doi:10.3969/j.issn.1672-1292.2026.03.003

基于分层动态掩码视觉 Transformer 的长尾图像识别研究

汪世豪[1,2],毛　煜[1,2],林耀进[1,2]

(1.闽南师范大学计算机学院,福建 漳州 363000)
(2.闽南师范大学数据科学与智能应用福建省高校重点实验室,福建 漳州 363000)

[**摘要**]　现实世界中的许多分类任务天然具有层次化语义结构,当这些层次化分类任务同时面临长尾分布时,识别难度进一步加剧. 现有的长尾学习方法大多将类别视为扁平结构,未能充分利用层次标签中固有的语义关联. 传统的层次分类方法通常假设类别分布均衡,对长尾分布缺乏鲁棒性. 为解决这一问题,本文提出了一种基于分层动态掩码视觉 Transformer 的长尾图像识别算法(hierarchical dynamic masked vision transformer,HDMViT),针对标签分布极不均衡问题,对数据集的父类层标签进行了重新构建以优化层次结构的表达能力. 同时,在视觉 Transformer 模型中引入层次化双重分类头结构,两个分类头均基于 Transformer 编码器输出的同一特征向量进行决策:粗粒度分类头(父类)通过全连接层直接建模父类分布,细粒度分类头(子类)则进一步设计动态 Logit 掩码机制来过滤无关特征干扰,确保模型能够更专注地学习目标类别的关键特征,从而更有效地利用细粒度和粗粒度的语义信息. 实验结果表明,该方法在多个长尾数据集上均表现优异.

[**关键词**]　深度学习,长尾学习,不平衡数据,视觉 Transformer,层次分类
[**中图分类号**]TP391.41　[**文献标志码**]A　[**文章编号**]1672-1292(2026)03-0022-09

Hierarchical Dynamic Masked Vision Transformer for Long-Tail Image Recognition

Wang Shihao[1,2],Mao Yu[1,2],Lin Yaojin[1,2]

(1.School of Computer Science,Minnan Normal University,Zhangzhou 363000,China)
(2.Key Laboratory of Data Science and Intelligent Application of Fujian Province,Minnan Normal University,Zhangzhou 363000,China)

Abstract:Many classification tasks in the real world inherently possess hierarchical semantic structures. When these hierarchical classification tasks simultaneously face long-tailed distributions, the difficulty of recognition is further exacerbated.Existing long-tailed learning methods mostly treat categories as flat structures and fail to fully utilize the inherent semantic correlations in hierarchical labels. Traditional hierarchical classification methods,on the other hand, usually assume balanced category distributions and lack robustness to long-tailed distributions.To address this problem, this paper proposes a long-tailed image recognition algorithm based on a Hierarchical Dynamic Masked Vision Transformer(HDMViT). Aiming at the extremely imbalanced label distribution,the parent-level labels of the dataset are reconstructed to enhance the expressive capability of the hierarchical structure. Proposed methodology reconstructs parent-level labels to strengthen hierarchical representation,specifically targeting extreme label imbalance. Meanwhile, a hierarchical dual-classification-head structure is introduced into the Vision Transformer model, in which both classification heads make decisions based on the same feature vector output by the Transformer encoder:the coarse-grained(parent-level)head directly models the parent-class distribution through fully connected layers,while the fine-grained(child-level)head further incorporates a dynamic logit masking mechanism to filter out irrelevant feature interference,ensuring that the model learns the key features of the target classes in a more focused manner,thereby

收稿日期:2025-08-02.
基金项目:国家自然科学基金资助项目(62576158)、福建省自然科学基金重点项目(2026J002035)、福建省自然科学基金面上项目(2026J001987).
通讯作者:林耀进,博士,教授,研究方向:数据挖掘、机器学习.E-mail:zzlinyaojin@163.com

making more effective use of both fine-grained and coarse-grained semantic information. Experimental results show that the proposed method achieves excellent performance on multiple long-tailed datasets.

Key words: deep learning, long-tail learning, imbalanced data, vision transformer, hierarchical classification

随着计算机视觉领域的迅速发展,深度学习技术已经在图像分类、目标检测和视频分析等任务中取得了显著的成功[1]. 然而,当数据集呈现出长尾分布时,其性能就不够完美. 在长尾分布中,少数类(即头部类)包含了大量的样本,而多数尾部类仅包含少量样本. 在模型训练过程中,头部类主导了模型参数优化过程,导致模型对头部类表现出过强的偏好. 而由于尾部类包含的样本数量过少,模型对尾部类学习不充分,难以学到良好的特征表示. 尤其值得注意的是,现实世界中的许多分类任务本身就蕴含着天然的层次化语义结构,例如,生物分类学中的"界、门、纲、目、科、属、种",场景识别中的"室内、厨房、餐厅". 在这些任务中,类别间的层次关系为缓解长尾分布提供了宝贵的先验知识,但现有研究方法在这一交叉领域存在明显不足.

传统的长尾学习方法主要基于 ResNet 及其变体,通过设计特定的训练策略来缓解类别不平衡问题. 有研究基于再平衡策略增强模型对尾部类别的关注[2],但在头部类别上的性能却会相对下降. 有研究采用多专家网络[3],每个专家专注于不同的数据子集,并通过融合其预测结果来提升整体性能. 但这些研究主要集中在卷积神经网络体系结构上,对于视觉 Transformer 在长尾分类任务中的探索仍然有限.

近年来,视觉 Transformer 因其卓越的性能和可扩展性在平衡数据集上的表现令人瞩目. 尽管视觉 Transformer 在处理全局依赖关系方面具有显著优势,但在长尾数据集上的表现却不尽如人意. 视觉 Transformer 需要更长时间的训练来学习归纳偏置,且易受到标签统计偏差的影响,这进一步加剧了其在长尾数据集上性能的下降. 现实世界中的类别通常具有内在的层次结构,这种层次结构可以提供更丰富的语义信息并有助于提升模型的泛化能力. 尽管现有方法通过掩蔽预训练和再平衡策略缓解了标签偏差问题,但目前对层次化视觉 Transformer 的探索依然有限,其潜力在长尾学习中尚未得到充分挖掘.

如何在层次化标签的框架下有效应对长尾分布,是一个尚未得到充分探索的研究方向. 本文提出一种基于分层动态掩码视觉 Transformer 的长尾图像识别算法(HDMViT),首先对父类标签进行重新构建,以减少标签偏差并增强模型对于尾部类别数据的学习能力;其次为视觉 Transformer 设计双重分类头,分别预测父类标签和子类标签,使父类与子类预测任务既共享底层特征表示,又保持层次化决策的独立性,从而更好地捕捉标签空间中的层次依赖关系,提高模型在不同层次上的分类性能;最后,为了进一步提升子类分类头的表现,采用一种动态 Logit 掩码机制来屏蔽无关类别的 logits 值,确保模型专注于学习与分层结构密切相关的特征,不仅减少了不必要的干扰,而且提高了特征学习的有效性.

1 相关工作

随着深度学习规模的不断扩大,大规模长尾数据集的普及性日益显著[4]. 针对此类数据集,研究者们开发了多种技术手段.

(1)再平衡策略:该方法通过调整每个类的贡献来减少模型对多数类的偏倚,主要包括重采样、重加权以及 Logit 调整. 重采样方法通过调整训练数据的类别分布来平衡样本数量,具有较高的通用性与灵活性,如 DeepSMOTE[5]利用插值生成合成样本. 但是,在类内重叠严重或数据呈现多模态分布时,DeepSMOTE 的效果有限. 相比之下,生成对抗网络(GANs)[6]能够生成更逼真的样本,特别是在高分辨率图像生成任务中表现优异,但数据层面的重采样方法仍可能引入信息丢失. 重加权策略通过调整损失函数中的类别权重,以降低类别不均衡对模型训练的影响. 例如,IB Loss[7]通过衡量每个样本对复杂且偏倚的决策边界的影响程度自适应地调整其权重,从而提升模型的泛化能力. ImbSAM[8]通过引入类别先验信息,限制平滑优化的范围至尾部类别,从而有效缓解尾部类别因训练样本不足而导致的严重过拟合. Logit 调整方法通过根据类别样本数量对 logits 进行修正,使模型更加关注少数类,提高长尾类别的识别能力. 例如,平衡 Softmax[9]通过在训练过程中动态调整 logits 分布,以减少类别偏倚. 然而,再平衡策略在提升尾部类性能的同时,都会导致头部类的性能下降.

(2)多专家网络:该方法通过采用多个专家模型来分别处理不同的样本子集,并通过某种形式的知识

聚合来提升整体的表现,有效缓解了长尾分类任务在头部类上的性能退化. 这种方法在长尾数据集上特别有效,能够在面对高度不平衡的数据时灵活地调整学习策略,从而更好地捕捉长尾类别的特征. Rangwani 等[10]提出 DeiT-LT,用经过 SAM 训练的卷积神经网络教师进行蒸馏,促使所有 Transformer 块中 DIST 标记学习低秩泛化特征,有效地学习与多数类和少数类相对应的特征. 多专家网络虽然在长尾识别等任务中表现出色,但其设计和实现复杂,训练和推理过程资源消耗极高,这些问题限制了多专家网络在实际应用中的广泛使用.

(3)多阶段训练:该方法采用解耦特征学习与分类器优化的训练范式,通过分阶段参数更新策略显著提升了训练效率,有效缓解了数据分布不平衡带来的偏倚问题. Cao 等[11]提出首先学习特征,在第二阶段通过重加权策略优化分类器学习. Kang 等[12]进一步将表示学习和分类器学习分离为两个独立的阶段,并证明了阶段性训练对长尾学习的重要性. 随着技术演进,研究者通过引入混合增强(如 mixup、remix)与对比学习(如监督对比学习[13-14]、概率对比学习[15])强化特征学习的泛化能力. 当前研究多基于两阶段框架,但如何平衡网络权重以协调头部与尾部类的表征冲突仍是亟待突破的关键挑战.

2　分层动态掩码视觉 Transformer

本文提出一种用于不平衡数据的视觉 Transformer 模型,在 LiVT[16]的基础上对视觉 Transformer 模型进行改进,如图 1 所示. 模型基于 LiVT 架构,采用两阶段训练策略:预训练阶段通过掩蔽式生成预训练增强视觉 Transformer 的特征提取能力,随后在平衡微调阶段使用平衡二元交叉熵损失(Bal-BCE)以优化微调阶段的训练过程,通过引入额外的分类头使其能够有效处理层次化标签,并学习上层标签的语义特征.

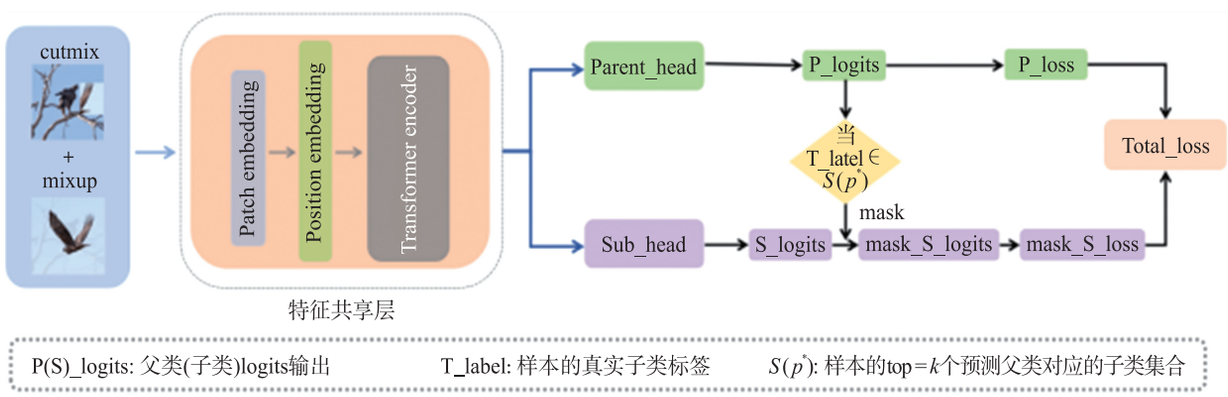


图 1　HDMViT 模型

Fig. 1　HDMViT model

2.1　重新构建父类标签

根据 iNaturalist 2018 数据集的层次分类任务分析显示,父类标签分布存在显著的长尾现象. 具体而言,样本数量最多的父类(Aves)包含 143 950 个样本,而样本数量最少的父类(Bacteria)仅包含 16 个样本,两者之间的样本数量差异高达 8 997 倍. 这种极端的类别不平衡分布使得现有模型难以充分学习父类层次的特征表示,进而影响分类性能. 同样,在通过 WordNet 工具从 ImageNet-LT 数据集和 Places-LT 数据集中提取父类标签后,发现这些数据集的父类层次也存在显著的类别不平衡问题. 针对这一问题,本研究提出父类标签重构策略,核心是通过阈值 α 筛选并分离“头部子类”(样本量过高的子类),避免其在父类层持续主导分布. 阈值 α 的设定需同时满足两个核心目标:(1)分离样本量过多的“头部子类”,避免其在父类层仍占据主导地位导致父类重构后仍存在长尾;(2)保证剩余子类在父类层的样本量足以支撑模型学习粗粒度语义特征(避免父类层出现“超尾部”类别). 由于不同数据集的长尾程度、类别数量及语义结构存在差异,α 的取值无法通过统一公式直接计算,因此本研究通过经验性实验探索逐步确定各数据集的合理 α 范围. 父类标签重构的具体操作流程如图 2 所示,对于子类别样本数量超过 α 的情况,将其从父类集合中分离出来,并将其父类标签重新设置为子类标签,以优化类别分布的平衡性. 通过上述操作,将样本数量超过阈值 α 的子类别(如“0”、“1”、“3”)从原有父类别中分离出来,并将其父类标签更新为子类标签. 例如,“0”不再归属于类别“A”,而是作为一个独立的父类别. 该调整方法优化了类别分布的平衡性,缓解了长尾分布问题,使模型能够更好地学

习样本数量较多的类别特征,同时减少对样本稀少类别的过拟合.

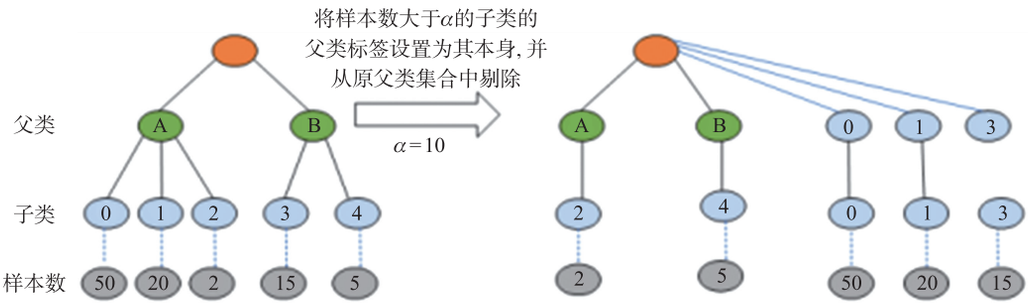


图 2　重构父类标签

Fig. 2　Rebuilt parent class labels(RPL)

2.2　双重分类头

尽管视觉 Transformer 在图像识别任务中表现出强大的全局建模能力,但其全局自注意力机制对所有图像区域平等处理的特性与长尾数据中的尾部类别依赖局部特征相矛盾. 此外,传统视觉 Transformer 采用单一的线性分类头,无法同时建模父类(粗粒度)和子类(细粒度)的语义层级关系,无法利用父类标签的信息指导子类的特征学习.

针对上述问题,本文提出双重分类头架构,其核心思想是:在共享视觉 Transformer 全局特征表示的基础上,通过双路径分类架构实现父类与子类特征的层次化解耦与协同优化. 具体而言,给定共享特征向量 $\boldsymbol{z}\in\mathbf{R}^{d}$(来自视觉 Transformer 主干),父类和子类分类头分别计算:

$$\boldsymbol{o}_{\text{parent}}=\boldsymbol{\omega}_{\text{parent}}^{\mathrm{T}}\boldsymbol{z},\boldsymbol{p}_{\text{parent}}=\text{Softmax}(\boldsymbol{o}_{\text{parent}});\tag{1}$$

$$\boldsymbol{o}_{\text{sub}}=\boldsymbol{\omega}_{\text{sub}}^{\mathrm{T}}\boldsymbol{z},\boldsymbol{p}_{\text{sub}}=\text{Softmax}(\boldsymbol{o}_{\text{sub}}).\tag{2}$$

式中,$\boldsymbol{\omega}_{\text{parent}}^{\mathrm{T}}\in\mathbf{R}^{d\times C_{\text{parent}}}$ 和 $\boldsymbol{\omega}_{\text{sub}}^{\mathrm{T}}\in\mathbf{R}^{d\times C_{\text{sub}}}$ 是独立的可训练权重矩阵,分别对应父类和子类任务;$\boldsymbol{o}_{\text{parent}}$ 和 $\boldsymbol{o}_{\text{sub}}$ 是输出的 logits;$\boldsymbol{p}_{\text{parent}}$ 和 $\boldsymbol{p}_{\text{sub}}$ 是预测概率分布;监督信号分别来自粗粒度的父类标签 y_{parent} 和细粒度的子类标签 y_{sub}.

相较于传统单分类头设计,双重分类头架构通过共享视觉 Transformer 主干网络提取的全局特征,在统一特征空间中实现层次化语义建模,既保留了父类分类头对粗粒度全局特征的捕捉能力,又赋予子类分类头聚焦局部判别特征的优势,为后续的动态掩码设计提供了关键的架构基础.

2.3　Logit 掩码机制

为进一步提升模型的全局特征表示,本文提出 Logit 动态掩码机制. 该方法在获得父类分类头的 logit 输出后,基于父类分类头的预测结果对子类分类头的 logit 输出实施掩码操作,如图 3 所示. 该操作保留与

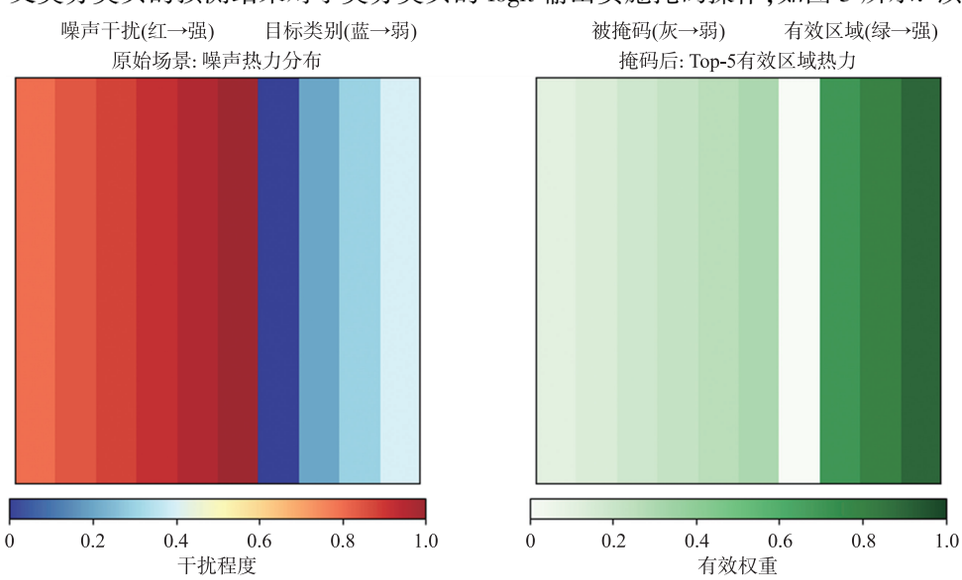


图 3　子类 logit 动态掩码

Fig. 3　Dynamic subclass logit masking(DSLM)

预测父类标签相关的子类别,并过滤掉其他无关类别的干扰,从而增强模型对相关类别特征的学习能力. 然而,在实验过程中,若仅使用父类 logit 向量的 top-1 预测标签作为掩码依据,易导致模型过拟合. 因此,本文采用 top-k 个父类预测标签作为掩码的参考范围,以增加预测的鲁棒性.

此外,考虑到即使在父类分类头的 logit 向量中选取 top-k 个标签作为父类预测结果,仍可能存在预测错误的情况,因此引入了一种动态掩码机制:当且仅当图像的真实子类标签属于 top-k 个父类预测标签中的某一类别时,才对子类分类头的 logit 向量施加掩码;否则,保持子类分类头的 logit 向量不受掩码影响. 这一设计不仅能够有效减少错误预测对模型训练的干扰,还能进一步提升模型在复杂层次分类任务中的泛化能力.

2.4　损失函数

2.4.1　预训练阶段

受 LiVT 的启发,本文同样采用掩蔽式生成预训练(MGP),以无监督方式学习图像的全局特征,从而有效降低长尾标签带来的负面影响. MGP 的核心思想是通过高比例掩蔽的输入图像训练特征编码器 F_{θ_f},并联合优化轻量级解码器 D_{θ_d}以重建原始像素空间. 具体流程为:

给定输入图像 $x\in\mathbf{R}^{H\times W\times 3}$,首先生成随机掩码 $M\in\{0,1\}^{H\times W}$,其中掩蔽比例设定为 75%(遵循 MAE 中的设置). 掩码后的图像块 $M\odot x$ 经编码器 F_{θ_f}提取特征后,由解码器 D_{θ_d}重建为完整图像,其优化目标为最小化原始图像与重建图像间的均方误差,具体公式为:

$$\mathcal{L}_{\mathrm{MGP}}=E_{x\sim D}\parallel x-D_{\theta_d}(F_{\theta_f}(M\odot x))\parallel_2^2. \tag{3}$$

式中,D 为训练数据分布. 通过端到端联合优化编码器参数 θ_f 与解码器参数 θ_d,最小化重建损失 $\mathcal{L}_{\mathrm{MGP}}$.

该算法采用 MGP 的原因有两个:(1)直接使用标签监督训练 ViT 面临收敛困难与高计算开销问题. 即使采用更强的数据增强(如 RandAug)或扩展模型规模(如 DeiT III),传统监督方法在长尾识别(LTR)任务中仍难以达到最先进水平. MGP 由于遮蔽 token 机制的引入,训练速度得到了提升,同时保持了可接受的内存占用. (2)与基于混合增强或对比学习的监督方法不同,MGP 的特征提取能力受类别样本数量影响较小. LiVT 中提出:尽管标签分布不均,只要训练图像的总数量相当,就能获得类似的特征提取能力,这在很大程度上缓解了长尾标签带来的负面影响. 这一特性使其能够为后续的平衡微调阶段提供无偏的初始化特征.

2.4.2　平衡微调

在平衡微调(balanced fine-tuning,BFT)阶段,通过层级分类任务优化模型对长尾分布的适应性. 本阶段采用双分类头架构(父类分类头 parent_head 与子类分类头 sub_head),并基于 LiVT 提出的平衡二元交叉熵损失(balanced BCE,Bal-BCE)对两个分类头进行联合优化. Bal-BCE 主要通过 Logit 调整缓解类别不平衡问题,其定义如式(4)所示:

$$\mathcal{L}_{\mathrm{Bal\text{-}BCE}}=-\sum_{y_i\in C}\omega_i\left[1(y_i)\log\sigma\left(z_{y_i}+\log\frac{\pi_{y_i}}{1-\pi_{y_i}}\right)+(1-1(y_i))\log\left(1-\sigma\left(z_{y_i}+\log\frac{\pi_{y_i}}{1-\pi_{y_i}}\right)\right)\right]. \tag{4}$$

式中,$\pi_{y_i}=\dfrac{n_{y_i}}{N}$为类别 y_i 的样本概率;z_{y_i}为分类器输出的原始 logit 值;$\sigma(\cdot)$表示 Sigmoid 函数;ω_i 为类别权重系数. Bal-BCE 通过偏置项 $\log\pi_{y_i}-\log(1-\pi_{y_i})$动态调整 logit 分布,抑制对头部类的关注度,同时增强尾部类别的梯度信号. 基于此,本文设计父类损失和掩码子类损失,分别如式(5)、(6)所示:

$$\mathcal{L}_{\mathrm{parent}}=\mathcal{L}_{\mathrm{Bal\text{-}BCE(Parent)}}, \tag{5}$$

$$\mathcal{L}_{\mathrm{mask_sub}}=\mathcal{L}_{\mathrm{Bal\text{-}BCE}}(z_{i,s}\odot m_i,y_{i,s}),\ m_i=\begin{cases}1, & s\in S(\hat{p}_i);\\ 0, & \text{其他}.\end{cases} \tag{6}$$

式中,$z_{i,s}$表示样本在子类分类头输出的原始 logits 向量;$y_{i,s}$表示样本 i 在子类 s 上的真实标签;$S(\hat{p}_i)$表示样本 i 的预测父类 $\hat{p}_i$ 包含的子类集合.

为协调父类与子类特征学习的稳定性,本文设计了一种渐进式训练策略,将平衡微调分为 3 个部分:

(1)父类特征学习阶段(epoch $\leqslant$ 15):仅启用父类分类头(parent_head),总损失函数为:

$$\mathcal{L}_{\mathrm{total}}=\mathcal{L}_{\mathrm{parent}}. \tag{7}$$

此阶段通过聚焦父类任务,使编码器充分捕获层级语义特征,避免过早引入子类任务的噪声干扰.

(2)联合优化阶段(15<epoch ≤ 32):同步启用父类与子类分类头,总损失函数为:

$$\mathcal{L}_{\text{total}}=(1-\lambda)\mathcal{L}_{\text{parent}}+\lambda\mathcal{L}_{\text{mask_sub}}. \tag{8}$$

式中,λ 为动态权重参数,随训练轮次从 0 线性增长至 1. 子类损失 $\mathcal{L}_{\text{mask_sub}}$ 通过掩码机制过滤与父类预测无关的子类 logit,以减少冗余梯度干扰. 此阶段通过逐步增加子类损失权重,使模型从粗粒度(父类)向细粒度(子类)特征平稳过渡,缓解训练的不稳定性.

(3)子类强化阶段(epoch>32):仅优化子类分类头(sub_head),总损失函数为:

$$\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{sub}}. \tag{9}$$

此阶段通过冻结父类分类器参数,集中优化子类决策边界,进一步提升细粒度分类性能.

3 实验

3.1 数据集

为验证所提方法在层次化长尾识别任务上的有效性,本文在 3 个具有天然层次结构的标准长尾数据集上进行评估. 这些数据集的共同特点是同时具备层次化标签体系和长尾分布特性,包括 ImageNet-LT[17]、iNaturalist 2018[18]以及 Places-LT[19]. ImageNet-LT 源自 ImageNet 数据集,其标签体系构建于 WordNet 语义层次网络. 每个 Synset(同义词集)在 WordNet 中都有明确的上下位关系,构成了树状的语义层次. 本文基于 WordNet 提取了父类-子类关系,涵盖 1 000 个类别共计 115.8 k 张图像,其类别样本数量遵循 Pareto 分布(参数 $\gamma=256$). iNaturalist 2018 是一个典型的细粒度层次分类数据集,其标签体系基于完整的生物学分类层次:界、门、纲、目、科、属、种. 本文利用其完整的层次结构,构建了父类(纲级别)和子类(种级别)的对应关系. 该数据集包含 8 142 个类别共计 437.5 k 张图像,呈现出极端的长尾分布特性($\gamma=512$). 本文从中构建了父类-子类关系,用于层次分类任务. Places-LT 是 Places 数据集的合成长尾版本,具有内在的语义层次结构,如‘餐厅-快餐店’,涵盖 365 个类别共计 62.5 k 张图像,每个类别的样本数量分布范围为 5~4 980($\gamma=996$). 为确保实验的公平性与可比性,实验采用各数据集的官方验证集进行评估.

3.2 实验设置

在主要基准数据集上的图像分类实验中,所有实验都采用了改进版的 ViT-base-16 模型作为骨干网络,基于原始视觉 Transformer 结构增加了一个分类头,实现层次分类. 所有模型均采用 AdamW 优化器进行训练,其中 β 参数设置为 $\{0.9,0.95\}$. 由于视觉 Transformer 模型需要较长的训练时间才能收敛,HDMViT 进行了 800 个 epoch 的掩蔽式生成预训练,此阶段参数量为 111 M,掩码比例为 0.75,并在 ViT-B 上微调了 100 个 epoch. 训练过程中,所有模型均使用了 RandAug(9,0.5)、mixup(0.8)和 cutmix(1.0)数据增强技术. 模型性能评估包括 Top-1 准确率,并根据样本数量将结果细分为 3 类:Many(>100 张图像)、Medium(20~100 张图像)和 Few(<20 张图像).

3.3 对比实验

表 1 展示了在 ImageNet-LT 数据集上本文方法与当前最先进方法的实验对比结果.

表 1 在 ImageNet-LT 数据集上的 Top-1 准确率

Table 1 Top-1 accuracy on ImageNet-LT %

方法	Many	Medium	Few	Overall
GCL(CVPR 22)	63.0	52.7	37.1	54.5
TSC(CVPR 22)	63.5	49.7	30.4	52.4
NCL(CVPR 22)	67.3	55.4	39.0	57.7
SAFA(ECCV 22)	63.8	49.9	33.4	53.1
ImbSAM(ICCV 23)	63.2	53.7	38.3	55.3
ProCo(TPAMI 24)	70.6	57.4	**40.8**	60.2
GBG(AAAI 24)	—	—	—	58.7
DeiT-LT(CVPR 24)	65.2	54.0	37.1	55.6
LiVT(CVPR 23)	75.2	57.0	33.7	60.8
HDMViT(ours)	74.7	57.9	36.1	61.4
HDMViT * (ours)	**77.8**	**60.2**	36.0	**63.8**

注:* 表示使用 384 分辨率进行训练.

本方法在预训练与微调阶段均采用 224×224 输入分辨率,并设置阈值 $\alpha=20$ 筛选头部子类重构父类标签,同时选取 Top-5 父类预测标签指导子类掩码生成. 实验表明,基于多专家集成的方法(如 RIDE、TADE)通过扩展模型规模提升性能,但需付出显著的计算代价;对比学习方案(如 ProCo、TSC)则依赖长周期优化与大批量训练实现性能增益. 相比之下,HDMViT 在未引入额外数据的情况下,在所有评测指标上均优于基线方法,验证了其高效性与鲁棒性. HDMViT 在重构父类标签时通过阈值 $\alpha=20$ 有效平衡类别分布,结合 Top-5 父类预测的容错机制,显著缓解了过拟合问题,且在平衡微调阶段采用更高分辨率训练时,HDMViT 取得了更高的性能,这一结果验证了 Liu 等[20]工作中提出的视觉 Transformer 对分辨率敏感性的结论一致.

表 2 展示了在 iNaturalist 2018 数据集上本文方法与最先进方法的实验对比结果. 本方法在预训练与微调阶段分别采用 128×128 和 224×224 的输入分辨率,通过阈值 $\alpha=10$ 重构父类标签,并基于 Top-7 父类预测结果生成子类掩码. 实验表明,HDMViT 在所有评测指标(Top-1 准确率、分组准确率)上均显著优于当前 SOTA 方法. 与现有长尾识别方法不同,HDMViT 在提升少样本类别(Few-shot,<20 样本)性能的同时,未牺牲多样本类别(Many-shot,>100 样本)的准确率,表明其层级掩码机制能均衡捕捉头部与尾部特征.

表 2　在 iNaturalist 2018 数据集上的 Top-1 准确率

Table 2　Top-1 accuracy on iNaturalist 2018　　%

方法	Many	Medium	Few	Overall
GCL(CVPR 22)	67.5	71.3	71.5	71.0
TSC(CVPR 22)	72.6	70.6	67.8	69.7
NCL(CVPR 22)	72.0	74.9	73.8	74.2
LiVT(CVPR 23)	74.2	73.8	72.4	73.3
ImbSAM(ICCV 23)	68.2	72.5	72.9	71.1
ProCo(TPAMI 24)	74.0	76.0	76.0	75.8
GBG(AAAI 24)	—	—	—	71.9
DeiT-LT(CVPR 24)	69.0	73.3	73.3	72.9
HDMViT(ours)	75.1	74.5	73.5	74.2
HDMViT * (ours)	**80.9**	**78.9**	**77.5**	**78.5**

注:* 表示使用 384 分辨率进行训练.

表 3 总结了 Places-LT 数据集上的实验结果. 为公平对比,所有长尾识别方法均采用 ImageNet-1K 预训练的 ResNet-152 作为主干网络. 实验表明,HDMViT 通过阈值 $\alpha=20$ 重构父类标签并基于 Top-6 父类预测生成子类掩码,在 Top-1 准确率上优于当前 SOTA 方法. 值得注意的是,Places-LT 数据量远低于 iNaturalist 2018 和 ImageNet-1K,在数据较少时视觉 Transformer 会受到限制. 然而,即便是在数据稀缺的情况下,HDMViT 依旧展现出最佳性能.

表 3　在 Places-LT 数据集上的 Top-1 准确率

Table 3　Top-1 accuracy on Places-LT　　%

方法	Many	Medium	Few	Overall
ALA(AAAI 22)	43.9	40.1	32.9	40.1
CKT(CVPR 22)	41.6	**41.4**	**35.1**	40.2
DeiT(ECCV 22)	**51.6**	31.0	9.4	34.2
Bread(ECCV 22)	40.6	41.0	33.4	39.3
LiVT(CVPR 23)	48.1	37.7	24.2	38.8
HDMViT(ours)	49.3	39.0	26.3	40.2
HDMViT * (ours)	50.5	40.4	25.6	**41.2**

注:* 表示使用 384 分辨率进行训练.

4　分析与讨论

本节通过系统性消融实验验证 HDMViT 核心模块与关键参数设计的有效性,所有实验均在 ImageNet-LT 数据集上进行,基线模型为未引入任何改进的 LiVT 模型(Top-1 准确率 60.8%).

4.1 HDMViT 模块消融分析

为探究 HDMViT 中层级掩码机制各模块的独立贡献与协同作用,依次引入以下组件进行对比:重构父类标签(rebuilt parent class labels,RPL)、双分类头设计(dual classification heads,DCH)、子类 Logit 掩码(dynamic subclass logit masking,DSLM). 如表 4 所示,仅添加 DCH 模块对性能提升微乎其微,表明单纯任务解耦难以应对长尾分布. 进一步引入 RPL 模块或 DSLM 模块后,性能显著改善,验证了层次化标签以及掩码机制的必要性. 最终,三模块联合使用时准确率达 61.4%,较基线提升 0.6%,证明模块间存在协同优化效应.

表 4 HDMViT 组件消融实验

Table 4 Ablations for components in HDMViT

RPL	DCH	DSLM	Overall
×	×	×	60.8
×	✓	×	60.9
✓	✓	×	61.1
×	✓	✓	61.0
✓	✓	✓	61.4

4.2 参数敏感性分析

为进一步探究样本数筛选阈值 α 与父类预测标签数 k 对性能的影响,实验中固定 $\alpha=20$,调整 k 值:当 $k=5$ 时性能最优,k 增大至 7 时准确率下降 0.3%. 这表明过度扩展父类预测范围会引入噪声干扰,当 k 取值在 5~7 间波动较小. 随后,固定 $k=7$,调整 α 值:当 $\alpha=20$ 时性能最优,降低 α 至 18 时准确率下降 0.3%. 过低的 α 会导致误将尾部子类提升为父类,破坏层级语义一致性. 表 5 结果表明,$\alpha=20$ 与 $k=5$ 的组合在容错性与判别性间达到最佳平衡:$\alpha=20$ 时能有效区分头部子类(样本数>20),避免层级结构过载;$k=5$ 时能覆盖潜在关联父类,同时抑制噪声传播.

表 5 参数敏感性分析

Table 5 Parameter sensitivity analysis

参数		Accuracy/%	参数		Accuracy/%
$\alpha=20$	$k=5$	61.4	$k=7$	$\alpha=18$	61.0
	$k=6$	61.3		$\alpha=19$	61.2
	$k=7$	61.1		$\alpha=20$	61.3

5 结论

本文针对层次化长尾图像识别这一问题,提出了一种高效的 HDMViT 算法. 该算法的核心思想是显式利用标签的层次结构来引导模型在长尾分布下的学习过程. 训练主要分为两个阶段:掩蔽式生成预训练和平衡微调. 掩蔽式生成预训练大大降低了不平衡标签对特征学习的有害影响,使得视觉 Transformer 能够在无需额外数据的情况下从头训练. 随后,在平衡微调阶段对视觉 Transformer 模型进行结构优化,引入额外的分类头,使其能够捕捉层次化标签的语义信息,并增强对上层类别特征的学习. 此外,提出动态 Logit 掩码机制,通过屏蔽目标 logit 向量中的无关类别信息,进一步减少类别不均衡对模型决策的影响. HDMViT 在不依赖额外数据的情况下成功实现了视觉 Transformer 的训练,并在长尾数据场景下显著提升了分类性能. 尽管 HDMViT 在长尾图像识别中表现出色,但仍存在一些局限性:一方面,HDMViT 采用的掩蔽式生成预训练需 800 个 epoch,计算成本较高,且 111M 参数量进一步加剧了硬件资源消耗;另一方面,父类标签重构的阈值 α 需基于不同数据集通过经验性实验确定,缺乏自适应调整能力. 未来工作可探索轻量预训练方法、层级标签处理方法,并扩展至动态增量学习场景;同时优化动态掩码模块的计算逻辑以降低开销,设计结合数据集分布特性的自适应阈值策略.

[参考文献](References)

[1] 石泽男,陈海鹏,张冬,等. 预训练驱动的多模态边界感知视觉 Transformer[J]. 软件学报,2022,34(5):2051-2067.

[2] Li M K,Cheung Y M,Lu Y. Long-tailed visual recognition via gaussian clouded logit adjustment[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition.New Orleans,LA,USA:IEEE,2022.

[3] Li J,Tan Z C,Wan J,et al. Nested collaborative learning for long-tailed visual recognition[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition.New Orleans,LA,USA:IEEE,2022.

[4] 韩佳艺,刘建伟,陈德华,等. 深度长尾学习研究综述[J]. 自动化学报,2024,51(5):985-1020.

[5] Dablain D,Krawczyk B,Chawla N V. DeepSMOTE:Fusing deep learning and SMOTE for imbalanced data[J]. IEEE Transactions on Neural Networks and Learning Systems,2023,34(9):6390-6404.

[6] Goodfellow I,Pouget-Abadie J,Mirza M,et al. Generative adversarial networks[J]. Communications of the ACM,2020,63(11):139-144.

[7] Park S,Lim J,Jeon Y,et al. Influence-balanced loss for imbalanced visual classification[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal,QC,Canada:IEEE,2021.

[8] Zhou Y X,Qu Y,Xu X,et al. ImbSAM:A closer look at sharpness-aware minimization in class-imbalanced recognition[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris,France:IEEE,2023.

[9] Ren J W,Yu C J,Sheng S N,et al. Balanced meta-Softmax for long-tailed visual recognition[J]. Advances in Neural Information Processing Systems,2020,33:4175-4186.

[10] Rangwani H,Mondal P,Mishra M,et al. DeiT-LT:Distillation strikes back for vision transformer training on long-tailed dataset[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,WA,USA:IEEE,2024.

[11] Cao K D,Wei C,Gaidon A,et al. Learning imbalanced datasets with label-distribution-aware margin loss[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver BC,Canada:NIPS,2019.

[12] Kang B Y,Xie S N,Rohrbach M,et al. Decoupling representation and classifier for long-tailed recognition[PP/OL]. V2. arXiv (2020-02-19)[2025-08-02]. https://doi.org/10.48550/arXiv.1910.09217.

[13] Chen T,Kornblith S,Norouzi M,et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of the 2020 International Conference on Machine Learning.PmLR,2020.

[14] He K M,Fan H Q,Wu Y X,et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,WA,USA:IEEE,2020.

[15] Lin T Y,Goyal P,Girshick R,et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,42(2):318-327.

[16] Xu Z Z,Liu R K,Yang S,et al. Learning imbalanced data with ViTs[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver,BC,Canada:IEEE,2023.

[17] Liu Z W,Miao Z Q,Zhan X H,et al. Large-scale long-tailed recognition in an open world[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,CA,USA:IEEE,2019.

[18] Van Horn G,Mac Aodha O,Song Y,et al. The inaturalist species classification and detection dataset[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT,USA:IEEE,2018.

[19] Zhou B L,Lapedriza A,Khosla A,et al. Places:A 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,40(6):1452-1464.

[20] Liu Z,Lin Y T,Cao Y,et al. Swin transformer:Hierarchical vision transformer using shifted windows[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal,QC,Canada:IEEE,2021.

[责任编辑:严海琳]