软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software [doi: 10.13328/j.cnki.jos.007052]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于跨模态特权信息增强的图像分类方法*

李象贤, 郑裕泽, 马浩凯, 齐 壮, 闫晓硕, 孟祥旭, 孟 雷

(山东大学 软件学院, 山东 济南 250101)

通信作者: 孟雷, E-mail: lmeng@sdu.edu.cn

摘 要: 图像分类算法的性能受限于视觉信息的多样性和背景噪声的影响, 现有研究通常采用跨模态约束或异构特征对齐算法学习可判别力强的视觉表征. 然而, 模态异构带来的特征分布差异等问题限制了视觉表征的有效学习. 针对该问题, 提出一种基于跨模态语义信息推理和融合的图像分类框架 (CMIF), 引入图像语义描述及统计先验知识作为特权信息, 使用特权信息学习范式在模型训练阶段指导图像特征从视觉空间向语义空间映射, 提出类感知的信息选择算法 (CIS) 学习图像的跨模态增强表征. 针对表征学习中的异构特征差异性问题, 使用部分异构对齐算法 (PHA) 实现视觉特征与特权信息中提取的语义特征的跨模态对齐. 为进一步在语义空间中抑制视觉噪声带来的干扰, 提出基于图融合的 CIS 算法选取重构语义表征中的关键信息, 从而形成对视觉预测信息的有效补充. 在跨模态分类数据集 VireoFood-172 和 NUS-WIDE 上的实验表明, CMIF 能够学习鲁棒的图像语义特征, 并且能够作为通用框架在基于卷积的 ResNet-50 和基于 Transform 架构的 ViT 图像分类模型上取得稳定的性能提升.

关键词: 图像分类; 跨模态学习; 特权信息; 特征对齐; 图卷积网络

中图法分类号: TP391

中文引用格式: 李象贤, 郑裕泽, 马浩凯, 齐壮, 闫晓硕, 孟祥旭, 孟雷. 基于跨模态特权信息增强的图像分类方法. 软件学报. http://www.jos.org.cn/1000-9825/7052.htm

英文引用格式: Li XX, Zheng YZ, Ma HK, Qi Z, Yan XS, Meng XX, Meng L. Image Classification Method Based on Cross-modal Privileged Information Enhancement. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7052.htm

Image Classification Method Based on Cross-modal Privileged Information Enhancement

LI Xiang-Xian, ZHENG Yu-Ze, MA Hao-Kai, QI Zhuang, YAN Xiao-Shuo, MENG Xiang-Xu, MENG Lei

(School of Software Engineering, Shandong University, Jinan 250101, China)

Abstract: The performance of image classification algorithms is limited by the diversity of visual information and the influence of background noise. Existing works usually apply cross-modal constraints or heterogeneous feature alignment algorithms to learn visual representations with strong discrimination. However, the difference in feature distribution caused by modal heterogeneity limits the effective learning of visual representations. To address this problem, this study proposes an image classification framework (CMIF) based on cross-modal semantic information inference and fusion and introduces the semantic description of images and statistical knowledge as privileged information. The study uses the privileged information learning paradigm to guide the mapping of image features from visual space to semantic space in the training stage, and a class-aware information selection (CIS) algorithm is proposed to learn the cross-modal enhanced representation of images. In view of the heterogeneous feature differences in representation learning, the partial heterogeneous alignment (PHA) algorithm is used to achieve cross-modal alignment of visual features and semantic features extracted from privileged information. In order to further suppress the interference caused by visual noise in semantic space, the CIS algorithm based on graph fusion is selected to reconstruct the key information in the semantic representation, so as to form an effective supplement to the visual prediction information. Experiments on the cross-modal classification datasets VireoFood-172 and NUS-WIDE show that CMIF can learn robust semantic features of images, and it has achieved stable performance improvement on the convolution-based ResNet-50 and

* 基金项目: 山东省优秀青年科学基金 (海外) 计划 (2022HWYQ-048); 济南市科技局"新高校 20 条"资助项目引进创新团队计划 (2021GXRC073); 国家重点研发计划 (2021YFC3300203)
收稿时间: 2022-12-06; 修改时间: 2023-03-21, 2023-06-22; 采用时间: 2023-09-04; jos 在线出版时间: 2024-01-31

Transform-based ViT image classification models as a general framework.

Key words: image classification; cross-modal learning; privileged information; feature alignment; graph convolution network (GCN)

图像分类作为计算机视觉领域重要的基础任务之一, 广泛运用在人脸识别[1,2], 智能驾驶[3,4]等领域. 传统的图像识别方法往往使用单一的图像信息, 通过人工手动提取特征[5,6]或以深度学习端到端的方式[7,8]训练模型实现分类预测, 然而仅依靠视觉模态学习图像分类容易在图像主体不明确、拍摄环境混杂的情况下, 让模型关注到图像背景信息从而产生错误的分类预测. 近年来, 随着大量多模态数据被上传到社交媒体和网络平台, 图像分类领域的研究也尝试将特征空间中维度更低的图像描述信息加入模型的学习当中[9], 但由于实际应用中的图像数据通常缺少准确而直接的文本描述, 因此该方向的研究更多采用特权信息学习 (LUPI) 的范式[10–12], 将跨模态的图像描述信息作为仅在训练过程中能够使用的特权信息加入模型学习当中.

基于 LUPI 范式提升图像分类的现有工作主要分为两类. 一类是基于跨模态约束的方法, 这类方法将模型对图像整体描述信息的预测[13−17]或对图像局部对应单词的预测[18−21]作为图像分类之外的一个附加任务, 从而隐式地约束图像表征的学习过程[22]; 但由于多任务约束的过程难以控制, 导致优化过程不稳定, 因此跨模态约束带来的提升往往有限. 另一类方法是通过基于特征对齐的方式将从图像中学习到的视觉表征向着从特权信息中提取到的语义表征进行显式地相似性约束[23,24]或分布对齐[25−27], 并进一步尝试关注到视觉表征中关键的分类信息[28,29], 这类方法中的文本和图像表征属于异构模态, 它们在特征分布和值域范围等方面存在差异, 导致跨模态映射实效有限. 因此, 为从跨模态信息中获取到更多有利于图像分类的知识, 需要首先缓解模态异构带来的跨模态信息获取的问题, 并进一步将获取到的知识充分融合到图像分类任务中.

针对上述问题, 本文提出了基于跨模态信息推理和融合的图像分类框架 (CMIF), 引入图像的语义描述及其统计信息作为特权信息, 来指导模型将特征从视觉空间向语义空间映射, 从而实现从图像到描述信息的推理, 将推理出的语义信息作为补充, 缓解模态异构问题对现有的基于跨模态特征对齐的视觉表征学习效果的限制; 上述过程中, 由于跨模态映射存在着错误传播的问题, 本文设计了类感知信息选择算法 CIS, 通过先验知识对关键语义信息实现筛选, 从而提升跨模态信息对视觉表征学习的补充效果. CMIF 的总体思路如图 1 所示, CMIF 通过引入图像描述信息作为特权信息, 一方面用特征对齐强化了对视觉表征的抽取, 另一方面通过跨模态推理挖掘视觉中的语义信息, 二者融合实现信息的互补, 提升模型的图像分类效果. 在任务通道中, CMIF 提取图像的视觉表征, 并在训练中通过部分异构对齐算法 (PHA) 使模型关注到视觉表征中的分类关键信息; 在特权通道中, CMIF 通过特权信息辅助模型学习将视觉表征跨模态迁移为语义空间的视觉-语义表征, 并产生相应的语义预测. 由于跨模态迁移受到视觉噪声带来的误差传播影响, 本文提出在特权通道中使用类感知信息选择算法 (CIS), CIS 通过由视觉潜在类选择的类别-语义先验知识和模型的语义预测构建关系图, 从而筛选并融合关键的跨模态语义信息. 最终通过视觉信息与跨模态信息的跨通道融合, CMIF 有效实现了跨模态的信息互补, 提升了分类预测的效果. CMIF 在 LUPI 学习范式下增强了模型对图像表征的学习能力和对语义信息的推理能力, 相较于先前的跨模态约束方法能够在跨模态预测时减少错误传播问题, 同时相较于先前的跨模态对齐方法能够有效缓解模态异构对表征学习的影响, 为多模态数据缺乏时的图像分类任务提供了有效的学习框架.

为验证提出方法的有效性, 本文在两个真实世界的数据集 VireoFood-172[13]和 NUS-WIDE[30]上进行了实验, 其中 NUS-WIDE 为多标签分类数据集, 图像样本包含在 81 个类别中, 对应文本描述信息分为 1 000 类; VireoFood-172 是包含 172 种类别的单标签分类数据集, 含有 353 种食材文本信息. 对比实验表明, CMIF 框架应用在基于卷积和基于 Transform 架构[31]的视觉骨干网络后, 都取得了明显的性能提升. 进一步的消融实验证明了 CMIF 框架中各模块有信息互补的效果, 同时也展现出类感知信息选择对分类性能的提升. 本文还通过深入分析探究了 CIS 算法中关键参数对信息选择结果的影响, 以及不同融合策略下性能提升的效果. 最后, 本文通过案例分析展示了 CMIF 框架中, 模型学习不同模态表征时关注点的变化, 从而体现各模块协作的优势.

综上, 本文的主要贡献如下.

(1) 提出了基于特权信息学习的跨模态图像分类框架 CMIF, 通过引入跨模态推理, 实现了对视觉表征学习的信息互补, 有效缓解了模态异构对特征对齐的限制, 从而提升了图像分类性能.

(2) 提出了类感知信息选择算法 CIS, 对跨模态推理中的噪声实现有效抑制, 并通过构建关系图充分融合模型的预测信息和先验知识, 实现跨模态表征的增强.

(3) 通过使用不同架构网络, 在不同领域数据的实验验证 CMIF 具有良好的泛化能力, 为多模态数据缺乏时的图像分类任务提供了有效的学习方法.

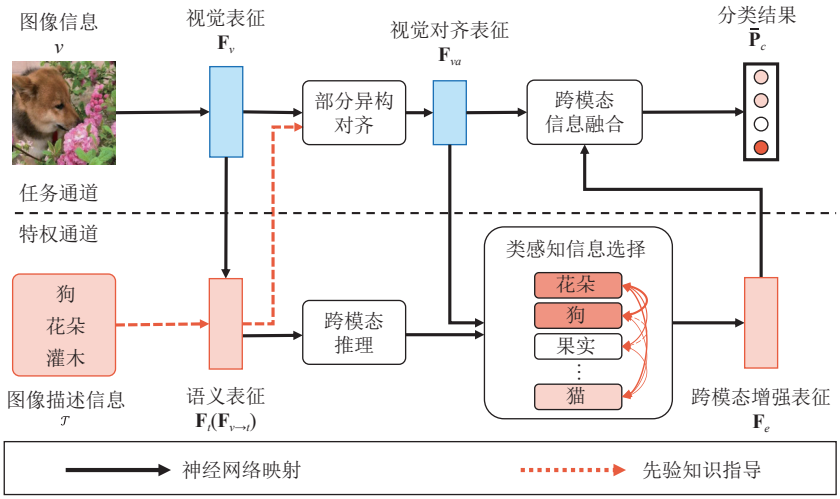


图 1 跨模态图像分类框架 CMIF

本文第 1 节将介绍特权信息学习范式的含义以及使用跨模态语义描述信息提升图像分类的相关工作. 第 2 节为问题陈述, 包括跨模态图像分类的普遍方法和本文提出方法的描述. 第 3 节介绍本文提出的跨模态图像分类框架 CMIF 的具体内容. 第 4 节通过实验验证所提模型的有效性. 最后总结全文并提出展望.

1 相关工作

1.1 特权信息学习范式

利用特权信息学习 (learning using privileged information, LUPI) 范式由 Vapnik 等人[10,11]提出, 特权信息是仅在模型的训练阶段可用而在测试阶段则不可用的信息. 特权信息学习范式对于多模态学习有着重要的现实意义, 如今的大规模数据通常来自互联网, 但由于用户和平台等因素的多样性, 多模态信息可能存在缺失和质量参差不齐的问题. 引入特权信息学习范式则能够在个别模态数据不足的情况下, 支撑模型在训练中用多种模态之间的互补性, 筛除模态之间的冗余, 从而提升模型的表征提取等能力. 在基于特权信息增强的图像分类相关工作中, 图像的描述文本、主体的边界框等信息常被作为特权信息引入[12,28], 随着社交媒体的发展, 图像及其对应描述信息获得成本降低, 且描述信息中的文本具有维度更低、特征空间中可分性更好的特点因此被广泛采用, 这类工作通常使用文本信息对于特征空间中表征更复杂的图像特征的学习进行正则化约束从而提升模型的表征提取和分类能力.

1.2 基于跨模态约束的图像分类

在使用跨模态语义描述信息作为特权信息的图像分类方法中, 一类重要的方法是基于跨模态约束的方法, 该类方法将模型对跨模态图像描述信息的预测作为除图像类别预测之外的附加任务, 以此对视觉特征的提取实现正则化约束. 这类方法可以分为跨模态全局约束和跨模态局部约束两种策略, 跨模态全局约束的策略是将图像的整体描述信息作为语义预测目标, 将视觉特征映射为跨模态信息[13−16]作为图像分类之外的另一任务目标, 引入这种约束的主要思路是促进模型在提取视觉信息的同时兼顾特征中包含的语义信息, 从而在表征过程中过滤干扰信息; 但对图像的全局约束缺少对图中语义目标的针对性, 实际的提升效果并不稳定, 因此跨模态局部约束的策略[17−22]则是通过人为或语义驱动的方式提取原始图像中不同尺度的区域, 如通过特征图的池化操作获得子区域[18,19], 或

通过空间变换网络 (STN)[20]搜寻图像中可能的语义区域等. 在找到相应区域后, 模型对每一个区域分别进行语义描述信息的预测, 这类方法通常比整体约束方法能够产生更有鲁棒性的预测结果, 但通常需要设计更复杂的模型结构, 且仍然受限于视觉错误传播的影响, 视觉中错误的关注点仍继续延续到跨模态的预测中, 从而使得对分类的提升效果受限.

1.3 基于异构特征对齐的图像分类

使用特权信息学习的另一类图像分类方法是基于异构特征对齐的方法, 这类方法主要可以分为特征相似性约束、特征分布对齐、特征解耦后对齐这 3 种主要方式. 特征的相似性约束方法是在不同模态的特征间使用 KL 散度、L2 范数等距离度量函数, 计算不同模态特征间的差距并作为优化目标[23,24], 这类方法能够一定程度上减少不同模态的特征在某种度量空间中的距离, 然而在与分类任务结合时可能会影响目标任务的优化, 因此对分类性能的提升有限; 特征分布对齐则是进一步计算跨模态特征在特定分布假设下的分布差异, 如高斯分布中均值和方差的分布距离[25−27]; 上述两类方法, 即特征的相似性约束方法和分布对齐方法由于模态异构导致的特征值域差异等因素的影响, 跨模态对齐的效果和对分类性能的影响仍不足; 因此近年来的工作针对模态异构问题, 尝试特征解耦后对齐的方式, 减少模态异构带来的负面影响, 这类方法将特征中包含的信息分解为关键信息 (目标任务需要的信息) 和非关键信息 (如模态的风格化信息等), 并设计算法在跨模态对齐当中通过模态间信息的互补剔除非关键信息, 如仅对齐部分关键特征的部分异构对齐方法[28], 和对特征中非重要部分乱序进而使模型学习到关键信息的方法[29], 这类方法一定程度上缓解了模态异构对表征对齐的不利影响, 但解耦的效果受到表征提取效果的限制, 一些重要的信息也可能在结构当中损失.

2 问题陈述

基于特权信息学习的图像分类方法借助跨模态信息在特征表达方面的优势, 如特征的维度更低、在特征空间中的可分性更好等, 来约束模型的视觉表征学习过程, 提升所学视觉特征在特征空间中的可分辨性, 进而提升模型最终的分类效果. 为便于理解和参照, 我们将本文所使用的关键符号及其定义在表 1 中列出. 包含跨模态特权信息的图像分类数据集 $\mathcal{D}$ 通常可以表示为 $\mathcal{D}=\{\mathcal{V},\mathcal{T}^*,\mathcal{Y}\}$, 其中 $\mathcal{V}=\{v_n|n=1,2,\ldots,N\}$ 表示图像集合, $\mathcal{T}=\{t_n|n=1,2,\ldots,N\}$ 表示图像对应的描述信息, $\mathcal{Y}=\{y_c|c=1,2,\ldots,C\}$ 表示图像类别标签集合, N 表示图像样本数量, C 为标签类别数, * 指某一信息作为特权信息, 仅在训练数据中出现. 在基于跨模态特征对齐的方法中, 通常在训练时先分别从图像 $\mathcal{V}$ 和对应描述信息 $\mathcal{T}$ 中提取视觉表征 $\mathbf{F}_v$ 和语义表征 $\mathbf{F}_t$, 为了使模型在视觉表征的学习中学到 $\mathbf{F}_t$ 的特性, 一般而言首先将视觉表征和语义表征映射到跨模态对齐的特征空间中, 而后通过与语义对齐表征 $\mathbf{F}_{ta}$ 的相似性约束提升视觉对齐表征 $\mathbf{F}_{va}$ 的表征能力, 并基于 $\mathbf{F}_{va}$ 获得模型对类别的预测 $\tilde{\mathbf{P}}_{va}$, 该过程可描述为 $\mathcal{V}\rightarrow\mathbf{F}_v\rightarrow\mathbf{F}_{va}\rightarrow\tilde{\mathbf{P}}_{va}$.

表 1 本文中所使用的主要符号和对应定义

符号	定义
$\mathcal{D}$	跨模态图像分类数据集
$\mathcal{V}$	数据集中的图像
$\mathcal{T}$	数据集中图像对应的描述信息
$\mathcal{Y}$	数据集中的标签
$\mathbf{F}$	模型提取的表征, 通过下标区分信息来源
$\tilde{\mathbf{P}}$	模型输出的类别预测, 通过下标区分信息来源
$\tilde{\boldsymbol{T}}$	跨模态预测的单词信息, 通过下标区分信息来源
$\mathbf{K}$	来自训练集统计的先验知识, 通过下标区分知识的类型
$\mathbf{w}$	基于先验信息计算的权重矩阵, 通过下标区分类型
$\mathbf{A}$	表示跨模态预测单词之间语义关系的邻接矩阵, 通过下标区分类型

然而在上述过程中, 由于视觉表征 $\mathbf{F}_v$ 和语义表征 $\mathbf{F}_t$ 在特征的空间分布、取值范围等方面存在着模态异构差异, 使得映射到对齐特征空间的表征 $\mathbf{F}_{va}$ 和 $\mathbf{F}_{ta}$ 通常对齐效果不佳, 因而视觉预测的提升也受到限制. 为了缓解特征

对齐方法中模态异构带来的问题, CMIF 框架在提取视觉对齐表征 $\mathbf{F}_{va}$ 之外, 还进一步将视觉表征向着语义空间映射为视觉-语义表征 $\mathbf{F}_{v\to t}$, 之后通过语义预测 $\tilde{\mathbf{P}}_{v\to t}$ 推理图像对应的语义信息 $\tilde{\boldsymbol{T}}_{v\to t}$, 为缓解在此过程中由视觉噪声带来的错误信息, CMIF 中引入了先验知识 $\mathbf{K}$, 包括每个类别中各种文本出现的概率分布信息、文本共同出现的概率分布信息, 并将这些信息经过类感知关系图合成跨模态的增强表征 $\mathbf{F}_e$ 和跨模态的类别预测 $\tilde{\mathbf{P}}_e$, 此过程可描述为 $\mathbf{F}_v \to \mathbf{F}_{v\to t} \to \tilde{\mathbf{P}}_{v\to t} \to \tilde{\boldsymbol{T}}_{v\to t}$, $\mathbf{K} \to \mathbf{F}_e \to \tilde{\mathbf{P}}_e$. 为了充分结合跨模态对齐与跨模态推理的优势, CMIF 将跨模态语义推理获得的增强表征 $\mathbf{F}_e$ 与视觉对齐表征 $\mathbf{F}_{va}$ 结合为融合表征 $\mathbf{F}_f$, 融合表征产生的预测信息 $\tilde{\mathbf{P}}_f$ 结合跨模态类别预测 $\tilde{\mathbf{P}}_e$ 获得最终预测分类结果 $\tilde{\mathbf{P}}_c$.

3 关键技术

CMIF 的整体框架如图 2 所示, 包括 3 个主要模块, 分别是视觉表征学习模块, 跨模态语义信息推理模块, 和跨模态信息融合模块, 图中黑色实线箭头表示神经网络映射的前向传播线路, 点划线箭头为先验知识指导的前向传播线路, 虚线箭头表示共享信息传递线路, 圆点虚线箭头表示视觉表征学习中对齐损失和分类损失的梯度回传. 首先在视觉表征学习模块中, 模型借助从特权信息中提取的对齐信息通过部分异构对齐算法增强视觉对齐表征 $\mathbf{F}_{va}$ 的提取并获得视觉类别预测 $\tilde{\mathbf{P}}_{va}$; 在跨模态语义信息推理模块中学习将视觉特征 $\mathbf{F}_v$ 迁移至语义空间, 推理视觉-语义信息 $\tilde{\boldsymbol{T}}_{v\to s}$, 并结合先验知识 $\mathbf{K}$ 通过类感知信息选择算法筛选关键信息生成跨模态增强表征 $\mathbf{F}_e$, 之后获得相应跨模态类别预测 $\tilde{\mathbf{P}}_e$; 在跨模态信息融合模块中, 前两个模块产生的表征和预测信息经过特征层面及决策层面融合为最终的类别预测结果 $\tilde{\mathbf{P}}_c$, 提升模型分类的性能和稳定性. 模型借助特权信息在视觉表征学习模块学习抽取视觉对齐表征和对应潜在类别信息, 同时在跨模态语义信息推理模块通过跨模态推理和类感知信息选择有效提取图像中的语义信息, 从而在跨模态信息融合模块中实现多视角的互补和融合, 提升模型的分类效果.

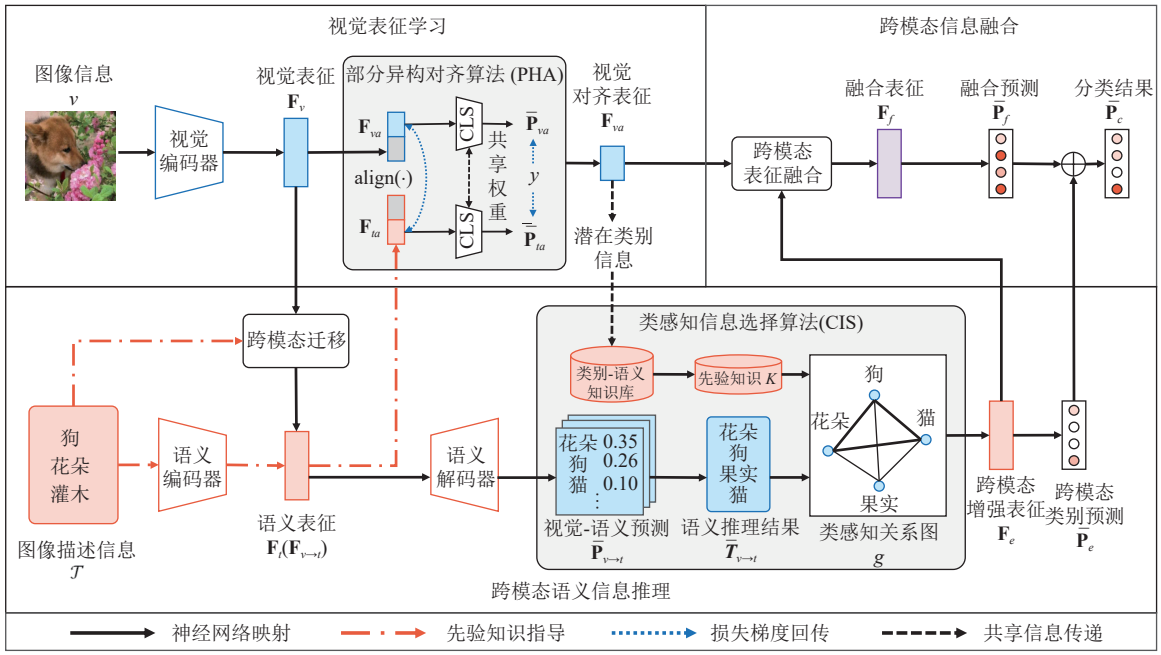


图 2 基于跨模态信息推理和融合的图像分类框架 CMIF

3.1 基于跨模态对齐的视觉表征学习

由于图像信息 $\mathcal{V}$ 映射到视觉特征空间的维度较高且受到背景噪声的干扰导致可分性较差, 因此 CMIF 中引入跨模态的图像描述信息 $\mathcal{T}$ 作为特权信息, 并通过特征对齐引导视觉表征的学习过程.

在视觉表征学习模块中, 初始的视觉表征 $\mathbf{F}_v = E_v(\mathcal{V})$ 和语义表征 $\mathbf{F}_t = E_t(\mathcal{T})$ 通过部分异构对齐算法 (PHA)[28] 实现增强, 其中 $E_v(\cdot)$ 和 $E_t(\cdot)$ 分别是视觉编码器和语义编码器. PHA 算法使得 $\mathbf{F}_v$ 和 $\mathbf{F}_t$ 寻找共享的对齐特征空间, 并在约束下使得两个模态的表征分布相似. 由于不同模态存在独特的风格化特征, 因此 PHA 保留两个模态表征中对分类更关键的部分, 最终生成视觉对齐表征 $\mathbf{F}_{va}$ 和语义对齐表征 $\mathbf{F}_{ta}$, 该过程描述为:

$$\mathbf{F}_{va} = \mathcal{M}_{va}(p_v(\mathbf{F}_v)) \tag{1}$$

$$\mathbf{F}_{ta} = \mathcal{M}_{ta}(p_t(\mathbf{F}_t)) \tag{2}$$

其中, $p_v(\cdot)$ 和 $p_t(\cdot)$ 是对视觉和语义特征的部分选取操作, $\mathcal{M}_{va}(\cdot)$ 和 $\mathcal{M}_{ta}(\cdot)$ 分别是视觉表征和语义表征向对齐空间的映射, 映射网络为线性层后接 LeakyReLU 函数激活, 映射后 $\mathbf{F}_{va}$ 及 $\mathbf{F}_{ta}$ 维度一致.

上述过程通过显式的相似性约束和隐式的分类任务约束共同控制, 显式的相似性约束首先通过 KL 散度引导共享空间中的视觉表征和语义表征的分布相似, 而后 L2 范数约束表征之间的距离. 显式约束的损失定义为:

$$\mathcal{L}_{\text{explict}} = \alpha_{kl} \cdot Softmax(\mathbf{F}_{ta}) \log\left(\frac{Softmax(\mathbf{F}_{ta})}{Softmax(\mathbf{F}_{va})}\right) + \alpha_{l2} \cdot \|\mathbf{F}_{va}, \mathbf{F}_{ta}\|_2 \tag{3}$$

其中, $Softmax(\cdot)$ 为特征的 *Softmax* 归一化, α_{kl} 和 α_{l2} 分别是控制第 1 项 KL 散度和第 2 项 L2 距离的系数.

隐式的分类任务约束将对齐后的视觉和语义表征输入同一个线性分类器, 并用共同的标签提供约束信息, 从而实现二者在特征提取中对关键分类信息的聚焦, 隐式约束的损失函数定义为:

$$\mathcal{L}_{\text{implict}} = \mathcal{L}_{\text{cls}}(f_a(\mathbf{F}_{va}), \mathcal{Y}) + \alpha_{\text{semantic}} \cdot \mathcal{L}_{\text{cls}}(f_a(\mathbf{F}_{ta}), \mathcal{Y}) \tag{4}$$

其中, $\mathcal{L}_{\text{cls}}$ 根据分类任务的不同, 在单标签分类任务中为交叉熵损失, 在多标签分类任务中为二元交叉熵损失, $f_a(\cdot)$ 是从对齐特征到类别预测的映射, α_{semantic} 是语义分类损失的权重系数.

显式和隐式的损失共同组成特征对齐损失约束 PHA 的训练过程:

$$\mathcal{L}_{\text{aligned}} = \mathcal{L}_{\text{implict}} + \mathcal{L}_{\text{explict}} \tag{5}$$

3.2 基于类感知信息选择的跨模态语义信息推理

模型通过视觉表征学习模块学习了视觉对齐表征 $\mathbf{F}_{va}$ 的提取并相应产生了视觉预测 $\tilde{\mathbf{P}}_{va} = f_a(\mathbf{F}_{va})$, 然而由于模态异构的问题, 跨模态对齐主要增强了视觉表征中关键分类信息的提取, 未能有效挖掘视觉表征中的语义信息, 因此 CMIF 中将视觉表征向语义空间映射缓解上述问题.

3.2.1 跨模态语义重构

从视觉向语义空间映射从迁移学习[32]的角度而言, 是将源域视觉空间的表征迁移到目标域语义空间中, 即通过跨模态迁移映射 $\mathcal{M}_{v\to t}(\cdot)$ 将视觉表征 $\mathbf{F}_v$ 转换为视觉-语义表征 $\mathbf{F}_{v\to t} = \mathcal{M}_{v\to t}(\mathbf{F}_v)$, 使得 $\mathbf{F}_{v\to t}$ 能够在特征空间中接近由跨模态图像描述信息 $\mathcal{T}$ 编码的语义表征 $\mathbf{F}_t$, 这一过程通过相似性损失约束:

$$\mathcal{L}_{\text{transfer}} = \|\mathbf{F}_{v\to t}, \mathbf{F}_t\|_2 \tag{6}$$

与特征对齐类似, 这一过程中仍存在模态异构带来的问题, 因此 CMIF 提出进一步使用视觉-语义表征预测 $\mathcal{T}$, 使模型学习预测“图片中可能存在哪些语义信息”, 从而实现 $\mathcal{V} \to \mathbf{F}_v \to \mathbf{F}_{v\to t} \to \mathcal{T}$ 的信息流动, 由公式 (7) 约束:

$$\mathcal{L}_{\text{decode}} = \mathcal{L}_{\text{ce}}(\mathcal{D}(\mathbf{F}_{v\to t}), \mathcal{T}) \tag{7}$$

其中, $\mathcal{L}_{\text{ce}}$ 为交叉熵损失, $\mathcal{D}(\cdot)$ 是基于长短时记忆网络 LSTM 的语义解码器[33], $\mathcal{D}$ 在每一个单词位 (即 LSTM 单元) 接受 $\mathbf{F}_{v\to t}$ 输入和上一个单词位的信息, 并输出当前单词位的描述信息.

综合公式 (6) 和公式 (7), 跨模态语义重构的损失为:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{decode}} + \alpha_{\text{trans}} \cdot \mathcal{L}_{\text{transfer}} \tag{8}$$

其中, α_{trans} 为迁移损失的权重系数.

语义解码器 $\mathcal{D}(\cdot)$ 解码出语义信息预测 $\tilde{\mathbf{P}}_{v\to t} = \{p_{l,m} \mid l = 1,2,\ldots,L,\ m = 1,2,\ldots,M\}$, 其中 L 为语义解码器的单元数量, M 为语义描述信息类别总数, $p_{l,m}$ 表示解码器在第 l 个单词位对第 m 个描述信息的预测结果. 根据公式 (7) 的约束, 准确的描述信息往往属于预测概率前几的语义信息预测 $\tilde{\mathbf{P}}_{v\to t}$ 中, 由此从预测中推理可能的语义结果:

$$\tilde{\boldsymbol{T}}_{v\to t}=\mathcal{I}\left(\tilde{\mathbf{P}}_{v\to t},l',m'\right) \tag{9}$$

其中, $\mathcal{I}(\cdot,\cdot,\cdot)$ 为语义推理操作, 即从语义预测 $\tilde{\mathbf{P}}_{v\to t}$ 中挑选出前 l' 个单词位的前 m' 个语义预测.

3.2.2 类感知信息选择

跨模态的语义信息的推理受到视觉噪声的影响容易产生错误预测, 为了避免进一步的误差传播, 本文提出类感知信息选择算法 (CIS) 从混杂的语义预测信息中过滤错误信息, 算法过程如图 3 所示, 跨模态语义信息推理信息中的语义关系和类别先验知识中的语义关系融合为类感知关系图, 通过图卷积融合为跨模态增强表征, 进一步获得相应的类别预测. 为选取推理结果 $\tilde{\boldsymbol{T}}_{v\to t}$ 中的重要信息, CIS 从训练集中统计并构建为类别-语义知识库, 其中包括类别-语义频率信息 $\mathbf{K}_{\text{freq}}$ 和类别-语义共同出现频率信息 $\mathbf{K}_{\text{co}}$, 分别定义为:

$$\mathbf{K}_{\text{freq}}=\{P(t_m\,|\,y_c)\,|\,c=1,2,\ldots,C,\ m=1,2,\ldots,M\} \tag{10}$$

$$\mathbf{K}_{\text{co}}=\{P(t_m,t'_{m'}\,|\,y_c)\,|\,c=1,2,\ldots,C,\ m=1,2,\ldots,M,\ m'=1,2,\ldots,M\} \tag{11}$$

其中, P 为概率分布, C 为数据集标签类别总数, M 是数据集中语义信息的类别总数. $\mathbf{K}_{\text{freq}}$ 表示训练集中各类别 y_c 的语义描述信息 t_m 出现的概率分布, $\mathbf{K}_{\text{co}}$ 表示训练集中各类别 y_c 中语义信息 t_m 和 t'_m 共同出现概率的分布.

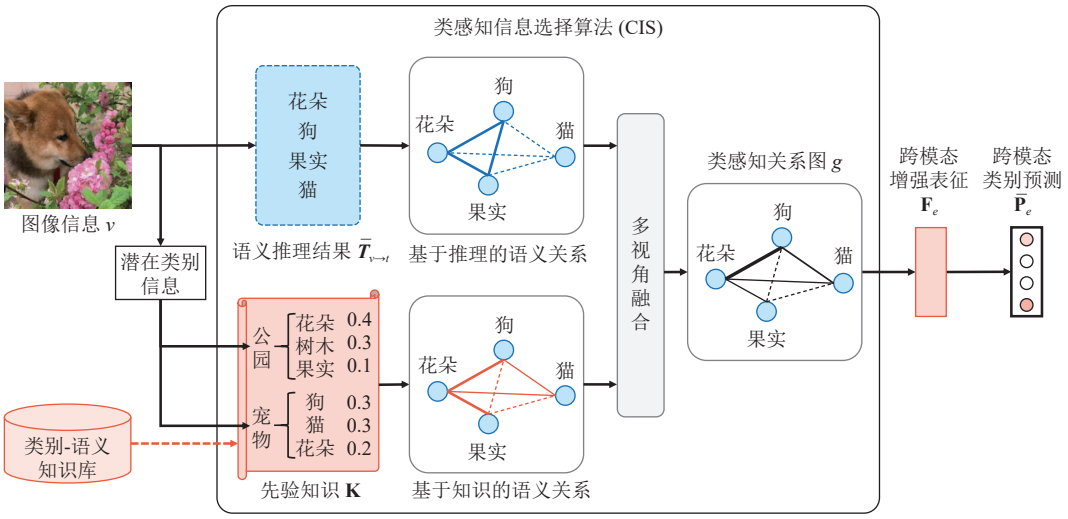


图 3 类感知信息选择算法 CIS 的整体流程

由于类别-语义知识库中包含的是由类主导的语义分布知识, 因此 CIS 算法在选取知识时使用了视觉预测产生的类别排序信息. 由于公式 (4) 的约束, 准确的类别信息常位于预测前几的类别 (潜在类别) 中, 因此选择先验知识的过程可描述为:

$$\mathbf{K}'_{\text{freq}}=\{P(t_m|y_{c'})|c'\in top\left(\tilde{\mathbf{P}}_{va},J\right),m=1,2,\ldots,M\} \tag{12}$$

$$\mathbf{K}'_{\text{co}}=\{P(t_m,t'_{m'}|y_{c'})|c'\in top\left(\tilde{\mathbf{P}}_{va},J\right),m=1,2,\ldots,M,m'=1,2,\ldots,M\} \tag{13}$$

其中, $top(\cdot,\cdot)$ 表示从预测中挑选出前 J 个类别的操作算子, $y_{c'}$ 为的潜在类别.

公式 (9)–公式 (13) 阐述了跨模态推理结果 $\tilde{\boldsymbol{T}}_{v\to t}$ 和先验知识 $\mathbf{K}=\left\{\mathbf{K}'_{\text{freq}},\mathbf{K}'_{\text{co}}\right\}$ 的形成过程, CIS 进一步将已有的预测和知识通过文本间的关系连接和融合, 构建类感知的关系图 $\mathcal{G}=\{\mathcal{N},\mathcal{E}\}$. $\mathcal{G}$ 的节点 $\mathcal{N}$ 从语义上表示推理结果所对应的语义单词, $\mathcal{N}$ 对应的表征除了通过语义嵌入 $Embed(\cdot)$ 获得的语义推理结果表征 $\mathbf{F}_i=Embed\left(\tilde{\boldsymbol{T}}_{v\to t}\right)$ 以外还融入了视觉表征 $\mathbf{F}_v$, 从而补足在跨模态推理中可能舍弃的有效信息, 并使得视觉表征及语义表征通过语义关系实现交互和融合[34], 提升表征学习效果的稳定性.

边 $\mathcal{E}$ 的权重则由 $\tilde{\boldsymbol{T}}_{v\to t}$ 和先验知识 $\mathbf{K}$ 共同参与计算, 并通过视觉分类预测提供的类别引导, 表示为邻接矩阵 $\mathbf{A}$.

由基于语义推理和基于类别先验知识的语义关系融合获得, 其中基于语义推理的邻接关系矩阵为:

$$\mathbf{A}_{v\to t}=\boldsymbol{I}+\beta_{v\to t}\cdot\boldsymbol{I}' \tag{14}$$

其中, $\boldsymbol{I}$ 为单位矩阵, $\boldsymbol{I}'$ 为主对角线为 0 其他位置为 1 的矩阵, $\boldsymbol{I},\boldsymbol{I}'\in\mathbb{R}^{B\times B},B=l'\cdot m'$, 即 B 为 $\bar{\boldsymbol{T}}_{v\to t}$ 包含的单词数, $\beta_{v\to t}$ 用于控制语义预测单词间的连接强度.

基于潜在类别信息选择的类别-语义频率信息 $\mathbf{K}'_{\text{freq}}$, 其语义权重通过类感知融合计算:

$$\boldsymbol{w}_{\text{seq}}=\sum_{j}^{J}\left(\mathbf{K}_{\text{seq}}^{C'(j)}\cdot\tilde{\mathbf{P}}_{va}^{C'(j)}\right) \tag{15}$$

$$\mathbf{A}_{\text{seq}}=\sigma\left(\boldsymbol{w}_{\text{seq}}^{\text{T}}\times\boldsymbol{w}_{\text{seq}},\bar{\boldsymbol{T}}_{v\to t}\right) \tag{16}$$

其中, $\boldsymbol{w}_{\text{seq}}$ 是根据类别频率信息和潜在类别信息计算的权重向量, $\boldsymbol{w}_{\text{seq}}^{\text{T}}$ 表示向量 $\boldsymbol{w}_{\text{seq}}$ 的转置, $C'(j)$ 作为上标表示根据第 j 个潜在类别选择出的对应信息, σ 表示从矩阵 $\boldsymbol{w}_{\text{seq}}^{\text{T}}\times\boldsymbol{w}_{\text{seq}}$ 中抽取出 $\bar{\boldsymbol{T}}_{v\to t}$ 中对应单词的操作.

类似的, 基于潜在类别信息选择的类别-语义共同出现频率信息 $\mathbf{K}'_{\text{co}}$ 的语义权重通过类感知融合计算:

$$\mathbf{A}_{\text{co}}=\sigma\left(\sum_{j}^{J}\left(\mathbf{K}_{\text{co}}^{C'(j)}\cdot\tilde{\mathbf{P}}_{va}^{C'(j)}\right),\bar{\boldsymbol{T}}_{v\to t}\right) \tag{17}$$

其中, σ 表示从矩阵 $\sum_{j}^{J}\left(\mathbf{K}_{\text{co}}^{C'(j)}\cdot\tilde{\mathbf{P}}_{va}^{C'(j)}\right)$ 中抽取出 $\bar{\boldsymbol{T}}_{v\to t}$ 中对应单词的操作.

基于预测的语义关系 $\mathbf{A}_{v\to t}$ 描述了节点 $\mathcal{N}$ 中语义信息彼此之间的连接关系, 是从当前样本的角度出发, 解答图像当中可能存在哪些语义信息; 而基于先验知识的语义关系 $\mathbf{A}_k=\mathbf{A}_{\text{seq}}\oplus\mathbf{A}_{\text{co}}$ 则通过统计信息描绘了潜在类当中哪些语义单词通常会出现或共同出现, 哪些单词几乎不出现或不同时出现. 通过 $\mathbf{A}_k$ 对 $\mathbf{A}_{v\to t}$ 的补充, 从类别-语义连接的角度抑制了对分类无正面意义的语义连接, 提升后续语义表征对类别信息的表达能力, 二者从推理的视角和先验知识的视角共同得出更为可靠的语义关系构建, 公式描述为:

$$\mathbf{A}=\theta(\mathbf{A}_k,\mathbf{A}_{v\to t}) \tag{18}$$

其中, $\theta(\cdot,\cdot)$ 是基于线性网络映射的矩阵融合操作.

3.2.3 跨模态增强信息生成

CIS 算法形成类感知关系图 $\mathcal{G}$, 其中节点包含了图像中提取的多模态表征, 边则体现了从跨模态预测和先验知识中共同学习的语义关系, 在此基础上, CMIF 通过图卷积的计算将其融合为跨模态增强表征 $\mathbf{F}_e$, 过程如下:

$$\mathbf{F}_e^{(h)}=\delta^{(h)}(Concat(\mathbf{F}_i,\mathbf{F}_v),Softmax(\mathbf{A})) \tag{19}$$

其中, $\delta(\cdot)$ 为图卷积融合操作[35], h 为图卷积计算的阶数, *Concat* 是特征拼接操作.

过程中, 由分类损失约束预测 $\tilde{\mathbf{P}}_e=f_e(\mathbf{F}_e)$, 其中 $f_e(\cdot)$ 是从跨模态增强特征到类别预测的映射, 损失函数为:

$$\mathcal{L}_{\text{cis}}=\mathcal{L}_{\text{cls}}\left(\tilde{\mathbf{P}}_e,\mathcal{Y}\right) \tag{20}$$

其中, $\mathcal{L}_{\text{cls}}$ 根据分类任务的不同, 在单标签分类任务中为交叉熵损失, 在多标签分类任务中为二元交叉熵损失.

3.3 跨模态信息融合

在跨模态信息融合模块中, CMIF 使用了特征层面融合和决策层面融合的策略. 特征层面的融合通过从视觉表征学习模块中学习到的 $\mathbf{F}_{\text{va}}$ 和在跨模态语义信息推理模块中学习到的 $\mathbf{F}_{\text{e}}$ 融合, 实现不同模态信息的互补:

$$\mathbf{F}_f=(\mathcal{M}_{fv}(\mathbf{F}_{va})\otimes\mathcal{M}_{fe}(\mathbf{F}_e))\oplus\mathcal{M}_{fv}(\mathbf{F}_{va})\oplus\mathcal{M}_{fe}(\mathbf{F}_e) \tag{21}$$

其中, $\mathcal{M}_{fv}$ 和 $\mathcal{M}_{fe}$ 分别为对于视觉对齐表征和跨模态增强表征的维度映射, $\otimes$ 为对应元素相乘操作.

特征融合过程通过分类任务对基于融合特征的预测 $\tilde{\mathbf{P}}_f=f_f(\mathbf{F}_f)$ 约束, 其中 $f_f(\cdot)$ 是从跨模态融合特征到类别预测的映射, 对应的损失函数为:

$$\mathcal{L}_{\text{fuse}}=\mathcal{L}_{\text{cls}}\left(\tilde{\mathbf{P}}_f,\mathcal{Y}\right) \tag{22}$$

决策层面的融合中, 通过跨模态语义推理模块提供的类别预测 $\tilde{\mathbf{P}}_e$ 有助于进一步增强分类的稳定性, 由此计算

的最终分类结果 $\tilde{\mathbf{P}}_c$ 定义如下:

$$\tilde{\mathbf{P}}_c = Softmax\left(\tilde{\mathbf{P}}_e\right) + Softmax\left(\tilde{\mathbf{P}}_f\right) \tag{23}$$

3.4 训练策略

为提升 CMIF 的使用效率, 本节将介绍分别是分步与单步的训练策略. 分步训练的方式主要分为以下步骤.

(1) 基础表征提取和对齐: 训练视觉编码器 $E_v(\cdot)$、语义编码器 $E_t(\cdot)$、对齐映射网络 $\mathcal{M}_{va}$ 和 $\mathcal{M}_{ta}$, 和类别映射网络 f_a, 通过对齐损失 $\mathcal{L}_{\text{aligned}}$ 约束, 使模型从图像 $\mathcal{V}$ 中学习视觉对齐表征 $\mathbf{F}_{va}$ 及其对应预测 $\tilde{\mathbf{P}}_{va}$.

(2) 跨模态推理: 冻结上述步骤网络, 训练跨模态迁移映射网络 $\mathcal{M}_{v\to t}$, 语义解码器 $\mathcal{D}$, 通过跨模态语义重构损失 $\mathcal{L}_{\text{recon}}$ 约束, 使得模型能够从视觉表征 $\mathbf{F}_v$ 中推理对应语义信息 $\tilde{\boldsymbol{T}}_{v\to t}$.

(3) 类感知信息选择: 冻结上述步骤网络, 训练语义嵌入映射网络 $Embed$, 线性融合网络 θ, 图卷积融合网络 δ, 和类别映射网络 f_e, 通过跨模态增强的预测损失 $\mathcal{L}_{\text{cis}}$ 约束, 使得模型从语义推理 $\tilde{\boldsymbol{T}}_{v\to t}$ 中选择关键信息并形成跨模态增强表征 $\mathbf{F}_e$ 及其对应类别预测 $\tilde{\mathbf{P}}_e$.

(4) 跨模态融合: 冻结上述步骤网络, 训练维度映射网络 $\mathcal{M}_{fv}$ 和 $\mathcal{M}_{fe}$, 分类映射网络 f_f, 由特征融合预测的分类损失 $\mathcal{L}_{\text{fuse}}$ 约束, 形成融合表征 $\mathbf{F}_f$ 和对应预测 $\tilde{\mathbf{P}}_c$.

单步训练需要控制不同损失函数之间的比例, 总体的损失函数如下:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{aligned}} + \gamma_{\text{recon}} \cdot \mathcal{L}_{\text{recon}} + \gamma_{\text{cis}} \cdot \mathcal{L}_{\text{cis}} + \gamma_{\text{fuse}} \cdot \mathcal{L}_{\text{fuse}} \tag{24}$$

其中, γ_{recon}, γ_{cis} 和 γ_{fuse} 分别为损失权重系数, 其取值范围详见第 4.1.3 节.

4 实 验

为验证 CMIF 的有效性并进一步解释其原因, 我们在本节设计了以下几个研究问题 (research questions, RQ) 并开展了对应实验进行验证.

- RQ1: CMIF 在跨模态图像分类数据集上是否比当前方法能更有效提升分类效果? (见第 4.2 节)
- RQ2: CMIF 框架中设计的不同模块对分类效果产生了怎么样的影响? (见第 4.3 节)
- RQ3: 关键模块 CIS 中选择不同数量的潜在类别会是否影响模型的预测效果? (见第 4.4.1 节)
- RQ4: 在融合视觉和跨模态信息时, 第 3.3 节中的融合方法是否比常见融合方法更有效? (见第 4.4.2 节)
- RQ5: 如何解释 CMIF 中视觉和跨模态融合的设计, 这种设计存在哪些优点和局限性? (见第 4.5 节)

4.1 实验设置

4.1.1 数据集

本文的实验结果来自以下两个真实世界数据集, 表 2 展示了两个数据集的统计数据.

表 2 数据集的统计信息

数据集	数据总量	标签类别数	文本类别数	训练样本数	测试样本数
VireoFood-172	99 225	172	353	66 071	33 154
NUS-WIDE	203 598	81	1 000	121 962	81 636

VireoFood-172 数据集[13]: 含有 110 241 张菜品图片的单标签分类数据集, 样本共分 172 类. 数据集中包含食材文本 353 种, 平均每个图像样本对应 3 个食材文本. 我们按照原论文的设置划分数据集[10], 其中 66 071 张图像用于模型训练, 33 154 张图像分别用于测试.

NUS-WIDE 数据集[30]: 多标签分类数据集, 原始数据集共包含 269 648 个从互联网收集的图像样本, 图像主体包括人、动物、建筑等, 共对应 81 种分类. 每个图像样本对应若干文本, 共有 1 000 类文本, 平均每个图像样本对应 7 个描述文本. 我们参考原设定[30]及相关论文[24,36]设定划分训练集和测试集, 并对数据集进行清洗, 去除缺少标签或文本的数据后, 剩余图像样本 203 598 个, 其中训练集 121 962 个样本和测试集 81 636 个样本.

4.1.2 评估标准

在 VireoFood-172 数据集上的实验采用了准确率 (*Acc*) 评估模型对单标签分类的预测性能. 在 NUS-WIDE 数据集上采用精确率 (*Pre*) 和召回率 (*Recall*) 评估多标签分类任务中的模型预测性能. 公式如下:

准确率公式:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{25}$$

精确率公式:

$$Pre = \frac{TP}{TP + FP} \tag{26}$$

召回率公式:

$$Recall = \frac{TP}{TP + FN} \tag{27}$$

其中, *TP*, *FP*, *TN*, *FN* 分别是正阳样本数, 负阳样本数, 正阴样本数, 和负阴样本数. 上述的指标均计算其 top-1 和 top-5 预测的平均值作为最终结果.

4.1.3 实现细节

在骨干网络选择部分, 实验中使用的视觉网络均加载了在 ImageNet 数据集上预训练的参数. 为了算法对比时的公平起见, 对于跨模态约束以及跨模态对齐的对比算法, 本文实现时均使用了 ResNet-50[37]作为视觉骨干网络. 对于 CMIF 算法, 为了验证其在基于卷积的神经网络和基于 Transform 架构神经网络上的通用性, 实现了基于 ResNet-50[37]和基于 ViT[31]的版本, 语义编码器/解码器采用了门控神经网络.

模型结构和训练当中使用的超参数如表 3 所示, 其中“通用超参数”为实验中各类对比方法及本文提出方法共同使用的超参数, “对比方法的超参数”包括了跨模态约束和跨模态对齐方法使用的超参数, “CMIF 方法的超参数”为本文提出方法在实验中所涉及的超参数. 需要额外说明的是, 对于实验中的每个模型, 每训练 4 轮之后, 学习率衰减为之前的 0.1 倍, 共衰减 3 次并额外训练 1 轮, 另外对于 CMIF 方法, 在一步训练的方法中 $\gamma_{\text{recon}}, \gamma_{\text{cis}}$ 和 γ_{fuse} 按照 3:1:1 的比例选择.

表 3 实验中超参数设置的细节

类别	超参数	符号	取值/取值范围
通用超参数	数据批次大小	—	64
	学习率	—	[5E–5, 1E–2]
	学习率衰减间隔轮次	—	4
	学习率衰减率	—	0.1
	优化器权重衰减参数	—	1E–3
	BCE损失中正样本的损失权重	—	[40, 100]
对比方法的超参数	语义编解码器的潜空间维度	—	256, 512, 1 024
	跨模态方法中的语义损失权重	—	1.0, 1.5, 2.0, 2.5
	跨模态方法中的对齐损失权重	—	1E–2, 1E–1, 1
CMIF方法的超参数	跨模态迁移的权重系数	α_{trans}	1E–2, 1E–1, 1
	跨模态对齐的权重系数	α_{kl}, α_{l2}	1E–2, 1E–1, 1
	CIS中保留的单词位取值	l'	[2, 5]
	CIS中各单词位保留的语义预测	m'	3, 5, 8, 10
	CIS中潜在类别信息的类别数	J	[1, 5]
	CIS中预测单词间的连接强度	$\beta_{v\to t}$	[0, 0.6]

4.2 对比实验结果 (RQ1)

为了验证 CMIF 对分类效果提升的有效性, 本文进行了对比实验, 对比方法主要分为 3 类: 第 1 类是仅基于视觉信息的分类方法, 包括基础的骨干网络 ResNet-18[37]、ResNet-50[37]和 VGG19[38]及基于 ResNet-50 改进的

WRN[39]、WISeR[40]网络, 以及近年来提出的骨干网络 RepVGG[41], RepMLPNet[42]和 ViT[31]; 第 2 类是基于跨模态约束的方法, 包括基于全局约束的 ARCH-D[13], CMFL[16], 和 MSMVFA[22], 和基于局部约束的 CMRR[18], IG-CMAN[21]; 第 3 类是基于跨模态对齐的方法, 包括基于特征分布对齐的 Deep-CORAL[23]和 SSAN[27], 特征解耦后对齐的方法 Disentangle-VAE[29]和 ATNet[28]. 通过 t-test 显著性检测结果表明, 在 VireoFood-172 数据集上, 提出方法 CMIF 的预测性能相对于视觉信息分类方法表现出了显著性, 其中 CMIF(ResNet-50) 在对于对应的骨干网络 ResNet-50 在测试集上的预测 $p\text{-}value \approx 0.031 < 0.05$, CMIF(ViT) 对于对应的视觉模型 ViT 在测试集上的预测 $p\text{-}value \approx 0.095 < 0.1$, 同时对于其他的跨模态方法 *p-value* 也介于 0.061 至 0.347 之间; 在 NUS-WIDE 数据集上 CMIF 的预测性能对于视觉信息分类方法同样表现出显著性, 其中 CMIF(ResNet-50) 在对于对应的骨干网络 ResNet-50 在测试集上的预测 $p\text{-}value \approx 0.045 < 0.05$, CMIF(ViT) 对于对应的视觉模型 ViT 在测试集上的预测 $p\text{-}value \approx 0.098 < 0.1$, 同时对于其他的跨模态方法 *p-value* 介于 0.122–0.386 之间. 总体的性能对比结果如表 4 所示, 其中 VireoFood-172 数据集用 top-1 和 top-5 的精确度 (*Acc*) 作为衡量指标, NUS-WIDE 数据集通过 top-1 和 top-5 的准确率 (*Pre*) 和召回率 (*Recall*) 衡量其多分类任务的表现性能. 从表 4 中有以下发现.

表 4　对比实验结果

类型	算法	VireoFood-172			NUS-WIDE		
		Acc@1	*Acc*@5	*Pre*@1	*Pre*@5	*Recall*@1	*Recall*@5
视觉信息分类	ResNet-18	0.773	0.932	0.785	0.391	0.439	0.846
	ResNet-50	0.816	0.949	0.786	0.391	0.440	0.846
	VGG19	0.811	0.950	0.789	0.393	0.442	0.851
	WRN	0.824	0.965	0.787	0.394	0.440	0.853
	WISeR	0.828	0.965	0.789	0.395	0.441	0.855
	RepVGG	0.835	0.963	0.797	0394	0.448	0.856
	RepMLPNet	0.833	0.962	0.801	0.405	0.448	0.877
	ViT	0.836	**0.966**	0.796	0.395	0.453	0.855
跨模态约束	CMRR	0.819	0.954	0.802	0.403	0.454	0.874
	ARCH-D	0.825	0.956	0.797	0.397	0.448	0.864
	IG-CMAN	0.831	0.963	0.800	0.405	0.443	0.868
	CMFL	0.831	0.958	0.809	0.402	0.456	0.871
	MSMVFA	0.836	**0.966**	0.798	0.395	0.443	0.855
跨模态对齐	Deep-CORAL	0.833	0.963	0.803	0.396	0.449	0.855
	SSAN	0.836	0.955	0.803	0.396	0.450	0.855
	Disentangle-VAE	0.838	0.950	0.811	0.402	0.459	0.872
	ATNet	0.838	0.946	0.805	0.404	0.463	0.870
信息选择和融合	CMIF(ResNet-50)	0.846	0.951	0.817	0.407	0.467	0.875
	CMIF(ViT)	**0.856**	0.958	**0.824**	**0.411**	**0.468**	**0.882**

(1) CMIF 是一种不局限于特定骨干网络的跨模态图像增强框架. 在 VireoFood-172 数据集和 NUS-WIDE 数据集上, CMIF 相比于基于卷积的 ResNet-50 网络和基于 Transform 架构的 ViT 网络, 均带来了 4%–7% 的稳定提升, 且在大部分性能指标上超过了各类对比方法.

(2) CMIF 框架在不同领域数据集中具有更强的泛化能力. 由于数据集的分布和标签类型的差异, 一些在 VireoFood-172 数据集上表现良好 (提升约 3%) 的方法如 MSMVFA 和 SSAN, 在 NUS-WIDE 上则提升有限 (提升 0.3%, 远低于各类跨模态对齐方法), CMIF 则在不同数据集上均取得了明显提升效果.

(3) 基于模型架构改进的骨干网络显著提升了分类准确性. 与在基线网络基础上增加分支的改进方法 WRN 和 WISeR 相比, 基于 Transform 架构的 ViT 使用自注意力模块取代了先前图像分类中卷积模块堆叠的方式提升了对图像的整体感知, 而基于重参数方法的 RepVGG, RepMLPNet 则强调了训练中使用多分支信息和信息交互, 同样带来较为显著的性能提升.

(4) 基于跨模态对齐的方法总体比基于跨模态约束的方法提升更显著. 特征对齐方法在 VireoFood-172 和 NUS-WIDE 数据集上总体效果更好, 验证了特征对齐中显式的相似性约束对跨模态学习的重要性. 尽管基于局部约束策略的跨模态约束方法 IG-CMAN 和多尺度约束的 MSMVFA 在 VireoFood-172 上表现出了和跨模态对齐方法总体接近的性能, 但在 NUS-WIDE 上则表现不佳, 这也表现了跨模态约束方法隐式约束的不稳定性.

(5) 在单标签分类中, 基于特征解耦后对齐的方法可能带来 top-5 性能下降. 基于特征解耦后对齐的 ATNet 和 Disentangle-VAE 通过关键分类信息的保留能够在模型关注到图像主体的情况下, 从视觉表征中进一步获取主体信息, 因而在 VireoFood-172 上提升了 top-1 准确率; 然而当模型关注发生偏差时, 则容易进一步偏差到其他类别当中, 从而削弱了 top-5 准确率的提升. CMIF 在视觉表征学习中使用了特征解耦方法 PHA 削弱了 top-5 准确率, 但后续通过跨模态信息融合使得 top-5 准确率重新回升.

4.3 消融实验 (RQ2)

为了进一步探究 CMIF 中各模块对模型分类提升的有效性及模块间相互协同的效果, 本文进一步开展了消融实验. 消融实验的第 1 部分是在基线模型上添加了视觉表征模块中的 PHA 算法增强原始视觉表征; 第 2 部分是在第 1 部分基础上将原始视觉表征通过跨模态推理模块中的 CIS 算法增强, 同时探究了跨模态预测信息和先验知识的组合带来的增强效果. 由于跨模态推理性能的提升并不一定带来特征融合后的相应提升 (见第 4.4.1 节), 因此第 2 部分为跨模态增强表征和视觉增强表征融合后的分类结果; 第 3 部分实验是在表征融合结果之上加入决策融合. 消融实验结果如表 5 所示, Baseline 为以 ResNet-50 为骨干网络的基础分类算法, PHA 为部分异构迁移算法, CIS 为类感知信息选择算法, CR 为跨模态推理, FF 为特征融合, CK 为类别先验知识, DF 为决策融合. 主要可以得到以下结论.

表 5 消融实验结果

模型	VireoFood-172			NUS-WIDE		
	Acc@1	*Acc*@5	*Pre*@1	*Pre*@5	*Recall*@1	*Recall*@5
Baseline	0.815	0.949	0.786	0.391	0.440	0.846
+PHA	0.829	0.945	0.797	0.393	0.448	0.853
+PHA+CIS(CR) +FF	0.835	0.937	0.811	0.396	0.454	0.865
+PHA+CIS(CK) +FF	0.839	0.941	0.812	0.399	0.458	0.869
+PHA+CIS(CR+CK) +FF	0.841	0.941	0.814	0.405	0.459	0.871
+PHA+CIS(CR+CK) +FF+DF	**0.846**	**0.951**	**0.817**	**0.407**	**0.467**	**0.875**

(1) 视觉表征学习模块中, 部分异构对齐算法 PHA 对于模型预测的 top-1 性能提升明显, 但可能降低 top-5 性能. 加入部分异构对齐 (+PHA) 算法, 使得模型学习视觉表征时保留了关键分类信息, 视觉模型关注到图像的主体部分时不容易被其他信息干扰, 从而增大了从潜在类中选出最佳预测结果 (即 top-1 结果) 的可能, 但同时也在模型关注出现偏差时, 由于缺少辅助信息参与表征学习, 从而降低对正确类别的预测可能性 (top-5 性能降低), 这一点对于单标签预测的 VireoFood-172 数据集更为明显.

(2) 在融合跨模态语义信息时, 基于先验知识和基于模型预测的关系融合有效提升了跨模态表征融合的效果. 在类感知关系图的构建中, 基于模型预测的 CIS(CR) 和基于先验知识的 CIS(CK) 都能与视觉表征融合 (+FF) 取得 top-1 性能提升. 然而 CIS(CR) 由于模型视觉关注的影响, 可能未能推理出类别中的关键语义信息, 从而降低了 top-5 准确率. 通过两种关系的融合 CIS(CR+CK), CMIF 一方面借助先验知识抑制语义预测的错误连接, 另一方面通过预测信息增强关键语义的选择, 从而进一步提升了模型性能.

(3) 在跨模态信息融合模块中, 多种决策信息的融合有利于提升模型预测的稳定性. 如前文分析, 跨模态表征融合 (+FF) 提升了模型的 top-1 预测能力, 而决策融合 (+DF) 则是在融合表征映射的类别预测之外, 又引入了跨模态推理后的增强表征映射的类别预测, 从而补充了使得一些由于表征融合而平滑或损失的信息, 从而使得模型的 top-1 和 top-5 性能稳定提升.

4.4 深入分析

4.4.1 类感知信息选择算法中潜在类的选择对预测的影响分析 (RQ3)

消融实验展现了 CIS 算法对于信息互补效果的重要影响, 本节将深入分析在 CIS 类感知关系图构建中, 潜在类的选择对预测的影响. 实验在 VireoFood-172 上进行, 使用 ResNet-50 作为视觉骨干网络, 结果如表 6 所示. 在公式 (4) 的约束下, 准确的类别信息常位于前几的预测类 (即潜在类) 当中, 随着潜在类的增加 (top 从 1 至 5), 更多视觉预测中的潜在类信息被融合到跨模态增强表征的构建中. 当 top 从 1 至 3 时, 准确的类别-语义先验知识更大概率被包含在其中, 但随着 top 的继续上升, 虽然准确类别被包含在潜在类中的可能性进一步提升, 但其他类别在潜在类中的比重也同时增加, 此时跨模态预测仍比较平稳, 但融合预测相对于 top 为 3 时反而降低, 这可能是由于引入了过多类别的先验知识使得关系图权重分布变得相对均匀, 导致生成的跨模态增强表征为视觉提供的互补效果削弱.

表 6 在 CIS 算法中选取前 1–5 个潜在类时跨模态预测及融合预测的结果

top	视觉预测 $\bar{\mathbf{P}}_{va}$		跨模态预测 $\bar{\mathbf{P}}_{e}$		融合预测 $\bar{\mathbf{P}}_{f}$	
	Acc@1	*Acc*@5	*Acc*@1	*Acc*@5	*Acc*@1	*Acc*@5
1			0.826	0.921	0.837	0.939
2			0.827	0.928	0.838	0.939
3	0.829	0.945	0.828	0.930	**0.841**	**0.941**
4			0.828	0.930	0.839	**0.941**
5			**0.829**	**0.931**	0.839	0.940

4.4.2 特征融合方式对融合性能的影响分析 (RQ4)

本节将深入分析跨模态信息融合模块中特征融合方式对性能的影响, 实验在 VireoFood-172 上进行, 使用 ResNet-50 作为视觉骨干网络, 将模型通过视觉表征学习模块学习的视觉对齐表征与通过跨模态推理模块学习的跨模态增强表征进行融合, 并将融合表征映射到类别, 结果如表 7 所示, 其中 Con 为特征拼接, Add 为特征相加, Mul 为特征相乘, Mix 为公式 (21) 中相加与相乘结合的融合方式, Max 和 Min 为特征对应位置取最大最小值. 总体而言, 特征融合后的预测效果优于单一模态特征, 其中, 基于特征拼接 (Con) 的方法效果相对最差, 这可能由于特征拼接缺少特征间的交互, 而在特征间产生交互的方法中, 特征最大 (Max) 最小 (Min) 值选择只保留视觉或语义表征中的部分特征, 削弱了表征中的语义连贯性, 因此提升也不明显. 基于相加 (Add), 相乘 (Mul), 和 CMIF 中采用的加乘结合的方式 (Mix) 则能够使得两部分特征得到充分融合, 提升最终分类效果.

表 7 特征融合结果 (*Acc*@1)

视觉预测 $\bar{\mathbf{P}}_{va}$	跨模态预测 $\bar{\mathbf{P}}_{e}$	融合预测 $\bar{\mathbf{P}}_{f}$					
		Con	Add	Mul	Mix	Min	Max
0.829	0.828	0.830	0.839	0.838	**0.841**	0.835	0.832

4.5 案例分析 (RQ5)

本节将通过 GradCAM[43]可视化展示模型学习不同表征时的关注点, 从而分析不同模态间的互补机制. 如图 4 所示是模型在 VireoFood-172 数据集随机样本上的关注点变化情况. 图 4 中视觉表征对应模型学习表征 $\mathbf{F}_{va}$ 时的关注点, 语义表征对应学习 $\mathbf{F}_{v\to l}$ 时的关注点, 融合表征对应学习 $\mathbf{F}_{f}$ 时的关注点. 图 4(a)、图 4(b) 展示了当图像中主体清晰时, 对视觉和语义表征的学习均能关注到主体部分, 因而融合后也准确关注到了主体; 图 4(c)、图 4(d) 展示视觉学习时模型一定程度上关注到了主体部分, 但由于视觉噪声干扰, 跨模态推理总体置信度不高, 除图像主体外, 图 4(c) 中对于叶片, 图 4(d) 中对于文字都产生了关注, 因此在跨模态迁移时的误差传播对视觉-语义表征的学习产生了干扰, 基于 CIS 的信息选择有效避免了错误语义预测对模型关注的干扰, 并进一步通过特征融合扩大了对图像信息的感受野, 提升对图像主体的整体关注; 图 4(e)、图 4(f) 中存在的背景噪声误导了模型对视觉表征的

学习, 在 CMIF 中通过跨模态语义推理关注到图像中存在语义信息的区域, 从而通过跨模态信息融合辅助模型减少对背景噪声的关注, 提升对主体的识别能力; 图 4(g)、图 4(h) 中, 在图像主体不清晰或背景特别混乱的情况下, 视觉和语义关注都偏离主体, 最终的融合表征也很难关注到主体部分, 这说明 CMIF 中跨模态语义表征的效果一定程度上受限于视觉表征的学习.

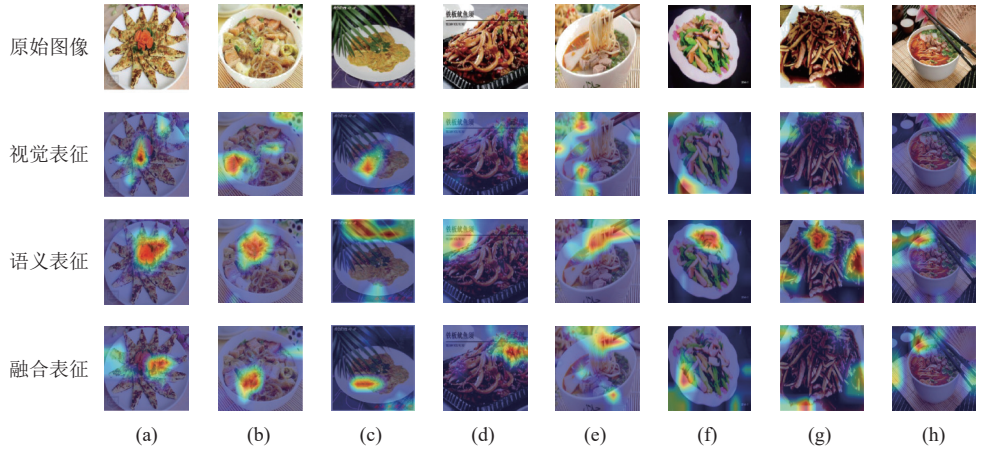


图 4 CMIF 中各模态特征关注点变化

5 总 结

本文提出了一种跨模态语义表征学习和融合的框架 CMIF, 通过跨模态的异构表征对齐和语义重构增强了模型表达能力. 为解决跨模态语义表征学习中的问题, 本文提出使用类感知信息选择方法 CIS, 基于跨模态预测和先验知识的融合选择视觉信息中包含的关键语义信息, 提升语义信息的挖掘和对视觉表征的补充效果, 从而有效提升模型分类表现.

CMIF 有效缓解了视觉识别中多模态数据缺乏的问题, 下一步工作将尝试扩展这一框架, 通过研究视觉-语义信息互补时的特点, 来缓解数据分布问题带来的影响, 如样本缺乏[44,45]和数据分布不平衡[46,47]时模型的表征学习问题. 同时, 下一步工作还将尝试把类感知信息选择算法与因果推理[48]相结合, 指导模型在重构语义信息时获得更准确的对应关系, 从而提升模态间的互补能力.

References:

[1] Yu Q, Gao Y, Huo J, Zhuang YK. Discriminative joint multi-manifold analysis for video-based face recognition. Ruan Jian Xue Bao/Journal of Software, 2015, 26(11): 2897–2911 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4894.htm [doi: 10.13328/j.cnki.jos.004894]

[2] Ding CX, Tao DC. Robust face recognition via multimodal deep face representation. IEEE Trans. on Multimedia, 2015, 17(11): 2049–2058. [doi: 10.1109/TMM.2015.2477042]

[3] Zhang MH, Du DH, Zhang MZ, Zhang L, Wang Y, Zhou WT. Spatio-temporal trajectory data-driven autonomous driving scenario meta-modeling approach. Ruan Jian Xue Bao/Journal of Software, 2021, 32(4): 973–987 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6226.htm [doi: 10.13328/j.cnki.jos.006226]

[4] Sun QY, Wang C, Fu R, Guo YS, Yuan W, Li Z. Lane change strategy analysis and recognition for intelligent driving systems based on random forest. Expert Systems with Applications, 2021, 186: 115781. [doi: 10.1016/j.eswa.2021.115781]

[5] Wang Y, He Y, Zhu FQ, Boushey C, Delp E. The use of temporal information in food image analysis. In: Proc. of the 2015 Int'l Conf. on Image Analysis and Processing. Genoa: Springer, 2015. 317–325. [doi: 10.1007/978-3-319-23222-5_39]

[6] Zhu FQ, Bosch M, Khanna N, Boushey CJ, Delp EJ. Multiple hypotheses image segmentation and classification with application to dietary assessment. IEEE Journal of Biomedical and Health Informatics, 2015, 19(1): 377–388. [doi: 10.1109/JBHI.2014.2304925]

[7] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

[8] Li JY, Ma HK, Li XX, Qi Z, Meng L, Meng XX. Unsupervised contrastive masking for visual haze classification. In: Proc. of the 2022 Int'l Conf. on Multimedia Retrieval. Newark: ACM, 2022. 426–434. [doi: 10.1145/3512527.3531370]

[9] Du PF, Li XY, Gao YL. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 327–348 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[10] Vapnik V, Vashist A. A new learning paradigm: Learning using privileged information. Neural Networks, 2009, 22(5–6): 544–557. [doi: 10.1016/j.neunet.2009.06.042]

[11] Vapnik V, Izmailov R. Learning using privileged information: Similarity control and knowledge transfer. The Journal of Machine Learning Research, 2009, 16(1): 2023–2049.

[12] Yan Y, Nie FP, Li W, Gao CQ, Yang Y, Xu D. Image classification by cross-media active learning with privileged information. IEEE Trans. on Multimedia, 2016, 18(12): 2494–2502. [doi: 10.1109/TMM.2016.2602938]

[13] Chen JJ, Ngo CW. Deep-based ingredient recognition for cooking recipe retrieval. In: Proc. of the 24th ACM Int'l Conf. on Multimedia. Amsterdam: ACM, 2016. 32–41. [doi: 10.1145/2964284.2964315]

[14] Motiian S, Piccirilli M, Adjeroh DA, Doretto G. Information bottleneck learning using privileged information for visual recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1496–1505. [doi: 10.1109/CVPR.2016.166]

[15] Yang H, Zhou JT, Cai JF, Ong YS. MIML-FCN+: Multi-instance multi-label learning via fully convolutional networks with privileged information. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5996–6004. [doi: 10.1109/CVPR.2017.635]

[16] George A, Marcel S. Cross modal focal loss for RGBD face anti-spoofing. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 7878–7887. [doi: 10.1109/CVPR46437.2021.00779]

[17] Wen KY, Xia J, Huang YY, Li LY, Xu JY, Shao J. COOKIE: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 2188–2197. [doi: 10.1109/ICCV48922.2021.00221]

[18] Chen JJ, Ngo CW, Chua TS. Cross-modal recipe retrieval with rich food attributes. In: Proc. of the 25th ACM Int'l Conf. on Multimedia. Mountain View: ACM, 2017. 1771–1779. [doi: 10.1145/3123266.3123428]

[19] Chen JJ, Zhu B, Ngo CW, Chua TS, Jiang YG. A study of multi-task and region-wise deep learning for food ingredient recognition. IEEE Trans. on Image Processing, 2021, 30: 1514–1526. [doi: 10.1109/TIP.2020.3045639]

[20] Wang ZL, Min WQ, Li Z, Kang LP, Wei XM, Wei XL, Jiang SQ. Ingredient-guided region discovery and relationship modeling for food category-ingredient prediction. IEEE Trans. on Image Processing, 2022, 31: 5214–5226. [doi: 10.1109/TIP.2022.3193763]

[21] Min WQ, Liu LH, Luo ZD, Jiang SQ. Ingredient-guided cascaded multi-attention network for food recognition. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 1331–1339. [doi: 10.1145/3343031.3350948]

[22] Jiang SQ, Min WQ, Liu LH, Luo ZD. Multi-scale multi-view deep feature aggregation for food recognition. IEEE Trans. on Image Processing, 2020, 29(1): 265–276. [doi: 10.1109/TIP.2019.2929447]

[23] Sun B, Saenko K. Deep CORAL: Correlation alignment for deep domain adaptation. In: Proc. of the 2016 European Conf. on Computer Vision. Amsterdam: Springer, 2016. 443–450. [doi: 10.1007/978-3-319-49409-8_35]

[24] Tang JH, Shu XB, Li ZC, Qi GJ, Wang JD. Generalized deep transfer networks for knowledge propagation in heterogeneous domains. ACM Trans. on Multimedia Computing, Communications, and Applications, 2016, 12(4S): 68. [doi: 10.1145/2998574]

[25] Chen SM, Xie GS, Liu Y, Peng QM, Sun BG, Li H, You XG, Shao L. HSVA: Hierarchical semantic-visual adaptation for zero-shot learning. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems, 2021. 16622–16634.

[26] Theodoridis T, Chatzis T, Solachidis V, Dimitropoulos K, Daras P. Cross-modal variational alignment of latent spaces. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2020. 960–961. [doi: 10.1109/CVPRW50498.2020.00488]

[27] Li S, Xie BH, Wu JS, Zhao Y, Liu CH, Ding ZM. Simultaneous semantic alignment network for heterogeneous domain adaptation. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 3866–3874. [doi: 10.1145/3394171.3413995]

[28] Meng L, Chen L, Yang X, Tao DC, Zhang HW, Miao CY. Learning using privileged information for food recognition. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 557–565. [doi: 10.1145/3343031.3350870]

[29] Li XY, Xu Z, Wei K, Deng C. Generalized zero-shot learning via disentangled representation. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 1966–1974. [doi: 10.1609/aaai.v35i3.16292]

[30] Chua TS, Tang JH, Hong RC, Li HJ, Luo HJ, Zheng YT. NUS-WIDE: A real-world Web image database from National University of Singapore. In: Proc. of the ACM Int'l Conf. on Image and Video Retrieval. Santorini: ACM, 2009. 48. [doi: 10.1145/1646396.1646452]

[31] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2020.

[32] Day O, Khoshgoftaar TM. A survey on heterogeneous transfer learning. Journal of Big Data, 2017, 4: 29. [doi: 10.1186/s40537-017-0089-0]

[33] Graves A. Long short-term memory. In: Graves A, ed. Supervised Sequence Labelling with Recurrent Neural Networks. Berlin: Springer, 2012. 37–45. [doi: 10.1007/978-3-642-24797-2_4]

[34] Hu JW, Liu YC, Zhao JM, Jin Q. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing (Vol. 1: Long Papers). Association for Computational Linguistics, 2021. 5666–5675. [doi: 10.18653/v1/2021.acl-long.440]]

[35] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[36] Tang JH, Shu XB, Qi GJ, Li ZC, Wang M, Yan SC, Jain R. Tri-clustered tensor completion for social-aware image tag refinement. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016, 39(8): 1662–1674. [doi: 10.1109/TPAMI.2016.2608882]

[37] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[38] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015. 1–14.

[39] Zagoruyko S, Komodakis N. Wide residual networks. In: Proc. of the 2016 British Machine Vision Conf. York: BMVA Press, 2016.

[40] Martinel N, Foresti GL, Micheloni C. Wide-slice residual networks for food recognition. In: Proc. of the 2018 IEEE Winter Conf. on Applications of Computer Vision (WACV). Lake Tahoe: IEEE, 2018. 567–576. [doi: 10.1109/WACV.2018.0006]

[41] Ding XH, Zhang XY, Ma NN, Han JG, Ding GG, Sun J. RepVGG: Making VGG-style convNets great again. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13728–13737. [doi: 10.1109/CVPR46437.2021.01352]

[42] Ding XH, Chen HH, Zhang XY, *et al*. RepMLPnet: Hierarchical vision MLP with re-parameterized locality. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 568–577. [doi: 10.1109/CVPR52688.2022.00066]

[43] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 618–626. [doi: 10.1109/ICCV.2017.74]

[44] Lü TG, Hong RC, He J, Hu SJ. Multimodal-guided local feature selection for few-shot learning. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2068–2082 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6771.htm [doi: 10.13328/j.cnki.jos.006771]

[45] Ma HK, Qi Z, Dong XX, Li XX, Zheng YZ, Meng XX, Meng L. Cross-modal content inference and feature enrichment for cold-start recommendation. In: Proc. of the 2023 Int'l Joint Conf. on Neural Networks (IJCNN). Gold Coast: IEEE, 2023. 1–8. [doi: 10.1109/IJCNN54540.2023.10191979]

[46] Tang KH, Huang JQ, Zhang HW. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 128.

[47] Li XX, Ma HK, Meng L, Meng XX. Comparative study of adversarial training methods for long-tailed classification. In: Proc. of the 1st Int'l Workshop on Adversarial Learning for Multimedia. ACM, 2021. 1–7. [doi: 10.1145/3475724.3483601]

[48] Wang YQ, Li XX, Ma HK, QI Z, Meng XX, Meng L. Causal inference with sample balancing for out-of-distribution detection in visual classification. In: Proc. of the 2nd CAAI Int'l Conf. on Artificial Intelligence. Beijing: Springer, 2022. 572–583. [doi: 10.1007/978-3-031-20497-5_47]

附中文参考文献:

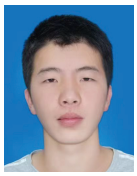
[1] 于谦, 高阳, 霍静, 庄韫恺. 视频人脸识别中判别性联合多流形分析. 软件学报, 2015, 26(11): 2897–2911. http://www.jos.org.cn/1000-9825/4894.htm [doi: 10.13328/j.cnki.jos.004894]

[3] 张梦寒, 杜德慧, 张铭苗, 张雷, 王耀, 周文韬. 时空轨迹数据驱动的自动驾驶场景元建模方法. 软件学报, 2021, 32(4): 973–987. http://www.jos.org.cn/1000-9825/6226.htm [doi: 10.13328/j.cnki.jos.006226]

[9] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述. 软件学报, 2021, 32(2): 327–348. http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[44] 吕天根, 洪日昌, 何军, 胡社教. 多模态引导的局部特征选择小样本学习方法. 软件学报, 2023, 34(5): 2068–2082. http://www.jos.org.cn/1000-9825/6771.htm [doi: 10.13328/j.cnki.jos.006771]

李象贤(1995－), 男, 博士生, CCF 学生会员, 主要研究领域为长尾图像分类, 跨模态学习.



闫晓硕(2000－), 男, 硕士生, CCF 学生会员, 主要研究领域为长尾分类, 因果推断.

郑裕泽(1998－), 男, 硕士生, 主要研究领域为跨模态特征对齐.

孟祥旭(1962－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为人机交互与图形学理论及方法, 虚拟现实与虚拟样机, 网格计算与服务计算, 制造业信息化等领域的理论研究与系统开发.

马浩凯(1997－), 男, 博士生, CCF 学生会员, 主要研究领域为冷启动推荐, 序列推荐, 跨域推荐.

孟雷(1987－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为多媒体知识挖掘, 内容表征的机器学习理论与技术研究.

齐壮(1997－), 男, 博士生, CCF 学生会员, 主要研究领域为图像分类, 联邦学习.