

长尾分布下图像分类方法综述

杨春晖

(安庆职业技术学院信息技术学院, 安徽 安庆 246003)

摘要 长尾分布下的图像数据会显著降低分类模型性能, 如何缓解该问题是图像分类研究中的关键课题。本文从类别重平衡、信息增强与模块改进三大技术路径展开研究, 系统梳理了近年来长尾图像分类方法的研究进展, 并进一步将结合大模型的方法作为第四大类进行归纳分析, 对 Logit 调整、集成学习等各类代表性方法进行清晰解读, 剖析其核心思想。最后探讨当前领域面临的主要挑战, 并对大模型融合方法探索、具体应用领域适配等未来研究方向进行合理展望, 以期相关人员提供参考。

关键词 长尾分布; 图像分类; 类别不平衡; 大模型

中图分类号: TP391.41

文献标志码: A

DOI: 10.3969/j.issn.2097-3365.2026.24.007

0 引言

图像分类任务是计算机视觉领域的关键任务, 在工业质量检测、医学影像分析、野生动物识别等诸多现实领域中发挥着重要作用。然而长久以来, 在实际应用中却始终面临数据不平衡这一基本问题, 尤其是真实世界的图像数据分布往往呈现出明显的长尾分布特征。这种极度不平衡的分布会导致模型训练过程被头部类别主导, 决策边界向头部偏移, 最终使得整体分类准确率尤其是对尾部类别识别的准确率显著下降。因此, 长尾分布下图像分类方法的研究成为计算机视觉领域的重要课题, 受到人们的广泛关注。本文聚焦长尾分布下图像分类问题, 围绕类别重平衡、信息增强、模块改进与结合大模型四个大类进行研究, 系统综述各类方法与典型方法, 并对未来发展趋势进行合理展望。

1 长尾分布问题概述

1.1 长尾分布定义

长尾分布即指在数据集中, 少数类别拥有大量图像样本, 而大多数类别则只有极少量样本的一种极端类别不平衡现象^[1]。用幂律分布描述, 设训练集包含 c 个类别, 各类别样本数量记为 $\{n_1, n_2, \dots, n_c\}$, 按降序排列 $n_1 \geq n_2 \geq \dots \geq n_c$ 。则有 $n_c \propto c^{-\alpha}$, $\alpha > 0$, 其中 α 控制分布的倾斜程度, α 值越大, 尾部越“长”。不平衡因子 $\rho = n_1/n_c$ 用来衡量数据平衡程度, $\rho=1$ 代表完全平衡, 通常 $\rho \geq 100$ 时表示当前数据分布重度不平衡。

1.2 基准数据集与评价指标

主流的长尾分类基准数据集包括 CIFAR-100-LT、ImageNet-LT、iNaturalist 2018 及 Places-LT 等。其中既包含人造数据也包含真实数据, 能够满足细粒度与粗粒度多种长尾分布场景。为了反映评估结果, 在评价指标方面, 目前通常采用分层评估: 即对头部类、中部类、尾部类各自的准确率进行综合评估, 以及宏观平均 F1 分数等指标也被广泛使用。

2 四类技术范式的梳理与研究

2.1 基于类别重平衡的方法

类别重平衡的主要思路是, 通过在不同层面施加直接或间接的人为干预, 对模型训练过程中各类别的影响力进行重新平衡。即拉近头部类和尾部类对训练的影响力差异, 缓解头部类别主导训练这一问题。依据具体实现方式, 该大类可划分为重采样、类敏感学习和 Logit 调整三个细类。

重采样方法是最简单直观类别重平衡: 包括过采样、欠采样和类平衡采样, 在训练中直接对数据入手, 分别通过复制尾部类、丢弃头部类以及平衡采样概率的方式, 使各类别在训练过程中获得的采样机会或出现频次趋于均衡^[2]。类敏感学习方法主要包括重加权和边际损失, 两类方法均以损失函数作为切入点。重加权的核心思想是根据各类别的样本数量来分配损失权重, 使尾部类别获得更大的梯度。典型重加权包括按类别样本数倒数直接加权, 以及基于有效样本数

作者简介: 杨春晖 (1998-), 男, 硕士研究生, 助教, 研究方向: 人工智能和计算机视觉。

等更精细的理论进行权重设计等。边际损失方法通过引入类别相关的边际来增大尾部类别的决策边界,将类别边际与样本数量关联,使样本数量越少的图像类别获得越大的边际,从而在特征空间中为尾部类别赋予更大的分类边际。

Logit 调整方法则是从贝叶斯决策理论出发推导,在交叉熵损失中引入基于类先验相关的校正项,从统计层面缓解了头部类对决策边界的挤压,以此提升尾部类别识别能力。该方法具备较高的灵活性,既可用于训练阶段,也可在推理时后处理Logit 输出。例如:赵亮^[3]等在延续了Logit 层面进行校正的基础上,借鉴批归一化的思想,进一步提出Logit 归一化(LogN)方法。该方法通过对各类别输出Logit 进行标准化校准来消除长尾偏差,同时针对目标检测任务中前景与背景样本严重不平衡的特有问题,引入了背景校准机制以恢复背景判别力。LogN 方法具备无需额外训练、不改动模型结构且即插即用的特点,在ImageNet-LT 等长尾图像分类数据集上展现出优异性能。

2.2 基于信息增强的方法

长尾分布下图像分类问题的根源在于数据本身分布的严重不平衡,因此信息增强类别的方法直接干预数据本身,为尾部类别补充信息量。该类别包括数据增强与迁移学习两类方法。

数据增强通过对原始图像施加各类变换产生新样本,以此扩充尾部类别的数据规模,增加特征多样性。常见的变换方式包括几何变换、颜色变换、图像混合等。例如:以Mixup 为基础衍生出诸多改进方法,通过赋予尾部类别图像更高的混合权重,使尾部样本在以头部样本为主导的混合过程中获得更多曝光机会,从而强化其在训练中的影响力。此外,基于生成对抗网络和扩散模型的生成式数据增强方法近些年也逐渐受到关注,进一步突破了传统变换的局限。

迁移学习认为尽管不同类别在视觉外观上差异显著,但在深度特征空间中,许多基础且通用的视觉模式(边缘、形状、纹理等)具有跨类别共享的特性。因此其核心思想是借用现有的从样本充足的源域中已学到的稳定特征表示,迁移到信息稀缺的尾部类别上。根据知识来源的不同,迁移学习可细分为:从外部数据集迁移的模型预训练、从教师模型向轻量学生模型迁移的模型蒸馏,以及从头部类别向尾部类别迁移特征表示的头尾知识迁移等。三种迁移策略知识来源各不相同,但核心思想一致:通过知识复用来弥补尾部类别的信息缺口,为模型提供更丰富的特征表示基础。

2.3 基于模块改进的方法

模块改进类方法的思路与前两种不同,既不尝试平衡不同类别的影响力,也不从数据层面补充尾部类别的信息量,而是改造模型内部组件或训练流程,让模型本身就具备对抗长尾分布不平衡的能力。这类思路可分为解耦训练、表征学习、分类器设计和集成学习四种方法。

解耦训练^[4]的提出是因为在研究中发现:即使在长尾分布下,特征提取器可以学到高质量的视觉表示,但分类器会受到头部类别的严重干扰,最终导致决策边界偏移。因此解耦训练将二者分离:第一阶段在原始长尾数据上端到端训练,获得泛化能力良好的特征提取器;第二阶段冻结特征提取器参数,采用类别均衡采样策略,重新训练分类器。在具体应用领域中,尹乾等将解耦训练策略引入脉冲星识别任务。该方法经过在HTRU_bands 和 HTRU_ints 数据集测试,召回率相较基准方法得到显著提升,验证了解耦训练策略在跨领域长尾分类任务中的通用性与可移植性。

表征学习聚焦改进特征提取器,旨在获得更均衡的特征表示。其中,对比学习通过拉近同类样本、推远异类样本的方式增强特征空间的类别区分性;原型学习则为每个类别在特征空间里学出一个或几个能代表该类别语义中心的特征向量作为原型,强制样本向其所归属类别的原型靠拢,防止尾部类特征被头部类淹没。分类器设计则是旨在通过权重归一化、原型分类器等方式,消除分类器层在长尾训练中产生的结构性偏见。例如:权重归一化通过将所有类别的分类器权重模长强制约束至同一水平,使决策不再受权重模长差异的影响,而仅依赖于特征与权重向量之间的方向相似度,从而有效纠正了分类器层的结构性偏见。

集成学习的研究思路是构建多个差异化学习模块,先分而治之再综合集成以此获得比单一模型更稳健的结果。这类方法最具代表性的就是多专家模型:通过构建多分支网络,在一个完整网络内部构造多个专家,通过共享底层特征和专业化学习缓解长尾分布带来的决策偏向。例如:崔国恒等近期提出的多专家动态协作模型^[5],包含多专家网络和动态学习策略两部分。在多专家网络中设计了一个传统分支和两个再平衡分支,通过多专家协作实现头部尾部平衡兼顾的效果。动态学习策略在训练过程中动态调整两个再平衡分支对模型的影响力,防止因过度关注尾部类而损害模型对头部类的判别能力。经过实验测试,该模型在多个长尾声呐图像数据集上表现出非常优秀的整体准确率。

2.4 基于多模态大模型的新方法

随着近几年视觉—语言模型（VLM）与大语言模型（LLM）等大规模预训练模型的蓬勃发展，越来越多的研究者开始尝试结合大模型，利用大模型蕴含的跨模态知识来弥补尾部视觉样本不足这一基本矛盾。这一类研究思路中，根据大模型结合的层面不同，大体可划分为检索与生成式数据增强、语义表征引导和跨模态知识迁移的三类方法。

检索与生成式数据增强的思路是将跨模态知识转化为补充的训练样本，直接在输入端补充数据以缓解尾部类数据匮乏的根本矛盾。这一方法充分发挥了大模型优势，从语义层面创造或检索出新的样本，打破了传统数据增强仅能进行像素级变换的局限，将跨模态知识系统性地引入信息增强范式。

语义表征引导的思路是利用 LLM 生成高质量的文本描述，再由 VLM 的文本编码器编码为文本特征原型，以此作为训练时额外的监督信号，在特征或损失层面引导视觉模型学习更均衡的类别表示，从而强化对尾部类的监督。例如：李象贤等针对长尾分类中尾部类表征学习不足的问题，提出了一种基于跨模态特权信息增强的图像分类框架^[6]。该方法将图文描述词作为跨模态特权信息，仅在训练阶段引入，通过类感知信息选择（CIS）筛选高质量文本描述词，再由部分异构对齐（PHA）机制将筛选后的文本特征映射至视觉特征空间，引导图像特征从视觉空间向语义空间靠拢。该方法在 NUS-WIDE 和 VireoFood-172 等长尾跨模态数据集上的实验结果均证明了该框架的有效性。

与语义表征引导不同，跨模态知识迁移则通常采用跨模态知识蒸馏范式，以 VLM 为教师模型，引导下游的视觉学生模型学习跨模态对齐能力，将 VLM 已掌握的图文对齐能力直接迁移至视觉模型。该类方法建立在 CLIP 等预训练模型的基础之上——CLIP 首次证明了在海量图文匹配学习的基础上，模型可以获得对开放概念的跨模态理解能力。例如：史江明等提出的 CLIP2FL 方法^[7]，在联邦学习框架下构建了完整的跨模态知识迁移方案。该方法通过知识蒸馏将 CLIP 的跨模态知识迁移至各客户端模型，以提升其在异构长尾数据上特征表示的质量。同时以 CLIP 文本编码器生成的类别语义原型为监督信号，引入原型对比学习机制，生成类别均衡的联邦特征并重训分类器。这种将大规模预训练的跨模态先验注入联邦学习流程的方法，有效缓解了长尾分布导致的决策偏见。

3 结束语

本文从类别重平衡、信息增强与模块改进经典三范式出发，系统梳理长尾分布下图像分类的各种经典方法及其核心思想。在此基础上结合先进技术发展，引入结合大模型方法作为第四范式，梳理现有相关方法，进一步归纳了检索与生成式数据增强、跨模态知识迁移和语义表征引导的三个主要分支。类别重平衡着力于平衡训练过程中各类别明显的影响力差异，纠正头部类的主导地位；信息增强聚焦数据本身，通过补足尾部样本信息量缓解数据分布不均；模块改进致力于优化模型结构或训练流程，使模型对不平衡数据具备更强的适应性。结合大模型的第四范式则突破传统视觉数据局限，引入丰富的跨模态语义知识，为尾部类别提供了更强大的先验信息或监督支撑，在尾部样本极度稀缺甚至部分零样本场景下展现出显著优势。当前长尾分布下的图像分类研究在各类方法层面均已取得丰富成果，但在实践中仍面临诸多挑战：现有研究大多假设训练集与测试集具备相似的静态分布，但对真实场景中开放、动态的长尾分布图像的适应能力仍显不足；诸多先进方法的计算开销与资源受限场景的部署需求之间存在显著差距；大模型结合等新路径的发展也带来基准污染与数据泄露、现有评估体系不完备不统一等新的挑战。未来，多类方法的有机融合与协同互补，尤其是结合多模态大模型的研究方兴未艾，相关研究仍有巨大深挖潜力。结合大模型范式下的长尾图像分类新方法研究，以及对医学影像、遥感监测、工业质检等具体细分领域的适配性研究和对应落地方案的探索实验，具有广阔的发展前景。

参考文献：

- [1] 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, 2025, 51(05): 985-1020.
- [2] 魏秀参. 长尾分布图像识别: 途径与进展 [J]. 中国基础科学, 2023, 25(05): 48-55.
- [3] 赵亮, 滕宇, 王亮. 长尾目标检测中的 Logit 归一化 [J]. 国际计算机视觉杂志, 2024, 132(06): 2114-2134.
- [4] 尹乾, 车润琪, 杨如意, 等. 基于解耦训练的脉冲星识别算法 [J]. 光学技术, 2023, 49(06): 680-684, 698.
- [5] 崔国恒, 周浩, 王超, 等. 基于多专家动态协作学习的长尾声呐图像识别 [J]. 山西大学学报(自然科学版), 2026, 49(20): 232-243.
- [6] 李象贤, 郑裕泽, 马浩凯, 等. 基于跨模态特权信息增强的图像分类方法 [J]. 软件学报, 2024, 35(12): 5636-5652.
- [7] 史江明, 郑珊珊, 尹星博, 等. 异构及长尾数据下的 CLIP 引导联邦学习 [C]// 第 38 届 AAAI 人工智能会议论文集. 2024.