

融合YOLOv8n与BERT多模态特征提取的多 标签图文分类研究

余仁杰, 姜 军, 张少华, 任艺昊

(西藏大学信息科学技术学院, 西藏拉萨 850000)

摘要: 针对传统单一模态方法在图像与文本信息利用上的局限性以及其他多模态模型的不足, 提出一种融合YOLOv8n与BERT多模态特征提取的多标签分类方法。利用YOLOv8n从图像中提取多尺度特征(P3、P4、P5), 并通过全局均值池化获得固定维度的图像表示。同时, 采用预训练BERT模型对图像对应的文本描述进行编码, 抽取[CLS]向量作为文本语义表示。之后利用投影层将图像特征映射至与文本特征相同的维度, 并采用跨模态注意力机制进行特征融合, 最终通过分类头实现多标签预测。在COCO 2014数据集上, 融合YOLOv8n与BERT的多模态在多标签分类任务中, 精确率为81.64%, 召回率为56.71%, F1值为66.90%。该方法通过深度融合图像的多尺度视觉特征与文本的语义信息, 有效提升了多标签分类任务的性能, 为跨模态信息协同利用在复杂场景分类中的应用提供了实践参考。

关键词: 多模态; 特征提取; 跨模态注意力; 多标签分类

DOI: 10.11907/rjdk.251654

中图分类号: TP391.4

文献标识码: A

文章编号: 1672-7800(2025)012-0197-08

扫描二维码阅读全文:



Research on Multi-Label Image-Text Classification with Multi-Modal Feature Extraction by Integrating YOLOv8n and BERT

YU Renjie, JIANG Jun, ZHANG Shaohua, REN Yihao

(School of Information Science and Technology, Xizang University, Lhasa 850000, China)

Abstract: In view of the limitations of traditional single-modal methods in the utilization of image and text information, as well as the limitations of other multi-modal models, a multi-label classification method that integrates multi-modal feature extraction of YOLOv8n and BERT is proposed. YOLOv8n is used to extract multi-scale features (P3, P4, P5) from images. Then, a fixed-dimensional image representation is obtained through global average pooling. At the same time, the pre-trained BERT model is adopted to encode the text descriptions corresponding to the images, and the [CLS] vector is extracted as the text semantic representation. After that, a projection layer is utilized to map the image features to the same dimension as the text features. Moreover, a cross-modal attention mechanism is employed for feature fusion. Finally, multi-label prediction is realized through a classification head. On the COCO 2014 dataset, the multi-modal approach integrating YOLOv8n and BERT achieved a precision of 81.64%, a recall of 56.71%, and an F1-score of 66.90% in the multi-label classification task. This method effectively improves the performance of multi-label classification tasks by deeply integrating the multi-scale visual features of images and the semantic information of text. It provides a practical reference for the application of cross-modal information collaborative utilization in complex scene classification.

Key Words: multi-modal; feature extraction; cross-modal attention; multi-label classification

0 引言

以YOLO系列为代表的检测模型(如YOLOv8n^[1])依

托CSPNet骨干网络^[3]、多尺度特征金字塔与端到端回归机制,在实时检测场景中表现卓越,然而在复杂背景、遮挡严重和细粒度目标识别等挑战性场景中仍存在明显局限^[2]。另外,自然语言处理中的预训练模型(如BERT^[4])凭借其

收稿日期: 2025-09-19

基金项目: 国家自然科学基金项目(62261051); 西藏自治区科技计划项目(XZ202501ZY0060, XZ202402ZY0002)

作者简介: 余仁杰(1998-), 男, CCF会员, 西藏大学信息科学技术学院硕士研究生, 研究方向为多模态图文分类; 姜军(1980-), 男, 西藏大学信息科学技术学院教授, 研究方向为图像处理; 张少华(2001-), 男, 西藏大学信息科学技术学院硕士研究生, 研究方向为情感识别; 任艺昊(2001-), 男, 西藏大学信息科学技术学院硕士研究生, 研究方向为语音转换。本文通讯作者: 姜军。

强大的上下文语义建模能力,为引入文本模态以增强视觉理解提供了新的可能。尽管多模态学习已取得广泛进展,现有研究如 CLIP^[5]和 FLAVA^[6]等模型主要聚焦于图像—文本检索或描述生成任务,而在面向多标签分类的视觉—语言特征对齐与融合方面,仍缺乏深入、系统的探索。

近年来,一些研究尝试跨越视觉与语言之间的模态鸿沟。例如,CLIP 通过大规模图像—文本对进行对比学习,实现了粗粒度的跨模态对齐^[5]。ALBEF 引入动量蒸馏策略,以增强表示学习^[7]。MDETR 将检测任务与文本调制相结合,实现了端到端的多模态检测,但其计算开销较大,且未能充分挖掘多尺度视觉结构信息^[8]。此外,主流方法多采用特征拼接或均值融合等简单策略,难以实现模态间细粒度、自适应的交互,导致信息整合效率低下,尤其难以应对多标签场景中固有的复杂语义关联^[9]。

针对上述挑战,本文致力于研究一种融合 YOLOv8n 与 BERT 的多模态特征提取与表示方法,重点解决细粒度视觉—语言对齐、动态融合与标签结构化建模等关键问题。本文的主要贡献包括:

(1)提出多尺度视觉—语言特征动态对齐机制。基于 YOLOv8n 提取的多层级视觉特征(P3 - P5),结合跨模态注意力实现图像区域与 BERT 文本表示(如[CLS]向量)间的精细化语义关联,有效抑制背景干扰。

(2)设计一种轻量化的门控融合模块。通过双线性投影与门控权重自适应调节多模态特征贡献,避免冗余与信息损失,提升模型表达效率^[19]。

(3)引入标签共现先验知识以及门控共现增强机制。通过结构化语义建模增强类别间的相关性推理,显著提升长尾及多标签情境下的分类性能。

实验结果显示,在 COCO 2014 多标签分类任务中,本文方法相比基线模型和传统多模态模型在 F1、Precision 与 Recall 等指标上都有明显提升,表明该方法为复杂场景下的多模态感知提供了可行的技术路径^[10]。

1 基于 YOLOv8n 与 BERT 的多模态融合方法

1.1 YOLOv8n 网络结构

YOLOv8n 的网络结构如图 1 所示,主要由 Backbone、Neck 和 Head 三部分构成,各部分之间通过高效的信息传递和特征融合,实现从图像预处理到目标检测结果预测的全流程优化。其中,每部分都由不同的小组件组成,如图 2 所示。

(1)Backbone。主要负责从输入图像中提取底层到高级的视觉特征。该部分不仅包含传统的卷积(Conv)层和 Bottleneck 模块,而且引入了改进的 C2f 模块以提高特征融合效率。

(2)Neck。位于 Backbone 与 Head 之间,负责多尺度特征的融合和增强。YOLOv8n 的 Neck 部分采用类似特征金字塔网络(Feature Pyramid Network)的结构,通过多尺度池化操作在不同层次上整合信息,以增强特征图的表达能力,从而提升检测精度。

(3)Head。Head 部分负责实际的目标检测和边界框回归,其包含多个输出层,每个输出层对应不同尺度的特征图。Head 通过卷积层对 Neck 传递的特征图进行处理,输出目标类别概率、边界框坐标和置信度,最终整合成完整的检测结果。

1.2 BERT 网络结构

BERT (Bidirectional Encoder Representations from Transformers) 是一种基于 Transformer 架构的深度学习模型,能够利用双向注意力机制进行上下文建模。与传统的单向语言模型不同,BERT 通过双向 Transformer 编码器从文本的左侧和右侧同时学习语义信息,从而提升了对自然语言的理解能力。BERT 的整体结构主要包括输入层(Embedding 层)、多层 Transformer 编码层(Encoder)、输出层 3 部分,如图 3 所示。

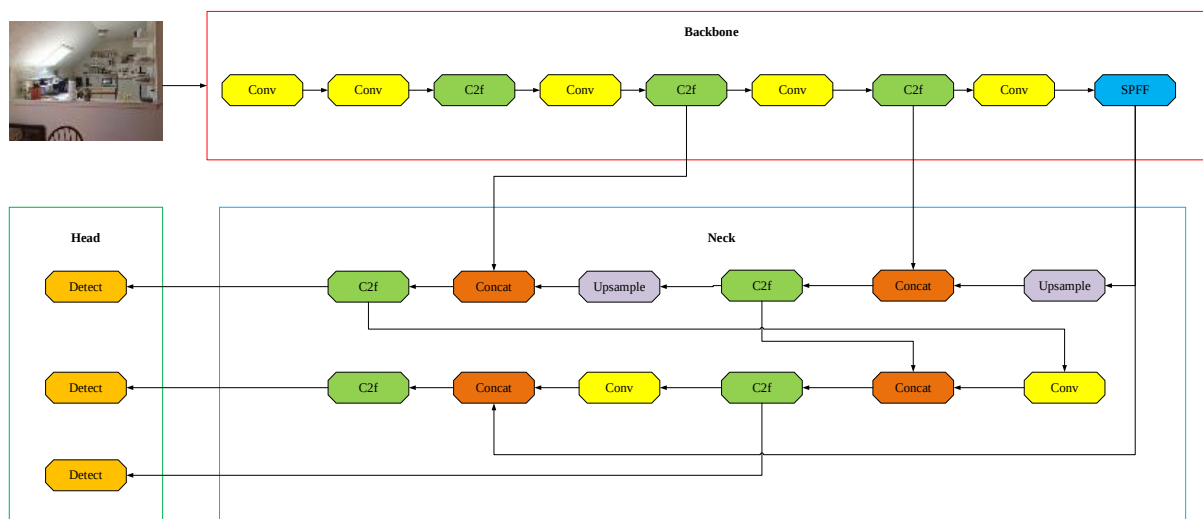


Fig. 1 Architecture of YOLOv8 network

图 1 YOLOv8 网络结构

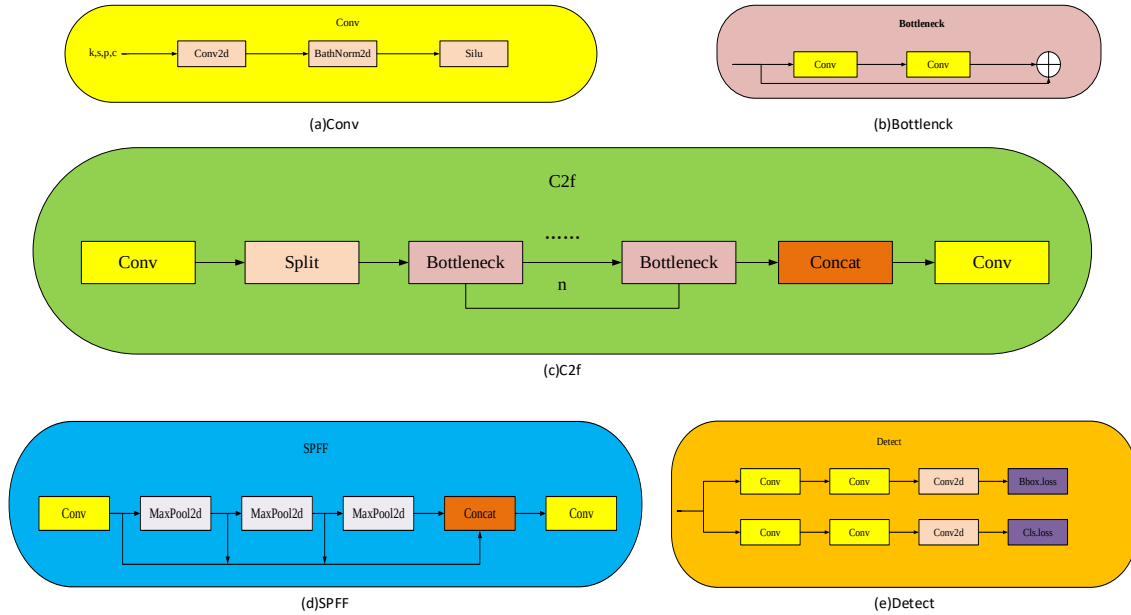


Fig. 2 Structural components of YOLOv8 network

图 2 YOLOv8 网络的小组件

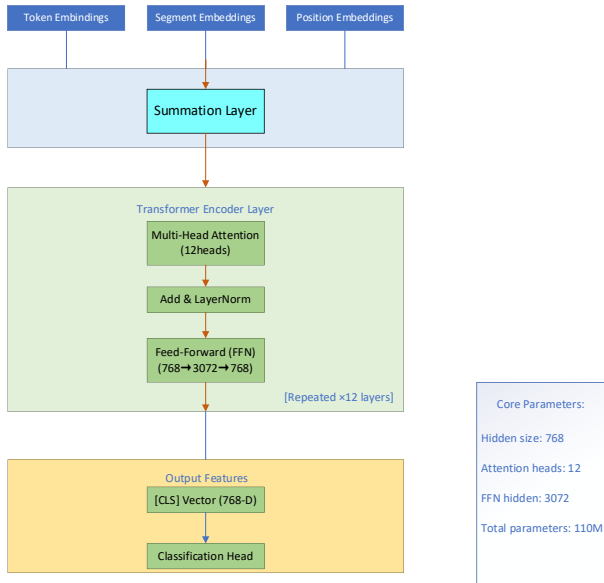


Fig. 3 Architecture of the BERT network

图 3 BERT 网络结构

(1) 输入层。输入层由 Token Embedding、Positional Embedding 和 Segment Embedding 组成。其中,Token Embedding 负责将输入文本转换为词向量,Positional Embedding 通过固定位置编码提供单词位置信息,而 Segment Embedding 用于区分不同的句子,如在 NSP(Next Sentence Prediction)任务中区分句子 A 和句子 B。最终,将这 3 种嵌入向量相加后作为 Transformer 编码层的输入。

(2) 多层 Transformer 编码层。BERT 采用多个 Transformer 层进行深度语义建模,BERT-base 版本包含 12 层 Transformer 编码器,BERT-large 版本则包含 24 层 Transformer 编码器。每个 Transformer 层由自注意力机制(Self-Attention)和前馈神经网络(Feed Forward Network, FFN)

组成。其中,自注意力机制通过计算 Query、Key 和 Value 之间的相关性,捕获不同词语之间的依赖关系。前馈神经网络则用于对注意力加权后的表示进行非线性变换,以进一步增强特征表达能力。

(3) 输出层。在 BERT 的输出层,不同任务可以选择不同的方式进行特征提取。例如,在文本分类任务中,通常使用特殊的 [CLS] token 对整个输入序列进行表示,并将其送入全连接层进行分类;在命名实体识别(Named Entity Recognition)任务中,则对所有 token 进行标注。此外,BERT 还支持问答任务(Question Answering),可直接提取 [SEP] 分隔的两个句子中的答案位置。

1.3 融合 YOLOv8n 与 BERT 的多模态特征提取及表示方法

本文方法通过层次化视觉语义提取、跨模态动态交互和标签语义增强 3 阶段实现多模态信息融合与多标签分类,整体架构如图 4 所示。

在视觉特征提取方面,本文选用 YOLOv8n 作为视觉骨干网络,原因在于其多尺度特征金字塔(P3、P4、P5)能够同时捕获细节纹理、中层结构及全局语义信息,相较于仅输出单一高层特征的分类网络(如 ResNet),更适用于与文本特征进行细粒度区域对齐。文本特征提取则采用 BERT 模型,依托其深层双向自注意力机制,生成能够充分表达词语上下文关系的语义表示,为后续融合提供高质量文本嵌入。为克服传统特征拼接或均值池化等静态融合策略的局限性,本文引入双向跨模态注意力机制。该模块实现了图像区域与文本词符之间的软对齐,通过动态计算跨模态关联权重,实现内容自适应的特征融合。此外,为进一步利用标签间固有的语义相关性,本文提出一种门控共现增强机制,将从训练数据中统计得到的标签共现先验,通过

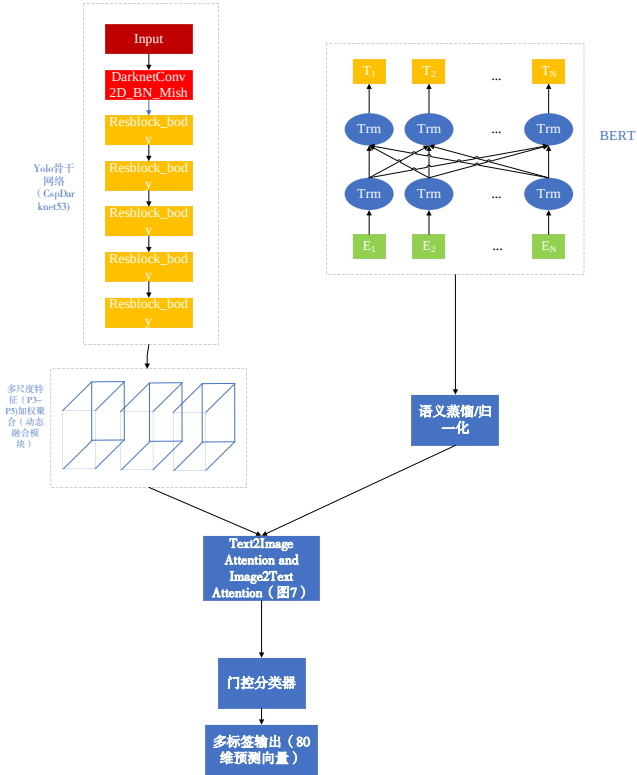


Fig. 4 Overall architecture of the fusion framework

图4 融合整体架构

可学习门控单元自适应地注入分类过程,从而显著提升模型对长尾类别的识别能力。

(1)输入处理。原始图像经YOLOv8n骨干网络提取多尺度特征(P3-P5),其空间分辨率分别为80×80、40×40、

20×20,通道数分别为256、512、1 024,生成局部细节、中层语义与全局上下文表示。同时,图像的文本描述通过动态实体保留策略(保留名词短语、随机掩码修饰词)输入BERT模型^[4]。

(2)特征表达。视觉特征采用动态门控融合(Dynamic Gated Fusion, DGF)模块(见图5)将P3-P5特征加权聚合为 $f_v \in R^{1024}$ 。具体公式为:

$$\alpha_i = \sigma(W_g \cdot \text{CONCAT}(f_{P3}, f_{P4}, f_{P5})) \quad (1)$$

$$f_v = \sum_{i=3}^5 \alpha_i \cdot f_{P_i} \quad (2)$$

其中, W_g 是门控权重矩阵,为可学习参数; f_{P_i} 为第*i*层(P3、P4、P5)特征图提取的视觉特征向量; CONCAT 为通道维度拼接操作; σ 为Sigmoid激活函数; α_i 为对应尺度的门控系数,表示第*i*层特征在融合中所占权重; f_v 为融合后的视觉特征向量,可综合不同尺度特征的加权结果。

文本特征采用BERT输出的[CLS]标记,经语义蒸馏层(移除停用词干扰)得到 $f_t \in R^{768}$ 。

(3)多模态交互。在多模态交互阶段,首先通过轻量级线性投影实现特征空间对齐,视觉特征 f_v 经线性变换 W_v 与GELU(Gaussian Error Linear Unit)激活函数映射为 $h_v = \text{GELU}(W_v f_v + b_v)$,文本特征 f_t 通过归一化得到 $h_t = \text{LayerNorm}(f_t)$ 。随后构建双向注意力交互机制流程(Bidirectional Attention Interaction Mechanism Process),如图6所示^[16-17]。

视觉到文本的注意力权重为:

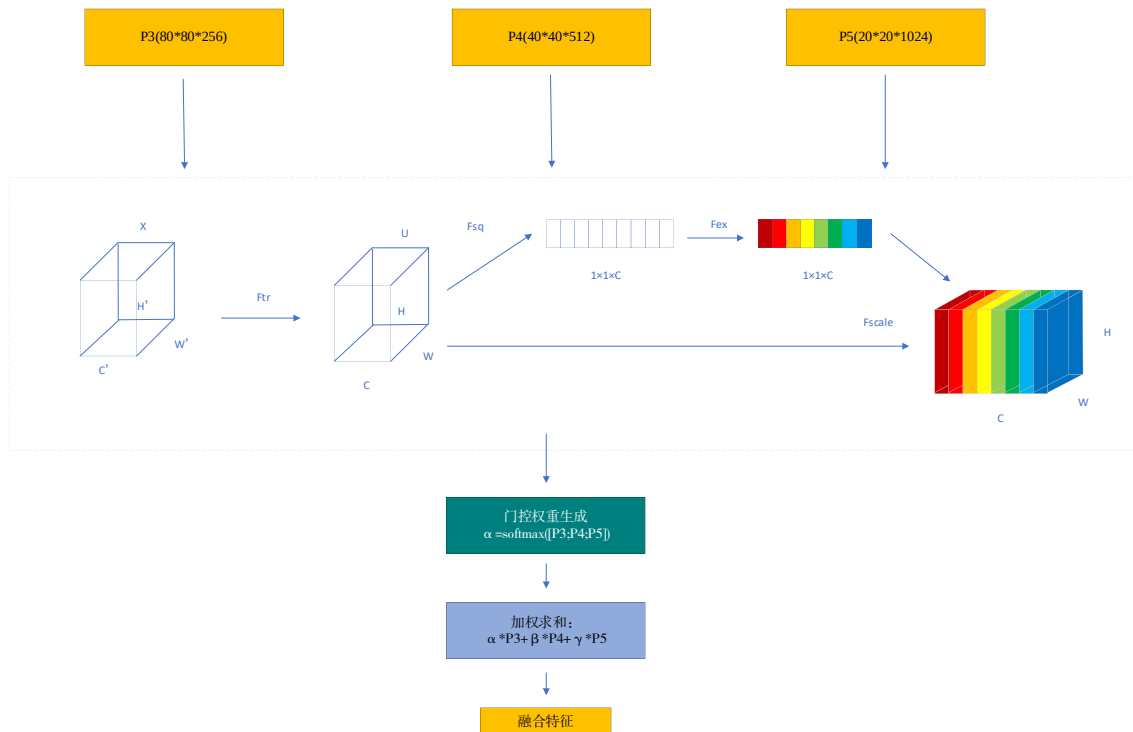


Fig. 5 Dynamic gated fusion module

图5 动态门控融合模块

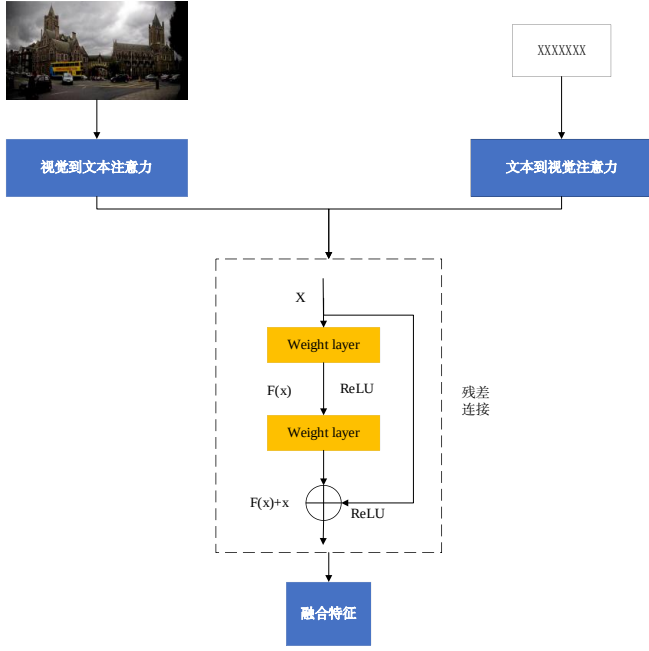


Fig. 6 Flow of the bidirectional attention mechanism

图 6 双向注意力机制流程

$$A_{v \rightarrow t} = \text{Softmax} \left(\frac{h_v W_Q (h_t W_K)^T}{\sqrt{d}} \right) h_t W_V \quad (3)$$

聚焦图像区域与关键词的关联, 文本到视觉的注意力权重为:

$$A_{t \rightarrow v} = \text{Softmax} \left(\frac{h_t W'_Q (h_v W'_K)^T}{\sqrt{d}} \right) h_v W'_V \quad (4)$$

该权重将聚焦后的视觉信息注入文本表征, 实现反向增强语义^[11]。最终通过残差链接融合多模态信息: $h_{\text{fusion}} = h_v + h_t + A_{v \rightarrow t} + A_{t \rightarrow v}$ 。其中, W_Q/W'_Q 、 W_K/W'_K 、 W_V/W'_V 为可学习的查询、键、值投影矩阵, d 为特征维度^[12]。

图 7 展示了本文在图文多模态融合中所采用的双向注意力机制(Bidirectional Attention)^[18]。该机制包含两个方向: Text-to-Image Attention 与 Image-to-Text Attention,

分别建模文本对图像的注意力与图像对文本的注意力。相较于常规自注意力机制仅在同一模态内进行特征交互, 存在一定的不足, 而本文提出的图像到文本的注意力(Image-to-Text Attention)机制, 使得每个图像区域能够聚焦于最相关的文本词符, 进而增强视觉特征的语义表达能力。与此同时, 文本到图像的注意力(Text-to-Image Attention)机制则反向强化了文本对图像关键区域的关注, 从而提升了语义表示的准确性。这种双向交互不仅有效弥补了单模态注意力在跨模态语义鸿沟上存在的不足, 而且能抑制模态间的噪声干扰, 最终实现更细粒度的语义对齐。

如图 7(左图)所示, 文本特征为 U_1, U_2, U_3 与图像特征为 P_1, P_2, P_3 构成一个匹配相似度矩阵。通过对每一列(即每个图像区域)进行 softmax 操作, 得到 attention 权重。利用该权重对文本特征进行加权平均, 得到与每个图像区域对应的上下文文本表示。

如图 7(右图)所示, 从相似度矩阵中提取每一行的最大值构成向量, 该向量表示文本整体对图像各区域的注意力强度。将其作用于所有图像特征, 得到全局图像上下文向量。将每个图像区域特征 P 与其对应的文本上下文表示及全局图像表示进行拼接, 融合后得到的 G_t 表示图像区域融合了与其最相关的文本信息与整体图像语义信息。

(4) 分类标签。在多标签预测阶段, 提出门控共现增强机制^[15]。基于 COCO 训练集统计的标签共现矩阵 $M_{\text{co-occur}} \in R^{80 \times 80}$, 通过可学习门控权重动态融合原始特征与先验知识。具体而言, 门控系数为: $g = \sigma(W_g[h_{\text{fusion}}; M_{\text{co-occur}} h_{\text{fusion}}])$, 该系数用于动态权衡多层感知机(Multilayer Perceptron, MLP)的预测结果与标签共现先验信息在最终输出中的贡献比例。最终输出预测向量为:

$$y = g \cdot \text{MLP}(h_{\text{fusion}}) + (1 - g) \cdot (M_{\text{co-occur}} h_{\text{fusion}}) \quad (5)$$

其中, σ 为 Sigmoid 函数, 用于将门控值约束在 $(0, 1)$;

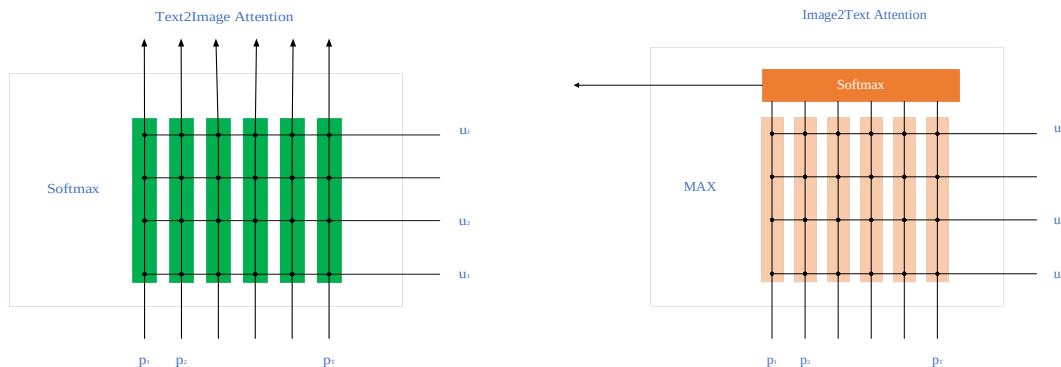


Fig. 7 Structure of the bidirectional attention mechanism

图 7 双向注意力机制结构

$W_g \in R^{2d \times 1}$ 为门控权重矩阵。模型采用多任务联合优化策略,损失函数定义为:

$$L = L_{det} + \lambda_1 L_{cont} + \lambda_2 L_{cls} \quad (6)$$

其中, λ_1, λ_2 为平衡超参数(实验中分别设置为 0.5 和 1.0)。通过统计各类别在训练阶段的门控权重变化趋势发现,门控单元在高共现类别(如“fork - knife”“dog - person”)间会自动分配更高权重,而在低相关类别间保持低耦合,说明模型能够准确地根据标签语义关系自适应调整特征贡献。这种动态调节机制使得模型在长尾分布数据中表现更加稳健^[13]。

2 实验与结果分析

2.1 数据集

本文采用 MS COCO 2014 数据集作为实验平台。MS COCO(Common Objects in Context)数据集由 Lin 等^[10]提出,包含 82 783 张训练图像和 40 504 张验证图像,涵盖 80 个物体类别。每张图像提供 5 条人工标注的文本描述。采用多标签生成策略对每张图片的实例标注进行类别映射,生成 80 维二值向量。当某类别存在至少一个实例时,对应位置为 1。

2.2 实验参数

本文的模型训练基于 PyTorch 框架实现,需重点调控以下 6 个具有强关联性的核心调节要素与工程技术节点:迭代轮数(Epochs)、批量样本数(Batch Size)、输入维度(Imgsz)、学习率(Learning Rate)、并行载入进程(Workers)、预训练权重初始化策略(Pretrained Weights)。优化器采用 AdamW 算法,初始学习率设置为 $1e-4$,并耦合 $1e-5$ 权重衰减系数,视觉分支加载 YOLOv8n 官方 COCO 预训练权重,文本编码器冻结 BERT-base 参数以保留语言表征能力。图像输入尺寸保持与 YOLOv8n 默认训练分辨率一致,数据增强包含随机水平翻转与标准化处理。通过调整参数,以实现最佳的训练效果和模型性能。

2.3 实验结果对比

为评估本文提出的 YOLOv8n+BERT 多模态图像—文本分类框架的性能,本节将其与单模态模型 YOLOv8n、BERT 以及主流多模态模型(包括 CLIP、VILT 和 BLIP)进行比较。评估指标包括精确率(Precision)、召回率(Recall)和 F1 分数(F1-score)。表 1、表 2 展示了各多模态模型的性能表现,表 3 展示了模型的消融实验结果。

CLIP 的 F1 分数为 50.20%(精确率 62.40%,召回率

Table 1 Performance comparison of multi-modal models

表 1 多模态模型性能比较 (%)

模型	Precision	Racall	F1
CLIP	62.40	41.80	50.20
VILT	78.89	35.20	48.50
BLIP	74.20	46.30	57.10
YOLOv8n+BERT	81.64	56.71	66.90

Table 2 Result comparison of models on multi-label metrics

表 2 模型在多标签指标上的结果比较

模型	mAP	AUC(macro)	Hamming Loss
CLIP	0.39	0.75	0.24
VILT	0.37	0.73	0.26
BLIP	0.47	0.79	0.20
YOLOv8n+BERT	0.53	0.83	0.15

Table 3 Module ablation experiment results

表 3 模块消融实验结果 (%)

模型	Precision	Recall	F1
YOLOv8n	65.20	48.51	55.79
BERT	74.17	46.29	57.12
+ BERT (拼接融合)	76.80	50.93	61.15
+ 动态门控融合(DGF)	78.52	52.49	62.71
+ 双向注意力	80.49	54.77	65.30
+ 门控共现(完整模型)	81.64	56.71	66.90

41.80%)。尽管其通过对比学习实现了较好的图像—文本对齐效果,但在细粒度分类任务中表现不佳。这主要因为 CLIP 的预训练目标更偏向于泛化匹配,而非特定任务的优化,导致在复杂场景下语义对齐不足。同时召回率较低,表明模型在捕获多样化正例样本时存在局限。

VILT 的精确率较高(78.89%),但召回率仅为 35.20%,F1 分数为 48.50%。其主要问题包括:①模型对图像—文本交互的建模较为简单,难以处理复杂场景下的跨模态关联;②在长尾分布数据上表现不佳,导致召回率严重不足;③训练过程中缺乏对细粒度类别的充分优化。

BLIP 在基线模型中表现最佳,F1 分数为 57.10%(精确率 74.20%,召回率 46.30%)。然而,其性能仍存在不足。虽然 BLIP 在视觉—语言预训练方面表现较好,但其跨模态融合机制较为浅层,未能充分利用文本的深层语义信息。召回率偏低,表明模型在处理多样化图像—文本对时存在遗漏。

最后在模型融合阶段逐步加入模块进行消融实验,并对结果进行比较。引入 BERT 文本特征并与视觉特征简单拼接,F1 提升至 58.20%,表明引入文本模态信息的巨大优势,但简单的拼接融合效率有限。替换简单的全局特征,采用动态门控融合(DGF)模块融合多尺度特征后,F1 提升至 61.51%。值得注意的是,DGF 模块在设计中采用了轻量化策略,仅引入少量可学习参数,却使 F1 提升了约 1.6%,说明轻量化设计不仅降低了计算复杂度,而且在特征聚合阶段提升了有效信息比例,验证了该模块设计在保持性能的同时兼顾效率的优势。用双向注意力机制替换特征拼接,实现动态融合,F1 值提升至 65.30%,证明了多尺度特征对于丰富视觉表示的重要性。用双向注意力机制替换特征拼接,实现动态融合,F1 值显著提升至 65.30%。这充分验证了跨模态注意力在建模模态间交互方面的优越性,其是性能提升的关键。在分类头中加入门控共现先验,F1 分数达到最终的 66.90%,表明显式地利用标签相关性知识,特别是对提升长尾类别的 Recall 有积极效果。

同时,本文在原有 Precision、Recall、F1 指标基础上,补充计算了 mAP、AUC(macro)和 Hamming Loss 3 项指标,如表 2 所示。结果表明,本文提出的 YOLOv8n+BERT 模型在 3 个指标上均优于其他模型,表现出更强的多标签预测能力和更低的错误率。从表 2 可以看出,YOLOv8n+BERT 在 mAP 与 AUC 指标上分别达到 0.54 和 0.83,较 BLIP 提升了约 7% 与 5%,表明其在整体排序与正样本区分能力上更强。同时,Hamming Loss 降至 0.145 9,较基线模型减少约 30%,说明该模型在标签预测的准确性和鲁棒性方面均有所提升。结合 P、R、F1 指标可知,本文方法在多模态特

征融合中取得了兼顾全局精度与细粒度语义捕获的平衡效果。

消融实验结果表明,本文提出的每一个模块都对最终性能有正向贡献,其中双向注意力机制带来的增益最大,验证了该机制在细粒度语义对齐和噪声抑制方面相比常规注意力的显著优势,表明其是解决核心模态融合问题的关键。

为进一步验证本文提出多模态融合的多标签图文分类算法的有效性,对 4 个典型场景下的图文标签预测结果进行比较,如图 8 所示。

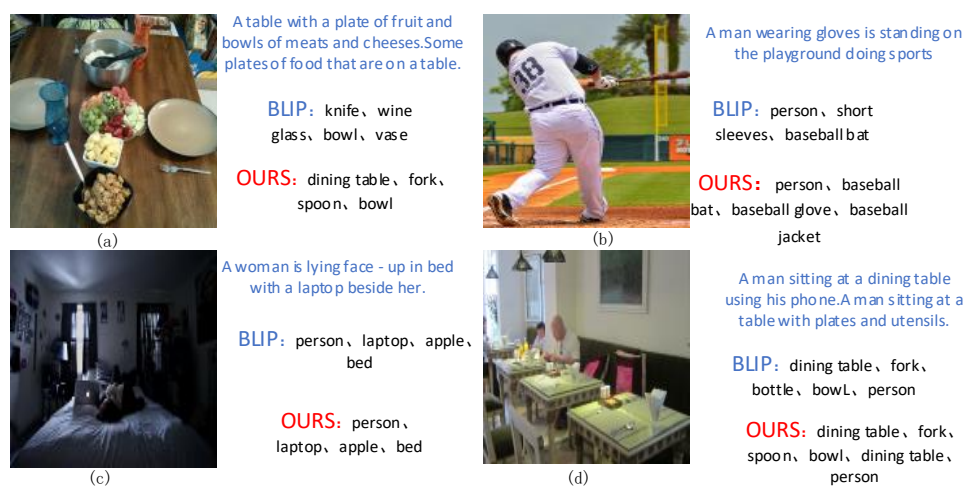


Fig. 8 Comparison results of models

图 8 模型对比结果

将本文算法与 BLIP 模型进行详细对比分析,从中获得以下关键发现:首先,在场景理解能力上显著提升,例如在第一个餐桌场景中,BLIP 模型识别出了“knife, wine glass, bowl, vase”等物品,而本文方法能够更准确地识别出“dining table, fork, spoon, bowl”等核心餐具,体现了对用餐场景的更深层理解。这一优势源于本文提出的层次化视觉语义提取机制和双向注意力交互设计。通过 YOLOv8n 多尺度特征提取(P3-P5)结合动态门控融合模块,能够同时捕获局部细节和全局上下文信息;而双向注意力机制使得文本语义能够引导图像区域关注,利用图像特征反向增强文本理解,从而实现了比 BLIP 单向融合更深层的语义关联。其次,门控共现增强机制通过引入 COCO 标签共现先验,有效捕捉了物品间的固有关联(如“刀叉”组合),使模型能够理解物品间的内在关联性,而非仅仅依赖视觉表面特征进行独立识别。同时,上下文关联性增强,细节捕获能力突出。与 BLIP 等模型相比,本文方法通过跨模态注意力与门控融合策略有效抑制了背景噪声干扰。例如,在复杂场景(如图 8 的街景与餐桌场景)中,模型能自动削弱无关背景区域(如天空、桌布)的注意力响应,从而提升主要目标的识别精度,体现出模型在噪声抑制方面的显著优势。这些定性结果充分表明,本文提出的多模态融合机制不仅提升了量化指标,更重要的是,在实际的多标签图文分类应用场景中,能够实现更加全面、准确的图

文内容理解。基于 4 组典型样例的扩展分析显示,模型在自然场景、交通和室内等不同语义复杂度的环境下均能保持较高的分类稳定性。尤其在低光照与部分遮挡等实际部署中常见的极端条件下,仍能准确捕获关键目标区域,在环境中体现出较强的鲁棒性。

综合表 1、表 2 的量化结果与图 8 的定性表现,本文方法在多标签识别任务中对精确率与召回率实现了较好的平衡,具备良好的泛化潜力,为智能图文分析系统的实际部署提供了坚实有力的技术支撑。

3 结语

本文针对图像—文本多模态多标签分类任务,提出一种基于 YOLOv8n 与 BERT 的融合架构,通过跨模态注意力机制实现视觉与语言特征的深度交互,显著提升了多标签分类性能。

本文的主要贡献包括:提出层次化多模态特征融合框架,有效结合 YOLOv8n 的多尺度视觉特征与 BERT 的语义表示,通过双向跨模态注意力机制实现图像区域与文本标记间的精细化对齐,在增强语义关联的同时有效抑制了噪声干扰;设计门控共现增强机制,创新性地引入标签共现先验,通过可学习门控权重实现数据驱动特征与语义先验的动态融合,显著提升了模型对于长尾类别的识别能力。

在 MS COCO 2014 等多标签分类基准上的实验表明,本文方法在精确率、召回率和 F1 分数等指标上均显著优于基线模型与主流多模态方法,验证了多模态融合在复杂场景下的优势。实验分析进一步揭示,本文方法在保持较高精确率的同时,显著提升了召回率与 F1 值,证明其在不同样本分布与复杂场景下具备更强的泛化能力。可视化结果也表明,该方法能够更准确地捕捉图像—文本间的语义关联,在细粒度物体识别与场景上下文建模方面表现出色。

本文实验主要基于 COCO 2014 数据集,尽管该数据集广泛用于多模态研究,但仍存在一定局限性:一是场景类型集中,部分类别样本有限,难以反映真实环境中多样的视觉与语义分布;二是文本描述质量参差,标注者语言表达差异导致语义一致性不足,可能影响模型对真实场景的泛化能力。未来将在跨域数据集(如 Visual Genome、Flickr30K 及自建场景图文数据集)上进行验证,以评估模型的鲁棒性与迁移能力。同时探索弱标注或自监督学习,减少对高质量标注文本的依赖,提升对开放环境的适应性。

尽管本文研究效果良好,但模型仍有改进空间。如对图像—文本对中的噪声敏感,尤其在文本描述存在语义偏差时表现不稳定;参数量较大,推理效率待提升;预训练语言模型未进行任务特定微调,可能限制其视觉—语言任务的特征适配能力。针对上述问题,未来将引入课程学习策略分阶段优化视觉与语言模块,提升困难样本的鲁棒性;结合自监督对比学习构建跨模态对齐损失,增强特征的判别性;探索模型轻量化与知识蒸馏技术,以达成性能与推理效率的平衡,满足实际部署的需求。

综上,本文提出的跨模态注意力与门控融合机制在语义对齐、噪声抑制和特征自适应调节方面优于传统注意力机制,为多模态特征高效融合提供了新思路,也为后续多标签分类与跨模态理解研究提供了有益参考。

参考文献:

- [1] WANG G, CHEN Y, AN P, et al. UAV-YOLOv8: a small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios[J]. *Sensors*, 2023, 23(16): 7190.
- [2] YANG J, LI C, ZHANG P, et al. Unified contrastive learning in image-text-label space[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 19163-19173.
- [3] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 390-391.
- [4] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.
- [5] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning, 2021: 8748-8763.
- [6] SINGH A, HU R, GOSWAMI V, et al. Flava: a foundational language and vision alignment model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 15638-15650.
- [7] LI J, SELVARAJU R, GOTMARE A, et al. Align before fuse: vision and language representation learning with momentum distillation[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 9694-9705.
- [8] KAMATH A, SINGH M, LECUN Y, et al. Mdetr-modulated detection for end-to-end multi-modal understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 1780-1790.
- [9] ZHANG L, LIU Z, ZHU X, et al. Weakly aligned feature fusion for multi-modal object detection[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2025, 36: 4145-4159.
- [10] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context[C]//European Conference on Computer Vision, 2014: 740-755.
- [11] LU J, BATRA D, PARIKH D, et al. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks [DB/OL]. <https://arxiv.org/abs/1908.02265>.
- [12] TAN H, BANSAL M. LXMERT: learning cross-modality encoder representations from transformers[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 5100-5111.
- [13] CHEN Y C, LI L, YU L, et al. UNITER: universal image-text representation learning[C]//European Conference on Computer Vision, 2020: 104-120.
- [14] CHEN Z M, WEI X S, WANG P, et al. Multi-label image recognition with graph convolutional networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5177-5186.
- [15] AREVALO J, SOLORIO T, MONTES-Y-GÓMEZ M, et al. Gated multimodal units for information fusion [DB/OL]. <https://arxiv.org/abs/1702.01992>.
- [16] LI Y, LU T, ZHENG H T, et al. MESED: a multi-modal entity set expansion dataset with fine-grained semantic classes and hard negative entities[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 8697-8706.
- [17] JI J, XU C, ZHANG X, et al. Spatio-temporal memory attention for image captioning[J]. *IEEE Transactions on Image Processing*, 2020, 29: 7615-7628.
- [18] CAO P, YANG Z, SUN L, et al. Image captioning with bidirectional semantic attention-based guiding of long short-term memory [J]. *Neural Processing Letters*, 2019, 50(1): 103-119.
- [19] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension [DB/OL]. <https://arxiv.org/abs/1611.01603>.

(责任编辑:黄健)