中国科学:信息科学 2026 年 第 56 卷 第 7 期: 1601–1619
SCIENTIA SINICA Informationis
论文

基于多类混合适配的细粒度视觉识别统一模型

卢志颖, 刘传彬, 谢洪涛, 张勇东*

中国科学技术大学信息科学技术学院, 合肥 230026
* 通信作者. E-mail: zhyd73@ustc.edu.cn

收稿日期: 2024–12–26; 修回日期: 2025–07–31; 接受日期: 2026–01–12; 网络出版日期: 2026–06–17

国家重点研发计划 (批准号: 2022YFB3104700) 和国家自然科学基金 (批准号: 62425114, 62121002, U23B2028) 资助项目

摘要 细粒度视觉识别任务旨在区分每个特定大类别下高度相似的子类别. 受制于不同大类别间的差异性, 当前研究方法需要对不同大类训练单独的模型, 缺乏一个能对所有大类中所有子类别进行识别的统一模型. 通过简单合并数据集训练统一模型需要大量的计算成本, 且模型缺乏对未见大类的迁移能力, 导致模型的实际应用受限. 本文提出了一个多类混合适配的细粒度视觉识别统一模型 MoCA (mixture of category adapters), 实现高精度、计算轻量、可扩展的统一识别. 具体而言, 该模型通过冻结的预训练模型通用知识, 结合轻量化的单类适配器进行大类知识的学习, 实现每个大类上的高精度识别; 再通过本文设计的多类混合适配策略 (mixture of adapters, MoA), 将多类知识独立且完备地与通用知识融合, 从而实现多类统一的细粒度视觉识别. 此外, 本文模型具有可扩展性, 能够在保持原有知识的同时, 通过插入适配器的方式, 持续扩展到新的识别任务. 实验结果充分表明, 本文模型在识别精度、计算开销和任务可扩展性方面具有显著优势, 为细粒度视觉识别统一模型的理论和方法提供了新的基准和参照.

关键词 细粒度视觉识别, 统一模型, 预训练, 混合专家模型, 微调

1 引言

细粒度视觉识别 (fine-grained visual categorization, FGVC) 是计算机视觉领域中一个日益重要且富有挑战性的研究方向, 其目标是在每个特定大类别中准确区分视觉上高度相似的子类别, 这一任务远比传统的粗粒度大类识别复杂 [1]. 例如, 粗粒度识别只需要区分出物体属于鸟、汽车、飞机中的哪一个大类, 而 FGVC 需要模型能够在鸟类中区分出具体子类, 在汽车和飞机中区分出具体品牌型号等 [2]. 由于不同大类之间的差异较大, 模型学习到的知识难以共享, 例如识别鸟类的模型与识别汽车的模型无法兼容, 不同大类的细粒度视觉识别模型难以统一, 因此需要在每个特定大类上单独训练模型, 再将多个特定大类模型构成模型库, 造成了模型参数的冗余和存储空间的浪费, 如图 1(a) 所示. 此外, 单独训练模型虽然能够在每个大类上实现较高的识别精度, 但缺少统一的 FGVC 模型也带来了一系列的问题: 首先, 每个模型都需要单独的存储空间和计算资源, 随着需要识别的大类的增加, 单独训练和维护大量特定大类模型变得越来越困难和耗时. 同时, 当需要识别新的大类时, 往往需

引用格式: 卢志颖, 刘传彬, 谢洪涛, 等. 基于多类混合适配的细粒度视觉识别统一模型. 中国科学: 信息科学, 2026, 56: 1601–1619, doi: 10.1360/SSI-2024-0389
Lu Z Y, Liu C B, Xie H T, et al. A unified model for fine-grained visual categorization with mixture of category adapters. Sci Sin Inform, 2026, 56: 1601–1619, doi: 10.1360/SSI-2024-0389

© 2026《中国科学》杂志社 www.scichina.com infocn.scichina.com
https://www.sciengine.com/doi/10.1360/SSI-2024-0389

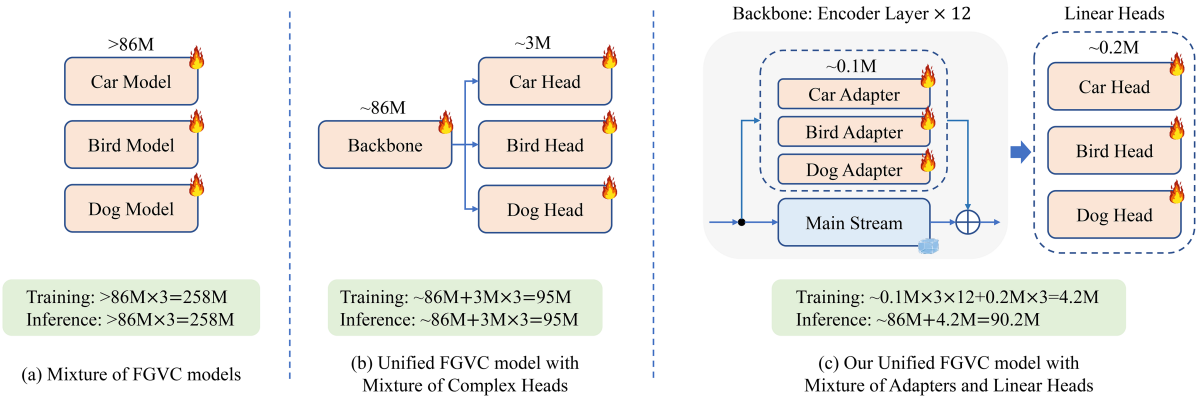


图 1　(网络版彩图) 细粒度视觉识别统一模型框架对比 (以 ViT 为基础模型). (a) 多个大类模型的集合; (b) 全参数微调的统一模型, 训练基础模型和复杂设计的头的集合; (c) 本文提出的统一模型框架, 冻结基础模型中的主干部分, 训练多个适配器和线性头的集合, 训练参数量和最终的推理参数量更低.

Figure 1　(Color online) Comparison of unified FGVC frameworks (based on ViT). (a) Mixture of FGVC models on different general categories; (b) full-finetuned unified FGVC model with mixture of complex heads; (c) our unified FGVC model with mixture of adapters and linear heads, which requires less cost for training and inference.

要从头开始训练新模型, 难以有效利用已有模型的知识. 另外, 在实际场景中可能需要同时识别多个特定大类的对象, 使用多个独立模型会大大增加系统的复杂度.

针对以上问题, FGVC 研究人员开始探索更加高效、通用的 FGVC 解决方案, 旨在克服传统独立模型方法的局限性, 努力构建能够处理多个大类的统一 FGVC 模型. Chang 等人 [3] 提出的方法是构建细粒度视觉识别统一模型的首次尝试. 如图 1(b) 所示, 该方法探讨了多个大类数据集混合训练下, 各个大类相互作用对精度的影响, 并设计了空间注意力模块、特征解耦模块、特征融合模块和元学习策略等方法, 初步实现了 4 个大类统一识别的模型. 然而, 该方法存在下列问题: 该方法只采用一种预训练模型 ResNet50 [4], 缺乏对更广泛预训练模型的实验; 该方法训练时需要微调网络全部参数, 计算消耗大; 训练好的模型仅能实现 4 个大类上的识别, 无法进一步扩展. 由于上述缺陷的存在, 该方法构建的统一模型与实际需求仍存在距离. 在基础模型微调研究中, LoRA 系列 [5,6] 和 Adapter 系列 [7,8] 方法为高效参数微调提供了新思路, 使得具有通用知识的预训练模型仅通过微调少量参数就能学习到下游任务的知识. 另一方面, 混合专家模型是一种常见的多域知识融合策略, 使得单一模型显式地学习到不同领域的知识 [9], 或者同一领域知识的不同部分 [10]. 然而, 细粒度视觉识别领域通常对模型采用全参数微调和单一大类训练的策略, 缺乏对参数高效微调和多类知识融合的探索.

为了进一步构建识别精度高、计算轻量化、知识可扩展的细粒度视觉识别统一模型, 本文提出基于多类混合适配的细粒度视觉识别统一模型 mixture of category adapters (MoCA), 其框架图如图 1(c). 首先, MoCA 模型以预训练的视觉 Transformer 模型 (vision Transformer, ViT) [11] 架构为基础模型, 在训练过程中通过冻结参数维持其通用知识. 为选取最优的基础模型, 我们系统性地评估了 6 种主流视觉预训练模型在 4 种不同微调策略下的性能, 从而筛选出最适合作为 FGVC 统一模型通用知识库的预训练模型. 其次, 本文采用单类适配器 (single category adapter, SCA) 进行每个大类知识的学习, 将其与主干网络 (ViT 的自注意力和多层感知机) 特征相加后实现通用知识与大类知识. SCA 的参数量仅有基础模型参数量的不足百分之一, 可以在每个大类中实现计算轻量化的高精度识别. 本文设计了多类混合策略 (mixture of adapters, MoA), 将训练完成的多个 SCA 模块通过路由模块 (router, RT) 进行权重分配, 加权求和后与主干特征相结合, 实现多类知识与通用知识的融合, 初步实现细粒度视觉识别统一模型. 最后, 本文模型采用两阶段的训练法, 在每个大类上达到最高精度后再进行多类知识融合, 保证融合后的统一模型在每个大类上的精度. 同时为了实现多类知识独立且完备的融合, 本文设计了辅助损失函数对 RT 模块进行约束. 本文模型多类知识融合的方法允许模型不断添加新的 SCA, 赋予统一模型对新的大类的识别能力, 从而不断扩展模型的识别任务的场景.

本文最终构建了融合 6 个大类的细粒度视觉识别统一模型. 由于 SCA 具有参数高效性, 本文方法训练和推

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

理的参数规模都远小于基线方法, 并且在低成本训练的情况下达到了与传统方法相当的精度. 大量的消融实验证明了本文方法的可行性与有效性, 为细粒度视觉识别领域的研究设计提供全新思路. 综上所述, 本文提出的多类混合适配的细粒度视觉识别统一模型, 以其识别高精度、计算轻量化、知识易扩展的优势, 为细粒度视觉识别的研究提供新的思路与基准.

2　相关研究现状

2.1　细粒度视觉识别方法

传统的基于卷积神经网络 (convolutional neural network, CNN) 的细粒度视觉识别方法针对细粒度视觉识别提出了两大解决路径. 第 1 类方法 [12~15] 主要关注于挖掘图像中具有区分性的局部区域, 引导模型关注该区域, 利用其精细特征进行细粒度视觉识别. 通常来说, 这一类方法采用 Faster-RCNN [16] 中的区域建议网络自动地进行图像中潜在区分性区域的挖掘 [14,15], 或采用空间注意力机制从物体特征中进行区分性部位的筛选 [12,13]. 另一类方法 [17,18] 则将双线性池化操作作用于物体特征, 从而获得具有高阶语义的特征. 进一步地, 一些方法 [19,20] 充分利用网络的多层特征, 利用其互补性进行精确的细粒度视觉识别.

ViT 的发展提供了一种具有更广泛预训练特征的基础模型, 基于 ViT 的细粒度模型超越传统 CNN 系列的模型, 达到了当今最优的结果. ViT 固有的多层结构使得 ViT 每一层都编码了不同的物体表征. 因此, 一系列利用 ViT 网络的多层特征来改进细粒度视觉识别的 FGVC 方法 [21~24] 被提出. TransFG [21] 使用多层累积注意力权重来选择具有判别性的部分. FFVT [22] 在每一层编码器收集最具信息量的图像块元素并将它们送入最终的编码器层进行交互. DCAL [23] 使用多层累积注意力分数来生成局部查询向量, 并通过交叉注意力机制将这些局部特征融入全局特征.

与传统的细粒度视觉识别方法不同, 本文提出的细粒度视觉识别统一模型采用具有参数高效性的 SCA 模块进行大类知识的学习, 并未引入复杂的多层特征提取、融合和池化等模块, 整体架构简洁、计算高效, 具有实际应用前景.

2.2　视觉预训练模型

早期的计算机视觉模型通常将随机初始化的 CNN 在下游数据集上直接进行训练. 由于下游数据规模较小, 提供的特征不足, 使得网络训练难以收敛或导致过拟合, 导致识别精度难以提升. DeCAF [25] 是第 1 个提出利用 ImageNet [26] 预训练模型权重运用于下游任务的工作, 采用大规模数据集预训练后的模型特征具有通用性与泛化性, 可以帮助模型在下游任务上快速收敛. FGVC 作为计算机视觉的重要任务, 在后续的工作中也采用了上述训练范式. 早期常用的视觉预训练模型以卷积神经网络为主, 包括 ResNet 系列 [4,27]、DenseNet [28] 等, 这些模型往往只在 ImageNet-1k 和 ImageNet-21k 上进行预训练, 以获得较大规模的预训练特征. 基于 Transformer [29] 架构的视觉模型 ViT [11] 具有更强的通用表征能力, 成功推动视觉预训练研究进入新时代. 基于 ViT 的视觉预训练模型可以分为 3 类: 强监督的原始 ViT 和引入局部先验信息的 Swin Transformer [30]; 视觉语言对比学习预训练的 CLIP [31] 和改进损失函数的 SigLIP [32]; 自监督预训练的方法有基于掩码重建的 MAE [33] 和结合全局与局部自监督训练的 DINOv2 [34] 等. 通过大规模的预训练数据集和有效的预训练方式, 这些模型在下游任务中展现出了强大的泛化能力.

本文在细粒度视觉识别任务中采用冻结大部分参数、微调局部参数的方式, 保持了视觉预训练模型的通用特征, 同时综合评估了各个视觉预训练模型的在各个数据集上的性能, 综合选出了对 FGVC 最优的视觉预训练模型, 从而借助其通用视觉特征构建细粒度视觉识别统一模型.

2.3　混合专家架构

混合专家架构 (mixture-of-expert, MoE) 最初由 Jacobs 等人 [35] 提出, 它采用了多个独立的专家模块, 每个专家学习一个完整数据集的子集, 通过多个专家融合构建最终的系统, 使其能同时处理不同子域的数据. SMoE [36]

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

表 1　预训练模型详细介绍, 包含模型参数量、预训练数据集和预训练分辨率.

Table 1　Details of pretraining models, including model parameters, pretrain dataset, and pretraion resolution.

Type	Variant	#Params	Dataset	Pretrain resolution
Fully-supervised	ViT-B/16 [11]	86M	ImageNet-21k	224
	Swin-B [30]	88M	ImageNet-21k	224
Language-supervised	CLIP ViT-B/16 [31]	86M	WIT-400M	224
	SigLIP ViT-B/16 [32]	86M	WebLI	224
Self-supervised	MAE ViT-B/16 [33]	86M	ImageNet-1k	224
	DINOv2 ViT-B/16 [34]	86M	LVD-142M	518

提出了稀疏激活模的混合专家架构, 它为模型扩充了多个专家模块, 并通过 Top-K 操作稀疏地激活其中 K 个专家, 使得模型容量增大的同时, 总体计算量与原模型相近. 在现代的 Transformer 架构中, MoE 架构往往用于替换其中的负责知识表征的多层感知机模块, 其代表工作为 Gshard [37] 和 SwitchTransformer [38], 分别采用 Top-1 和 Top-2 方式进行稀疏激活, 从而扩展模型的上限. Komatsuzaki 等人 [39] 提出将预训练的密集模型参数用于初始化稀疏 MoE 模型的各个专家, 加快在下游任务上的收敛速度. 最新的大语言模型 [40, 41] 以及多模态大模型 [10, 42] 也广泛采用了 MoE 架构, 从而在计算负担均衡下构建表征能力上限更高的大模型.

本文提出的多类混合策略借助 MoE 的思想, 将训练好的各个 SCA 融合到冻结的主干网络上, 并通过 RT 模块对输入特征进行大类的划分, 激活其正确对应的适配器, 从而使得模型可以同时识别多个大类的视觉对象, 构建出细粒度视觉识别统一模型.

3　方法

本节将详细介绍基于多类混合适配的细粒度视觉识别统一模型 MoCA. 3.1 小节介绍本文实验所基于的 6 个通用视觉预训练模型以及对各个模型所采用的微调评估基准, 从而选出最优的通用视觉模型. 3.2 小节将介绍本文所设计的 SCA 结构和 MoA 融合策略, 基于上一步的预训练模型, 本文通过 SCA 实现特定大类知识的学习, 并采用 MoA 策略进行多个 SCA 的融合, 初步构建细粒度识别统一模型. 3.3 小节将介绍模型的训练策略、辅助损失函数以及模型的可持续拓展能力.

3.1　通用视觉特征评估

本文构建统一模型采用通用知识和特定大类知识相结合, 通过冻结预训练模型大部分的主干参数、仅微调其中小部分参数, 从而在训练过程中维持模型的通用知识并注入大类知识. 预训练模型所具有的通用知识对统一模型的识别能力至关重要, 因此本文设计多种微调方式对预训练模型知识通用性和泛化性进行评估, 从而得到通用知识泛化性最强的预训练视觉模型. 本文所评估的预训练模型均采用先进的 ViT 架构, 模型经过充分的大规模数据集预训练, 具有比传统 CNN 模型更强的知识通用性和泛化性. 本文评估了 6 个常用的 ViT 模型, 各个模型的详细参数如表 1 所示. 6 个模型可以分为 3 大类: 基于全监督预训练的 ViT [11] 和 Swin Transformer [30], 基于对比学习视觉语言预训练的 CLIP [31] 和 SigLIP [32], 以及基于视觉自监督预训练的 MAE [33] 和 DINOv2 [34]. 为了预训练模型之间的对比公平, 本文选取了相近参数量的预训练模型.

全监督训练采用 ImageNet 系列数据集, 其中 ImageNet-1k 包含 1000 个类别共 1.4M 张图像, ImageNet-21k 包含 21841 个类别共 14M 张图像. 作为基础的大规模视觉预训练数据集, 这两个数据集能为下游任务提供丰富的预训练特征. 传统 FGVC 方法通常以全监督的 ViT 和 Swin 作为基础模型, 并和额外添加的模块一起进行全参数的训练 [21~24]. 这样的训练方式只适合为各个大类训练模型, 无法构建细粒度视觉识别统一模型.

在视觉语言对比学习预训练模型方面, 本文采用两种主流模型 CLIP 和 SigLIP. 前者采用了 400M 大小的图文对数据集 WIT-400M 进行预训练, 后者采用了 12B 大小的图文对数据集 WebLI 进行预训练. 视觉语言预

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

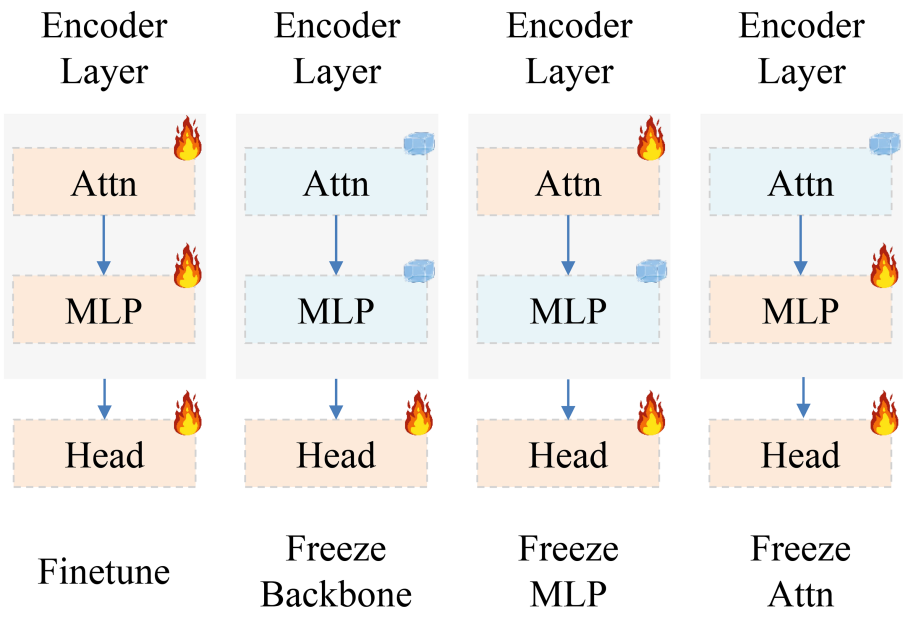


图 2　(网络版彩图) 视觉预训练模型微调评估方式.

Figure 2　(Color online) Finetune methods for vision pretraining model evaluation.

训练赋予了 CLIP 和 SigLIP 充分的视觉和语言特征对齐能力. 传统视觉任务中通常采用冻结 CLIP 和 SigLIP 参数且只训练线性头的方式进行微调. 本文额外探索了多种微调 CLIP 和 SigLIP 的方式, 从而衡量两种模型参与构建细粒度视觉识别统一模型的可行性.

基于自监督预训练的模型无需标签或者文本作为监督, 而是通过掩码重建图像像素或者保持单一图像在不同变化下一致性的方式, 使得模型能学习到更充分和细致的图像细节. 自监督方法的早期采用 ImageNet-1k 进行预训练, 后续工作则提出了更大的数据集 LVD-142M. 该数据集包含 142M 张图像数据, 覆盖并扩充了现有大部分视觉数据集, 能提供足够丰富的视觉数据. MAE 通过随机掩盖图像中部分区域, 再进行像素重建, 使得 ViT 模型具有更强的像素级建模能力. 但由于其预训练规模不够大, 不足以作为统一模型的基础模型. DINOv2 结合全局的模型蒸馏目标以及局部的图像块级别重建目标, 同时保证了模型的全局和局部建模能力. 通过 LVD-142M 对其进行充分的预训练, 构建出了当今综合泛化性最强的视觉预训练模型. 现有工作使用 DINOv2 时一般采用冻结特征或者全参数微调用于构建多模态大模型等方式, 本文则通过更多的微调方式对 DINOv2 的特征通用性进行更全面的评估.

本文所采用的微调评估方式如图 2 所示, 包含 4 种: 全参数微调 (finetune, FT)、冻结主干网络 (freeze backbone, FB)、冻结多层感知机模块 (freeze MLP, FM)、冻结注意力模块 (freeze Attn, FA). 训练过程中, 固定参数的部分保持预训练模型原有的知识, 通过微调其他部分的参数来对特定大类知识进行学习. FT 用于评估模型整体的知识是否具有迁移性, FB 评估的是预训练模型存储的知识能否直接用于细粒度视觉识别. 在 ViT 模型中, 注意力模块负责进行输入特征的空间关系建模, 多层感知机模块负责进行输入特征与模型知识库的交互. 因此, FA 方式评估的是固定原有空间建模能力情况下是否需要学习新的知识, FM 方式评估的是固定原有知识库的情况下是否要学习新的空间建模能力. 本文对 6 种视觉预训练模型在 6 个细粒度大类上利用上述 4 种微调方式进行评估实验, 从而综合衡量各个视觉预训练模型的整体知识、空间建模能力和知识库的丰富等. 评估实验的设置如实验部分所述, 具体评估实验结果如表 2 所示. 综合考虑微调精度与训练成本, DINOv2 是最适合构建细粒度视觉识别统一模型的预训练模型. 本文具体归纳出以下结论.

- FT 方法训练的参数量最多, 理论上能更好地实现迁移, 但该方法并非对所有模型有效. 对于 CLIP 和 DINOv2 而言, FT 会破坏整体的特征关联性, 导致模型崩塌. 对于所有模型在所有数据集上, FT 方法的精度对比其他方法并不具有一致的最优性. 因此从精度和训练成本上看, 预训练模型在迁移到 FGVC 任务时需要

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

表 2　6 种预训练模型以 4 种微调评估方式在 6 个大类数据集上的微调结果 (%). 加粗和下划线分别表示最高数值与次高数值.

Table 2 Results of six pretrain models with four finetune methods on six general category datasets (%). The bold and underlined values represent the best and second results, respectively.

	CUB-200-2011 [43]				Stanford Cars [44]				Oxford Flowers [45]			
	FT	FB	FM	FA	FT	FB	FM	FA	FT	FB	FM	FA
# Training parameters	86M	0.2M	28M	56M	86M	0.2M	28M	56M	86M	0.2M	28M	56M
Swin-B [30]	**91.01**	89.39	**90.99**	90.99	91.39	71.68	91.58	91.17	99.54	99.24	99.54	99.40
ViT-B-16 [11]	89.01	82.77	89.40	88.93	90.04	59.10	90.90	89.46	**99.56**	98.39	99.64	99.61
CLIP ViT-B-16 [31]	64.27	78.99	87.85	86.80	13.23	81.32	93.76	93.20	53.31	90.73	98.19	97.21
SigLIP ViT-B-16 [32]	86.03	37.62	87.12	86.53	**94.66**	37.99	**94.77**	94.70	98.42	43.18	98.88	98.28
MAE ViT-B-16 [33]	78.96	11.34	76.02	75.34	85.68	2.69	78.00	78.12	85.73	14.25	86.47	83.38
DINOv2 ViT-B-16 [34]	16.65	**90.21**	90.61	**91.20**	10.27	**87.61**	94.45	**94.73**	50.97	**99.69**	**99.67**	**99.71**
	Stanford Dogs [46]				FGVC Aircraft [47]				DF24 [48]			
	FT	FB	FM	FA	FT	FB	FM	FA	FT	FB	FM	FA
# Training parameters	86M	0.2M	28M	56M	86M	0.2M	28M	56M	86M	0.2M	28M	56M
Swin-B [30]	**88.08**	87.40	88.48	88.62	88.14	50.94	86.85	86.98	**70.31**	64.47	70.41	71.01
ViT-B-16 [11]	86.86	82.77	87.49	86.91	88.06	56.35	87.16	86.53	69.74	52.26	69.90	69.35
CLIP ViT-B-16 [31]	79.92	71.59	84.97	85.72	67.21	51.48	90.49	89.47	63.12	45.82	70.39	70.06
SigLIP ViT-B-16 [32]	85.01	24.52	84.94	85.13	**90.67**	25.53	92.17	91.27	68.65	22.54	70.49	70.20
MAE ViT-B-16 [33]	82.37	6.95	81.24	81.37	85.03	6.12	79.89	79.57	66.12	9.84	64.77	65.71
DINOv2 ViT-B-16 [34]	17.96	**88.46**	**88.79**	**89.01**	11.88	**79.11**	**93.33**	**93.51**	30.02	**66.52**	**73.43**	**73.95**

保持一部分的模型的通用知识, 为知识迁移到各个大类提供指导.

• FB 方法只训练一个线性分类头, 因此精度与预训练模型原有知识直接相关. 在 6 种模型中, MAE 表现最差, DINOv2 表现最佳. MAE 性能不佳的原因在于, 其自监督目标 (像素重建) 并未引导模型学习与分类任务直接相关的语义特征; 同时由于其预训练数据集 ImageNet-1k 相对较小, 提供的图像样本不够丰富. DINOv2 虽然没有直接对齐到图像标签, 但其通过知识蒸馏方法, 使得模型能够隐含地学习到图像与标签的对应; 同时由于其预训练数据集足够丰富, 即使仅微调线性分类头, 也可以利用其原有的丰富知识实现高精度识别.

• FM 和 FA 微调方式精度大部分情况下高于 FT 和 FB, 说明在训练过程中以模型原有的通用知识为微调剩余部分提供指导是更好的选择. FM 和 FA 在各个指标表现没有显著区别, 但由于 MLP 的参数更多, 因此 FM 方式在精度和训练成本上优于 FA 方法. 在 Stanford Cars 和 Aridcraft 数据集上, 各个模型的预训练知识难以直接迁移, 所以 FM 和 FA 方法能够弥补原有知识的不足, 显著提高微调精度.

• 在强监督模型中, Swin 更优的空间局部先验的设计, 使得模型能够捕捉更细致的图像特征. 对于两种视觉语言预训练模型 CLIP 和 SigLIP, 本文设计 FM 和 FA 微调方法来评估这两种模型性能, 并证明只需要微调部分参数即可达到和全参数微调相近的精度. 对于两种自监督模型, 本文实验发现 MAE 和 DINOv2 在微调上表现出相反的特性: 对于 MAE, 全参数微调是提升其性能的必要手段, 否则精度会大幅下降; 而 DINOv2 恰恰相反, 全参数微调会破坏其精心学习的表征, 导致模型性能崩溃, 因此不适合进行全参数微调.

• 综合多种微调评估结果, 受限于训练方式和训练数据集规模的 MAE 模型表现最差. 尽管 CLIP 和 SigLIP 采用了对比学习预训练获得了充分的表征, 但其对齐的文本中与视觉识别无关的噪声将会干扰迁移时的精度, 因此并非最优的预训练模型选择. 强监督的 ViT 和 Swin 能够达到最优的全参数微调精度, 但在部分参数微调情况下精度低于其他类型的模型, 计算成本偏高, 并非是最优的预训练模型选择. DINOv2 虽然在全参数微调情况下会导致模型崩塌, 但在部分参数微调情况下, 其精度在所有数据集上几乎一致性地取得了最优的结果, 从计算成本和精度上考虑是最优的选择.

综上所述, 考虑微调精度和计算成本, 本文最终采用 DINOv2 模型作为构建统一模型的基础模型. 通过冻结

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

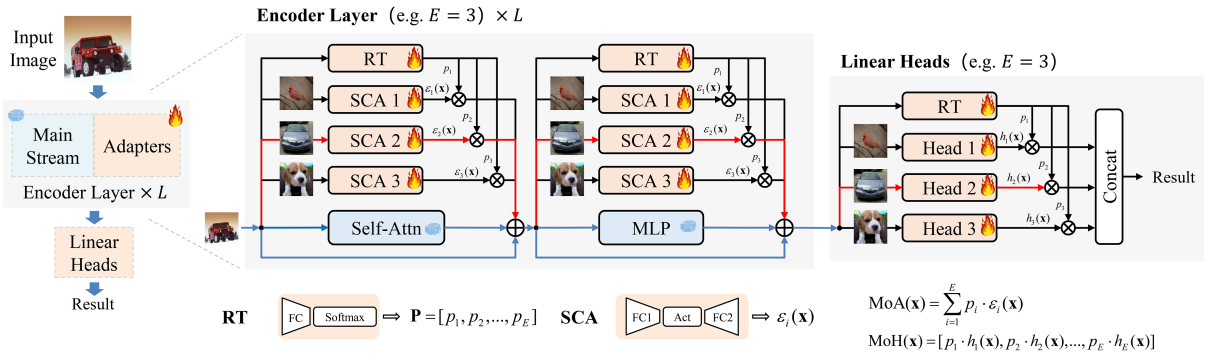


图 3　(网络版彩图) 细粒度视觉识别统一模型 MoCA 的框架图. 本文在 DINOv2 模型的主干 (自注意力和多层感知机) 之外通过多个 SCA 模块进行特定大类知识的学习, 并通过 RT 模块将多个 SCA 特征加权融合到主干特征, 在线性头部分也采用类似方式融合多个大类的线性头, 将最终输出并联得到识别结果.

Figure 3　(Color online) Illustration of our proposed unified FGVC model MoCA. Multiple SCA modules for category-specific learning are applied parallelly along the main stream (Self-Attn and MLP). We use RT modules to weight and sum the SCA features into the main stream features. This operation is similar in the linear heads part. The output logits of all heads are concatenated to form the final recognition results.

DINOv2 的预训练特征, 维持通用知识, 再引入多个单类适配器并进行融合, 最终构建具有高精度、低成本、可持续的具有实际应用价值细粒度视觉识别统一模型.

3.2　多类混合适配

由上一部分实验综合得出, DINOv2 是目前最优的通用预训练模型, 适用于对模型参数进行部分微调, 以固定参数部分的知识来指导剩余需要训练的部分. 通过微调 DINOv2 的部分参数, 可以实现从通用知识到特定大类知识的迁移. 虽然 FM 和 FA 可以对 DINOv2 进行高精度微调, 但需要训练的参数量仍然远超 FB 方法, 导致统一模型需要进行多个大类的识别时, 最终整体的训练和推理参数量仍然较大, 因此 FM 和 FA 在计算量上不是最优选择. 针对该问题, 本文旨在设计一个训练参数量少且精度高的微调方法, 实现高效的大类知识学习, 进而简化统一模型的总参数量.

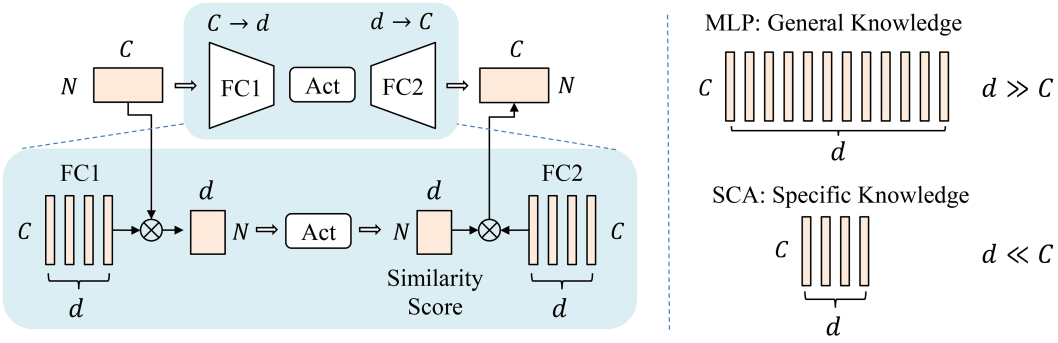
将通用模型在特定大类上进行迁移时, 线性分类头是必须要微调的部分. 另外, 由前文实验观察可知, 微调 Attn 和微调 MLP 都能带来精度的提升. 因此, 本文提出了单类适配器 SCA 微调方法, 将大类新知识以知识库嵌入的方式, 与冻结的 DINOv2 主干融合, 实现高精度轻量化的大类知识学习. 方法总体如图 3 所示, 本文将预训练模型每层编码器中的 Self-Attn 和 MLP 视为主干网络, 在训练过程中冻结这两种模块的参数 (包含 ViT 结构中自带的 LayerNorm 层). 在主干网络之外, 输入特征会经过各个 SCA 处理, 经过路由模块 RT 分配权重后, 输出特征与主干网络的特征相加, 实现通用知识与大类知识的融合, 赋予模型在大类进行识别的能力.

SCA 的结构如图 3 所示, 由一个将输入维度减小到隐藏维度的线性层, 一个非线性激活函数, 和一个将隐藏维度扩大回输入维度的线性层, 三者串联组成. 将维度减小再扩大的设计考虑到了 SCA 的参数量不宜过大, 否则会导致难以训练或过拟合; 同时考虑到大类知识相比于通用知识来说, 表征的维度不需要太大, 因此只需要少量参数即可实现大类知识的学习. 由于不同特定大类知识学习的难度不同, 因此不同 SCA 的隐藏维度可以设置成不同大小, 以实现在各个大类上的最佳性能, 确保统一模型在各个大类上的精度上限足够高. 设输入特征为 $\mathbf{x} \in \mathbb{R}^{N\times C}$, 其中 N 为 token 的数量, C 为模型的维度. 统一模型可以对 E 个大类进行细粒度视觉识别, 则对应一共有 E 个 SCA, 第 i 个 SCA 的输出 $\varepsilon_i(\mathbf{x}), i \in [1, 2, \ldots, E]$ 可以公式化表示为

$$\varepsilon_i(\mathbf{x}) = \mathbf{W}_2 \cdot (\mathrm{Act}(\mathbf{W}_1 \cdot \mathbf{x})), \tag{1}$$

其中两个线性层 FC1 和 FC2 的权重 (不含 bias 项) 为 $\mathbf{W}_1 \in \mathbb{R}^{C\times d}$ 和 $\mathbf{W}_2 \in \mathbb{R}^{d\times C}$. 采用知识库的角度来解释具有相似结构的 SCA 和 MLP, 如图 4 所示, 输入特征与 FC1 的权重相乘可以看作与 d 个知识向量组成的知识库一一相乘, 再经过非线性激活函数后得到相似度分数, 即输入特征与哪些知识最相关. 通过该分数与 FC2 的权

https://www.sciengine.com/doi/10.1360/SSI-2024-0389



(a) Explain SCA and MLP as Knowledge Base　　(b) Knowledge Difference

图 4　(网络版彩图) SCA 和 MLP 的知识库解释.

Figure 4　(Color online) Knowledge base explanation of SCA and MLP.

重相乘, 加权组合得到 SCA 的输出特征, 作为学到的大类知识与主干的通用知识合并. 在主干网络的 MLP 中, 隐藏维度远大于输入维度 $d \gg C$, 而在 SCA 中, 隐藏维度远小于输入维度 $d \ll C$. 激活函数采用 GELU 函数. SCA 适用于作为 Self-Attn 和 MLP 的旁路, 两者的 SCA 不共享参数, 同时隐藏维度可以不同, 以实现最优的特定大类精度.

本文设计了多类混合适配 (mixture of adapters, MoA) 以实现多类知识的融合. 多个 SCA 在各自的大类上训练好后, 通过 RT 分配权重进行融合. RT 模块由简单的线性层和 Softmax 函数组成, 将输入特征在空间维度取平均后, 利用线性层进行简单的线性分类, 预测出分配到不同 SCA 的权重, 再通过 Softmax 和 Top-1 操作选择出最高权重对应的 SCA, 其余权重置 0, 最终利用加权求和将所有 SCA 的特征融合到主干的通用特征. 对于平均输入特征 $\bar{\mathbf{x}} = \mathrm{mean}(\mathbf{x}) \in \mathbb{R}^C$, 路由模块 RT 的流程公式化表示如下:

$$\mathrm{RT}(\bar{\mathbf{x}}) = \mathrm{Softmax}(\mathbf{W} \cdot \bar{\mathbf{x}}) \in \mathbb{R}^E, \tag{2}$$

$$p_i = \begin{cases} \mathrm{RT}_i(\bar{\mathbf{x}}), & \text{if } \mathrm{RT}_i(\bar{\mathbf{x}}) \text{ is Top-1}(\mathrm{RT}(\bar{\mathbf{x}})), \\ 0, & \text{otherwise}, \end{cases} \quad i \in [1, 2, \ldots, E], \tag{3}$$

其中线性层权重为 $\mathbf{W} \in \mathbb{R}^{C \times E}$, 实现对输入特征的简单线性分类; Softmax 函数对分类分数进行重加权, 从而帮助选出最高的权重值. $\mathbf{P} = [p_1, p_2, \ldots, p_E] \in \mathbb{R}^E$ 代表了输入特征对 E 个 SCA 的权重, 利用该权重对 E 个 SCA 的输出进行重加权, 得到最终的多类混合特征. 公式化表示如下:

$$\mathrm{MoA}(\mathbf{x}) = \sum_{i=1}^{E} p_i \cdot \varepsilon_i(\mathbf{x}). \tag{4}$$

MoA 融合特征最终与主干网络的特征相加, 实现通用特征与多个大类特征的融合. 因为 Top-1 操作将分数最高外的 SCA 权重设置为 0, 融合的特征只包含与输入特征正确匹配的 SCA, 因此多个 SCA 之间不会相互干扰, 保证了无论输入特征属于哪一个大类, 都能在每一个编码器层正确匹配到对应的 SCA, 从而实现任意大类的细粒度视觉识别, 进而实现细粒度视觉识别统一模型. 在实际应用中, DINOv2 的前几层编码器保持原状, 不引入 SCA 和 RT, 避免初级低级语义特征的混淆导致 RT 无法正确预测权重, 详见 4.4 小节的可视化实验.

MoA 方法也可以迁移到网络的线性头部分, 改进为多头混合策略 (mixture of heads, MoH). 在各个大类上训练完成的线性头也采用类似的方式利用 RT 模块进行加权. 与多个 SCA 的融合特征与主干特征相加不同, 线性头部分里每个线性头的输出会进行拼接, 输出分数的长度维度等于所有大类的子类别数之和. 最终统一模型的识别结果为拼接后的 MoH 分数中分数最高的维度对应的类别, 它可以是模型可识别的任意大类中的任意子类别, 从而实现细粒度的多类统一识别. 设每个线性头的输出为 $h_i(\mathbf{x}), i \in [1, 2, \ldots, E]$, 则 MoH 可以表示为

$$\mathrm{MoH}(\mathbf{x}) = [p_1 \cdot h_i(\mathbf{x}), p_2 \cdot h_2(\mathbf{x}), \ldots, p_E \cdot h_E(\mathbf{x})]. \tag{5}$$

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

3.3　模型训练策略

本小节将详细介绍统一模型的训练策略. 与 Chang 等人 [3] 提出的端到端训练不同, 本文采用两阶段训练法. Chang 等人主要考虑的是端到端的模型训练时, 不同大类数据混合能否促进整体精度的提升; 本文考虑的是在各个大类单独训练完成的 SCA 实现独立且完备的融合从而构建统一模型. 本文在各个大类上训练 SCA 时已经获得了各自的最高精度, 进而通过优化 RT 模块实现更好的多类融合, 赋予统一模型在多个大类上进行高精度细粒度视觉识别的能力. 因此本文的设计无需考虑混合数据能否互相促进精度, 而是考虑不同 SCA 融合到主干后能否独立. 相互独立的 SCA 使得统一模型能够在原有模型基础上不断扩展, 使得统一模型具有实际应用价值. 本小节将介绍模型训练的策略, 包括路由模块的辅助损失函数、两阶段训练法、模型可扩展训练.

3.3.1　辅助损失函数

本文提出的 MoA 方法与传统的混合专家策略 MoE 不同. MoE 考虑的是将原有 MLP 扩展为多个专家, 并利用路由模块对每个 token 独立进行专家分配, 使得总激活参数量不变情况下提高模型的知识上限; 同时由于需要均衡每个专家分配到的 token 数量, 各个专家的参数量需保持一致, 降低了模型设计的灵活性. 本文的 MoA 方法则是将输入 token 全部正确分配到同一个 SCA, 只激活对应的 SCA 模块, 实现多个 SCA 融合到主干模型中互不干扰. 这要求 RT 模块能够对输入特征正确预测权重值, 且为了使大类知识完整地融合到主干, 正确预测的权重值必须接近 1. 本文设计无需考虑每个 SCA 的负载, 因此每个 SCA 可以采取不一样的设置, 这样有利于融合在不同大类上取得最高精度的 SCA, 提高最终统一模型的识别能力上限. 设训练数据为三元组 (X, Y, C), 对应图像本身、图像的细粒度类别标签, 图像的粗粒度大类标签. $C \in [1, 2, \ldots, E]$ 即表示对于每一个大类, 都有一个 SCA 对应. 本文设计了对 RT 模块的辅助损失函数公式表示如下:

$$L_{aux} = -\frac{1}{L}\sum_{l=1}^{L}\sum_{e=1}^{E} P_e^{RT} \log P_{l,e}, \tag{6}$$

$$P_e^{RT} = \begin{cases} 1, & \text{if } e = C, \\ 0, & \text{otherwise}, \end{cases} \quad e \in [1, 2, \ldots, E], \tag{7}$$

其中 $P_{l,e}$ 为 RT 模块在第 l 层输出的第 e 维数值, L 为编码器层数, E 为每层的 SCA 数量. 公式中 $P^{RT} = [P_1^{RT}, P_2^{RT}, \ldots, P_E^{RT}] \in \mathbb{R}^E$ 为 one-hot 向量, 只有输入图像的大类对应位置为 1, 其余位置为 0. 例如当输入图像为鸟类, 总共有 3 个 SCA 模块, 鸟类的 SCA 模块在第 2 个位置时, 监督标签为 $P^{RT} = [0, 1, 0]$. 利用该公式约束 RT 模块的输出 P 逼近一个正确位置 SCA 为 1, 其余位置为 0 的分布, 从而保证各个 SCA 独立且完备地融合到主干. 为实现每层编码器都能正确选择 SCA, P^{RT} 在所有层的值保持一致, 同时对于 Self-Attn 和 MLP, 都利用 P^{RT} 来监督其 RT 模块的输出. 辅助损失函数同样可以监督线性头部分的 RT 模块, 对应层数 $L = 1$ 的情况.

值得一提的是, 我们采用的 “Top-1 + L_{aux}” 路由策略与其他可微路由方法 (如 Gumbel-Softmax) 的对比. Gumbel-Softmax 能够实现端到端的可微采样, 允许来自最终任务损失的梯度流向所有专家, 但它依赖于这种隐式信号驱动的探索性学习. 在我们的特定场景中, 每个训练样本对应的 “正确专家” (即其所属大类) 是已知的. 因此, 通过 L_{aux} 施加一个直接且强大的监督信号是更高效和稳定的选择. L_{aux} 不仅为正确的专家提供了正向激励信号, 更关键的是, 它为所有不正确的专家提供了明确的抑制信号, 从而弥补了由 Top-1 操作导致的最终任务损失梯度稀疏的问题. 这种确定性的强监督方法避免了 Gumbel-Softmax 等随机方法可能引入的训练方差, 也更契合我们为每个大类训练一个专属适配器的设计目标.

3.3.2　两阶段训练法

两阶段训练法如图 5(a) 和 (b) 所示. 第 1 阶段训练为特定大类训练, 输入图像为特定大类的图像, 冻结主干网络的参数, 训练每层编码器的 SCA 模块, 以及最终网络输出端的线性分类头. 第 1 阶段的训练目标为标准

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

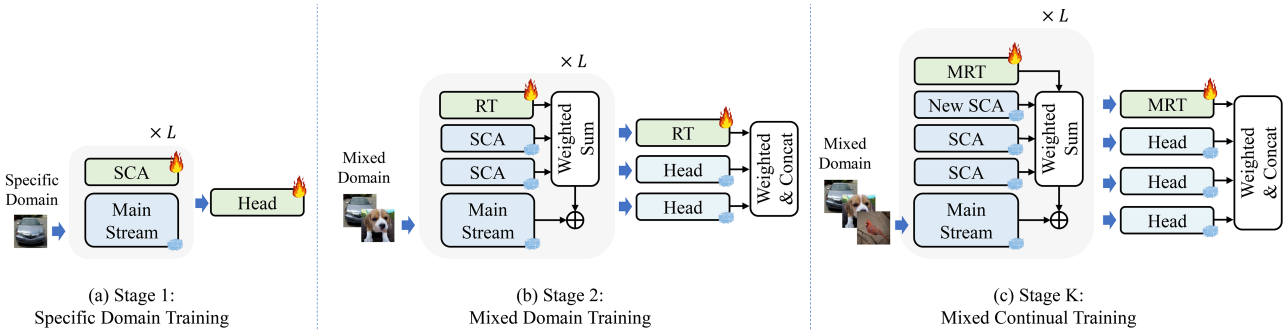


图 5　(网络版彩图) 两阶段训练法以及模型可扩展训练.

Figure 5　(Color online) Two-stage training and model continual training.

的在特定大类内进行细粒度视觉识别, 损失函数仅包含交叉熵损失:

$$\mathcal{L} = \mathcal{L}_{CE}. \tag{8}$$

第 2 阶段为多类混合训练, 输入图像为多类混合图像, 冻结第 1 阶段训练好的多个 SCA 模块和多个线性分类头, 仅训练每个编码器层的 RT 模块和最终网络输出端的 RT 模块, 实现多个 SCA 和 Head 与主干的加权融合或合并. 在混合训练时, 本文采用了 Chang 等人 [3] 提出的数据均衡采样策略, 使得每一个训练 batch 中各大类数据量均衡, 避免了大类之间的长尾分布问题. 第 2 阶段的损失函数包含交叉熵, 以及分别对自注意力、多层感知机和线性头的 RT 模块的辅助损失:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{aux}^{Attn} + \mathcal{L}_{aux}^{MLP} + \mathcal{L}_{aux}^{Head}. \tag{9}$$

第 2 阶段的交叉熵损失监督的是拼接后的分类头输出, 该部分设计参考了 Chang 等人 [3] 的方法. 在第 2 阶段训练中, 利用 4 个损失函数对 RT 模块进行监督, 实现更好的多类融合效果.

3.3.3　模型可扩展训练

模型可扩展训练如图 5(c) 所示. 当训练好的统一模型需要扩展新的大类时, 可以先根据第 1 阶段训练法, 得到对应的 SCA 和 Head, 再对原有的 RT 模块进行扩展. 设需要扩展新的 E' 个 SCA, 则需要新初始化一个大小为 $\mathbb{R}^{C\times E'}$ 的 RT 模块, 再与原来训练好的 RT 模块参数合并得到合并的 RT 模块 (merged router, MRT), 大小为 $\mathbb{R}^{C\times(E+E')}$. 如图 5(c) 所示的混合持续训练阶段, 将 MRT 和新的 SCA 与原有统一模型结合, 并将新的大类数据与原有数据混合, 按照上述数据均衡采样策略进行训练. 损失函数与阶段二相同, 保证新加入的 SCA 也与原有 SCA 相互独立且完备融合.

由于原有的 RT 模块只在原有的 SCA 之间进行权重预测, 合并后的 MRT 需要在添加新的 SCA 后所有的 SCA 之间进行权重预测, 因此在 MRT 中原有 RT 部分的参数也需要重新训练. 但由于 MRT 采用了原有 RT 的参数进行初始化, 因此收敛速度比随机初始化更快. 综上所述, 本文提出的统一模型可以实现新类数据的持续扩展, 成为具有实际应用意义的细粒度视觉识别统一模型.

4　实验结果

本节首先介绍统一模型所基于的大类数据集、实验设置, 然后介绍本文统一模型的识别效果并与基准方法进行对比. 本节接着通过大量的消融实验证明本文提出方法的有效性, 并利用可视化实验展示路由模块的权重分布.

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

表 3　细粒度视觉识别数据集统计及对应 SCA 设置.

Table 3　Statistics of FGVC datasets, and the corresponding SCA settings.

Dataset	Train size	Test size	Classes	Attn SCA		MLP SCA	
				Hidden dim	#Params	Hidden dim	#Params
FGVC Aircraft [47]	6667	3333	100	32	0.6M	64	1.2M
Stanford Cars [44]	8144	8041	196	32	0.6M	64	1.2M
Stanford Dogs [46]	12000	8580	120	32	0.6M	32	0.6M
CUB-200-2011 [43]	5994	5794	200	32	0.6M	32	0.6M
Oxford Flowers [45]	2040	6149	102	32	0.6M	32	0.6M
DF24 [48]	32692	3701	182	32	0.6M	32	0.6M

4.1　数据集与实验设置

本文所采用的数据集如表 3 所示, 包含了传统细粒度视觉识别广泛采用的 4 个数据集: 飞机数据集 FGVC Aircraft [47]、汽车数据集 Stanford Cars [44]、狗类数据集 Stanford Dogs [46]、鸟类数据集 CUB-200-2011 [43]. 除此之外, 本文还采用了识别难度较低的花类数据集 Oxford Flowers [45] 和识别难度较高的真菌数据集 Danish Fungi 2024 [48], 进一步丰富统一模型的识别范围, 并充分展示本文方法的通用性. 其中 Danish Fungi 2024 数据集为原文 Danish Fungi 2020 数据集的修订版, 在训练集和测试集的分布上有所变化, 使得识别的难度更高.

如上文所述, 本文可以采用异构的 SCA 以取得在每个大类上的最高精度. 对于 6 个大类, 本文最终采用的 SCA 设置如表 3 中所示, 添加在网络的第 4∼12 层编码器. SCA 的总参数量远小于预训练模型和传统方法的参数量, 体现了本文方法的参数高效性. RT 模块中的 Softmax 采用 0.2 的温度系数, 使得最高的分数值更突出, 重加权后最高值足够逼近 1. 统一模型中 6 个数据集的 SCA 排列顺序从第 1 位到第 6 位为 Cars, CUB, Air, Dogs, Flowers 和 DF24, 与表 4 中的数据集顺序对应.

本文提出的统一模型按照两阶段训练法构建. 预训练模型采用 DINOv2 中的 ViT-B-16 模型. 第 1 阶段中, 对于所有大类数据集, 训练图像随机缩放并裁剪到 224×224 分辨率, 仅采用随机水平翻转作为数据增强. 训练采用 SGD 优化器, 以余弦退火方式进行学习率的衰减. 初始学习率为 0.001, SGD 的动量设置为 0.9, 权重衰减设置为 0.0005, 训练 50 个 epoch, batch size 设置为 16. 上文方法部分评估实验中表 2 也采用第 1 阶段的实验设置. 第 2 阶段的训练将加载预训练模型参数和第 1 阶段训练好的各个 SCA 和线性头. 第 2 阶段训练中, 每个 batch 里采样每个数据集的 8 张图像, 每个 epoch 里反复采样所有数据集直至最大规模的数据集遍历完毕. 训练图像的分辨率、数据增强、优化器、SGD 动量和权重衰减均与第 1 阶段相同. 第 2 阶段的初始学习率设置为 0.0005, 训练 10 个 epoch. 对于持续扩展训练, 仅需训练 5 个 epoch, 其余设置均与第 2 阶段相同.

4.2　结果与分析

本文最终构建的统一模型实验结果如表 4 所示, 基准方法为 3.1 小节所述的冻结主干网络 (FB)、冻结多层感知机 (FM) 和冻结自注意力 (FA) 的结果. 可以看到, FB 方法仅微调线性分类头, 训练参数量低, 但由于未注入各个大类的新知识, 导致最终精度偏低. FA 和 FM 方法虽然取得较高的精度, 但训练参数量过大导致在计算高效性方面有所不足. FB, FM 和 FA 这 3 种方法需要为各个大类单独训练模型, 导致最终的多个模型总参数量过大, 不利于实际应用. FB+2SCA 代表第 1 阶段训练时冻结了骨干网络, 针对每个大类对每层的 Attn 和 MLP 这两个模块分别训练一个 SCA 作为知识偏置 (还包含一个最终的分类头, 此处略写). SCA 模块以较低参数量达到较高的大类识别精度, 通过将 SCA 模块学习大类知识, 可以实现逼近大参数量微调法 FA 和 FM 的效果, 证明了 SCA 方法的高精度轻量化. FB+2SCA 的结果为单个类别的识别性能上限, 因此第 2 阶段的目标即为最小化从单一模型到统一模型时的性能损失. 本文通过 RT 模块将多个大类的 SCA 融合到主干中, 构建出细粒度视觉识别统一模型 MoCA, 简化了训练和推理参数量, 同时保证各个大类上的精度逼近最优结果, 实现了高精度、轻量化的细粒度视觉识别统一模型.

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

表 4　统一模型结果与基准结果对比.

Table 4　Comparison with unified models and baselines.

Variant	Training #params	Inference #params	Cars Acc. (%)	CUB Acc. (%)	Air Acc. (%)	Dogs Acc. (%)	Flowers Acc. (%)	DF24 Acc. (%)	Avg. Acc. (%)
FB	0.7M	514.9M	87.61	90.21	79.11	88.46	99.69	66.52	85.27
FM	168.7M	514.9M	94.45	90.61	93.33	88.79	99.67	73.43	90.04
FA	336.0M	514.9M	94.73	91.20	93.51	89.01	99.71	73.95	90.35
FB+2SCA	10.3M	524.5M	94.35	90.61	92.59	89.58	99.72	73.22	90.01
MoCA (ours)	10.4M	96.1M	94.21	90.42	92.32	89.38	99.69	71.19	89.54

表 5　本文统一模型 (4 个大类) 与基准统一模型结果对比.

Table 5　Comparison between our unified model (on 4 categories) with baseline unified model.

Variant	Base model	Training #params	Inference #params	Cars Acc. (%)	CUB Acc. (%)	Air Acc. (%)	Flowers Acc. (%)	Avg. Acc. (%)
Chang et al. [3]	ResNet-50	33.8M	32.5M	–	–	–	–	88.08
MoCA-4 categories (ours)	DINOv2 ViT-B	6.6M	92.3M	94.23	90.40	92.35	99.68	94.17

表 6　混合持续扩展结果 (%).

Table 6　Results of mixed continual training (%).

Variant	Cars Acc.	CUB Acc.	Air Acc.	Dogs Acc.	Flowers Acc.	DF24 Acc.
Cars+CUB	94.17	90.30	–	–	–	–
Cars+CUB+Air+Dogs	94.09	90.58	91.75	89.28	–	–
Cars+CUB+Air+Dogs+Flowers+DF24	94.12	90.38	91.69	89.32	99.68	71.03
Air+Dogs	–	–	91.80	89.40	–	–
Air+Dogs+Cars+CUB	94.15	90.37	91.75	89.28	–	–
Air+Dogs+Cars+CUB+Flowers+DF24	94.16	90.35	91.73	89.32	99.67	71.08
Cars+CUB+Air	94.23	90.35	91.71	–	–	–
Cars+CUB+Air+Dogs+Flowers+DF24	94.18	90.33	91.62	89.35	99.59	71.13
Dogs+Flowers+DF24	–	–	–	89.20	99.65	71.20
Cars+CUB+Air+Dogs+Flowers+DF24	94.17	90.35	91.66	89.38	99.68	71.08

与 Chang 等人 [3] 的统一模型 (该方法只融合了 4 个大类) 进行对比, 如表 5 所示, 本文方法在同样的 4 个大类上取得了 6% 的平均精度优势, 且训练所需的参数量更少. 虽然本文采用的 DINOv2 作为基础模型而言参数量更多, 但其通用知识能够服务于构建性能更强、通用性更优的细粒度视觉识别统一模型, 推动细粒度视觉识别的实际应用.

模型可扩展训练的实验如表 6 所示. 前两组实验从两个大类混合结果开始, 每次加入两个大类; 后两组实验从 3 个大类混合结果开始, 再加入 3 个大类. 从整体上看, 在原有模型基础上加入新的大类, 对原来各个大类上的影响有正有负, 不具备一般规律, 但不存在精度剧烈下降的情况, 精度变化均在 0.2 以内. 这证明了本文方法可以在持续扩展新的 SCA 的情况下, 保证各个 SCA 之间互不干扰, 从而保持原有大类和新增大类上的识别精度, 实现持续扩展统一模型识别范围的效果. 持续扩展和所有大类直接混合训练对比, 精度略有下降, 但训练收敛需要的 epoch 数较少, 两种方式各有优劣, 若追求更高精度, 则可以直接采用两阶段的多类混合训练方法.

本文还在严格的持续学习设置下进行了实验, 即后续扩展更多大类训练时, 之前大类的数据不可访问. 在该设置下, 所有的 RT 模块参数需要进行扩展, 扩展方法为: 扩展 RT 模块的线性层输出维度, 加载已经训练好的

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

表 7　严格持续学习设置或少量数据可见设置下的结果 (%).

Table 7　Results under strict continual learning setting or few-shot setting (%).

Setting	Variant	Cars Acc.	CUB Acc.	Air Acc.	Dogs Acc.	Flowers Acc.	DF24 Acc.
Invisible	Cars+CUB	94.17	90.30	–	–	–	–
	Cars+CUB+Air+Dogs	90.47	82.38	76.21	86.63	–	–
1/16 visible	Cars+CUB	94.17	90.30	–	–	–	–
	Cars+CUB+Air+Dogs	94.09	90.34	89.75	87.45	–	–
Invisible	Cars+CUB+Air+Dogs	94.09	90.58	91.75	89.28	–	–
	Cars+CUB+Air+Dogs+Flowers+DF24	88.38	88.19	82.21	86.19	97.17	61.96
1/16 visible	Cars+CUB+Air+Dogs	94.09	90.58	91.75	89.28	–	–
	Cars+CUB+Air+Dogs+Flowers+DF24	93.74	90.09	90.46	88.09	99.46	68.23

表 8　SCA 模块消融实验 (%). 加粗表示最高数值.

Table 8　Ablation study of SCA module (%). Bold values represent the best results.

Variant	Cars Acc.	CUB Acc.	Air Acc.	Dogs Acc.	Flowers Acc.	DF24 Acc.
FB+2SCA (best dim)	94.35	90.61	92.59	89.58	99.72	73.22
FB+2SCA (dim = 16)	93.20	90.28	91.52	89.18	99.53	71.82
FB+2SCA (dim = 32)	93.43	**90.61**	91.78	**89.58**	**99.72**	**73.22**
FB+2SCA (dim = 64)	**94.35**	90.52	**92.59**	89.35	99.68	**73.22**
FB+Attn SCA (best dim)	94.16	90.30	92.35	88.75	99.69	72.77
FB+MLP SCA (best dim)	94.08	90.43	92.22	88.63	99.68	72.92

RT 模块线性层矩阵到新的线性层的前半部分, 后半部分对应新类别的扩展部分则从零初始化, 训练时候不冻结前半部分参数. 例如, 当前 4 个大类的线性层矩阵大小为 512×4, 需要扩展到 6 个大类的 512×6, 则前半部分 512×4 加载预训练参数, 后半部分 512×2 进行随机初始化. 实验超参数设置与混合持续扩展中相同, 结果如表 7 所示. 可以观察到在该设置下精度下降较为明显, 因此在后续阶段中如果之前数据不可见, 将会造成显著的性能退化. 由于当前阶段的 RT 模块在训练时辅助函数只考虑在当前多个大类中进行选择, 并未考虑输入特征不属于任何一个大类时的情况, 因此在后续扩展时候遇到新的大类时该差距难以弥补. 因此, 本文方法与严格的持续学习设置适配性较差, 未来工作中需要进一步改进.

另外, 本文还进一步实验了在后续阶段中让少量先前数据可见的设置, 如表 7 中所示, 本文按照先前数据与新数据 1:16 的配比进行持续学习训练, 观察到性能有明显的提升, 虽然略低于表 6 中的结果, 但远高于严格持续学习设置的结果, 证明了本文方法在少量数据可见设置下的可行性, 作为在严格持续学习和本文混合训练设置下的一种可行折中方案.

4.3　消融实验

本小节分别展示本文提出的两大主要模块 SCA 和 RT 的消融实验. 对于 SCA 模块的消融实验如表 8, 以冻结基础模型并在编码器层为 Self-Attn 和 MLP 添加 SCA 模块作为基准 (第 1 行), 其中 SCA 的隐藏维度采用表 3 中的最优维度. 本实验中不考虑利用 RT 模块进行多个 SCA 的融合, 只在单个大类上衡量 SCA 不同设置的效果. 可以看到, 针对不同大类数据集, 最优的隐藏维度不同, 这取决于通用特征与大类特征的差距, 以及大类自身知识的复杂程度. 表格中还展示了 SCA 模块添加在 Self-Attn 或 MLP 上的不同效果. 对于不同大类来说, SCA 的表现不尽相同, 但采用两个 SCA 的精度提升一致性地优于单个 SCA, 说明了同时在 Self-Attn 和 MLP 部分进行大类新知识的学习是更优的策略.

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

表 9　RT 模块消融实验 (%).

Table 9　Ablation study of RT modules (%).

Variant	Cars Acc.	CUB Acc.	Air Acc.	Dogs Acc.	Flowers Acc.	DF24 Acc.	Avg. Acc.
MoA-ViT full	94.21	90.42	91.78	89.38	99.69	71.19	89.45
w/o Aux. loss	93.52	90.23	91.58	89.13	99.03	70.47	88.99
Softmax → Sigmoid	92.85	89.87	89.74	89.15	99.48	70.20	88.57
Identity loss	91.35	89.58	88.13	89.02	–	–	–
Identity loss	90.71	89.26	86.41	87.42	99.58	65.39	86.46

对于 RT 模块, 消融实验如表 9 所示. 本文设计了辅助损失函数对 RT 模块进行监督, 使其对输入特征进行更准确的分类. 由表格数据可见, 如果不使用辅助损失函数, 可能会导致输出权值与 one-hot 分布偏移较远, 导致本该正确对应的 SCA 模块权重值过大或者过小, 或者导致对应不相关的 SCA 模块分数过高, 进而被错误激活. 多层编码器的其中一层特征出错容易形成误差累计, 导致模型最终的识别错误. 同理, Softmax 函数可以对线性层的输出进行重加权, 采用温度系数使得分布更为尖锐, 突出最高的特征, 并且将范围调整至 0∼1 区间; 如果采用 Sigmoid 函数对线性层的输出进行重加权, 可能会导致激活值出现多个峰值, 对应 SCA 的加权值不准确, 使得统一模型的识别精度严重下降. 综上所述, 本文所采用的 Softmax 函数重加权和辅助损失函数, 都是促进多类的 SCA 独立且完备地融合到主干网络, 实现通用知识与大类知识融合的有效手段.

此外, 本文还评估了基于独立性约束进行多个 SCA 融合的可行性. 本文提出的 MoA 结合 RT 的方法所采用的是 “功能级独立”, 并未考虑不同 SCA 输出的特征是否正交, 而是关心对于一个特定大类的输入, 只有与之对应的 SCA 被激活. 这是一种通过 RT 模块路由实现的 “稀疏激活” 式的独立, 使得多个 SCA 能够以简单高效的方式, 独立融合到主干网络, 并保持其在第 1 阶段单独训练时的性能. 另外可能的方法为 “特征级独立”, 不采用 MoA 结合 RT 模块来预测权重, 而是简单把特征经过所有 SCA 并且把所有输出完全相加到主干网络, 通过独立性约束作为损失函数, 要求各个 SCA 的输出特征是正交独立的. 对于该实验, 本文采用余弦相似度损失进行监督, 计算各个 SCA 输出特征之间的余弦相似度矩阵, 并使其趋近于单位矩阵. 该损失函数同样作用于统一模型训练的第 2 阶段, 保持第 1 阶段各个 SCA 单独训练不变. 实验结果如表 9 最后两行所示, 在 4 个大类和 6 个大类训练情况下, 结果均低于采用本文所提出的 MoA 结合 RT 的方法, 证明了本文方法的有效性. 同时, 本文方法在扩展数量数据集时仍适用, 而独立性约束在大类别更多的情况下性能损失加剧. 强制的特征正交约束可能是一种过强的约束, 对于视觉上相似的识别, 不同 SCA 学习到的知识偏置本身就存在一定的相关性, 强制它们正交可能会损害模型的性能. 本文提出的方法不限制特征本身, 仅通过路由预测对各个 SCA 进行灵活管理.

4.4　RT 预测可视化展示与分析

4.4.1　RT 预测可视化展示

本小节首先可视化展示针对不同大类, RT 模块为不同 SCA (Self-Attn 和 MLP 的 SCA) 和线性头分配的权重, 如图 6 所示. 以 Stanford Cars 和 DF24 数据集为例, 这两个大类应分别正确激活第 1 个和第 6 个 SCA. 可以看到, 无论是对于任意编码器层还是线性头, RT 模块的输出都能逼近 one-hot 向量, 最高分数趋近于 1, 其余分数接近 0, 使得输入图像能够与正确的 SCA 相对应, 同时使大类知识能完整地与主干特征相融合, 确保多个 SCA 在融合时互不干扰, 从而构建出细粒度视觉识别统一模型. 如果将 SCA 完整地用于每一个编码器, 则会如图 7 所示, 在较早的编码器层预测出错误的权重, 即正确的 SCA 不对应最高权重, 且该权重与数值 1 相差较远, 导致无法进行互不干扰且完整知识的融合. 由于网络的浅层特征语义层级低, 特征区分性不强, 使得不同大类的输入可能会获得相似的低层语义特征, 无法通过 RT 模块进行准确的线性分类, 从而获得错误的权重. 而到了网络的深层, RT 模块能够利用简单的线性层对高级语义进行准确划分, 从而获得正确的权重. 因此, 本文提出的在网络的前几层之后再加入 SCA 和 RT 模块是有效且合理的方法.

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

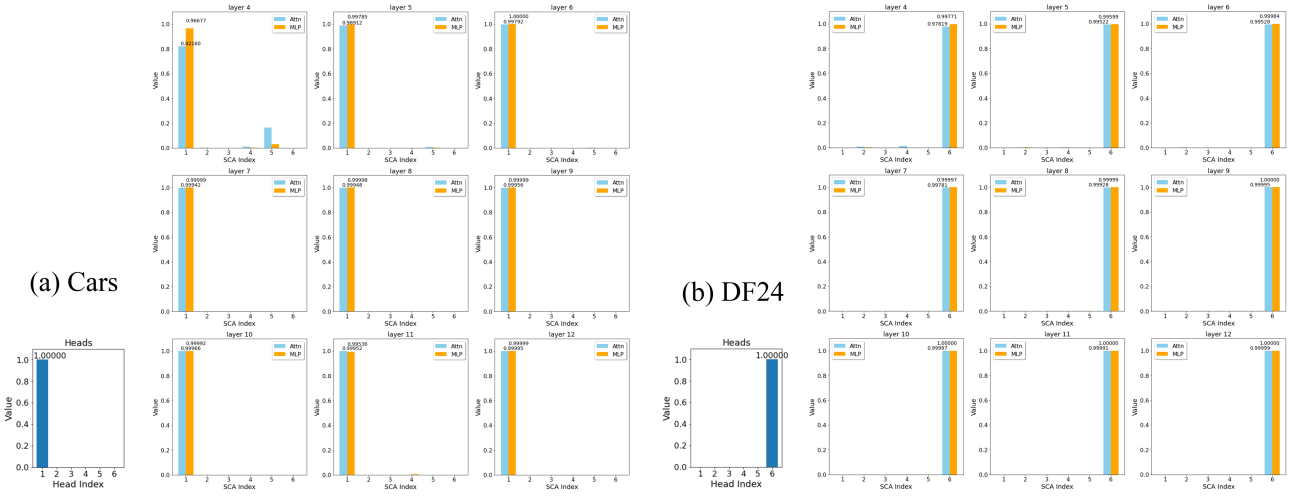


图 6　(网络版彩图) 第 4~12 层以及分类头中 RT 模块对于 6 个 SCA 的平均预测权重. (a) Stanford Cars 数据集的结果; (b) DF24 数据集的结果.

Figure 6　(Color online) Averaged weight distribution of RT modules for 6 SCAs from the 4th to 12th layers and for linear heads. (a) Results on Stanford Cars; (b) results on DF24.

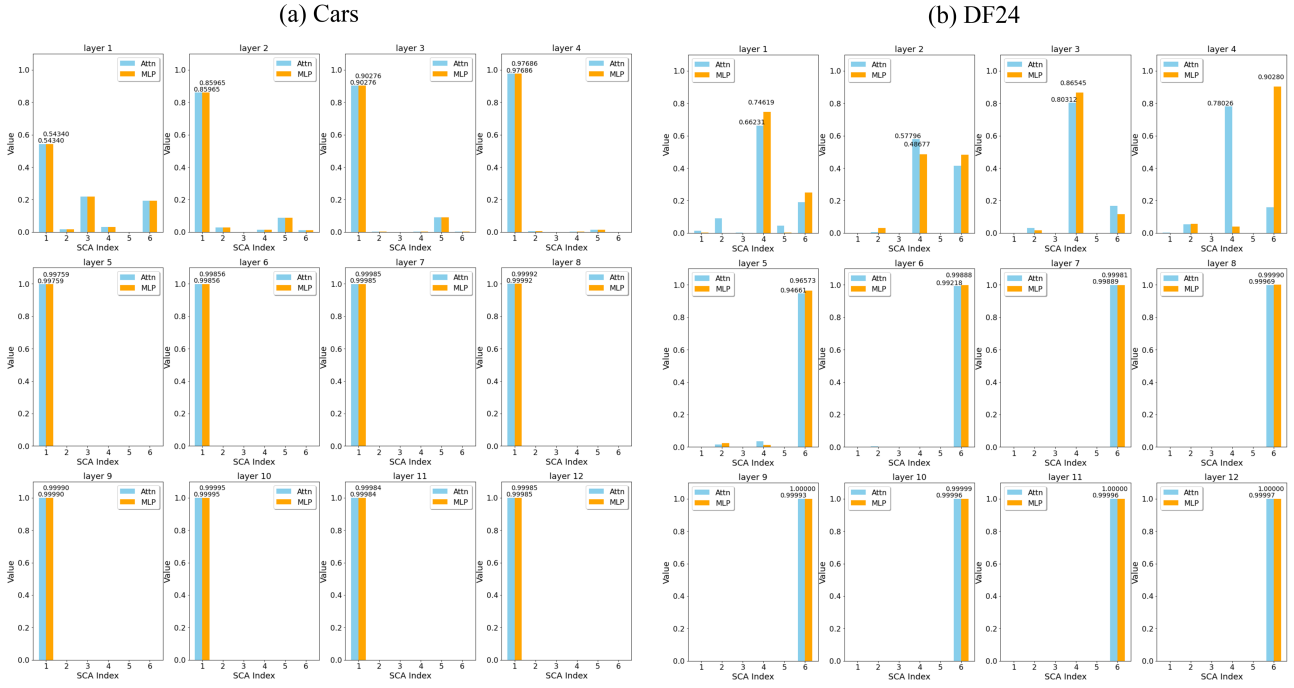


图 7　(网络版彩图) 对所有编码器层采用 SCA 和 RT 时的平均权重预测分布. (a) Stanford Cars 数据集的结果; (b) DF24 数据集的结果.

Figure 7　(Color online) Weight distribution of applying RT and SCA for all encoder layers. (a) Results on Stanford Cars; (b) results on DF24.

4.4.2　RT 预测数值分析

进一步地, 基于本文最终设置 (即在第 4~12 层采用 SCA 和 RT), 本小节通过数值分析对于不同大类, RT 模块在每一层为不同 SCA 预测的权重是否与大类一致 (例如 Cars 大类需要 RT 预测为第 1 个 SCA, 以此类推), 如表 10 所示. 可以得到以下 3 点结论: (1) 与可视化图类似, RT 预测与大类一致的比例从浅层到深层不断

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

表 10　不同层的 RT 预测与不同大类一致的比例 (%).

Table 10　Proportions (%) of the consistency between the RT predictions of different layers and the general categories.

Layers	Cars		CUB		Air		Dogs		Flowers		DF24	
	Attn	MLP	Attn	MLP	Attn	MLP	Attn	MLP	Attn	MLP	Attn	MLP
4	88.16	92.81	83.00	91.56	92.14	94.57	89.31	95.35	84.21	92.54	71.20	83.44
5	92.60	96.17	92.34	96.00	94.42	95.50	95.69	98.07	93.15	97.50	84.17	92.25
6	95.95	99.19	96.13	97.51	95.50	99.37	98.12	98.95	97.53	98.52	92.06	95.27
7	99.19	99.81	97.72	98.96	99.34	99.64	99.02	99.56	98.94	99.64	95.22	97.81
8	99.80	99.90	98.96	99.40	99.64	99.67	99.56	99.70	99.66	99.74	97.87	98.97
9	99.90	99.96	99.41	99.65	99.67	99.73	99.70	99.86	99.76	99.76	98.97	99.54
10	99.96	99.96	99.64	99.83	99.73	99.82	99.87	99.91	99.76	99.77	99.54	99.62
11	99.96	99.96	99.83	99.88	99.82	99.97	99.91	99.92	99.77	99.90	99.65	99.76
12	99.96	100.00	99.88	99.98	99.97	100.00	99.92	99.98	99.92	99.93	99.76	99.89
Head	100.0		99.95		99.97		99.95		99.95		99.78	

表 11　不同大类所有层 RT 预测一致性的分析 (%).

Table 11　Analysis (%) of the consistency on the RT predictions of all the layers, for different general categories.

Category	Cons. ratio	Incons. ratio	Acc. on Cons.	Acc. on Incons.	Acc. drop	MoCA Acc.	FB Acc.	FB+2SCA Acc.
Cars	87.84	12.16	94.38	93.08	1.30	94.21	87.61	94.35
CUB	81.53	18.47	90.53	89.91	0.62	90.42	90.21	90.61
Air	91.87	8.13	93.4	80.17	13.23	92.32	79.11	92.59
Dogs	88.65	11.35	89.43	89.01	0.42	89.38	88.46	89.58
Flowers	84.01	15.99	99.73	99.49	0.24	99.69	99.69	99.69
DF24	69.85	30.15	72.79	67.49	5.30	71.19	66.52	73.22

增加, 证明了在语义更深层中 RT 更容易区分不同的大类, 并正确预测权重; (2) 不同大类与 RT 预测的难易不同, 较为困难的是 CUB, Air 和 DF24, 这 3 个大类在第 4 层时候的一致性低于 85%, 尤其是 DF24 大类直到第 9 层才达到 99% 的一致性, 这一现象说明更困难的数据集需要更深层的语义才能正确预测; (3) 对 MLP 的 SCA 权重预测一致性均高于甚至显著高于对 Attn 的 SCA, 这一现象在浅层尤为明显. 一个可能原因是浅层的 Attn 需要在低层级的物体表征之间进行区分, 难度较大; 另一原因可能是, 我们设计的瓶颈式 SCA 在结构上与 MLP 更为相似 —— 两者都可被视为一种 “键 – 值” 式的知识库. 因此, 为 MLP 添加 SCA 作为知识偏置可能更容易匹配和优化, 因此需要在将来的工作中设计与 Attn 更加契合的知识偏置方法.

另外, 本文还分析了在各个大类预测一致性与识别准确率的关系, 衡量的是从第 4 层到第 12 层所有 RT 预测完全一致 (Cons. ratio) 和不完全一致 (Incons. ratio) 的比例, 以及在完全一致/不完全一致情况下的识别准确度差异 (Acc. on Cons. 和 Acc. on Incons. 以及 Acc. drop). 如表 11 所示. 可以看到, 这一结果与表 10 的观察一致, CUB, Air 和 DF24 是 3 个 RT 最难预测的大类, DF24 大类的 RT 预测在从头到尾完全相同的比例中只有 69.85%, 说明该大类在预测时最有可能会在不同层 RT 预测不同的 SCA. 通过观察在是否完全一致情况下的准确率差异可知, Cars, Air 和 DF24 是性能差异最大的, 在这 3 个大类上识别依赖更准确的 RT 预测.

最后, 为了便于分析在是否一致情况下的识别性能来源, 表 11 最右侧 3 列列出了表 4 中 MoCA, FB 和 FB+2SCA 的结果. (1) MoCA Acc. 可以由一致性比例与一致性准确率加权得到, 是本文最终模型的准确率. (2) FB Acc. 是固定预训练模型情况下的大类识别性能下限, 如果在 RT 不一致情况下准确率比该数值更低, 说明混用了其他大类的 SCA 有负面影响. 可以观察到: 对于 Cars, Air, DF24, Dogs 这 4 个大类, 混用了其他大类 SCA 有一定的正面效果, 因为它们可以从其他大类的 SCA 中学习到一定的共有知识进行辅助识别. 对于 CUB 和 Flowers, 它们更依赖于自身对应的 SCA, 因为其他大类的 SCA 会带来负面影响. (3) FB+2SCA Acc. 是在单

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

个类别上训练两个单独 SCA 的性能上限, 如果在 RT 完全一致的情况下性能超越了此上限, 则证明我们的 “软加权” 路由相比于简单的 “硬选择” (等价于权重为 1) 能够带来性能增益, 这可能是由于置信度门控机制提升了模型的鲁棒性. 在单独训练时两个 SCA 的权重都为 1, RT 权重通过辅助函数 L_{aux} 训练后逼近于 1, 能够带来一定的缩放效果, 在 Cars, Air, Flowers, DF24 这 4 个大类上有正面影响.

5　总结

针对细粒度视觉识别领域缺少高精度、轻量化、可扩展的统一模型的问题, 本文提出了多类混合适配的细粒度视觉识别统一模型 MoCA. 本文模型以预训练模型 DINOv2 为基础, 利用轻量化的单类适配器 SCA 进行大类知识学习, 并与冻结的通用特征融合实现在各个大类的高精度细粒度视觉识别. 多个 SCA 利用 MoA 策略、两阶段训练法和辅助函数, 实现多类知识与主干网络的独立且完备的融合, 构建出具有多类识别能力的统一模型. 本文方法可以将统一模型的识别能力持续扩展到新的大类, 最终构建出识别精度高、计算轻量化、知识可扩展的细粒度视觉识别统一模型, 推动细粒度视觉识别领域的实际应用.

参考文献

1 Hu Q H, Wang Y, Zhou Y C, et al. Review on hierarchical learning methods for large-scale classification task. Sci Sin Inform, 2018, 48: 487–500 [胡清华, 王煜, 周玉灿, 等. 大规模分类任务的分层学习方法综述. 中国科学: 信息科学, 2018, 48: 487–500]

2 Li H Y, Tu Z G, Xie W, et al. Fine-grained skeleton action recognition with pairwise motion salience learning. Sci Sin Inform, 2023, 53: 2440–2457 [李红艳, 涂志刚, 谢伟, 等. 学习成对样本运动显著性的细粒度人体骨架动作识别. 中国科学: 信息科学, 2023, 53: 2440–2457]

3 Chang D, Tong Y, Du R, et al. An erudite fine-grained visual classification model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023. 7268–7277

4 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778

5 Hu E J, Shen Y, Wallis P, et al. Lora: low-rank adaptation of large language models. In: Proceedings of International Conference on Learning Representations, 2022

6 Kopiczko D J, Blankevoort T, Asano Y M. Vera: vector-based random matrix adaptation. 2023. ArXiv:2310.11454

7 Chen S, Ge C, Tong Z, et al. Adaptformer: adapting vision transformers for scalable visual recognition. In: Proceedings of the Advances in Neural Information Processing Systems, 2022. 16664–16678

8 Zhang Y, Chen H, Zhang X, et al. Dyn-adapter: towards disentangled representation for efficient visual recognition. 2024. ArXiv:2407.14302

9 Lin X V, Shrivastava A, Luo L, et al. MoMa: efficient early-fusion pre-training with mixture of modality-aware experts. 2024. ArXiv:2407.21770

10 Lin B, Tang Z, Ye Y, et al. MoE-llava: mixture of experts for large vision-language models. 2024. ArXiv:2401.15947

11 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations, 2020

12 Zheng H, Fu J, Mei T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 5209–5217

13 Fu J, Zheng H, Mei T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4438–4446

14 Liu C, Xie H, Zha Z J, et al. Filtration and distillation: enhancing region attention for fine-grained visual categorization. AAAI, 2020, 34: 11555–11562

15 Liu C, Xie H, Zha Z, et al. Bidirectional attention-recognition model for fine-grained object classification. IEEE Trans Multimedia, 2019, 22: 1785–1795

16 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 2016, 39: 1137–1149

17 Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 1449–1457

18 Kong S, Fowlkes C. Low-rank bilinear pooling for fine-grained classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 365–374

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

19 Yu C, Zhao X, Zheng Q, et al. Hierarchical bilinear pooling for fine-grained visual recognition. In: Proceedings of the European Conference on Computer Vision, 2018. 574–589

20 Luo W, Yang X, Mo X, et al. Cross-x learning for fine-grained visual categorization. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 8242–8251

21 He J, Chen J N, Liu S, et al. TransFG: a transformer architecture for fine-grained recognition. AAAI, 2022, 36: 852–860

22 Wang J, Yu X, Gao Y. Feature fusion vision transformer for fine-grained visual categorization. In: Proceedings of the British Machine Vision Conference, 2021

23 Zhu H, Ke W, Li D, et al. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022. 4692–4702

24 Zhang Y, Cao J, Zhang L, et al. A free lunch from vit: adaptive attention multi-scale fusion transformer for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2022. 3234–3238

25 Donahue J, Jia Y, Vinyals O, et al. Decaf: a deep convolutional activation feature for generic visual recognition. In: Proceedings of the International Conference on Machine Learning, 2014. 647–655

26 Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248–255

27 Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1492–1500

28 Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4700–4708

29 Vaswani A. Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems, 2017

30 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision, 2021. 10012–10022

31 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning, 2021. 8748–8763

32 Zhai X, Mustafa B, Kolesnikov A, et al. Sigmoid loss for language image pre-training. In: Proceedings of the IEEE International Conference on Computer Vision, 2023. 11975–11986

33 He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022. 16000–16009

34 Oquab M, Darcet T, Moutakanni T, et al. Dinov2: learning robust visual features without supervision. 2023. ArXiv:2304.07193

35 Jacobs R A, Jordan M I, Nowlan S J, et al. Adaptive mixtures of local experts. Neural Comput, 1991, 3: 79–87

36 Shazeer N, Mirhoseini A, Maziarz K, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. In: Proceedings of the International Conference on Learning Representations, 2016

37 Lepikhin D, Lee H J, Xu Y, et al. GShard: scaling giant models with conditional computation and automatic sharding. In: Proceedings of the International Conference on Learning Representations, 2020

38 Fedus W, Zoph B, Shazeer N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. J Mach Learn Res, 2022, 23: 1–39

39 Komatsuzaki A, Puigcerver J, Lee-Thorp J, et al. Sparse upcycling: training mixture-of-experts from dense checkpoints. 2022. ArXiv:2212.05055

40 Jiang A Q, Sablayrolles A, Roux A, et al. Mixtral of experts. 2024. ArXiv:2401.04088

41 Dai D, Deng C, Zhao C, et al. DeepseekMoE: towards ultimate expert specialization in mixture-of-experts language models. 2024. ArXiv:2401.06066

42 Shen S, Yao Z, Li C, et al. Scaling vision-language models with sparse mixture of experts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023

43 Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset. 2011. https://www.vision.caltech.edu/datasets/cub_200_2011/

44 Krause J, Stark M, Deng J, et al. 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013. 554–561

45 Nilsback M E, Zisserman A. Automated flower classification over a large number of classes. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, 2008. 722–729

46 Khosla A, Jayadevaprakash N, Yao B, et al. Novel dataset for fine-grained image categorization: Stanford dogs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop on Fine-grained Visual Categorization, 2011

47 Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft. 2013. ArXiv:1306.5151

48 Picek L, Šulc M, Matas J, et al. Danish fungi 2020—not just another image recognition dataset. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2022. 1525–1535

https://www.sciengine.com/doi/10.1360/SSI-2024-0389

A unified model for fine-grained visual categorization with mixture of category adapters

Zhiying LU, Chuanbin LIU, Hongtao XIE & Yongdong ZHANG*

School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China
* Corresponding author. E-mail: zhyd73@ustc.edu.cn

Abstract The task of fine-grained visual categorization (FGVC) aims to distinguish highly similar sub-categories within specific general categories. Due to the significant disparities between these general categories, current methods necessitate training separate models for each, lacking a unified model capable of recognizing all sub-categories across all general categories. Training a unified model by simply merging datasets incurs substantial computational costs and lacks the ability to generalize to unseen categories, limiting its practical application. In this paper, we propose a unified FGVC model based on a mixture of category adapters, named MoCA (mixture of category adapters), to achieve high-accuracy, computationally efficient, and scalable unified recognition. Specifically, our model leverages the general knowledge from a frozen pretrained model and combines it with lightweight single-category adapters to learn category-specific knowledge, achieving high accuracy on each general category. Then, through our proposed mixture of adapters (MoA) strategy, multiple category-specific knowledge streams are independently and comprehensively fused with the general knowledge, enabling unified fine-grained visual categorization across multiple categories. Furthermore, our model is scalable, capable of continually extending to new recognition tasks by inserting new adapters while preserving existing knowledge. Experimental results demonstrate that our model holds significant advantages in terms of recognition accuracy, computational cost, and task scalability, providing a new benchmark and reference for the theory and methodology of unified FGVC models.

Keywords fine-grained visual categorization, unified model, pretrain, mixture-of-expert, finetune

https://www.sciengine.com/doi/10.1360/SSI-2024-0389