网络不同层的参数之间是相互独立的。

除了EWC以及其变体之外, 一些工作致力于提出新的参数重要性估计。SI(Synaptic Intelligence)^[68]给出如下的参数重要性

$$\Omega_t^i = \mathbb{E}_{(x,y) \sim \mathcal{D}_t} \left[\frac{|\partial \mathcal{L} / \partial \theta_t^i|}{(\theta_t^i - \theta_{t-1}^i)^2 + \varepsilon} \right]$$

其中, ε 是一个较小的正数以保证分母不为0。MAS(Memory Aware Synapses)^[69]使用了一种无监督的估计方法, 用网络输出的L2-范数平方来估计参数重要性, 即

$$\Omega_t^i = \mathbb{E}_{(x,y) \sim \mathcal{D}_t} \left[\frac{\partial \|\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}_t)\|_2^2}{\partial \theta_t^i} \right]$$

其中, $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}_t)$ 代表了样本 \mathbf{x} 属于各个类别的输出分布。RWALK(Riemannian Walk)^[70]结合了EWC和SI的正则项损失。IMM(Incremental Moment Matching)^[71]从贝叶斯视角对先前任务和当前任务的后验进行匹配。SOS(Second Order Synapses)^[72]指出MAS和SI的估计都可以理解为费希尔信息矩阵的近似, 为不同的参数正则化方法提供了一个统一的视角。

参数正则化方法, 仅仅从模型参数空间的角度保留知识, 根据参数重要性的不同, 对其施加不同程度的惩罚。然而, 该过程中过于理想化的假设和近似会导致估计偏差较大, 因此在持续学习任务上表现略有欠缺, 特别是在类增量学习的设定下表现较差^[34,39]。

3.1.2 数据正则化

模型学习到的知识不仅仅蕴含在模型参数空间中, 更直接地体现在模型的输出结果上。与参数正则化方法不同, 数据正则化方法通过迫使新旧模型对给定数据的输出一致, 来保证知识从旧模型转移到新模型。这类方法的本质是知识蒸馏。

知识蒸馏最早由Hinton等人^[73]提出, 用于将模型的知识从教师模型转移到学生模型。将其应用在持续学习任务上时, 通常以旧模型作为教师模型而当前模型作为学生模型。按照不同的蒸馏目标, 本节将其归纳为3种蒸馏策略: 输出分布蒸馏、特征蒸馏和样本关系蒸馏, 如图2。

输出分布蒸馏是指用旧模型的输出分布作为伪标签, 指导新模型的学习。Li等人提出无遗忘学习(Learning without Forgetting, LwF)^[74], 其蒸馏损失为

$$\mathcal{L}_{\text{KD}} = - \sum_{i=1}^{|\mathcal{Y}_{t-1}|} p_i^{t-1}(\mathbf{x}) \ln p_i^t(\mathbf{x})$$

其中, $p_i^{t-1}(\mathbf{x})$ 和 $p_i^t(\mathbf{x})$ 分别代表输入样本 \mathbf{x} 在旧模型和当前模型上的输出分布, $|\mathcal{Y}_{1:t-1}|$ 表示旧类别数量。输出分布揭示了输入样本与每个类别的语义相似性, 蒸馏损失强制旧模型和新模型的语义关系相同来抵抗遗忘。iCaRL(Incremental Classifier and Representation Learning)^[64]在其基础上回放部分样本, 进一步保留已有知识。

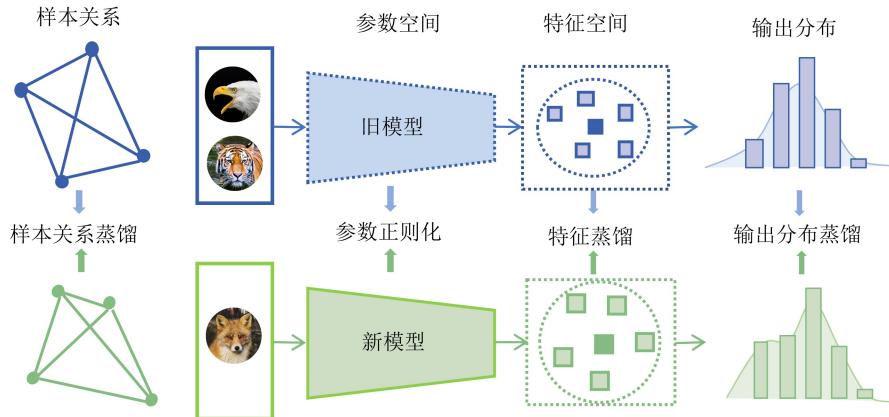


图 2 不同正则化方法

另一类做法是对模型提取的特征进行蒸馏。为了鼓励新旧模型的相似特征选取, LUCIR(Learning a Unified Classifier Incrementally via Rebalancing)^[75]对特征的余弦相似度进行惩罚

$$\mathcal{L}_{KD} = 1 - \langle \bar{\phi}^{t-1}(\mathbf{x}), \bar{\phi}^t(\mathbf{x}) \rangle$$

其中 $\bar{\phi}^{t-1}(\mathbf{x})$ 和 $\bar{\phi}^t(\mathbf{x})$ 分别表示新旧模型输出的样本 x 归一化后的特征, $\langle \cdot, \cdot \rangle$ 表示内积。LwM(Learning without Memorizing)^[76]引入注意力图来提取样本特征, 将旧类信息保存在注意力图中, 通过惩罚注意力图的变化来保留旧类的知识。AFC(Adaptive Feature Consolidation)^[77]在实施特征蒸馏时考虑了不同深度的特征重要性。POD(Pooled Outputs Distillation)^[78]在不同空间维度对网络的中间层特征进行池化, 来进行特征蒸馏。GeoDL(Geodesic-flow Distillation)^[79]沿着连接新旧特征空间的低维投影的路径来导出蒸馏损失。

还有一些工作认为, 样本之间的结构关系也保存了知识, 通过在新任务中保持样本间的结构关系也可以抵抗遗忘。Gao等人^[80]利用三元组来构造关系蒸馏损失。Tao等人^[81]通过Hebbian图对样本的相似性进行建模, 并利用Pearson相似性来惩罚顶点之间拓扑关系的变化。TOPIC(Topology-Preserving Knowledge Incrementer)^[82]用神经气体网络^[83]对样本关系进行建模, 在学习时惩罚拓扑结构中的样本特征的变化。MBP(Model Behavior Preserving)^[84]惩罚特征空间中相似性排序的变化。PRD(Prototype-sample Relation)^[85]将样本映射到一个潜在空间中, 并使用监督对比损失来学习特征表示。

此外, 由于旧的教师模型并未适应新任务样本, 因而教师模型的指导通常是不稳定的。一些工作使用更加稳定的模型作为教师模型进行知识蒸馏, 文献[86–88]使用历史模型的指数移动平均作为

教师模型, 来进行更稳定的知识蒸馏; PASS(Prototype augmentation and self-supervision)^[89]在表征上添加噪声来提高蒸馏的稳定性; TA(Teacher Adaptation)^[90]在学习新任务时, 同时更新教师模型和学生模型的批量归一化层(Batch Normalization, BN)统计量。

3.1.3 任务偏向修正

尽管参数正则化和数据正则化能够缓解模型在历史任务上的性能衰退, 然而, 模型还会遭受任务反应偏向问题, 即模型偏向于将样本识别为近期任务的类别。这主要是由于当前阶段新旧任务数据的严重不平衡导致的。使用Softmax分类器会加剧这一问题, 具体来讲, 在训练第 t 个任务时, 当前任务的损失可以分解为

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}_t; \mathbf{x}, y) = & -\ln \left(\frac{\exp(o_y)}{\sum_{j=m+1}^{m+n} \exp(o_j)} \right) \\ & - \ln \left(\frac{\sum_{j=m+1}^{m+n} \exp(o_j)}{\sum_{j=1}^{m+n} \exp(o_j)} \right) \end{aligned}$$

其中 m, n 分别是旧类别数量和新类别数量。式中第1项旨在学习样本类别, 而第2项则在学习区分新类别和旧类别从而导致任务偏向。

解决任务偏向性的一类方法是在不同的损失之间做出权衡。LODE(Loss Decoupling)^[91]在上述的两项损失中引入权衡参数来解决这一问题。BiC^[92]则在当前任务损失和蒸馏损失之间引入动态的权衡参数, 在学习任务过程中逐渐增加对旧任务的关注。

另一类方法在验证集上对输出进行修正。例如, BiC(Bias Correction)^[92]设计了偏正修正层如

$$q_k = \begin{cases} o_k, & k \leq m \\ \alpha_t o_k + \beta_t, & m < k \leq m+n \end{cases}$$

其中 α_t, β_t 是可学习的参数。WA(Weight Aligning)^[93] 将修正参数施加在新类别的权重矩阵上。通过在验证集上进行校准, 任务偏向问题能够得到缓解。然而在实际应用场景中, 模型通常无法获取验证集数据。

还有一些工作通过改进分类器来解决模型偏向问题。一种常用的分类器是最近类均值分类器^[64], 它存储每个类别的特征均值, 并采用最近邻策略进行分类。Goswami等人^[94]构建了一种特征协方差感知的最近类均值分类器, 能够更好地泛化到新类别上。Hou等人^[75]提出一种余弦分类器, 给定输入 x 模型预测其属于第 i 个类别的概率为概率为

$$p_i = \frac{\exp(\eta \langle \bar{\phi}(x), \bar{w}_i \rangle)}{\sum_j \exp(\eta \langle \bar{\phi}(x), \bar{w}_j \rangle)}$$

其中 $\bar{\phi}(x)$ 是归一化后的特征, \bar{w}_j 是归一化后的第 j 类的分类权重, η 是温度系数用来控制概率分布的峰度。Xiang等人^[95]将这种余弦分类器应用在少样本类增量学习中, 提出了一种由粗到细的学习方案, 并且证明了这种余弦分类器具有更强的稳固性。后续的许多工作也采用这种余弦归一化的分类器。Ahn等人^[96]使用分离的SoftMax和任务型的蒸馏损失。Yang等人^[97,98]受神经崩溃现象的启发, 提出等角紧结构的分类器, 能够有效地解决数据不平衡问题。Lyu等人^[99]基于贝叶斯策略动态地调整任务贡献并平衡BN层中的统计量, 从而减少任务偏向。

最后, 数据正则化的方法和任务偏向修正方法大都需要回放样本来提升任务性能, 回放数据的选取对这些方法有着至关重要的影响, 本文将在3.2节对这些回放的方法进行具体的介绍。

3.2 基于回放的方法

在人类认知体系中, 通过复习以前的知识可以克服遗忘。在持续学习的过程中, 网络也可以通过重新访问历史任务的数据来克服灾难性遗忘。然而持续学习的一个基本假定是模型在学习新任务时, 旧任务的数据不可获取。一些工作放宽了这个限制, 允许模型将旧任务的部分数据保存在内存缓存中, 本节将这一类基于保存、近似或恢复旧数据分布的方法归为基于回放的方法。

根据回放的内容和方式, 这一方法可以进一步分为3个子方向: 原始数据回放、原始特征回放、生成数据或生成特征回放。

3.2.1 原始数据回放

原始数据回放的方法直接将部分训练样本储存

在内存缓存区中, 由于存储空间有限, 该方法的挑战在于如何充分利用内存缓存区。

一些方法选取数据集的代表性样本组成核心集, 称为核心集选取^[100,101]。由于搜索性能最佳的子集是NP困难的, 早期方法大都通过启发式的搜索, 以保证所选取的子集和原始数据集之间的数据分布一致。“Herding”^[102,103]作为一种经典的贪婪策略, 它每次添加一个样本来最小化所选子集与类均值特征的距离, 即

$$s_k = \underset{x \in X}{\operatorname{argmin}} \left\| \mu - \frac{1}{k} \left(\phi(x) + \sum_{i=1}^{k-1} \phi(s_i) \right) \right\|$$

其中, μ 是类均值特征, $\phi(x)$ 是样本 x 的特征, s_i 是已挑选的样本。这一策略在后来的类别增量学习方法中得到了广泛的应用。Rebuffi等人^[64]使用该策略并结合知识蒸馏进一步缓解遗忘问题。Chaudhry等人^[104]每次挑选一个距离类均值特征最近的样本。Yoon等人^[105]提出最大化批量样本的相似性和多样性来进行核心集的选取。这些工作普遍认为应当保留最具代表性的样本, 而另一些工作则认为困难样本具有更大的保留价值, 通过回放“硬”样本, 模型将获得更高的泛化能力。例如, RWalk^[70]保留输出概率分布熵高的样本, MIR(Maximal Interfered Retrieval)^[106]则保留使损失变化最大的样本。RM(Rainbow Memory)^[107]对输入数据施加不同的数据增强来衡量样本的不确定程度, 进而保留不确定程度大的样本。

除了这些启发式方法之外, 核心集选取也可以描述为一个双层优化问题^[108], 目标是求解最佳的样本权重 w

$$\left. \begin{aligned} & \underset{\mathbf{w} \in \mathbb{R}_{+}^n, \|\mathbf{w}\|_0 \leq m}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\theta}^*(\mathbf{w})) = \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}; x_i, y_i) \\ & \text{s.t. } \boldsymbol{\theta}^*(\mathbf{w}) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n w_i \ell(\boldsymbol{\theta}; x_i, y_i) \end{aligned} \right\}$$

其中 n 和 m 分别代表全部样本量和选取样本量。该问题是一个非凸问题且通过贪婪方法求解代价高昂。Zhou等人^[109]放松了双层优化形式。Tiwari等人^[110]将内层优化问题替换为梯度近似, 所选子集的梯度应与完整数据集的梯度相近。Hao等人^[111]将样本权重限制在概率单纯形上求解双层优化问题。

实际上, 样本也可以通过优化生成。这类方法称为数据集蒸馏或数据集浓缩^[112,113], 其思想是将一个大规模的数据集浓缩为一个具有代表性的合成数据集。该方法同样可以描述为一个双层优化问题

$$\left. \begin{aligned} & S^* = \underset{S}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\theta}^*(S); \mathcal{D}) \\ & \text{s.t. } \boldsymbol{\theta}^*(S) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\theta}; S) \end{aligned} \right\}$$

其中 \mathcal{D} 是原始数据, S 是浓缩数据。Liu 等人^[114]提出 Mnemonics, 该方法实际上就是借鉴了数据集蒸馏的思想通过学习一组助记符来浓缩数据集, 其浓缩样本往往位于决策边界附近。Zhao 等人^[115,116]提出一种分布匹配的数据集蒸馏方法, 其合成样本与真实样本具有相近的特征分布。Yang 等人^[117]提出一种低秩的数据集浓缩方法 LoDC(Low-rank Dataset Consolidation), 将大型数据集压缩为低维流形上的紧凑合成数据集, 显著降低了内存消耗。如图3所示(引用自文献[117]), 浓缩后的数据可以视为原始数据集的一种压缩表示, 尽管两者在视觉上具有明显区别, 但在浓缩数据集上训练能取得在完整数据集上训练相近的结果。

还有一些工作使用其他的技术将数据进行压缩。例如, AQM(Adaptive Quantization Modules)^[118]用VQ-VAE^[119]将数据进行在线连续压缩、存储压缩数据以供新任务回放。MRDC(Memory Replay with Data Compression)^[120]通过行列式点过程^[121]压缩数据, 同时导出一种确定最佳压缩率的计算方法。Luo 等人^[122]学习图像的压缩掩码, 保留图像的关键区域, 而对其余区域进行降采样来提高存储效率。Zhai 等人^[123]则基于掩码自编码器(Mask Au-

toencoder, MAE)^[124]提出了一种双边架构, 模型同时进项表征学习和掩码重建, 在存储数据时只需要保存掩码图像, 模型能够将掩码图像重建为完整图像。

3.2.2 原始特征回放

除了直接回放数据之外, 也可以通过回放样本特征来进行知识重现。在存储效率上, 特征维度往往远小于原始样本, 而且经过特征提取器所得的特征, 一般被认为蕴含了样本的高级语义信息; 在应用方面, 保存特征而不是直接保存数据有利于保护用户隐私。这类方法面临的核心挑战是在学习新类别过程中旧类别表征偏移, 从而导致特征级的灾难性遗忘, 如图4。

为了应对这一问题, Iscen 等人^[125]设计了一个特征适应网络将保存的旧类别表征投影到新的特征空间。Belouadah 等人^[126]在数据回放的基础上保存了特征的一些统计值(如均值、协方差等)。Toldo 等人^[127]在每次任务中显式地估计表示偏移并更新保留特征。Wang 等人^[128]固定特征提取器的浅层, 并重构深层的表征。FeTriL^[129], C2F^[95]则使用从初始任务中学习到的固定特征提取器, 并结合一种伪特征生成器来回放旧类的特征。



图 3 数据集蒸馏方法示意图

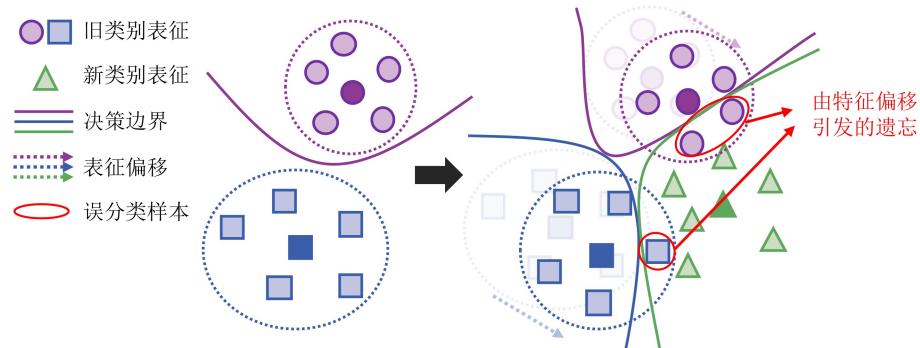


图 4 持续学习过程中旧类别表征偏移

还有一些工作保留每个类别在特征空间的代表向量, 称为类别原型。PASS^[89]将类别特征均值作为类别原型进行保存, 在学习新类别时, 对类别原型添加高斯噪声, 同时使用自监督学习策略来帮助模型生成更通用的表征。Zhu等人^[130]在保存类别均值和协方差基础上提出类别扩展, 即在原始输入空间中合成伪新类来增强模型的泛化性。SSRE(Self-sustaining Representation Expansion)^[131]通过上采样来对类别原型进行数据增强, 并且根据新类别与旧类别特征原型的相似性提出一种原型选择机制, 相似类别样本用于知识蒸馏, 不相似类别样本计算交叉熵损失。Shi等人^[132]将旧类原型与新类特征相结合, 以动态重塑旧的类特征分布。Jeeveswaran等人^[88]受人类视觉注意力机制的启发, 基于ViT架构有效回放了深层特征。

3.2.3 生成式回放

随着生成模型对图像分布的建模能力越来越强, 一些研究和工作利用生成模型, 如生成对抗网络(Generative Adversarial Network, GAN)^[133]、变分自编码器(Variational Auto-Encoders, VAE)^[134], 以及概率扩散模型(Denoising Diffusion Probabilistic Models, DDPM)^[135]等, 来对旧知识进行回放。

许多方法集中在GAN上, 因为GAN在细粒度生成上拥有优势。DGR(Deep Generative Replay)^[136]以GAN作为生成器, 利用生成的旧类别数据和新任务数据, 同时对分类器和生成器进行训练。在此基础上, MeRGANs(Memory replay GANs)^[137]对生成的数据进行了对齐, 强制新旧生成模型使用相同随机噪声进行采样来限制生成数据分布的偏移。ESGR(Exemplar-Supported Generative Reproduction)^[138]在每次增量任务都训练的单独的GAN, 导致更大的存储需求。此外, GAN在持续学习中还会遭遇到标签不匹配的问题。FearNet^[139]则利用自动编码器通过特征空间的类别均值生成旧类别的样本。L-VAE-GAN^[140]采用混合生成模型来实现更高质量的生成。

扩散模型是深度生成模型中的研究热点, 为生成式回放方法提供了新的技术手段。DDGR(Deep Diffusion-based Generative Replay)^[141]利用去噪扩散去噪模型对旧类别进行条件生成。具体来讲, 该方法利用扩散模型条件生成的特性, 在生成器和分类器之间构造了一个双向的指导: 生成器生成旧类样本帮助分类器进行训练, 而分类器对生成器的条件生成提供引导。SDDR(Stable Diffusion for Distillation and Replay)^[142]则使用预训练的文生图模型生成高质量的图像数据, 同时配合回放数据训练分类器。

由于图像数据通常具有较大的维度, 直接对其进行建模往往面临较大的挑战。相较之下, 样本特征更加容易建模和回放。Xiang等人^[143]使用条件对抗性训练的策略构建了一个生成器, 它以旧类的嵌入为条件, 生成感知卷积特征作为伪示例来回放旧类别的信息, 同时构造了一个判别器用于生成规范化嵌入, 该嵌入可以对多类别区分进行区分。Van De Ven等人^[144]受人脑启发, 利用VAE结合上下文调制反馈链接实现中间特征的回放。Liu等人^[145]探索了回放特征的最佳位置, 发现回放浅层特征会导致模型遗忘增强。

对于生成回放方法而言, 生成数据的质量直接影响分类器的性能。一方面, 高维数据可能导致生成失败或者生成质量较差; 另一方面, 生成数据的重复使用还会对生成模型的更新产生迭代影响, 进而导致生成数据偏移旧数据分布。此外, 生成模型还存在着灾难性遗忘问题, 即在持续学习的过程中, 生成质量往往逐渐恶化。为解决生成模型的遗忘问题, 一些研究者提出了持续的生成模型, 如LifelongGAN^[52], PiggybackGAN^[54]等。

3.3 基于梯度的方法

除了显式地在损失函数上添加正则项, 或使用历史数据来约束参数更新之外, 也可以直接对反向传播过程中的梯度进行修正, 从而缓解灾难性遗忘问题。本节将介绍这些直接作用于梯度的方法。根据不同的技术手段, 可以进一步细分为梯度情景记忆、子空间投影以平坦极小点。

3.3.1 梯度情景记忆

梯度情景记忆(Gradient Episodic Memory, GEM)^[21]基于回放样本的梯度构建约束, 以确保历史任务损失不会增加, 如图5。具体来讲, 该方法通过求解以下问题, 将训练阶段的参数梯度 \mathbf{g} 修正为 \mathbf{g}'

$$\begin{aligned} \min_{\mathbf{g}'} & \|\mathbf{g} - \mathbf{g}'\|_2^2 \\ \text{s.t. } & \langle \mathbf{g}', \mathbf{g}_{t-1} \rangle \geq 0 \end{aligned} \quad \left. \right\}$$

其中 $\mathbf{g} = \nabla_{\theta} \mathcal{L}(\theta, \mathcal{D}_t)$ 是当前任务数据 \mathcal{D}_t 的梯度, $\mathbf{g}_{t-1} = \nabla_{\theta} \mathcal{L}(\theta, \mathcal{M}_{1:t-1})$ 是历史数据 $\mathcal{M}_{1:t-1}$ 的梯度。约束条件保证了参数更新不会增加历史任务的损

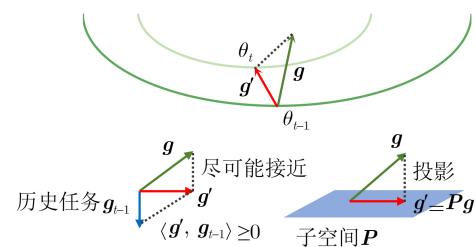


图 5 基于梯度的方法示意图

失；而优化目标迫使修正后的梯度 \mathbf{g}' 位于 \mathbf{g} 附近，来保证当前阶段训练目标是能够被优化的。该问题是一个二次规划问题具有明确的解析解

$$\mathbf{g}' = \mathbf{g} - \frac{\mathbf{g}^T \mathbf{g}_{t-1}}{\mathbf{g}_{t-1}^T \mathbf{g}_{t-1}} \mathbf{g}_{t-1}$$

该方法使用所有回放样本计算梯度，严重滞缓了训练进程。A-GEM^[146]提出从回放样本中抽样来提高训练效率。MER(Meta-Experience Replay)^[147]通过元学习促进任务迁移并减少任务干扰。OGD(Orthogonal Gradient Descent)^[148]将梯度投影到历史任务梯度的正交方向上。LOGD(Layerwise Optimization by Gradient Decomposition)^[149]将每个任务梯度分解为任务共享和任务特定两部分，以充分利用任务间的信息。

梯度情景记忆可以视为一类特殊的回放的方法，通过计算回放样本的梯度，来保证当前训练不干扰先前任务性能。该方法具有回放法的局限性：需要额外存储空间来保留样本，并且使用部分样本估计完整数据集存在偏差。

3.3.2 子空间投影

子空间投影法将梯度投影到某个子空间中，如图5右。投影矩阵可以蕴含历史任务信息从而保证当前阶段的训练不会干扰先前的任务，因此方法关键在于如何合理地设计与构建投影矩阵。

一类方法从优化的视角出发推导参数更新方式。NCL(Natural Continual Learning)^[150]将目标函数的优化问题限制在半径 r 的范围内，从而得到参数的更新方向为

$$\Delta\boldsymbol{\theta} = r\boldsymbol{\Lambda}_{k-1}^{-1}\mathbf{g} - r(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$$

其中 $\boldsymbol{\Lambda}_{k-1}$ 是历史任务损失的黑塞矩阵， $\boldsymbol{\theta}^*$ 是历史模型参数， r 是学习率。上式第1项表示用投影矩阵 $\boldsymbol{P}^{\text{NCL}} = -\boldsymbol{\Lambda}_{k-1}^{-1}$ 对梯度进行投影，以保证不会干扰历史任务，第2项可以视为隐式的L2范数正则来保证新模型参数 $\boldsymbol{\theta}$ 位于旧模型参数 $\boldsymbol{\theta}^*$ 附近。Liu等人^[151]提出RGO(Recursive Gradient Optimization)，通过迭代过程优化所有任务损失的上界，得到的投影矩阵为 $\boldsymbol{P}^{\text{RGO}} = -\tau\boldsymbol{\Lambda}_{k-1}^{-1}$ 。这两种方法都可以视为隐式的参数正则化方法，无需回放样本但需要计算黑塞矩阵，因而具有参数正则化方法的局限性：由于计算过程中的假设和近似，这类方法在类别增量的设定下表现略有欠缺。

另一类方法将梯度投影到历史任务特征空间的正交空间上。这类方法将神经网络视为一系列线性映射和非线性激活的组合，设 $\mathbf{X}_l^t \in \mathbb{R}^{n \times d_l}$ 是任务 t 在网络第 l 层的输入特征， $\mathbf{W}_l \in \mathbb{R}^{d_l \times d_{l+1}}$ 是该层的

参数矩阵。若投影矩阵 \boldsymbol{P}_l 位于先前任务特征 \mathbf{X}_l^{t-1} 的正交空间上，即 $\mathbf{X}_l^{t-1} \boldsymbol{P}_l \approx 0$ ，那么将梯度投影后，当前任务的参数更新不会影响先前任务，即

$$\mathbf{X}_l^{t-1}(\mathbf{W}_l + \Delta\mathbf{W}_l) = \mathbf{X}_l^{t-1}(\mathbf{W}_l - \eta\boldsymbol{P}_l\mathbf{g}) \approx \mathbf{X}_l^{t-1}\mathbf{W}_l$$

于是模型的稳固性得到了保证。尽管出发点都是将梯度投影到特征正交空间，但不同方法构造和计算投影矩阵的方式略有区别。OWM^[22]的投影矩阵具有形式

$$\boldsymbol{P}_l^{\text{OWM}} = \mathbf{I} - \mathbf{X}_l^T (\mathbf{X}_l \mathbf{X}_l^T + \alpha \mathbf{I})^{-1} \mathbf{X}_l$$

其中 α 为较小的正数来保证矩阵逆运算的稳定性。GPM^[152]的投影矩阵为

$$\boldsymbol{P}_l^{\text{GPM}} = \mathbf{I} - \widehat{\mathbf{X}}_l^T \widehat{\mathbf{X}}_l$$

其中 $\widehat{\mathbf{X}}_l$ 是 \mathbf{X}_l 经过SVD分解后保留的主成分。Adam-NSCL(Adam Null Space Continual Learning)^[153]利用特征空间和协方差特征空间具有相同正交空间这一性质，将梯度投影到特征协方差空间的正交空间。AdNS(Advanced Null Space)^[154]在其基础上考虑了前一个正交空间和当前正交空间的共享部分。TRGR(Trust Region Gradient Projection)^[155]在信任区域进行梯度投影，以促进相关任务带来的正向知识迁移。Lin等人^[156]在此基础上，选择性地对旧任务进行修改。

这些方法通过网络特征空间建模历史任务信息，约束参数更新的方向，从而使得模型具有持续学习的能力。不同于梯度情景记忆的方法需要存储历史任务的样本，该类方法需要存储历史任务的特征空间或协方差特征空间，且该空间通常也是随着任务而迭代更新的。

3.3.3 平坦极小点

除了利用历史信息来约束网络的训练之外，提升模型的泛化性能也对持续学习任务有直接帮助。近年来的理论分析和实验观察表明，损失景观的平坦程度^[157, 158]直接地影响了模型的鲁棒性和泛化性，进而影响模型的性能。对于持续学习而言，损失景观的平坦度对新旧任务性能也有着至关重要的影响。形式上，在训练完第 j 个任务后，模型对于第 i 个任务性能退化具有上界

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}_j^*, \mathcal{D}_i) - \mathcal{L}(\boldsymbol{\theta}_i^*, \mathcal{D}_i) &\approx \frac{1}{2} (\boldsymbol{\theta}_i^* - \boldsymbol{\theta}_j^*)^T \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_i) |_{\boldsymbol{\theta}=\boldsymbol{\theta}_i^*} \\ &\quad \cdot (\boldsymbol{\theta}_i^* - \boldsymbol{\theta}_j^*) \\ &\leq \frac{1}{2} \lambda_i^{\max} \|\boldsymbol{\theta}_i^* - \boldsymbol{\theta}_j^*\|_2^2 \end{aligned}$$

其中， λ_i^{\max} 是黑塞矩阵 $\nabla_{\boldsymbol{\theta}}^2 \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_i) |_{\boldsymbol{\theta}=\boldsymbol{\theta}_i^*}$ 的最大特征值，反映损失景观的平坦程度， λ_i^{\max} 越小意味着损

失景观越平坦, 任务性能退化的上界越紧。然而直接优化 λ_i^{\max} 比较困难, 一种可替的方案是锐度感知最小化(Sharpness Aware Minimization, SAM)^[159]

$$\operatorname{argmin}_{\theta} \mathcal{L}_s(\theta) + \mathcal{L}(\theta) = \operatorname{argmin}_{\theta} \max_{\|\delta\|_2 \leq \rho} \mathcal{L}(\theta + \delta)$$

其中, 内层优化问题可以用1阶泰勒展开近似, 近似解为 $\delta^* = \rho \nabla_{\theta} \mathcal{L}(\theta) / \|\nabla_{\theta} \mathcal{L}(\theta)\|_2$, 因而该算法可以理解为梯度先上升后下降。文献[160–162]将这种优化算法应用到持续学习任务中, 以获得平坦的局部极小值点来改善模型性能。相关的工作围绕SAM的计算效率^[163]或估计精度^[164]问题展开, 这些方法有望在持续学习任务上得到更进一步的应用。

此外, 使用额外的训练技巧也可以提升模型泛化性。例如, 使用数据重采样、数据变换和增强等技术手段能够对样本进行扩充, 或者采用自监督策略进行学习更加泛化的表征, 如文献[165–167]。

3.4 基于网络结构的方法

基于正则化、回放和梯度的方法均在共享的参数空间中进行学习, 而基于网络结构的方法为每个增量任务构建单独的参数空间进行学习, 在测试时只激活任务特定的神经元、参数或网络分支。由于不同任务的参数互相隔离, 该类方法在进行推理时需要先进行任务标识预测, 即判断输入样本属于哪个任务, 然后再调用该任务对应的参数和模块进行预测。

根据网络结构是否固定, 该方法可以细分为静态结构和动态结构两类, 如图6。

值得注意的是, 近年来的大规模预训练模型, 特别是基于ViT的预训练模型, 引起了各界的广泛关注, 这些模型往往使用大量进行有监督或自监督预训练, 能够提供丰富的先验知识从而改善和促进下游任务。基于参数高效微调的方法充分利用大规

模预训练模型强大的泛化能力, 在持续学习任务上取得了卓越的效果, 本节将其单独列为一小节进行介绍和讨论。

3.4.1 静态结构

在保持网络整体架构不变的情况下, 静态结构方法将网络的参数分配给每个任务。这等同于学习网络掩码或子网络搜索, 可以通过不同的技术手段实现。一类常用的技术是剪枝, PackNet^[23]根据权重绝对值大小对网络参数进行非结构化剪枝, 来学习有效的神经元连接从而得到任务特定掩码。UCL(Uncertainty-based Continual Learning)^[168]根据不确定性度量对贝叶斯网络进行剪枝。NISPA(Neuro-Inspired Stability-Plasticity Adaptation)^[169]使用神经元激活值衡量重要性来实施剪枝。

除了剪枝外, 掩码也可以通过直接优化获得。Jin等人^[170]直接为每个任务学习参数掩码; Xue等人^[171]基于预训练的ViT, 学习自注意力机制中的分词掩码。由于直接优化离散变量比较困难, 二者均通过Gumbel Softmax^[172]学习。Serra等人^[173]则通过施加硬注意力机制(Hard Attention to the Task, HAT)来学习基于神经元的掩码, 在反向传播时限制已用权重的更新。SupSup(Supermasks in Superposition)^[174]在一个随机初始化的网络上学习参数掩码。

由于网络容量有限, 随着任务数量的增多, 网络逐渐趋于饱和, 难以容纳新任务。因此静态结构的方法通常需要使用参数稀疏性约束, 并对冻结的旧参数进行选择性重用。Kang等人^[175]根据绝对值大小复用历史任务的参数。Jin等人^[170]用费希尔信息矩阵衡量历史任务参数对当前任务的敏感度, 并以此对参数进行重复使用。尽管重用参数能够一定程度上减缓模型容量有限的问题, 但会带来任务间干

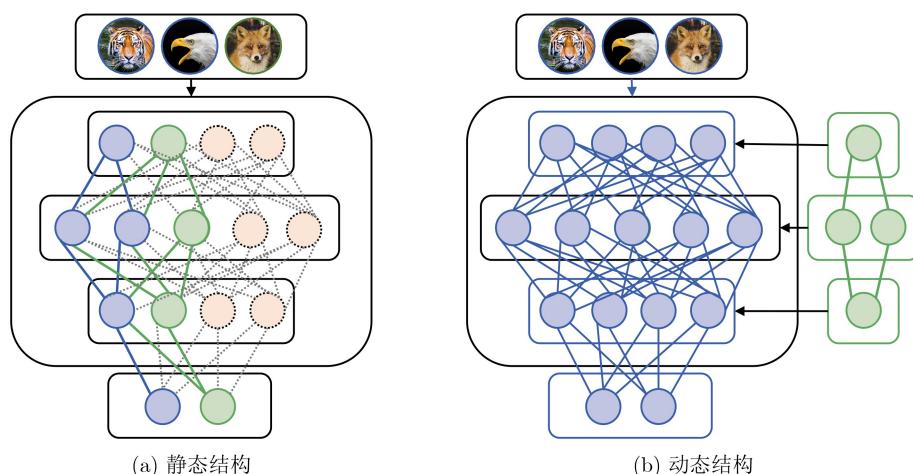


图 6 基于网络结构的方法

扰这一新问题。为了缓解这一困境，基于动态结构的方法不限制网络结构，动态地对网络进行扩张。

3.4.2 动态结构

基于动态结构的方法在训练新任务时动态地添加新的神经元、参数或网络模块。早期工作主要集中在神经元扩张上。DEN(Dynamic Expandable Networks)^[176]在训练新任务时，如果损失超过阈值则自下而上扩展新的神经元，并通过稀疏正则化消除无用的神经元。RCL(Reinforced Continual Learning)^[177]将网络扩展问题建模为一个强化学习问题，为每一个任务搜索最佳的神经结构。

由于扩展神经元带来的性能提升有限，一些工作通过扩展骨干网络来进行序列任务的学习。PNN(Progressive Neural Networks)^[178]为每一个任务引入相同的子网络，并通过适配器将知识在子网络之间进行转移；PathNet^[179]和RPSNet(Random Path Selection Network)^[180]设置多个并行结构，为每个任务选择最佳路径。ExpertGate^[181]在学习每个任务时扩展网络，同时设计了一个门控映射将输入映射到最适合的路径。DER(Dynamic Expandable Representation)^[24]在学习每个任务时添加任务特定的特征提取器，并将所有特征提取器提取的特征进行组合，一并输入分类器预测，在多个持续学习的实验中取得了优越的效果。然而，该模型所需的存储空间急剧增加，使用剪枝策略能够略微缓解这一问题。为了进一步解决参数存储问题，FOSTER(Feature Boosting and Compression)^[182]在每个学习阶段结束后使用自蒸馏，将多个网络蒸馏为单个网络，减少了所需存储空间。Zhou等人^[183]仅仅扩张网络的部分结构，在学习任务时微调任务特定的模块。

近年来，ViT模型引起了研究者的广泛关注，许多工作以ViT为骨干网络进行持续学习。DyTox^[25]利用ViT便于扩展的结构特性，在增量学习任务时逐步添加任务分词(task token)，相比于保存网络模块和骨干网络，能够更充分利用存储空间。此外，预训练的ViT模型中蕴含了丰富的先验知识和强大的泛化能力，一些工作充分利用预训练模型和参数高效微调技术来构建持续学习模型，本文将在下一小节进行介绍。

3.4.3 参数高效微调

参数高效微调(Parameter Efficient FineTuning, PEFT)^[27-32]是一种对预训练模型的微调方法，通过训练少部分参数从而使模型快速适配下游任务。基于参数高效学习的方法充分利用了预训练模型强大的表征能力和泛化能力，表现出卓越的效果。

在持续学习中，该类方法可以视为一种特殊的基于网络结构的方法，它从一个固定的预训练模型出发，为每个任务动态地构建参数高效模块。L2P(Learning to Prompt)^[26]首先借鉴了视觉提示学习的思想，利用提示学习(Prompt Tuning)^[32]对模型进行增量微调。在训练期间，预训练模型是被冻结的，模型只对提示嵌入进行微调来适应新任务。具体来讲，L2P定义了可学习的提示池

$$\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M\}$$

其中， M 是提示的个数， $\mathbf{p}_i \in \mathbb{R}^{l \times d}$ 是长度为 l 维度为 d 的提示嵌入，每个提示 \mathbf{p}_i 配有对应的键向量 \mathbf{k}_i 。给定输入 \mathbf{x} ，模型首先根据输入的预训练表征 $q(\mathbf{x})$ 在提示池中匹配与其最近邻的 N 个提示嵌入，随后将这些提示与输入的原始嵌入 \mathbf{h} 拼接在一起，共同输入到多头注意力机制(Multi-head Self-Attention, MSA)中。在L2P的基础上，DualPrompt^[184]以前缀学习(Prefix Tuning)^[31]的方式进行增量微调，将可学习的前缀插入到注意力机制的键和值向量中。提示和前缀的作用机制可以分别表示为

$$\mathbf{h}'_{\text{Prompt}} = \text{MSA}([\mathbf{p}; \mathbf{h}], [\mathbf{p}; \mathbf{h}], [\mathbf{p}; \mathbf{h}])$$

$$\mathbf{h}'_{\text{Prefix}} = \text{MSA}(\mathbf{h}, [\mathbf{p}_k; \mathbf{h}], [\mathbf{p}_v; \mathbf{h}])$$

CODA-Prompt(Continual Decomposed Attention-based Prompt)^[185]将不同提示进行组合，以结合不同模式的知识。除了提示学习和前缀学习，Gao等人^[186]和Zhou等人^[187]对比了其他类型的参数高效微调技术对持续学习任务的影响，例如适配器(Adapter)^[29]、低秩适配器(LoRA)^[30]等。

此外，与其他的网络结构方法相同，模型在推理时需要获取任务标识以选取任务特定的模块。L2P^[26]、DualPrompt^[184]利用预训练模型提取表征，在提示池中选取最近的 K 个任务提示。CODA-Prompt^[185]在此基础上额外引入了可训练的注意力向量。ESN(Energy Self-Normalization)^[188]则是一种基于能量的预测方法，在多个候选温度下进行投票以最大化输出的Helmholtz自由能来确定任务标识。Hide-Prompt(Hierarchical Decomposition of Prompt)^[189]将表征建模为高斯分布，通过回放表征自适应地更新任务标识预测器。

除了预训练的视觉模型之外，还有一些工作对预训练的视觉语言模型进行增量式微调。S-Prompt^[190]利用CLIP^[191]模型学习视觉和语言提示，来进行域增量任务的学习。PROOF(Projection Fusion)^[192]，SRPrompt (Self-Regulating Prompt)^[193]进一步挖掘了视觉-语言模型在持续学习任务中的潜力。

相较于不使用预训练模型的方法, 这些基于预训练模型的方法在持续学习任务上取得了明显的性能提升。这一结果很大程度上归功于预训练模型强大的表征能力和泛化能力。一些研究工作指出, 预训练模型本身就具有强大的零样本泛化能力^[191]; SimCIL^[187]在预训练模型的基础上, 仅学习分类器权重也获得优越的性能。Kim等人^[194]则认为, 预训练任务中的类别可能与下游任务重复而导致信息泄露, 因此应当使用排除了重复类别的预训练模型。Tang等人^[195]提出了自适应的提示生成, 设计了一种无需预训练模型的方法。但尽管如此, 是否使用预训练模型仍然对最终的结果有显著的影响。

随着预训练模型的进一步发展, 不同的预训练权重被相继提出。Wang等人^[189]发现, 通过不同方法获得的预训练模型, 对于下游持续任务的学习有着显著的影响。此外, 不同的参数高效微调技术, 在结构、参数量和推理速率上各有所异。在持续学习中, 如何合理地使用预训练模型和参数高效微调技术, 以及如何对相关方法进行客观公正的评估, 这些问题亟待解决。

3.5 小结

本节对持续学习的方法进行了整理, 将现有方法分为基于正则化、基于回放、基于梯度、基于网络结构4类。每类方法的特点以及优缺点, 见表3。

(1) 基于正则化的方法显式地添加正则损失来约束参数更新。其中, 参数正则化的方法通过参数重要性估计对参数施加惩罚, 由于估计的合理性和有效性难以保证, 因而在复杂任务上表现较差。数据正则化的方法从不同层面构建蒸馏关系, 来保持新模型与旧模型对给定数据的输出一致, 该方法通常使用回放样本或特征来进一步减缓遗忘, 取得显著的成效。同时, 针对任务偏向问题, 不同的工作提出了不同的见解。

(2) 基于回放的方法通过回放历史样本进行知

识巩固。其中, 数据回放的方法通过核心集选取、数据集蒸馏或数据压缩等手段, 回放部分原始数据或压缩数据; 特征回放的方法对样本特征进行回放, 面临着特征偏移问题; 而生成式回放的方法利用生成模型, 对旧任务数据或特征进行生成式回放, 生成质量难以保证。

(3) 基于梯度的方法, 利用历史任务样本进行梯度情景回忆; 或者将梯度进行子空间投影, 以保证历史任务免受当前任务的干扰。此外, 收敛到平坦极小点, 也能够提升模型对新旧任务的性能。

(4) 基于网络结构的方法为任务学习特定的参数。其中, 静态结构的方法, 将网络的参数分配给每个任务, 可以通过剪枝、优化不同方式实现; 而动态结构的方法在学习新任务时扩展网络结构。基于参数高效微调的方法在预训练模型上增量式地微调轻量模块, 取得了优越的性能。

4 理论研究进展

尽管持续学习在实验方面取得了显著进展, 但在理论分析方面的研究仍然相对初步。现有文献对理论工作的总结鲜有涉及。因此, 本节对近年来持续学习领域的理论研究进展进行了综述, 从不同的视角, 为持续学习问题提供更深入的理解和分析。通过剖析理论并联系具体方法, 本节旨在为持续学习领域的进一步发展奠定坚实的理论基础, 并促进理论与实践的有机结合。

在未做特殊声明的情况下, 本节的数学符号延续了2.1节的表示。

4.1 概率模型

一些工作^[17,65–68,196,197]在概率模型视角下对持续学习目标进行分析, 进而推导出网络的正则损失。根据贝叶斯公式, 所有已见任务的后验 $p(\theta|\mathcal{D}_{1:t})$ 满足

$$p(\theta|\mathcal{D}_{1:t}) \propto p(\theta) \prod_{k=1}^t p(\mathcal{D}_k|\theta) \propto p(\theta|\mathcal{D}_{1:t-1}) p(\mathcal{D}_t|\theta)$$

表3 持续学习方法分类及特点

类别	方法	方法特点	优缺点
基于正则化	参数正则化	通过参数重要性估计对参数进行保护	无需回放样本, 但难以有效估计参数重要性, 性能较差
	数据正则化	保持新旧模型对给定数据的输出一致性	简单有效, 但通常需要回放样本或特征以提高性能
	任务偏向修正	针对网络任务偏向问题提出不同的解决方案	需要额外的修正训练, 或额外的计算资源
基于回放	原始数据回放	回放部分任务的原始样本	简单有效, 但回放样本占据空间较大
	原始特征回放	回放样本特征或类别原型	节省存储空间, 面临特征偏移问题
	生成式回放	使用生成模型进行数据回放	生成数据的质量难以保证
基于梯度	梯度情景记忆	基于历史数据的梯度构建约束	需要回放样本, 并计算额外梯度
	子空间投影	将参数梯度投影到子空间	能有效减缓遗忘, 需要存储特征空间
	平坦极小点	获取平坦极小点	使用额外的技术手段, 增加训练成本
基于网络结构	静态结构	将网络参数分配给任务	模型容量有限, 难以解决长序列任务的学习问题
	动态结构	动态地扩张网络结构	能有效减缓遗忘, 但扩张网络带来额外存储和推理负担
	参数高效微调	对预训练模型进行增量式微调	能有效减缓遗忘, 但获取预训练模型需要成本

在该式下，先前任务的后验成为当前任务的先验。然而由于先前任务后验 $p(\boldsymbol{\theta}|\mathcal{D}_{1:t-1})$ 不可知，需要对其进行建模和估计。拉普拉斯估计(Laplace approximation)用高斯分布对其进行估计，即

$$p(\boldsymbol{\theta}|\mathcal{D}_{1:t-1}) \approx \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_{t-1}, \boldsymbol{A}_{1:t-1}^{-1})$$

其中均值 $\boldsymbol{\mu}_{t-1} = \boldsymbol{\theta}_{t-1}$ 表示历史模型参数，精度矩阵为

$$\begin{aligned} \boldsymbol{A}_{1:t-1} &= -\nabla_{\boldsymbol{\theta}}^2 \ln p(\boldsymbol{\theta}|\mathcal{D}_{1:t-1})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{t-1}} \\ &\approx \sum_{k=1}^{t-1} -\nabla_{\boldsymbol{\theta}}^2 \ln p(\mathcal{D}_k|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k} \end{aligned}$$

它表示先前任务负对数似然的黑塞矩阵的求和。然而由于网络参数量巨大，直接计算黑塞矩阵比较困难。可以用费希尔信息矩阵对其进行近似，即 $\boldsymbol{A}_{1:t-1} \approx \sum_{k=1}^{t-1} \boldsymbol{F}_k := \boldsymbol{F}_{1:t-1}$ ，其中费希尔信息矩阵定义为

$$\boldsymbol{F}_k := \mathbb{E} \left[\nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{\theta}|\mathcal{D}_k) \nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{\theta}|\mathcal{D}_k)^T \right]$$

其计算只依赖参数的1阶梯度。于是，当前任务的最大后验可以计算为

$$\begin{aligned} \boldsymbol{\theta}_t &= \operatorname{argmax}_{\boldsymbol{\theta}} \ln p(\boldsymbol{\theta}|\mathcal{D}_{1:t}) \\ &= \operatorname{argmax}_{\boldsymbol{\theta}} \ln p(\boldsymbol{\theta}|\mathcal{D}_{1:t-1}) + \ln p(\mathcal{D}_t|\boldsymbol{\theta}) \\ &\approx \operatorname{argmax}_{\boldsymbol{\theta}} \ln p(\mathcal{D}_t|\boldsymbol{\theta}) - (\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})^T \cdot \boldsymbol{F}_{1:t-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}) / 2 \end{aligned}$$

在训练时，网络的损失函数形式为

$$\mathcal{L}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathcal{D}_t) + \lambda(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})^T \boldsymbol{F}_{1:t-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}) / 2$$

其中第1项代表了当前任务的训练损失，第2项是正则损失， λ 是控制正则程度的超参数。

除了拉普拉斯近似，另一种近似方式是变分推断，即在一族推断函数 \mathcal{Q} 中最小化如下的KL散度(Kullback Leibler divergence)

$$q_t(\boldsymbol{\theta}) = \operatorname{argmin}_{q \in \mathcal{Q}} \text{KL}(q(\boldsymbol{\theta}) | q_{t-1}(\boldsymbol{\theta}) p(\mathcal{D}_t|\boldsymbol{\theta}))$$

VCL^[196]同样指定推断函数服从高斯分布，从而推导出与拉普拉斯估计类似的损失形式。此外，还有一些变分推断的扩展，例如Kapoor等人^[197]用变分自回归高斯过程改进后验更新。这些工作与参数正则化的方法密切相关。实际上，使用高斯分布进行近似等同于对先前任务的损失进行二阶泰勒近似，该近似的有效性受以下几个因素影响：

(1) 黑塞矩阵的估计。对黑塞矩阵进行更准确的估计能改善对损失函数的近似。如Ritter等人^[67]提出用Kronecker分解进行黑塞矩阵的估计，其效果优于基于费希尔信息矩阵估计的方法。

(2) 参数的移动。泰勒近似仅在展开点附近有

效，在训练新任务时，随着模型参数逐渐远离先前的参数，这种近似可能失效。

(3) 高阶导数项。根据泰勒展开的性质，泰勒展开和实际函数之间的误差受高阶导数影响。然而计算高阶导数是费时耗力的，难以应用到实际场景中。

这些问题的存在使得参数正则化的方法实际场景中的表现并不理想。

4.2 PAC学习

一些研究^[198-201]在PAC(Probably Approximately Correct)学习框架下，对持续学习的泛化误差进行界定。设 $h \in \mathcal{H}$ 是假设空间 \mathcal{H} 中的假设函数，假设空间 \mathcal{H} 的VC维度为 d ， $l: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ 是损失函数。任务 t 的数据分布记为 $\mathcal{D}_t = \mathcal{X}_t \times \mathcal{Y}_t$ ，训练数据集记为 $D_t = \{(\mathbf{x}_i^i, y_i^i)\}_{i=1}^{n_t}$ 。任务 t 的期望风险定义为 $\varepsilon_{\mathcal{D}_t}(h) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_t} [l(h(\mathbf{x}), y)]$ ，经验风险定义为 $\hat{\varepsilon}_{\mathcal{D}_t}(h) = \sum_{(\mathbf{x}, y) \in D_t} [l(h(\mathbf{x}), y)]$ 。持续学习的目标是最小化所有任务的期望风险，即 $h^* = \operatorname{argmin}_h \sum_{i=1}^t \varepsilon_{\mathcal{D}_i}(h)$ 。由于真实数据分布未知，实际中通常采用经验风险最小化，即 $h^* = \operatorname{argmin}_h \sum_{i=1}^t \hat{\varepsilon}_{\mathcal{D}_i}(h)$ 。

基于经验风险最小化(Empirical Risk Minimization, ERM)的泛化误差界描述了经验风险和期望风险的差距，即

$$\begin{aligned} \sum_{i=1}^t \varepsilon_{\mathcal{D}_i}(h) &\leq \sum_{i=1}^t \hat{\varepsilon}_{\mathcal{D}_i}(h) + \lambda_t, \\ \lambda_t &= \sqrt{\frac{d [\ln(n_{1:t}/d)] + \ln(1/\delta)}{2n_{1:t}}} \end{aligned}$$

以至少 $1 - \delta$ 的概率成立。影响该误差界的主要因素是样本数量 $n_{1:t}$ 和假设空间的大小 d 。在此基础上，Wang等人^[199]借鉴域适应(Domain Adaptation, DA)^[202]的思想，用 \mathcal{H} -散度对泛化误差进行界定。具体而言，任务 t 的期望风险以至少 $1 - \delta$ 的概率满足

$$\begin{aligned} \varepsilon_{\mathcal{D}_t}(h) &\leq \frac{1}{t-1} \sum_{i=1}^{t-1} \hat{\varepsilon}_{\mathcal{D}_i}(h) + \frac{1}{2(t-1)} \\ &\quad \cdot \sum_{i=1}^{t-1} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_t) + \lambda_{t-1} \end{aligned}$$

其中， \mathcal{H} -散度定义为 $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j) = 2 \sup_{h' \in \mathcal{H}} |\Pr_{\mathcal{D}_i}[\mathbb{I}(h' \neq h)] - \Pr_{\mathcal{D}_j}[\mathbb{I}(h' \neq h)]|$ ，反映了数据分布 \mathcal{D}_i 和 \mathcal{D}_j 的差异。影响该泛化误差界因素包含：

(1) 历史任务的经验风险。在已知历史任务数据的情况下，可以通过训练来降低该经验风险。然而，尽管基于回放的方法允许存储少量的历史任务样本，但回放样本数量过少会导致上式中的 λ_{t-1} 较

大, 从而使泛化误差界变得松散。这揭示了回放方法的局限性。

(2) 任务分布的差异。当任务之间的差异较大时, 泛化误差界相对较松, 学习历史任务对当前任务的影响较小; 相反, 当任务分布的差异较小时, 泛化误差界相对较紧, 通过优化历史任务的经验风险可以对当前任务产生正面效果。换句话讲, 当任务相近时不同任务之间的相互协同有助于促进模型的泛化性。

(3) 假设空间的大小。通常情况下, 较小的假设空间使得泛化误差边界更紧。然而, 如果假设空间过小, 可能导致经验误差难以被优化; 反之, 如果假设空间过大, 泛化误差边界可能会相对较松, 模型容易陷入过拟合。

此外, 任务的泛化误差还可以用模型间隔界定, 即

$$\varepsilon_{\mathcal{D}_t}(h) \leq \varepsilon_{\mathcal{D}_t}(h, h_{t-1}) + \varepsilon_{\mathcal{D}_t}(h_{t-1})$$

其中, h_{t-1} 为在任务 $t-1$ 结束后得到的模型。不等式右侧的第1项体现了蒸馏的思想, 即用先前模型指导当前模型的训练; 第2项则表示了先前模型在当前任务上的误差。

Shi等人^[201]提出域增量学习的统一框架, 通过自适应系数将不同的泛化误差界进行结合, 为不同的数据正则化方法提供了统一的视角。这些理论分析是与回放与蒸馏的方法密切相关的, 能够帮助理解不同任务如何相互作用从而共同影响网络学习。

4.3 神经正切核

还有一些工作旨在分析神经网络的遗忘性。然而由于神经网络结构复杂, 直接进行理论分析相当具有挑战性。因此, 许多理论工作侧重于简化模型, 以便更容易研究其学习过程。神经正切核(Neural Tangent Kernel, NTK)是一种有力的理论分析工具, 最初由Jacot等人^[203]于2018年提出, 刻画了无限宽的神经网络在动态演化过程中表现出的线性性质。对于持续学习而言, 给定第 t 个任务的训练数据 \mathbf{X}_t , 训练后的网络可以表示为

$$\begin{aligned} \mathbf{f}_t(\mathbf{x}) &= \mathbf{f}_{t-1}(\mathbf{x}) + \langle \phi(\mathbf{x}), \boldsymbol{\theta}_t^* - \boldsymbol{\theta}_{t-1}^* \rangle \\ &= \mathbf{f}_{t-1}(\mathbf{x}) + \mathcal{K}(\mathbf{x}, \mathbf{X}_t)(\mathcal{K}(\mathbf{X}_t, \mathbf{X}_t) + \lambda \mathbf{I})^{-1} \tilde{\mathbf{y}}_t \end{aligned}$$

其中, 核函数定义为 $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \phi(\mathbf{x}')^\top$, $\phi(\mathbf{x}) = \nabla_{\boldsymbol{\theta}} f_0(\mathbf{x})$ 是初始参数的梯度, $\tilde{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{f}_{t-1}(\mathbf{X}_t)$ 是先前模型对当前任务数据的预测残差。在此框架下, 当前模型对先前任务的遗忘可以用模型输出的变化衡量, 即

$$\begin{aligned} \Delta_{t-1} &:= \|\mathbf{f}_t(\mathbf{X}_{t-1}) - \mathbf{f}_{t-1}(\mathbf{X}_{t-1})\|_2^2 \\ &= \|\mathcal{K}(\mathbf{X}_{t-1}, \mathbf{X}_t)(\mathcal{K}(\mathbf{X}_t, \mathbf{X}_t) + \lambda \mathbf{I})^{-1} \tilde{\mathbf{y}}_t\|_2^2 \end{aligned}$$

导致遗忘产生的主要因素是 $\mathcal{K}(\mathbf{X}_{t-1}, \mathbf{X}_t) = \phi(\mathbf{X}_{t-1}) \phi(\mathbf{X}_t)^\top$ 与残差 $\tilde{\mathbf{y}}_t$ 。前者表示了先前任务梯度和当前任务梯度的相似性, 梯度相似性越小则模型的遗忘程度越小; 反之, 模型的遗忘程度越大。后者代表了先前模型对当前任务的预测残差, 残差越小代表先前模型对当前任务预测越精准, 因而遗忘越不容易发生。在神经正切核框架下, 梯度投影的方法可以解释为在梯度上施加投影矩阵从而导出新的核函数

$$\tilde{\mathcal{K}}(\mathbf{x}, \mathbf{x}') = \tilde{\phi}(\mathbf{x}) \tilde{\phi}(\mathbf{x}')^\top, \tilde{\phi}(\mathbf{x}) = \mathbf{P} \nabla_{\boldsymbol{\theta}} f_0(\mathbf{x})$$

其中投影矩阵 \mathbf{P} 的形式因方法而异。Bennami等人^[204]在NTK框架下研究了正交梯度投影方法的遗忘性, Doan等人^[205]将其扩展到梯度情景记忆方法中。除此之外, 在神经正切核框架下持续学习问题还可以表述为递归回归问题^[204]; Karakida等人^[206]研究了持续学习的一种变体, 每个任务具有相同目标, 当任务样本量平衡时泛化误差会单调地从任务减小到另一个任务, 而不平衡的样本量会恶化泛化误差。

此外, 一些研究者对线性模型的遗忘进行理论分析, 如文献[207–209]等。在转向更复杂的模型之前, 这些基础模式的理论分析工作是十分必要的。任何线性模型的结果都可以在神经正切核范式下应用到复杂的神经网络。然而, 神经正切核也具有明显的局限性: 其假定了网络是无限宽的, 当网络拥有较小的宽度时, 线性性质就不再成立。

4.4 任务分解

先前的理论研究大都集中于任务标识已知的设定, 即任务增量学习(TIL)设定。相比之下, 类别增量学习(CIL)设定下的理论工作较为有限。近年来, Kim等人^[210,211]指出, 类别增量学习问题可以分解为两个子问题: 任务内预测学习(Within-task Prediction, WP)和任务标识预测(Task-id Prediction, TP)。前者可以通过任务增量学习的方法解决, 而后者被证明与分布外检测密切相关。

在任务增量学习场景中, 基于网络结构的方法能够有效地解决任务内预测问题, 它们通常为任务学习独立的参数, 为便于表述, 本小节用 $\mathbf{x} \in \mathbf{X}_t$ 表示样本 \mathbf{x} 属于第 t 个任务, $\mathbf{x} \in \mathbf{X}_{t,j}$ 表示样本 \mathbf{x} 属于第 t 个任务的第 j 个类别, $\boldsymbol{\theta}^{(t)}$ 表示任务 t 的特定参数。不同任务的重叠部分 $\boldsymbol{\theta}^{(t)} \cap \boldsymbol{\theta}^{(t')}(t \neq t')$ 在训练过程中通常是被冻结的。在推理时, 给定任务标识, 只需用任务对应的参数进行推理, 即 $p(\mathbf{x} \in \mathbf{X}_{t,j} | \boldsymbol{\theta}^{(t)})$ 。而在任务标识未知时, 该过程可以表示为

$$\begin{aligned}
 p(\mathbf{x} \in \mathbf{X}_{t,j} | \boldsymbol{\theta}) &= \sum_{t'=1}^T p(\mathbf{x} \in \mathbf{X}_{t',j} | \mathbf{x} \in \mathbf{X}_{t'}, \boldsymbol{\theta}) \\
 &\quad \cdot p(\mathbf{x} \in \mathbf{X}_{t'} | \boldsymbol{\theta}) \\
 &= \sum_{t'=1}^T p(\mathbf{x} \in \mathbf{X}_{t',j} | \boldsymbol{\theta}^{(t')}) \\
 &\quad \cdot p(\mathbf{x} \in \mathbf{X}_{t'} | \boldsymbol{\theta}) \\
 &= p(\mathbf{x} \in \mathbf{X}_{t,j} | \boldsymbol{\theta}^{(t)}) p(\mathbf{x} \in \mathbf{X}_t | \boldsymbol{\theta})
 \end{aligned}$$

其中, 第1项代表在任务标识已知的前提下预测样本类别, 第2项表示预测样本的任务标识。该理论建立了任务增量学习和类别增量学习之间的联系, 现有的任务增量学习方法都可以配合任务标识预测或分布外检测方法来解决类别增量学习。然而, 由于任务标识预测仍然是依次学习的, 同样可能遭遇灾难性遗忘, 这可以通过采用其他策略来缓解, 例如数据回放^[64]。

在此基础上, Wang等人^[189]额外增加了自适应预测的目标。Kim等人^[194]则进一步证明了类别增量问题的可学习性。尽管如此, 对于类别增量学习的遗忘性和泛化性的分析工作仍较为有限, 需要进一步的研究和探索。

4.5 小结

如表4所示, 本节主要介绍了持续学习中的理论分析工作:

(1) 在概率模型视角下, 基于贝叶斯理论分析学习目标, 并对先前任务进行近似, 推导网络训练损失。这类理论工作对应了参数正则化方法。

(2) 在PAC学习框架下, 对持续学习的泛化误差进行界定。基于经验风险最小化(EMR)的泛化界解释了样本数量对泛化性能的重要性, 基于域适应的泛化界反映了任务差距对泛化的影响, 基于模型间隔的泛化界体现了数据正则化(蒸馏)在学习过程中的作用。

(3) 在神经正切核范式下, 网络可以近似为线性模型, 遗忘性主要受神经正切核矩阵 $\mathcal{K}(\mathbf{X}_{t-1}, \mathbf{X}_t)$ 的影响。基于梯度的方法可以理解为对梯度进行修正从而导出新的神经正切核 $\tilde{\mathcal{K}}$, 从而减少遗忘。

(4) 任务分解, 将类别增量学习分解为任务内

预测和任务标识预测两个子问题。前者可以用基于网络结构的方法解决, 而后者与分布外检测紧密相关。

现有的理论分析工作和不同类型方法具有密切的联系, 这为持续学习的实践和应用提供坚固的理论支撑和保证。然而其中的一些理论分析往往基于强的假设和前提、进行多种估计和近似, 或者简化网络模型, 这些条件在真实场景中很难成立。随着深度学习技术的快速发展, 持续学习的理论和实践之间仍然存在明显的差距, 因此需要进一步强化对持续学习理论的研究。

5 实验数据集与评价指标

本节以图像分类任务为例, 介绍持续学习的常用数据集和评价指标, 以公正客观地评判不同方法。

5.1 实验数据集

表5总结了持续学习中常用的图像分类数据集。其中, 每个单独的分类数据集都可按照类别进行划分, 从而构建持续学习任务。目前公认的划分方式有两种: 一种方式是将所有类别平均划分给每个任务; 另一种方式则是将一半数目的类别作为初始任务, 将剩下的类别平均划分给每个任务。在测试时可以是任务标识已知的, 即任务增量学习; 也可以是任务标识未知的, 即类别增量学习。此外, 遵循文献[64]中的协议, 在划分类别之前需对数据集的类别以随机数种子1993进行打乱。

早期的工作和实验主要集中在MNIST数据集上。MNIST数据集中包含了10类手写数字体图像, 图像分辨率为 32×32 。常用的任务设定是将MNIST按照类别划分成多个互不相交的子任务。此外, 为了增加任务难度, Kirkpatrick等人^[17]提出两种新的任务设定: (1)排列的MNIST数据集, 使用 T 个不同的排列将MNIST数据集中的图像像素进行重排, 从而构建 T 个任务; (2)旋转的MNIST数据集, 使用 T 个不同旋转角度对原始图像进行旋转, 以构建任务。这两种设定均属于域增量学习范畴。

CIFAR10, CIFAR100以及ImageNet数据集成

表4 持续学习理论工作总结

理论工作	主要结果	特点	对应方法
概率模型	$\theta_t \approx \text{argmax}_{\boldsymbol{\theta}} \log p(\mathcal{D}_t \boldsymbol{\theta}) - (\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})^T \mathbf{F}_{1:t-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}) / 2$	通过对先前任务进行近似和估计, 得到网络训练的正则损失	参数正则化
PAC学习	$\varepsilon_{\mathcal{D}_t}(h) \leq \frac{1}{t-1} \sum_{i=1}^{t-1} \hat{\varepsilon}_{\mathcal{D}_i}(h) + \frac{1}{2(t-1)} \sum_{i=1}^{t-1} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_t) + \lambda_{t-1}$	在PAC学习理论框架下, 对网络的泛化误差进行界定	数据正则化 基于回放的方法
神经正切核	$\Delta_{t-1} = \left\ \mathcal{K}(\mathbf{X}_{t-1}, \mathbf{X}_t) (\mathcal{K}(\mathbf{X}_t, \mathbf{X}_t) + \lambda I)^{-1} \tilde{\mathbf{y}}_t \right\ _2^2$	在神经正切核范式下, 分析神经网络的遗忘问题	基于梯度的方法
任务分解	$p(\mathbf{x} \in \mathbf{X}_{t,j} \boldsymbol{\theta}) = p(\mathbf{x} \in \mathbf{X}_{t,j} \boldsymbol{\theta}^{(t)}) p(\mathbf{x} \in \mathbf{X}_t \boldsymbol{\theta})$	将类别增量学习问题分解为任务内预测和任务标识预测两个子问题	基于网络结构的方法

表 5 持续学习常用数据集

数据集	年份	类别数	数据量
MNIST ^[212]	1998	10	60,000
CIFAR-10 ^[213]	2009	10	60,000
CIFAR-100 ^[213]	2009	100	60,000
CUB-200 ^[214]	2011	200	11,788
Tiny-ImageNet ^[215]	2015	200	120,000
Sub-ImageNet ^[216]	2009	100	60,000
Full-ImageNet ^[216]	2009	1,000	1,280,000
5-datasets ^[217]	2020	50	260,000
CORe50 ^[218]	2017	50	15,000
DomainNet ^[219]	2019	345	590,000
CCDB ^[220]	2023	2	--

为近年来普遍使用的分类数据集。CIFAR-10包含10个类别，每个类别包含60,000张图像，每张图像分辨率为 32×32 ；而CIFAR-100是CIFAR-10的扩展，包含了100个类别。Tiny-ImageNet包含了200个类别共120,000个样本，图像分辨率为 64×64 。ImageNet则包含了1 000个类别以及128万余张图像，Sub-ImageNet是ImageNet的子集，包含了100个类别，每个类别含有600张图像。在CIFAR-10, CIFAR-100, ImageNet或其子集上进行实验和性能评估，已成为目前主流持续学习方法的共识。

鉴于在单一数据集上进行划分容易导致任务之间相似性较大，Ebrahimi等人^[217]提出了多个数据集混合的测试基准：5-datasets，该基准包含了5个图像分类任务，分别是CIFAR-10, MNIST, FashionMNIST^[221], SVHN^[222]以及notMNIST^[223]。虽然模型对这些任务进行单独学习较为容易，但对它们进行连续学习却具有相当大的挑战。这是由于模型对于多样化的任务更容易产生灾难性遗忘问题。

除此之外，一些域增量任务也极具挑战性。CORe50是一个真实世界的连续物体识别数据集，包含了11个域的50类图像，该数据集可以按照域划分成11个任务进行域增量学习。DomainNet^[219]则是更具挑战的域增量数据集，包含了6个域，每个域具有345类图像，总计59万余张图像。CDDB是近年来Li等人^[220]提出的人脸伪造检测数据集，包含了多个生成模型生成的伪造人脸，每个生成模型的生成图像可以视为一个域，在学习过程中每个任务是判断人脸是否为伪造的，同样属于域增量学习任务。

5.2 评价指标

与单任务学习不同，持续学习在多个任务上依次地进行学习，在每个学习阶段结束后，模型对于

先前任务的性能可能发生变化。通常来讲，模型的持续学习性能可以从模型的整体性能和遗忘程度两方面进行评估。为便于表述，本节以分类任务为例，并用 $a_{t,i}$ 表示在第 t 个任务结束后模型在第 i 个任务测试集上的准确率。

模型的整体性能可以使用平均准确率以及平均增量准确率进行评估。在第 t 个任务结束后，模型的平均准确率以及平均增量准确率分别定义为

$$A_t = \frac{1}{t} \sum_{i=1}^t a_{t,i}$$

$$\bar{A}_t = \frac{1}{t} \sum_{i=1}^t A_i$$

其中， A_t 代表了模型在所有已见任务上的平均准确率， \bar{A}_t 则进一步反映了模型性能的历史变化。

模型的遗忘程度可以用负向迁移(Backward Transfer, BWT)以及遗忘性度量(Forgetting Measure)来衡量。在第 t 个任务结束后，模型对 i 个任务负向迁移和遗忘定义为

$$b_{t,i} = a_{t,i} - a_{i,i}$$

$$f_{t,i} = \max_{k \leq t} (a_{k,i} - a_{t,i})$$

由于第 i 个任务性能最大值通常就是 $a_{i,i}$ ，因此大多情况下，负向迁移和遗忘性度量之间仅有符号的差异。但在某些情况下则不然，这是由于新任务的学习可能会对旧任务具有促进作用。为了评估整体遗忘性，模型对所有已见任务的平均负向迁移以及平均遗忘性度量定义为

$$\text{BWT}_t = \frac{1}{t-1} \sum_{i=1}^{t-1} b_{t,i}$$

$$\text{FM}_t = \frac{1}{t-1} \sum_{i=1}^{t-1} f_{t,i}$$

以上指标只评估模型的学习性能，而在实际应用中，任务所需的额外存储空间、模型的规模以及模型的推理速率也应该被慎重考虑和评估。对于正则化方法，模型需要存储旧模型的参数，因此所需的存储空间与模型的规模相关。对于回放的方法，除了旧模型之外，历史数据或生成模型也占据额外的空间。基于梯度的方法，要么存储历史数据，要么存储网络特征空间。基于参数分配的方法，需要存储网络掩码，或者动态扩大网络规模。此外，一些方法使用多个模型的集成进行推理，尽管取得了优越的性能，但却降低了计算效率。故在评估方法时，计算效率也应被考虑在内。

因此，在评估不同的方法时，也应该比较其所

需的额外存储空间和推理速率，或者在相同的内存预算限制下进行测试。

6 应用

现实世界复杂多样，存在各种各样的任务和挑战。持续学习方法能够处理连续复杂任务，在计算机视觉、自然语言处理等多个领域都有广泛的应用。以下将介绍近年来持续学习在各个领域的应用，分析其应用价值以及应用潜力。

6.1 图像分类

在图像分类任务上，持续学习得到了广泛且成熟的应用。在任务增量场景下，正交投影的方法如 Adam-NSCL^[153], CUBER^[156]等，在梯度子空间更新参数，能够不影响历史任务；网络的方法如 SupSup^[174], HAT^[173]等，它们通过学习共享网络的子网络来为每个任务单独工作。这些现有的方法在任务增量场景下基本可以消除网络的遗忘问题，性能已经逼近上界。

类别增量学习则是应用范围更加广泛的一类设定，它可以被分解为任务内预测和任务标识预测两个子问题^[210]，前者可以通过已有的任务增量学习方法解决，而后者被证明与分布外数据(Out-of-distribution, OOD)检测紧密相关^[210]。为了克服模型在类别增量学习过程中的灾难性遗忘，主流方法大都是基于数据正则化和样本回放的；还有一些工作致力于不使用回放样本来解决遗忘，这类方法通常在特征空间中回放样本特征，如SSRE^[131], PASS^[89]等。预训练模型的发展和应用极大地提升了模型在类别增量任务上的表现，基于预训练模型微调的方法，如L2P^[26], DualPrompt^[184]等，一方面使用参数高效微调技术能够帮助模型快速适配当前任务，另一方面预训练模型表征空间能够更好地建模不同任务特性，从而改善任务标识的预测。

6.2 目标检测

持续学习在计算机视觉领域的另一个典型应用是目标检测，称为增量目标检测。与分类任务不同的是，目标检测中的单张图像往往包含了多个类别的目标，标注信息同时包含了目标的类别和边界框坐标。此外，标注信息是随任务依次到来的，在学习新类别时，旧类别的标注是不可知的，这为增量目标检测提供了挑战。在增量目标检测中，蒸馏是一种简单高效的手段，旧任务的标注信息可以自然地通过旧模型获取。Shmelkov等人^[43]基于Fast RCNN模型用旧模型对旧类别预测进行蒸馏。后续增量目标检测的方法，如RILOD^[44], SID^[45]等大多沿用蒸馏的策略。

6.3 语义分割

持续学习在语义分割任务上也有广泛应用。与增量目标检测类似，在持续语义分割任务上只有当前类别可见，而旧类别被标注为背景。蒸馏在持续语义分割任务上也是一种简单高效的手段，MiB^[47]利用旧模型的预测来校准当前模型的预测；PLOP^[48]、EM^[49]等方法显式地使用旧模型生成旧类的伪标注；SDR^[50]和UCD^[51]将对比学习引入到表征的蒸馏中。此外，这些方法不仅可以应用在二维的图像任务上，也可以部署到3维数据(如点云)以及视频任务上，如点云数据的持续语义分割^[224,225]、视频数据的持续动作识别^[226,227]等。

6.4 视觉预训练模型

视觉预训练模型旨在大规模数据集上学习图像的通用表征，以提高下游任务的性能。由于预训练所需数据量巨大，其收集和标注的过程往往是持续的。持续学习能够以更低的资源和成本获取大规模的预训练模型。具体来讲，模型增量地在多个任务数据上进行有监督或自监督学习，在测试时，通常使用线性探测性能或下游任务性能来评估模型的表征能力。

近年来针对视觉预训练模型的持续学习开始受到关注。一项研究^[56]发现，持续学习过程中模型的泛化性能在不断提升，随着模型所见任务越来越多，模型的泛化能力不断增强。Hu等人^[57]指出对于视觉任务的持续预训练学习，自监督预训练比有监督预训练更为有效，Cossu等人^[58]在视觉语言模型上也观察到一致的结论。Fini等人^[167]以自监督训练的范式进行持续学习，并采用蒸馏策略惩罚历史表征与当前表征一致。

6.5 自然语言处理

在自然语言处理领域，持续学习也具有十分广泛的应用。语言任务和视觉任务具有一些相同的设计，如任务增量、类别增量和域增量等；并且也相应地扩展了具有代表性的持续学习策略，如参数正则^[228]、输出正则^[229]、数据回放^[230,231]、生成回放^[232]和网络结构^[233]等。

大型语言模型(Large Language Model, LLM)逐渐成为自然语言处理领域的研究热点，其使用自监督技术在大量无标记的文本数据上进行训练，在多个自然语言处理任务上表现出色。然而，大型语言模型不仅要具备丰富知识，还应该能够适应和响应变化的世界。一些研究提出使用持续学习方法进行大语言模型的预训练，例如Chen等人^[234]提出Lifelong-MoE，用混合专家(Mixture-of-Experts, MoE)结构在多个数据分布上进行持续预训练学习。

相比于重新训练模型, 持续预训练能够大幅减少模型的训练成本。

另一方面, 在预训练模型适配到下游任务的过程中, 衍生了许多自适应方法, 如适配器^[29]、低秩适配器^[30]、提示^[32]和前缀^[31]等。尽管这些技术使得基础模型能够高效、快速地适配到下游任务, 但微调过程中仍然会不可避免地带来遗忘问题^[235], 即模型会忘记在预训练阶段所学习过的知识。Qi等人^[236]则发现经过指令微调的语言模型的安全性明显降低, 这也和灾难性遗忘有关。因此, 设计合适的持续学习方法来减缓微调模型时产生的遗忘, 这是持续学习未来的应用潜力。

6.6 生成模型

生成模型作为一种数据回放的方式, 能够改善判别式模型的持续学习能力。如DGR^[136]采用生成式回放的方式, 持续地训练判别模型和生成模型。然而, 生成式模型本身也面临着灾难性遗忘问题, 使用生成的样本来训练自身会导致生成能力会越来越差。一些研究工作将不同的持续学习方法部署到生成模型上, 包括正则化^[55]、知识蒸馏^[52]、生成式回放^[53]、网络结构^[54]等, 这些工作大多基于生成对抗网络进行生成。

扩散模型是一种新兴的生成模型, 它为输入逐步添加随机噪声, 并学习逐步去噪过程以恢复原始数据的分布。扩散模型展现出强大图像生成能力, 这为生成模型的持续学习提供了新的技术。Jodelet等人^[142]使用扩散模型的生成数据和回放数据共同促进判别模型的学习。Gao等人^[141]提出的DDGR使用分类器引导扩散模型进行条件生成。Smith等人^[237]使用自正则化低秩适配器增量式地微调扩散模型。扩散模型领域的最新进展有望进一步应用持续学习领域上, 作为生成器提供高质量的回放样本以促进判别器的学习; 另外, 将持续学习方法应用在扩散模型上从而使其具备持续学习和终生学习的能力, 这也是持续学习的应用潜力。

7 挑战和展望

深度学习的发展极大地促进了持续学习领域的相关进展, 但面对真实开放的环境, 持续学习从理论、方法到实践应用还面临一系列困难。本节针对持续学习所面临的挑战, 展望未来研究方向。

(1)存储和隐私问题: 近年来许多持续学习研究只关注提高性能, 而忽视了这些方法带来的存储和隐私问题。基于网络结构的方法需要动态扩张模型, 会增加存储负担; 而基于回放的方法在一些情况下需要保留历史样例, 可能会侵犯用户隐私。针对存储问题, 设计更加轻量高效的持续学习算法,

以便于更好地部署到一些边缘设备上, 是未来的一个重要研究方向; 针对用户隐私问题, 联邦持续学习(Federated Continual Learning, FCL)^[238]可以在不获取数据源的基础上进行训练, 从而避免隐私泄露, 也是一个极具价值的研究方向。

(2)稀疏的标注数据: 目前的持续学习设定大都假设每个任务都具有充足的标注数据, 而这些数据在现实中获取成本高昂。因此, 如何从稀疏的标注数据中进行学习是一个值得进一步探索的研究思路。Tao等人提出了少样本类增量学习(Few Shot Class Incremental Learning, FSCIL)^[82]的设定, 引起了研究者的广泛关注。FSCIL仍然是一个未被充分开发的领域, 需要进一步的研究来探索其潜在的应用和价值。此外, 考虑到现实世界中的数据分布通常具有长尾效应, Liu等人提出长尾类增量学习(Long Tail Class Incremental Learning, LT-CIL)^[239], 这一问题也值得更进一步的研究。

(3)在线学习: 当前的大部分研究通常对不同任务进行单独的离线训练, 然而真实世界中的数据通常以连续数据流的形式出现。因此对连续的数据流进行持续学习, 即在线学习, 是未来的一个重要研究方向。

(4)多模态持续学习: 目前的持续学习的方法往往只在单一类型任务上进行, 然而真实世界中任务类型多种多样, 任务数据也可能具备多种模态。一个优越的持续学习模型应该同时具备对不同类型任务和不同模态数据的处理能力。在未来应当充分探索这些交叉任务和多模态数据的持续学习, 以推动现有持续学习算法落实到实际应用。

(5)应用拓展: 尽管持续学习在视觉之外的其他领域中也有相关应用, 但这些应用大部分简单地使用回放数据进行蒸馏, 需要进一步将已有持续学习方法扩展到其余领域; 另一方面, 在一些前沿问题上, 如大语言模型、扩散模型等, 持续学习的价值和潜力也仍待进一步开发。

8 总结

持续学习旨在从连续的信息流中学习知识, 不断学习新知识的同时也能够不遗忘已有知识。和传统的机器学习方法相比, 持续学习更符合人和其他动物的学习过程, 是一个极具潜力的研究方向。本文首先介绍了持续学习的研究背景, 阐述了其问题定义、典型设定和问题关键。然后对现有的持续学习方法进行了综述, 将现有方法分为4类: 基于正则、基于回放、基于梯度和基于网络结构, 并分析了每类方法的特点和局限性。此外, 本文对持续学习领域的理论工作进行了整理, 建立理论与方法的

联系。最后介绍了持续学习在不同领域上的应用，并且讨论了持续学习面临的问题和挑战，展望了未来的研究方向。

参 考 文 献

- [1] KRIZHEVSKY A, SUTSKEVER I, and HINTON G E. ImageNet classification with deep convolutional neural networks[C]. The 25th International Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, USA, 2012: 1097–1105.
- [2] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 770–778. doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [3] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]. 9th International Conference on Learning Representations, Austria, 2021.
- [4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. The 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach, USA, 2017: 6000–6010.
- [5] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]. The 34th International Conference on Neural Information Processing Systems (NIPS), Vancouver, Canada, 2020: 159.
- [6] ABDEL-HAMID O, MOHAMED A R, JIANG Hui, et al. Convolutional neural networks for speech recognition[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, 22(10): 1533–1545. doi: [10.1109/TASLP.2014.2339736](https://doi.org/10.1109/TASLP.2014.2339736).
- [7] ZHOU Kaiyang, LIU Ziwei, QIAO Yu, et al. Domain generalization: A survey[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(4): 4396–4415. doi: [10.1109/TPAMI.2022.3195549](https://doi.org/10.1109/TPAMI.2022.3195549).
- [8] WANG Yi, DING Yi, HE Xiangjian, et al. Novelty detection and online learning for chunk data streams[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(7): 2400–2412. doi: [10.1109/TPAMI.2020.2965531](https://doi.org/10.1109/TPAMI.2020.2965531).
- [9] HOI S C H, SAHOO D, LU Jing, et al. Online learning: A comprehensive survey[J]. *Neurocomputing*, 2021, 459: 249–289. doi: [10.1016/J.NEUROCOMPUTING.2021.04.112](https://doi.org/10.1016/J.NEUROCOMPUTING.2021.04.112).
- [10] FRENCH R M. Catastrophic forgetting in connectionist networks[J]. *Trends in Cognitive Sciences*, 1999, 3(4): 128–135. doi: [10.1016/S1364-6613\(99\)01294-2](https://doi.org/10.1016/S1364-6613(99)01294-2).
- [11] MCCLOSKEY M and COHEN N J. Catastrophic interference in connectionist networks: The sequential learning problem[J]. *Psychology of Learning and Motivation*, 1989, 24: 109–165.
- [12] CICHON J and GAN Wenbiao. Branch-specific dendritic Ca^{2+} spikes cause persistent synaptic plasticity[J]. *Nature*, 2015, 520(7546): 180–185. doi: [10.1038/nature14251](https://doi.org/10.1038/nature14251).
- [13] ZENKE F, GERSTNER W, and GANGULI S. The temporal paradox of Hebbian learning and homeostatic plasticity[J]. *Current Opinion in Neurobiology*, 2017, 43: 166–176. doi: [10.1016/j.conb.2017.03.015](https://doi.org/10.1016/j.conb.2017.03.015).
- [14] POWER J D and SCHLAGGAR B L. Neural plasticity across the lifespan[J]. *WIREs Developmental Biology*, 2017, 6(1): e216. doi: [10.1002/wdev.216](https://doi.org/10.1002/wdev.216).
- [15] MCCLELLAND J L, MCNAUGHTON B L, and O'REILLY R C. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory[J]. *Psychological Review*, 1995, 102(3): 419–457. doi: [10.1037/0033-295x.102.3.419](https://doi.org/10.1037/0033-295x.102.3.419).
- [16] RATCLIFF R. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions[J]. *Psychological Review*, 1990, 97(2): 285–308. doi: [10.1037/0033-295x.97.2.285](https://doi.org/10.1037/0033-295x.97.2.285).
- [17] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2017, 114(13): 3521–3526. doi: [10.1073/pnas.1611835114](https://doi.org/10.1073/pnas.1611835114).
- [18] HINTON G E and PLAUT D C. Using fast weights to deblur old memories[C]. Proceedings of the 9th Annual Conference of the Cognitive Science Society, Seattle, USA, 1987: 177–186.
- [19] KAMRA N, GUPTA U, and LIU Yan. Deep generative dual memory network for continual learning[J]. arXiv: 1710.10368, 2017. doi: [10.48550/arXiv.1710.10368](https://arxiv.org/abs/1710.10368).
- [20] ROBBINS H and MONRO S. A stochastic approximation method[J]. *The Annals of Mathematical Statistics*, 1951, 22(3): 400–407. doi: [10.1214/aoms/1177729586](https://doi.org/10.1214/aoms/1177729586).
- [21] LOPEZ-PAZ D and RANZATO M A. Gradient episodic memory for continual learning[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6470–6479.
- [22] ZENG Guanxiong, CHEN Yang, CUI Bo, et al. Continual learning of context-dependent processing in neural networks[J]. *Nature Machine Intelligence*, 2019, 1(8): 364–372. doi: [10.1038/s42256-019-0080-x](https://doi.org/10.1038/s42256-019-0080-x).
- [23] MALLYA A and LAZEBNIK S. PackNet: Adding multiple tasks to a single network by iterative pruning[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7765–7773. doi:

- 10.1109/CVPR.2018.00810.
- [24] YAN Shipeng, XIE Jiangwei, and HE Xuming. DER: Dynamically expandable representation for class incremental learning[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 3013–3022. doi: [10.1109/CVPR46437.2021.00303](https://doi.org/10.1109/CVPR46437.2021.00303).
- [25] DOUILLARD A, RAMÉ A, COUAIRON G, et al. DyTox: Transformers for continual learning with DYnamic TOken eXpansion[C]. IEEE/CVF International Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 9275–9285. doi: [10.1109/CVPR52688.2022.00907](https://doi.org/10.1109/CVPR52688.2022.00907).
- [26] WANG Zifeng, ZHANG Zizhao, LEE C Y, et al. Learning to prompt for continual learning[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 139–149. doi: [10.1109/CVPR52688.2022.00024](https://doi.org/10.1109/CVPR52688.2022.00024).
- [27] HE Junxian, ZHOU Chunting, MA Xuezhe, et al. Towards a unified view of parameter-efficient transfer learning[C]. Tenth International Conference on Learning Representations, 2022.
- [28] JIA Menglin, TANG Luming, CHEN B C, et al. Visual prompt tuning[C]. The 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 709–727. doi: [10.1007/978-3-031-19827-4_41](https://doi.org/10.1007/978-3-031-19827-4_41).
- [29] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]. 36th International Conference on Machine Learning, Long Beach, USA, 2019: 2790–2799.
- [30] HU E J, SHEN Yelong, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[C]. Tenth International Conference on Learning Representations, 2022.
- [31] LI X L and LIANG P. Prefix-tuning: Optimizing continuous prompts for generation[C]. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021: 4582–4597. doi: [10.18653/v1/2021.acl-long.353](https://doi.org/10.18653/v1/2021.acl-long.353).
- [32] LESTER B, AL-RFOU R, and CONSTANT N. The power of scale for parameter-efficient prompt tuning[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 2021: 3045–3059. doi: [10.18653/v1/2021.emnlp-main.243](https://doi.org/10.18653/v1/2021.emnlp-main.243).
- [33] PARISI G I, KEMKER R, PART J L, et al. Continual lifelong learning with neural networks: A review[J]. *Neural Networks*, 2019, 113: 54–71. doi: [10.1016/j.neunet.2019.01.012](https://doi.org/10.1016/j.neunet.2019.01.012).
- [34] DE LANGE M, ALJUNDI R, MASANA M, et al. A continual learning survey: Defying forgetting in classification tasks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(7): 3366–3385. doi: [10.1109/tpami.2021.3057446](https://doi.org/10.1109/tpami.2021.3057446).
- [35] MASANA M, LIU Xialei, TWARDOWSKI B, et al. Class-incremental learning: Survey and performance evaluation on image classification[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(5): 5513–5533. doi: [10.1109/tpami.2022.3213473](https://doi.org/10.1109/tpami.2022.3213473).
- [36] BELOUADAH E, POPESCU A, and KANELLOS I. A comprehensive study of class incremental learning algorithms for visual tasks[J]. *Neural Networks*, 2021, 135: 38–54. doi: [10.1016/j.neunet.2020.12.003](https://doi.org/10.1016/j.neunet.2020.12.003).
- [37] 朱飞, 张煦尧, 刘成林. 类别增量学习研究进展和性能评价[J]. 自动化学报, 2023, 49(3): 635–660. doi: [10.16383/j.aas.c220588](https://doi.org/10.16383/j.aas.c220588).
- ZHU Fei, ZHANG Xuyao, and LIU Chenglin. Class incremental learning: A review and performance evaluation[J]. *Acta Automatica Sinica*, 2023, 49(3): 635–660. doi: [10.16383/j.aas.c220588](https://doi.org/10.16383/j.aas.c220588).
- [38] MERMILLOD M, BUGAISKA A, and BONIN P. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects[J]. *Frontiers in Psychology*, 2013, 4: 504. doi: [10.3389/fpsyg.2013.00504](https://doi.org/10.3389/fpsyg.2013.00504).
- [39] VAN DE VEN G M and TOLIAS A S. Three scenarios for continual learning[J]. arXiv: 1904.07734, 2019. doi: [10.48550/arXiv.1904.07734](https://doi.org/10.48550/arXiv.1904.07734).
- [40] BUZZEGA P, BOSCHINI M, PORRELLO A, et al. Dark experience for general continual learning: A strong, simple baseline[C]. The 34th International Conference on Neural Information Processing Systems (NIPS), Vancouver, Canada, 2020: 1335.
- [41] MAI Zheda, LI Ruiwen, JEONG J, et al. Online continual learning in image classification: An empirical survey[J]. *Neurocomputing*, 2022, 469: 28–51. doi: [10.1016/j.neucom.2021.10.021](https://doi.org/10.1016/j.neucom.2021.10.021).
- [42] GOODFELLOW I J, MIRZA M, XIAO Da, et al. An empirical investigation of catastrophic forgetting in gradient-based neural networks[J]. arXiv: 1312.6211, 2013. doi: [10.48550/arXiv.1312.6211](https://doi.org/10.48550/arXiv.1312.6211).
- [43] SHMELKOV K, SCHMID C, and ALAHARI K. Incremental learning of object detectors without catastrophic forgetting[C]. 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 3420–3429. doi: [10.1109/ICCV.2017.368](https://doi.org/10.1109/ICCV.2017.368).
- [44] LI Dawei, TASCI S, GHOSH S, et al. RILOD: Near real-

- time incremental learning for object detection at the edge[C]. The 4th ACM/IEEE Symposium on Edge Computing, Arlington, USA, 2019: 113–126. doi: [10.1145/3318216.3363317](https://doi.org/10.1145/3318216.3363317).
- [45] PENG Can, ZHAO Kun, MAKSOUD S, et al. SID: Incremental learning for anchor-free object detection via Selective and Inter-related Distillation[J]. *Computer Vision and Image Understanding*, 2021, 210: 103229. doi: [10.1016/j.cviu.2021.103229](https://doi.org/10.1016/j.cviu.2021.103229).
- [46] 商迪, 吕彦锋, 乔红. 受人脑中记忆机制启发的增量目标检测方法[J]. 计算机科学, 2023, 50(2): 267–274. doi: [10.11896/jsjkx.220900212](https://doi.org/10.11896/jsjkx.220900212).
- SHANG Di, LYU Yanfeng, and QIAO Hong. Incremental object detection inspired by memory mechanisms in brain[J]. *Computer Science*, 2023, 50(2): 267–274. doi: [10.11896/jsjkx.220900212](https://doi.org/10.11896/jsjkx.220900212).
- [47] CERMELLI F, MANCINI M, BULÒ S R, et al. Modeling the background for incremental learning in semantic segmentation[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 9230–9239. doi: [10.1109/CVPR42600.2020.00925](https://doi.org/10.1109/CVPR42600.2020.00925).
- [48] DOUILLARD A, CHEN Yifu, DAPOGNY A, et al. PLOP: Learning without forgetting for continual semantic segmentation[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 4039–4049. doi: [10.1109/CVPR46437.2021.00403](https://doi.org/10.1109/CVPR46437.2021.00403).
- [49] YAN Shipeng, ZHOU Jiale, XIE Jiangwei, et al. An EM framework for online incremental learning of semantic segmentation[C]. Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 2021: 3052–3060. doi: [10.1145/3474085.3475443](https://doi.org/10.1145/3474085.3475443).
- [50] MICHELI U and ZANUTTIGH P. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 1114–1124. doi: [10.1109/CVPR46437.2021.00117](https://doi.org/10.1109/CVPR46437.2021.00117).
- [51] YANG Guanglei, FINI E, XU Dan, et al. Uncertainty-aware contrastive distillation for incremental semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(2): 2567–2581. doi: [10.1109/TPAMI.2022.3163806](https://doi.org/10.1109/TPAMI.2022.3163806).
- [52] ZHAI Mengyao, CHEN Lei, TUNG F, et al. Lifelong GAN: Continual learning for conditional image generation[C]. 2019 IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 2019: 2759–2768. doi: [10.1109/ICCV.2019.00285](https://doi.org/10.1109/ICCV.2019.00285).
- [53] Zajac M, Deja K, Kuzina A, et al. Exploring continual learning of diffusion models[J]. arxiv:2303.15342, 2023. doi:10.48550/arXiv.2303.15342.
- [54] ZHAI Mengyao, CHEN Lei, HE Jiawei, et al. Piggyback GAN: Efficient lifelong learning for image conditioned generation[C]. The 17th European Conference on Computer Vision, Glasgow, UK, 2020: 397–413. doi: [10.1007/978-3-030-58589-1_24](https://doi.org/10.1007/978-3-030-58589-1_24).
- [55] WANG Liyuan, YANG Kuo, LI Chongxuan, et al. ORDisCo: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 5379–5388. doi: [10.1109/CVPR46437.2021.00534](https://doi.org/10.1109/CVPR46437.2021.00534).
- [56] YOON J, HWANG S J, and CAO Yue. Continual learners are incremental model generalizers[C]. 40th International Conference on Machine Learning, Honolulu, USA, 2023: 40129–40146.
- [57] HU Dapeng, YAN Shipeng, LU Qizhengqiu, et al. How well does self-supervised pre-training perform with streaming data?[C]. Tenth International Conference on Learning Representations, 2022.
- [58] COSSU A, CARTA A, PASSARO L, et al. Continual pre-training mitigates forgetting in language and vision[J]. *Neural Networks*, 2024, 179: 106492. doi: [10.1016/j.neunet.2024.106492](https://doi.org/10.1016/j.neunet.2024.106492).
- [59] CARUANA R. Multitask learning[J]. *Machine Learning*, 1997, 28(1): 41–75. doi: [10.1023/A:1007379606734](https://doi.org/10.1023/A:1007379606734).
- [60] HOSPEDALES T, ANTONIOU A, MICAELLI P, et al. Meta-learning in neural networks: A survey[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(9): 5149–5169. doi: [10.1109/TPAMI.2021.3079209](https://doi.org/10.1109/TPAMI.2021.3079209).
- [61] WEISS K, KHOSHGOFTAAR T M, and WANG Dingding. A survey of transfer learning[J]. *Journal of Big data*, 2016, 3(1): 9. doi: [10.1186/s40537-016-0043-6](https://doi.org/10.1186/s40537-016-0043-6).
- [62] PATEL V M, GOPALAN R, LI Ruonan, et al. Visual domain adaptation: A survey of recent advances[J]. *IEEE Signal Processing Magazine*, 2015, 32(3): 53–69. doi: [10.1109/MSP.2014.2347059](https://doi.org/10.1109/MSP.2014.2347059).
- [63] WANG Jindong, LAN Cuiling, LIU Chang, et al. Generalizing to unseen domains: A survey on domain generalization[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 35(8): 8052–8072. doi: [10.1109/TKDE.2022.3178128](https://doi.org/10.1109/TKDE.2022.3178128).
- [64] REBUFFI S A, KOLESNIKOV A, SPERL G, et al. iCaRL: Incremental classifier and representation learning[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 2017: 5533–5542. doi: [10.1109/CVPR.2017.587](https://doi.org/10.1109/CVPR.2017.587).

- [65] HUSZÁR F. Note on the quadratic penalties in elastic weight consolidation[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2018, 115(11): E2496–E2497. doi: [10.1073/pnas.1717042115](https://doi.org/10.1073/pnas.1717042115).
- [66] LIU Xialei, MASANA M, HERRANZ L, et al. Rotate your networks: Better weight consolidation and less catastrophic forgetting[C]. The 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 2018: 2262–2268. doi: [10.1109/ICPR.2018.8545895](https://doi.org/10.1109/ICPR.2018.8545895).
- [67] RITTER H, BOTEV A, and BARBER D. Online structured Laplace approximations for overcoming catastrophic forgetting[C]. The 32nd International Conference on Neural Information Processing Systems (NIPS), Montreal, Canada, 2018: 3742–3752.
- [68] ZENKE F, POOLE B, and GANGULI S. Continual learning through synaptic intelligence[C]. The 34th International Conference on Machine Learning (ICML), Sydney, Australia, 2017: 3987–3995.
- [69] ALJUNDI R, BABILONI F, ELHOSEINY M, et al. Memory aware synapses: Learning what (not) to forget[C]. The 15th European Conference on Computer Vision (ECCV), Munich, Germany, 2018: 144–161. doi: [10.1007/978-3-030-01219-9_9](https://doi.org/10.1007/978-3-030-01219-9_9).
- [70] CHAUDHRY A, DOKANIA P K, AJANTHAN T, et al. Riemannian walk for incremental learning: Understanding forgetting and intransigence[C]. The 15th European Conference on Computer Vision (ECCV), Munich, Germany, 2018: 556–572. doi: [10.1007/978-3-030-01252-6_33](https://doi.org/10.1007/978-3-030-01252-6_33).
- [71] LEE S W, KIM J H, JUN J, et al. Overcoming catastrophic forgetting by incremental moment matching[C]. The 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach, USA, 2017: 4655–4665.
- [72] BENZING F. Unifying importance based regularisation methods for continual learning[C]. The 25th International Conference on Artificial Intelligence and Statistics (ICAIS), 2022: 2372–2396.
- [73] HINTON G, VINYALS O, and DEAN J. Distilling the knowledge in a neural network[J]. arXiv: 1503.02531, 2015. doi: [10.48550/arXiv.1503.02531](https://doi.org/10.48550/arXiv.1503.02531).
- [74] LI Zhizhong and HOIEM D. Learning without forgetting[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(12): 2935–2947. doi: [10.1109/TPAMI.2017.2773081](https://doi.org/10.1109/TPAMI.2017.2773081).
- [75] HOU Saihui, PAN Xinyu, LOY C C, et al. Learning a unified classifier incrementally via rebalancing[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019: 831–839. doi: [10.1109/CVPR.2019.00092](https://doi.org/10.1109/CVPR.2019.00092).
- [76] DHAR P, SINGH R V, PENG Kuanchuan, et al. Learning without memorizing[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019: 5133–5141. doi: [10.1109/CVPR.2019.00528](https://doi.org/10.1109/CVPR.2019.00528).
- [77] KANG M, PARK J, and HAN B. Class-incremental learning by knowledge distillation with adaptive feature consolidation[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, 2022: 16050–16059. doi: [10.1109/CVPR52688.2022.01560](https://doi.org/10.1109/CVPR52688.2022.01560).
- [78] DOUILLARD A, CORD M, OLLION C, et al. PODNet: Pooled outputs distillation for small-tasks incremental learning[C]. The 16th European Conference on Computer Vision (ECCV), Glasgow, UK, 2020: 86–102. doi: [10.1007/978-3-030-58565-5_6](https://doi.org/10.1007/978-3-030-58565-5_6).
- [79] SIMON C, KONIUSZ P, and HARANDI M. On learning the geodesic path for incremental learning[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, USA, 2021: 1591–1600. doi: [10.1109/CVPR46437.2021.00164](https://doi.org/10.1109/CVPR46437.2021.00164).
- [80] GAO Qiankun, ZHAO Chen, GHANEM B, et al. R-DFCIL: Relation-guided representation learning for data-free class incremental learning[C]. The 17th European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 2022: 423–439. doi: [10.1007/978-3-031-20050-2_25](https://doi.org/10.1007/978-3-031-20050-2_25).
- [81] TAO Xiaoyu, CHANG Xinyuan, HONG Xiaopeng, et al. Topology-preserving class-incremental learning[C]. The 16th European Conference on Computer Vision (ECCV), Glasgow, UK, 2020: 254–270. doi: [10.1007/978-3-030-58529-7_16](https://doi.org/10.1007/978-3-030-58529-7_16).
- [82] TAO Xiaoyu, HONG Xiaopeng, CHANG Xinyuan, et al. Few-shot class-incremental learning[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2020: 12180–12189. doi: [10.1109/CVPR42600.2020.01220](https://doi.org/10.1109/CVPR42600.2020.01220).
- [83] MARTINETZ T M and SCHULTEN K J. A "neural-gas" network learns topologies[M]. KOHONEN T, MÄKISARA K, SIMULA O, et al. *Artificial Neural Networks*. Amsterdam: North-Holland, 1991: 397–402.
- [84] LIU Yu, HONG Xiaopeng, TAO Xiaoyu, et al. Model behavior preserving for class-incremental learning[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, 34(10): 7529–7540. doi: [10.1109/TNNLS.2022.3144183](https://doi.org/10.1109/TNNLS.2022.3144183).
- [85] ASADI N, DAVARI M R, MUDUR S, et al. Prototype-sample relation distillation: Towards replay-free continual

- learning[C]. The 40th International Conference on Machine Learning, Honolulu, USA, 2023: 1093–1106.
- [86] ARANI E, SARFRAZ F, and ZONOOZ B. Learning fast, learning slow: A general continual learning method based on complementary learning system[C]. The Tenth International Conference on Learning Representations (ICLR), 2022.
- [87] VIJAYAN P, BHAT P, ZONOOZ B, *et al.* TriRE: A multi-mechanism learning paradigm for continual knowledge retention and promotion[C]. The 37th Conference on Neural Information Processing Systems, New Orleans, USA, 2023: 3226.
- [88] JEEVESWARAN K, BHAT P S, ZONOOZ B, *et al.* BiRT: Bio-inspired replay in vision transformers for continual learning[C]. 40th International Conference on Machine Learning, Honolulu, USA, 2023: 14817–14835.
- [89] ZHU Fei, ZHANG Xuyao, WANG Chuang, *et al.* Prototype augmentation and self-supervision for incremental learning[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 5867–5876. doi: [10.1109/CVPR46437.2021.00581](https://doi.org/10.1109/CVPR46437.2021.00581).
- [90] SZATKOWSKI F, PYLA M, PRZEWIĘŻLIKOWSKI M, *et al.* Adapt your teacher: Improving knowledge distillation for exemplar-free continual learning[C]. IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, USA, 2024: 1966–1976. doi: [10.1109/WACV57701.2024.00198](https://doi.org/10.1109/WACV57701.2024.00198).
- [91] LIANG Yanshuo and LI Wujun. Loss decoupling for task-agnostic continual learning[C]. The 37th International Conference on Neural Information Processing Systems (NIPS), New Orleans, USA, 2023: 492.
- [92] WU Yue, CHEN Yinpeng, WANG Lijuan, *et al.* Large scale incremental learning[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 374–382. doi: [10.1109/CVPR.2019.00046](https://doi.org/10.1109/CVPR.2019.00046).
- [93] ZHAO Bowen, XIAO Xi, GAN Guojun, *et al.* Maintaining discrimination and fairness in class incremental learning[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2020: 13205–13214. doi: [10.1109/CVPR42600.2020.01322](https://doi.org/10.1109/CVPR42600.2020.01322).
- [94] GOSWAMI D, LIU Yuyang, TWARDOWSKI B, *et al.* FeCAM: Exploiting the heterogeneity of class distributions in exemplar-free continual learning[C]. The 37th Conference on Neural Information Processing Systems (NIPS), New Orleans, USA, 2023: 288.
- [95] XIANG Xiang, TAN Yuwen, WAN Qian, *et al.* Coarse-to-fine incremental few-shot learning[C]. The 17th European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 2022: 205–222. doi: [10.1007/978-3-031-19821-2_12](https://doi.org/10.1007/978-3-031-19821-2_12).
- [96] AHN H, KWAK J, LIM S, *et al.* SS-IL: Separated softmax for incremental learning[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 824–833. doi: [10.1109/ICCV48922.2021.00088](https://doi.org/10.1109/ICCV48922.2021.00088).
- [97] YANG Yibo, CHEN Shixiang, LI Xiangtai, *et al.* Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network?[C]. The 36th Conference on Neural Information Processing Systems (NIPS), New Orleans, USA, 2022: 2753.
- [98] YANG Yibo, YUAN Haobo, LI Xiangtai, *et al.* Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning[C]. The 11th International Conference on Learning Representations (ICLR), Kigali, Rwanda, 2023.
- [99] LYU Yilin, WANG Liyuan, ZHANG Xingxing, *et al.* Overcoming recency bias of normalization statistics in continual learning: Balance and adaptation[C]. The 37th Conference on Neural Information Processing Systems (NIPS), New Orleans, USA, 2023: 1108.
- [100] GUO Chengcheng, ZHAO Bo, and BAI Yanbing. DeepCore: A comprehensive library for coresnet selection in deep learning[C]. 33rd International Conference on Database and Expert Systems Applications, Vienna, Austria, 2022: 181–195. doi: [10.1007/978-3-031-12423-5_14](https://doi.org/10.1007/978-3-031-12423-5_14).
- [101] FELDMAN D. Introduction to core-sets: An updated survey[J]. arXiv: 2011.09384, 2020. doi: [10.48550/arXiv.2011.09384](https://doi.org/10.48550/arXiv.2011.09384).
- [102] CHEN Yutian, WELLING M, and SMOLA A J. Super-samples from kernel herding[C]. 26th Conference on Uncertainty in Artificial Intelligence, Catalina Island, USA, 2010: 109–116.
- [103] WELLING M. Herding dynamical weights to learn[C]. The 26th Annual International Conference on Machine Learning, Montreal, Canada, 2009: 1121–1128. doi: [10.1145/1553374.1553517](https://doi.org/10.1145/1553374.1553517).
- [104] CHAUDHRY A, ROHRBACH M, ELHOSEINY M, *et al.* On tiny episodic memories in continual learning[J]. arXiv: 1902.10486, 2019. doi: [10.48550/arXiv.1902.10486](https://doi.org/10.48550/arXiv.1902.10486).
- [105] YOON J, MADAAN D, YANG E, *et al.* Online coresnet selection for rehearsal-based continual learning[C]. Tenth International Conference on Learning Representations, 2022.
- [106] ALJUNDI R, CACCIA L, BELILOVSKY E, *et al.* Online continual learning with maximally interfered retrieval[C]. The 33rd International Conference on Neural Information Processing Systems (NIPS), Vancouver, Canada, 2019: 1063.

- [107] BANG J, KIM H, YOO Y J, *et al.* Rainbow memory: Continual learning with a memory of diverse samples[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 8214–8223. doi: [10.1109/CVPR46437.2021.00812](https://doi.org/10.1109/CVPR46437.2021.00812).
- [108] BORSOS Z, MUTNÝ M, and KRAUSE A. Coresets via bilevel optimization for continual learning and Streaming[C]. The 34th International Conference on Neural Information Processing Systems (NIPS), Vancouver, Canada, 2020: 1247.
- [109] ZHOU Xiao, PI Renjie, ZHANG Weizhong, *et al.* Probabilistic bilevel coresnet selection[C]. 39th International Conference on Machine Learning, Baltimore, USA, 2022: 27287–27302.
- [110] TIWARI R, KILLAMSETTY K, IYER R, *et al.* GCR: Gradient coresnet based replay buffer selection for continual learning[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, 2022: 99–108. doi: [10.1109/CVPR52688.2022.00020](https://doi.org/10.1109/CVPR52688.2022.00020).
- [111] HAO Jie, JI Kaiyi, and LIU Mingrui. Bilevel coresnet selection in continual learning: A new formulation and algorithm[C]. The 37th Conference on Neural Information Processing Systems, New Orleans, USA, 2023: 2220.
- [112] WANG Tongzhou, ZHU Junyan, TORRALBA A, *et al.* Dataset distillation[J]. arXiv: 1811.10959, 2018. doi: [10.48550/arXiv.1811.10959](https://doi.org/10.48550/arXiv.1811.10959).
- [113] YU Ruonan, LIU Songhua, and WANG Xinchao. Dataset distillation: A comprehensive review[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(1): 150–170. doi: [10.1109/TPAMI.2023.3323376](https://doi.org/10.1109/TPAMI.2023.3323376).
- [114] LIU Yaoyao, SU Yuting, LIU Anan, *et al.* Mnemonics training: Multi-class incremental learning without forgetting[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 12242–12251. doi: [10.1109/CVPR42600.2020.01226](https://doi.org/10.1109/CVPR42600.2020.01226).
- [115] ZHAO Bo, MOPURI K R, and BILEN H. Dataset condensation with gradient matching[C]. The 9th International Conference on Learning Representations, Austria, 2021.
- [116] ZHAO Bo and BILEN H. Dataset condensation with differentiable Siamese augmentation[C]. The 38th International Conference on Machine Learning, 2021: 12674–12685.
- [117] YANG Enneng, SHEN Li, WANG Zhenyi, *et al.* An efficient dataset condensation plugin and its application to continual learning[C]. The 37th Conference on Neural Information Processing Systems, New Orleans, USA, 2023: 2957.
- [118] CACCIA L, BELILOVSKY E, CACCIA M, *et al.* Online learned continual compression with adaptive quantization modules[C]. The 37th International Conference on Machine Learning, 2020: 1240–1250.
- [119] VAN DEN OORD A, VINYALS O, and KAVUKCUOGLU K. Neural discrete representation learning[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6309–6318.
- [120] WANG Liyuan, ZHANG Xingxing, YANG Kuo, *et al.* Memory replay with data compression for continual learning[C]. Tenth International Conference on Learning Representations, 2022.
- [121] KULESZA A and TASKAR B. Determinantal point processes for machine learning[J]. *Foundations and Trends® in Machine Learning*, 2012, 5(2/3): 123–286. doi: [10.1561/2200000044](https://doi.org/10.1561/2200000044).
- [122] LUO Zilin, LIU Yaoyao, SCHIELE B, *et al.* Class-incremental exemplar compression for class-incremental learning[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 11371–11380. doi: [10.1109/CVPR52729.2023.01094](https://doi.org/10.1109/CVPR52729.2023.01094).
- [123] ZHAI Jiangtian, LIU Xialei, BAGDANOV A D, *et al.* Masked autoencoders are efficient class incremental learners[C]. 2023 IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 19047–19056. doi: [10.1109/ICCV51070.2023.01750](https://doi.org/10.1109/ICCV51070.2023.01750).
- [124] HE Kaiming, CHEN Xinlei, XIE Saining, *et al.* Masked autoencoders are scalable vision learners[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 15979–15988. doi: [10.1109/CVPR52688.2022.01553](https://doi.org/10.1109/CVPR52688.2022.01553).
- [125] ISCEN A, ZHANG J, LAZEBNIK S, *et al.* Memory-efficient incremental learning through feature adaptation[C]. The 16th European Conference on Computer Vision, Glasgow, UK, 2020: 699–715. doi: [10.1007/978-3-030-58517-4_41](https://doi.org/10.1007/978-3-030-58517-4_41).
- [126] BELOUADAH E and POPESCU A. IL2M: Class incremental learning with dual memory[C]. 2019 IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 2019: 583–592. doi: [10.1109/ICCV.2019.00067](https://doi.org/10.1109/ICCV.2019.00067).
- [127] TOLDO M and OZAY M. Bring evanescent representations to life in lifelong class incremental learning[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 16711–16720. doi: [10.1109/CVPR52688.2022.01623](https://doi.org/10.1109/CVPR52688.2022.01623).
- [128] WANG Kai, VAN DE WEIJER J, and HERRANZ L. ACAE-REMIND for online continual learning with compressed feature replay[J]. *Pattern Recognition Letters*,

- 2021, 150: 122–129. doi: [10.1016/j.patrec.2021.06.025](https://doi.org/10.1016/j.patrec.2021.06.025).
- [129] PETIT G, POPESCU A, SCHINDLER H, *et al.* FeTrIL: Feature translation for exemplar-free class-incremental learning[C]. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, USA, 2023: 3900–3909. doi: [10.1109/WACV56688.2023.00390](https://doi.org/10.1109/WACV56688.2023.00390).
- [130] ZHU Fei, CHENG Zhen, ZHANG Xuyao, *et al.* Class-incremental learning via dual augmentation[C]. The 35th International Conference on Neural Information Processing Systems (NIPS), 2021: 1096.
- [131] ZHU Kai, ZHAI Wei, CAO Yang, *et al.* Self-sustaining representation expansion for non-exemplar class-incremental learning[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 9286–9395. doi: [10.1109/CVPR52688.2022.00908](https://doi.org/10.1109/CVPR52688.2022.00908).
- [132] SHI Wuxuan and YE Mang. Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning[C]. 2023 IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 1772–1781. doi: [10.1109/ICCV51070.2023.00170](https://doi.org/10.1109/ICCV51070.2023.00170).
- [133] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, *et al.* Generative adversarial nets[C]. The 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 2672–2680.
- [134] KINGMA D P and WELLING M. Auto-encoding variational Bayes[C]. 2nd International Conference on Learning Representations, Banff, Canada, 2014.
- [135] HO J, JAIN A, and ABBEEL P. Denoising diffusion probabilistic models[C]. The 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 574.
- [136] SHIN H, LEE J K, KIM J, *et al.* Continual learning with deep generative replay[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 2994–3003.
- [137] WU Chenshen, HERRANZ L, LIU Xialei, *et al.* Memory replay GANs: Learning to generate images from new categories without forgetting[C]. The 32nd International Conference on Neural Information Processing Systems, Montreal, Canada, 2018: 5966–5976.
- [138] HE Chen, WANG Ruiping, SHAN Shiguang, *et al.* Exemplar-supported generative reproduction for class incremental learning[C]. British Machine Vision Conference 2018, Newcastle, UK, 2018: 98.
- [139] KEMKER R and KANAN C. FearNet: Brain-inspired model for incremental learning[C]. 6th International Conference on Learning Representations, Vancouver, Canada, 2018: 1–15.
- [140] YE Fei and BORS A G. Learning latent representations across multiple data domains using lifelong VAEGAN[C]. The 16th European Conference on Computer Vision, Glasgow, UK, 2020: 777–795. doi: [10.1007/978-3-030-58565-5_46](https://doi.org/10.1007/978-3-030-58565-5_46).
- [141] GAO Rui and LIU Weiwei. DDGR: Continual learning with deep diffusion-based generative replay[C]. 40th International Conference on Machine Learning, Honolulu, USA, 2023: 10744–10763.
- [142] JODELET Q, LIU Xin, PHUA Y J, *et al.* Class-incremental learning using diffusion model for distillation and replay[C]. 2023 IEEE/CVF International Conference on Computer Vision Workshops, Paris, France, 2023: 3417–3425. doi: [10.1109/ICCVW60793.2023.00367](https://doi.org/10.1109/ICCVW60793.2023.00367).
- [143] XIANG Ye, FU Ying, JI Pan, *et al.* Incremental learning using conditional adversarial networks[C]. 2019 IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 2019: 6618–6627. doi: [10.1109/ICCV.2019.00672](https://doi.org/10.1109/ICCV.2019.00672).
- [144] VAN DE VEN G M, SIEGELMANN H T, and TOLIAS A S. Brain-inspired replay for continual learning with artificial neural networks[J]. *Nature Communications*, 2020, 11(1): 4069. doi: [10.1038/s41467-020-17866-2](https://doi.org/10.1038/s41467-020-17866-2).
- [145] LIU Xialei, WU Chenshen, MENTA M, *et al.* Generative feature replay for class-incremental learning[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, USA, 2020: 915–924. doi: [10.1109/CVPRW50498.2020.00121](https://doi.org/10.1109/CVPRW50498.2020.00121).
- [146] CHAUDHRY A, RANZATO M A, ROHRBACH M, *et al.* Efficient lifelong learning with A-GEM[C]. 7th International Conference on Learning Representations, New Orleans, USA, 2019.
- [147] RIEMER M, CASES I, AJEMIAN R, *et al.* Learning to learn without forgetting by maximizing transfer and minimizing interference[C]. 7th International Conference on Learning Representations, New Orleans, USA, 2019.
- [148] FARAJTABAR M, AZIZAN N, MOTT A, *et al.* Orthogonal gradient descent for continual learning[C]. 23rd International Conference on Artificial Intelligence and Statistics, Palermo, Italy, 2020: 3762–3773.
- [149] TANG Shixiang, CHEN Dapeng, ZHU Jinguo, *et al.* Layerwise optimization by gradient decomposition for continual learning[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 9629–9638. doi: [10.1109/CVPR46437.2021.00051](https://doi.org/10.1109/CVPR46437.2021.00051).
- [150] KAO T C, JENSEN K T, VAN DE VEN G M, *et al.* Natural continual learning: Success is a journey, not (just)

- a destination[C]. The 35th International Conference on Neural Information Processing Systems, 2021: 2150.
- [151] LIU Hao and LIU Huaping. Continual learning with recursive gradient optimization[C]. Tenth International Conference on Learning Representations, 2022.
- [152] SAHA G, GARG I, and ROY K. Gradient projection memory for continual learning[C]. 9th International Conference on Learning Representations, Austria, 2021.
- [153] WANG Shipeng, LI Xiaorong, SUN Jian, *et al.* Training networks in null space of feature covariance for continual learning[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 184–193. doi: [10.1109/CVPR46437.2021.00025](https://doi.org/10.1109/CVPR46437.2021.00025).
- [154] KONG Yajing, LIU Liu, WANG Zhen, *et al.* Balancing stability and plasticity through advanced null space in continual learning[C]. 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 219–236. doi: [10.1007/978-3-031-19809-0_13](https://doi.org/10.1007/978-3-031-19809-0_13).
- [155] LIN Sen, YANG Li, FAN Deliang, *et al.* TRGP: Trust region gradient projection for continual learning[C]. Tenth International Conference on Learning Representations, 2022.
- [156] LIN Sen, YANG Li, FAN Deliang, *et al.* Beyond not-forgetting: Continual learning with backward knowledge transfer[C]. The 36th Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 1176.
- [157] HOCHREITER S and SCHMIDHUBER J. Flat minima[J]. *Neural Computation*, 1997, 9(1): 1–42. doi: [10.1162/neco.1997.9.1.1](https://doi.org/10.1162/neco.1997.9.1.1).
- [158] KESKAR N S, MUDIGERE D, NOCEDAL J, *et al.* On large-batch training for deep learning: Generalization gap and sharp minima[C]. 5th International Conference on Learning Representations, Toulon, France, 2017.
- [159] FORET P, KLEINER A, MOBAHI H, *et al.* Sharpness-aware minimization for efficiently improving generalization[C]. 9th International Conference on Learning Representations, Austria, 2021.
- [160] HUANG Zhongzhan, LIANG Mingfu, LIANG Senwei, *et al.* AlterSGD: Finding flat minima for continual learning by alternative training[J]. arXiv: 2107.05804, 2021. doi: [10.48550/arXiv.2107.05804](https://doi.org/10.48550/arXiv.2107.05804).
- [161] SHI Guangyuan, CHEN Jiaxin, ZHANG Wenlong, *et al.* Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima[C]. The 35th Conference on Neural Information Processing Systems, 2021: 517.
- [162] DENG Danruo, CHEN Guangyong, HAO Jianye, *et al.* Flattening sharpness for dynamic gradient projection memory benefits continual learning[C]. 35th International Conference on Neural Information Processing System, 2021: 1430.
- [163] LIU Yong, MAI Siqi, CHEN Xiangning, *et al.* Towards efficient and scalable sharpness-aware minimization[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 12350–12360. doi: [10.1109/CVPR52688.2022.01204](https://doi.org/10.1109/CVPR52688.2022.01204).
- [164] WU Tao, LUO Tie, and WUNSCH II D C. CR-SAM: Curvature regularized sharpness-aware minimization[C]. Proceedings of the 38th AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2024: 6144–6152. doi: [10.1609/aaai.v38i6.28431](https://doi.org/10.1609/aaai.v38i6.28431).
- [165] MADAAN D, YOON J, LI Yuanchun, *et al.* Representational continuity for unsupervised continual learning[C]. Tenth International Conference on Learning Representations, 2022.
- [166] CHA H, LEE J, and SHIN J. Co²L: Contrastive continual learning[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 9496–9505. doi: [10.1109/ICCV48922.2021.00938](https://doi.org/10.1109/ICCV48922.2021.00938).
- [167] FINI E, DA COSTA V G T, ALAMEDA-PINEDA X, *et al.* Self-supervised models are continual learners[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 9611–9620. doi: [10.1109/CVPR52688.2022.00940](https://doi.org/10.1109/CVPR52688.2022.00940).
- [168] AHN H, CHA S, LEE D, *et al.* Uncertainty-based continual learning with adaptive regularization[C]. The 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 2019: 395.
- [169] GURBUZ M B and DOVROLIS C. NISPA: Neuro-inspired stability-plasticity adaptation for continual learning in sparse networks[C]. The 39th International Conference on Machine Learning, Baltimore, USA, 2022: 8157–8174.
- [170] JIN H and KIM E. Helpful or harmful: Inter-task association in continual learning[C]. The 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 519–535. doi: [10.1007/978-3-031-20083-0_31](https://doi.org/10.1007/978-3-031-20083-0_31).
- [171] XUE Mengqi, ZHANG Haofei, SONG Jie, *et al.* Meta-attention for ViT-backed continual learning[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 150–159. doi: [10.1109/CVPR52688.2022.00025](https://doi.org/10.1109/CVPR52688.2022.00025).
- [172] JANG E, GU Shixiang, and POOLE B. Categorical reparameterization with gumbel-softmax[C]. 5th International Conference on Learning Representations, Toulon, France, 2017.
- [173] SERRÀ J, SURIS D, MIRON M, *et al.* Overcoming catastrophic forgetting with hard attention to the task[C]. 35th International Conference on Machine Learning, Stockholm, Sweden, 2018: 4548–4557.

- [174] WORTSMAN M, RAMANUJAN V, LIU R, *et al.* Supermasks in superposition[C]. The 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 1272.
- [175] KANG H, MINA R J L, MADJID S R H, *et al.* Forget-free continual learning with winning subnetworks[C]. 39th International Conference on Machine Learning, Baltimore, USA, 2022: 10734–10750.
- [176] YOON J, YANG E, LEE J, *et al.* Lifelong learning with dynamically expandable networks[C]. 6th International Conference on Learning Representations, Vancouver, Canada, 2018: 1–15.
- [177] XU Ju and ZHU Zhanxing. Reinforced continual learning[C]. The 32nd International Conference on Neural Information Processing Systems, Montreal, Canada, 2018: 907–916.
- [178] RUSU A A, RABINOWITZ N C, DESJARDINS G, *et al.* Progressive neural networks[J]. arXiv: 1606.04671, 2016. doi: [10.48550/arXiv.1606.04671](https://doi.org/10.48550/arXiv.1606.04671).
- [179] FERNANDO C, BANARSE D, BLUNDELL C, *et al.* PathNet: Evolution channels gradient descent in super neural networks[J]. arXiv: 1701.08734, 2017. doi: [10.48550/arXiv.1701.08734](https://doi.org/10.48550/arXiv.1701.08734).
- [180] RAJASEGARAN J, HAYAT M, KHAN S, *et al.* Random path selection for incremental learning[J]. arXiv: 1906.01120, 2019. doi: [10.48550/arXiv.1906.01120](https://doi.org/10.48550/arXiv.1906.01120).
- [181] ALJUNDI R, CHAKRAVARTY P, and TUYTELAARS T. Expert gate: Lifelong learning with a network of experts[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 7120–7129. doi: [10.1109/CVPR.2017.753](https://doi.org/10.1109/CVPR.2017.753).
- [182] WANG Fuyun, ZHOU Dawei, YE Hanjia, *et al.* FOSTER: Feature boosting and compression for class-incremental learning[C]. The 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 398–414. doi: [10.1007/978-3-031-19806-9_23](https://doi.org/10.1007/978-3-031-19806-9_23).
- [183] ZHOU Dawei, WANG Qiwei, YE Hanjia, *et al.* A model or 603 exemplars: Towards memory-efficient class-incremental learning[C]. The 11th International Conference on Learning Representations, Kigali, Rwanda, 2023.
- [184] WANG Zifeng, ZHANG Zizhao, EBRAHIMI S, *et al.* DualPrompt: Complementary prompting for rehearsal-free continual learning[C]. The 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 631–648. doi: [10.1007/978-3-031-19809-0_36](https://doi.org/10.1007/978-3-031-19809-0_36).
- [185] SMITH J S, KARLINSKY L, GUTTA V, *et al.* CODA-Prompt: COntinual decomposed attention-based prompting for rehearsal-free continual learning[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 11909–11919. doi: [10.1109/CVPR52729.2023.01146](https://doi.org/10.1109/CVPR52729.2023.01146).
- [186] GAO Qiankun, ZHAO Chen, SUN Yifan, *et al.* A unified continual learning framework with general parameter-efficient tuning[C]. IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023: 11449–11459. doi: [10.1109/ICCV51070.2023.01055](https://doi.org/10.1109/ICCV51070.2023.01055).
- [187] ZHOU Dawei, CAI Ziwen, YE Hanjia, *et al.* Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need[J]. arXiv: 2303.07338, 2023. doi: [10.48550/arXiv.2303.07338](https://doi.org/10.48550/arXiv.2303.07338).
- [188] WANG Yabin, MA Zhiheng, HUANG Zhiwu, *et al.* Isolation and impartial aggregation: A paradigm of incremental learning without interference[C]. Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, USA, 2023: 10209–10217. doi: [10.1609/aaai.v37i8.26216](https://doi.org/10.1609/aaai.v37i8.26216).
- [189] WANG Liyuan, XIE Jingyi, ZHANG Xingxing, *et al.* Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality[C]. The 37th Conference on Neural Information Processing Systems, New Orleans, USA, 2023: 3022.
- [190] WANG Yabin, HUANG Zhiwu, and HONG Xiaopeng. S-prompts learning with pre-trained transformers: An Occam’s razor for domain incremental learning[C]. The 36th Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 411.
- [191] RADFORD A, KIM J W, HALLACY C, *et al.* Learning transferable visual models from natural language supervision[C]. 38th International Conference on Machine Learning, 2021: 8748–8763.
- [192] ZHOU Dawei, ZHANG Yuanhan, NING Jingyi, *et al.* Learning without forgetting for vision-language models[J]. arXiv: 2305.19270, 2023. doi: [10.48550/arXiv.2305.19270](https://doi.org/10.48550/arXiv.2305.19270).
- [193] KHATTAK M U, WASIM S T, NASEER M, *et al.* Self-regulating prompts: Foundational model adaptation without forgetting[C]. 2023 IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 15144–15154. doi: [10.1109/ICCV51070.2023.01394](https://doi.org/10.1109/ICCV51070.2023.01394).
- [194] KIM G, XIAO Changnan, KONISHI T, *et al.* Learnability and algorithm for continual learning[C]. 40th International Conference on Machine Learning, Honolulu, USA, 2023: 16877–16896.
- [195] TANG Yuming, PENG Yixing, and ZHENG Weishi. When prompt-based incremental learning does not meet strong pretraining[C]. 2023 IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 1706–1716. doi: [10.1109/ICCV51070.2023.00164](https://doi.org/10.1109/ICCV51070.2023.00164).
- [196] NGUYEN C V, LI Yingzhen, BUI T D, *et al.* Variational continual learning[C]. 6th International Conference on Learning Representations, Vancouver, Canada, 2018.

- [197] KAPOOR S, KARALETOS T, and BUI T D. Variational auto-regressive Gaussian processes for continual learning[C]. 38th International Conference on Machine Learning, 2021: 5290–5300.
- [198] RAMESH R and CHAUDHARI P. Model zoo: A growing brain that learns continually[C]. Tenth International Conference on Learning Representations, 2022.
- [199] WANG Liyuan, ZHANG Xingxing, LI Qian, *et al.* CoSCL: Cooperation of small continual learners is stronger than a big one[C]. 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 254–271. doi: [10.1007/978-3-031-19809-0_15](https://doi.org/10.1007/978-3-031-19809-0_15).
- [200] YE Fei and BORS A G. Task-free continual learning via online discrepancy distance learning[C]. The 36th International Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 1720.
- [201] SHI Haizhou and WANG Hao. A unified approach to domain incremental learning with memory: Theory and algorithm[C]. Thirty-Seventh Conference on Neural Information Processing Systems, New Orleans, USA, 2023: 660.
- [202] BEN-DAVID S, BLITZER J, CRAMMER K, *et al.* A theory of learning from different domains[J]. *Machine Learning*, 2010, 79(1/2): 151–175. doi: [10.1007/s10994-009-5152-4](https://doi.org/10.1007/s10994-009-5152-4).
- [203] JACOT A, GABRIEL F, and HONGLER C. Neural tangent kernel: Convergence and generalization in neural networks[C]. 32nd International Conference on Neural Information Processing Systems, Montreal, Canada, 2018: 8580–8589.
- [204] BENNANI M A, DOAN T, and SUGIYAMA M. Generalisation guarantees for continual learning with orthogonal gradient descent[J]. arXiv: 2006.11942, 2020. doi: [10.48550/arXiv.2006.11942](https://doi.org/10.48550/arXiv.2006.11942).
- [205] DOAN T, BENNANI M A, MAZOURE B, *et al.* A theoretical analysis of catastrophic forgetting through the NTK overlap matrix[C]. 24th International Conference on Artificial Intelligence and Statistics, 2021: 1072–1080.
- [206] KARAKIDA R and AKAHO S. Learning curves for continual learning in neural networks: Self-knowledge transfer and forgetting[C]. Tenth International Conference on Learning Representations, 2022.
- [207] EVRON I, MOROSHKO E, WARD R A, *et al.* How catastrophic can catastrophic forgetting be in linear regression?[C]. 35th Conference on Learning Theory, London, UK, 2022: 4028–4079.
- [208] LIN Sen, JU Peizhong, LIANG Yingbin, *et al.* Theory on forgetting and generalization of continual learning[C]. 40th International Conference on Machine Learning, Honolulu, USA, 2023: 21078–21100.
- [209] GOLDFARB D and HAND P. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime[C]. 26th International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 2023: 2975–2993.
- [210] KIM G, XIAO Changnan, KONISHI T, *et al.* A theoretical study on solving continual learning[C]. Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 366.
- [211] KIM G, LIU Bing, and KE Zixuan. A multi-head model for continual learning via out-of-distribution replay[C]. 1st Conference on Lifelong Learning Agents, Montreal, Canada, 2022: 548–563.
- [212] LECUN Y, BOTTOU L, BENGIO Y, *et al.* Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11): 2278–2324. doi: [10.1109/5.726791](https://doi.org/10.1109/5.726791).
- [213] KRIZHEVSKY A and HINTON G. Learning multiple layers of features from tiny images[J]. *Handbook of Systemic Autoimmune Diseases*, 2009, 1(4): 1–60.
- [214] WAH C, BRANSON S, WELINDER P, *et al.* The Caltech-UCSD birds-200-2011 dataset[R]. CNS-TR-2010-001, 2011.
- [215] LE Ya and YANG Xuan. Tiny ImageNet visual recognition challenge[J]. *CS 231N*, 2015, 7(7): 3.
- [216] DENG Jia, DONG Wei, SOCHER R, *et al.* ImageNet: A large-scale hierarchical image database[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 248–255. doi: [10.1109/CVPR.2009.5206848](https://doi.org/10.1109/CVPR.2009.5206848).
- [217] EBRAHIMI S, MEIER F, CALANDRA R, *et al.* Adversarial continual learning[C]. The 16th European Conference on Computer Vision, Glasgow, UK, 2020: 386–402. doi: [10.1007/978-3-030-58621-8_23](https://doi.org/10.1007/978-3-030-58621-8_23).
- [218] LOMONACO V and MALTONI D. CORe50: A new dataset and benchmark for continuous object recognition[C]. 1st Annual Conference on Robot Learning, Mountain View, USA, 2017: 17–26.
- [219] PENG Xingchao, BAI Qinxun, XIA Xide, *et al.* Moment matching for multi-source domain adaptation[C]. 2019 IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 2019: 1406–1415. doi: [10.1109/ICCV.2019.00149](https://doi.org/10.1109/ICCV.2019.00149).
- [220] LI Chuqiao, HUANG Zhiwu, PAUDEL D P, *et al.* A continual deepfake detection benchmark: Dataset, methods, and essentials[C]. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, USA, 2023: 1339–1349. doi: [10.1109/WACV56688.2023.00139](https://doi.org/10.1109/WACV56688.2023.00139).
- [221] XIAO Han, RASUL K, and VOLLMGRAF R. Fashion-MNIST: A novel image dataset for benchmarking machine

- learning algorithms[J]. arXiv: 1708.07747, 2017. doi: [10.48550/arXiv.1708.07747](https://doi.org/10.48550/arXiv.1708.07747).
- [222] NETZER Y, WANG Tao, COATES A, *et al.* Reading digits in natural images with unsupervised feature learning[C]. Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning, Granada, Spain, 2011.
- [223] BULATOV Y. Notmnist dataset[EB/OL]. <http://yaroslavvb.blogspot.it/2011/09/notmnist-dataset.html>, 2011.
- [224] YANG Yuwei, HAYAT M, JIN Zhao, *et al.* Geometry and uncertainty-aware 3D point cloud class-incremental semantic segmentation[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 21759–21768. doi: [10.1109/CVPR52729.2023.02084](https://doi.org/10.1109/CVPR52729.2023.02084).
- [225] CAMUFFO E and MILANI S. Continual learning for LiDAR semantic segmentation: Class-incremental and coarse-to-fine strategies on sparse data[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 2447–2456. doi: [10.1109/CVPRW59228.2023.00243](https://doi.org/10.1109/CVPRW59228.2023.00243).
- [226] CASTAGNOLO G, SPAMPINATO C, RUNDO F, *et al.* A baseline on continual learning methods for video action recognition[C]. IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 2023: 3240–3244. doi: [10.1109/ICIP49359.2023.10222140](https://doi.org/10.1109/ICIP49359.2023.10222140).
- [227] NAQUSHBANDI F S and JOHN A. Sequence of actions recognition using continual learning[C]. 2022 Second International Conference on Artificial Intelligence and Smart Energy, Coimbatore, India, 2022: 858–863. doi: [10.1109/ICAIS53314.2022.9742866](https://doi.org/10.1109/ICAIS53314.2022.9742866).
- [228] LI Dingcheng, CHEN Zheng, CHO E, *et al.* Overcoming catastrophic forgetting during domain adaptation of Seq2seq language generation[C]. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, USA, 2022: 5441–5454. doi: [10.18653/v1/2022.naacl-main.398](https://doi.org/10.18653/v1/2022.naacl-main.398).
- [229] MONAIKUL N, CASTELLUCCI G, FILICE S, *et al.* Continual learning for named entity recognition[C]. Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021: 13570–13577. doi: [10.1609/aaai.v35i15.17600](https://doi.org/10.1609/aaai.v35i15.17600).
- [230] LIU Qingbin, YU Xiaoyan, HE Shizhu, *et al.* Lifelong intent detection via multi-strategy rebalancing[J]. arXiv: 2108.04445, 2021. doi: [10.48550/arXiv.2108.04445](https://doi.org/10.48550/arXiv.2108.04445).
- [231] MARACANI A, MICHELI U, TOLDO M, *et al.* RECALL: Replay-based continual learning in semantic segmentation[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 7006–7015. doi: [10.1109/ICCV48922.2021.00694](https://doi.org/10.1109/ICCV48922.2021.00694).
- [232] WANG Rui, YU Tong, ZHAO Handong, *et al.* Few-shot class-incremental learning for named entity recognition[C]. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 2022: 571–582. doi: [10.18653/v1/2022.acl-long.43](https://doi.org/10.18653/v1/2022.acl-long.43).
- [233] GENG Binzong, YUAN Fajie, XU Qiancheng, *et al.* Continual learning for task-oriented dialogue system with iterative network pruning, expanding and masking[C]. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 517–523. doi: [10.18653/v1/2021.acl-short.66](https://doi.org/10.18653/v1/2021.acl-short.66).
- [234] CHEN Wuyang, ZHOU Yanqi, DU Nan, *et al.* Lifelong language PRETRAINING with distribution-specialized experts[C]. 40th International Conference on Machine Learning, Honolulu, USA, 2023: 5383–5395.
- [235] LUO Yun, YANG Zhen, MENG Fandong, *et al.* An empirical study of catastrophic forgetting in large language models during continual fine-tuning[J]. arXiv: 2308.08747, 2023. doi: [10.48550/arXiv.2308.08747](https://doi.org/10.48550/arXiv.2308.08747).
- [236] QI Xiangyu, ZENG Yi, XIE Tinghao, *et al.* Fine-tuning aligned language models compromises safety, even when users do not intend to![C]. Twelfth International Conference on Learning Representations, Vienna, Austria, 2024.
- [237] SMITH J S, HSU Y C, ZHANG Lingyu, *et al.* Continual diffusion: Continual customization of text-to-image diffusion with C-LoRA[J]. arXiv: 2304.06027, 2023. doi: [10.48550/arXiv.2304.06027](https://doi.org/10.48550/arXiv.2304.06027).
- [238] YANG Xin, YU Hao, GAO Xin, *et al.* Federated continual learning via knowledge fusion: A survey[J]. arXiv: 2312.16475, 2023. doi: [10.48550/arXiv.2312.16475](https://doi.org/10.48550/arXiv.2312.16475).
- [239] LIU Xialei, HU Yusong, CAO Xusheng, *et al.* Long-tailed class incremental learning[C]. 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 495–512. doi: [10.1007/978-3-031-19827-4_29](https://doi.org/10.1007/978-3-031-19827-4_29).

张东阳：男，博士生，研究方向为持续学习。

陆子轩：男，硕士生，研究方向为持续学习。

刘军民：男，教授，研究方向为多源图像融合与目标检测研究、深度学习的泛化性与可解释性研究。

李澜宇：男，博士，研究方向为智能遥感。

责任编辑：马秀强