智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

结合多面图像特征提取和门控融合机制的多模态方面级情感分析

赵雪峰, 狄恒西, 柏长泽, 仲兆满, 仲晓敏

引用本文:
赵雪峰, 狄恒西, 柏长泽, 等. 结合多面图像特征提取和门控融合机制的多模态方面级情感分析[J]. 智能系统学报, 2025, 20(6): 1461-1473.
ZHAO Xuefeng, DI Hengxi, BAI Changze, et al. Multimodal aspect-based sentiment analysis combining multifaceted image feature extraction and gated fusion mechanism[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(6): 1461-1473.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202503032

DOI: 10.11992/tis.202503032
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20251010.1108.004

结合多面图像特征提取和门控融合机制的多模态方面级情感分析

赵雪峰, 狄恒西, 柏长泽, 仲兆满, 仲晓敏
(江苏海洋大学 计算机工程学院, 江苏 连云港 222005)

摘 要: 针对现阶段多模态方面级情感分析 (multimodal aspect-based sentiment analysis, MABSA) 模型仅提取单一图像全局特征、忽略关键细节信息的问题, 提出一种结合多面图像特征提取和门控融合机制的网络模型。该模型通过构建多面图像特征提取模块, 采用跨模态翻译技术, 从图像中与情感相关的多个维度生成场景、人脸、物体和颜色文本描述, 实现细节信息提取与跨模态信息对齐; 设计门控融合交互模块, 引入门控机制与交互注意力实现特征间的高效融合交互; 为了弥补不同模态间的表示差距, 构建融合图片提示的序列信息, 将图像特征转换到预训练语言模型 (pre-trained language model, PLM) 的输入空间中, 实现更准确的情感分类。在 Twitter-2015 和 Twitter-2017 数据集上的实验表明, 该模型较现有模型在准确率和 F_1 上平均提高 0.93% 和 0.52%, 能有效改善情感分类效果。

关键词: 全局特征; 多模态; 方面级情感分析; 文本描述; 门控机制; 交互注意力; 图片提示; 预训练语言模型
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2025)06−1461−13

中文引用格式: 赵雪峰, 狄恒西, 柏长泽, 等. 结合多面图像特征提取和门控融合机制的多模态方面级情感分析 [J]. 智能系统学报, 2025, 20(6): 1461–1473.
英文引用格式: ZHAO Xuefeng, DI Hengxi, BAI Changze, et al. Multimodal aspect-based sentiment analysis combining multifaceted image feature extraction and gated fusion mechanism[J]. CAAI transactions on intelligent systems, 2025, 20(6): 1461–1473.

Multimodal aspect-based sentiment analysis combining multifaceted image feature extraction and gated fusion mechanism

ZHAO Xuefeng, DI Hengxi, BAI Changze, ZHONG Zhaoman, ZHONG Xiaomin
(College of Computer Engineering, Jiangsu Ocean University, Lianyungang 222005, China)

Abstract: Existing multimodal aspect-based sentiment analysis models only extract single global image features, thereby overlooking key detailed information. To address this issue, this study proposes a network model that combines multifaceted image feature extraction and a gated fusion mechanism. Specifically, a multifaceted image feature extraction module is constructed in the proposed model. By leveraging cross-modal translation technology, textual descriptions of scenes, human faces, objects, and colors are generated from multiple sentiment-related dimensions of the image. This process achieves detailed information extraction and cross-modal information alignment. Furthermore, a gated fusion interaction module has been developed, incorporating a gating mechanism and interactive attention to facilitate efficient fusion and interaction between features. In order to address the representation gap across different modalities, sequence information is integrated with image prompts to convert image features into the input space of the pre-trained language model (PLM). This facilitates more accurate sentiment classification. Experiments conducted on the Twitter-2015 and Twitter-2017 datasets demonstrate that compared with existing models, the proposed model achieves an average improvement of 0.93% in accuracy and 0.52% in F_1-score, effectively enhancing the performance of sentiment classification.

Keywords: global feature; multimodal; aspect-based sentiment analysis; text description; gating mechanism; cross attention; image-prompt; pre-trained language model

作为情感分析任务中的一个重要分支, 多模态方面级情感分析 (multimodal aspect-based sentiment analysis, MABSA) 旨在通过分析给定的文本和图像信息, 进一步推断文本中某个特定方面词实体的情感极性。随着社交媒体和在线评论平台的广泛使用, 用户越来越倾向于使用多种模态的数据[1] 表达自己的情绪, 如何从海量的生成内容中准确提取有价值的情感信息, 已经成为多模态情感分析任务中一个重要的研究课题, 具有广阔的研究前景和重要的社会意义[2]。

在多模态情感分析任务中, 不同的模态携带不同的情感信息, 相比单一文本情感分析, 图像与文本的相互作用能够为情感分析提供一个全面的情感视角。现阶段研究人员已经在 MABSA 任务中取得了一定成果, Xu 等[3] 通过注意力机制对方面词、文本和图像实现模态间的交互; Wang 等[4] 基于 BERT(bidirectional encoder representation from transformers) 和 ResNet(residual network)架构提出了一个面向目标的多模态情感分类模型;

收稿日期: 2025−03−24. **网络出版日期**: 2025−10−10.
基金项目: 国家自然科学基金项目 (72174079); 江苏省“青蓝工程”优秀教学团队项目 (2022-29).
通信作者: 赵雪峰. E-mail: zhaoxf@jou.edu.cn.

Khan 等[5] 通过 Transformer 将图像转换为描述性文本语句，将多模态任务转化为单模态任务；Yang 等[6] 针对图像中存在的面部情感引入 FITE (face-sensitive image-to-emotional-text translation) 方法，在文本模态中有选择地将其与目标方面词进行匹配和融合。这些方法证明了将图片与单一文本情感分析结合能够较好地提高整体情感分析任务性能[7]。但仍存在图像特征提取不充分。现有研究仅侧重关注视觉特征的某一方面，如 Xu 等[3]、Wang 等[4] 分别通过 CNN(convolutional neural network) 和 ResNet 提取图像的全局特征，Yang 等[6] 通过面部表情捕捉视觉情感线索，但忽略了其他相关的细节情感特征，如场景描述、物体分析、颜色背景等，导致图像中丰富的情感线索未能被充分挖掘，限制了情感分析任务的性能表现。在实际社交媒体内容中，生成的图像蕴含着丰富的情感特征，包括人脸、场景、物体和色彩等，其中整体场景描述在传达情感信息方面起着关键作用，人脸表情和物体也表现出相应的情绪信号，整体颜色也提供了重要的情绪特征。通过在情感分析任务中考虑场景、人脸、物体和颜色等特征元素，可以获得更为准确的分析结果。为此，设计一种结合多种视觉特征信息的提取框架是非常重要的。

特征提取后的多模态数据面临另一个重要的问题：如何增强情感特征的融合交互能力。Xu 等[3]、余本功等[8] 通过多层注意力实现特征的融合交互，曹银妮等[9] 利用多头交叉注意力指导模型关注与方面词相关的图像特征，都取得了不错的效果。而在具体的多模态特征融合过程中，鉴于输出特征的多样性与模态间特征的差异性，盲目地与不相关信息交互易导致信息错位和丢失，简单的合并操作往往无法较好地利用各种特征的优势，若针对多种特征不进行任何过滤可能会在学习过程中引入噪声，进而影响整体任务性能，因此如何更好地处理和融合多模态特征信息显得尤为关键。

基于上述问题，本文提出了一种结合多面图像特征提取和门控融合机制的 MABSA 模型。通过构建多面图像特征提取模块，生成场景、人脸、物体和颜色文本描述，实现跨模态间信息对齐与细节信息提取；设计门控融合交互模块，通过门控机制与交互注意力对特征进行融合交互，以充分利用各个情感特征的优势；最后构建融合图像提示编码的序列信息，传入预训练模型 (pre-trained language model, PLM) 中进行情感极性预测。本文的主要贡献如下：

1) 提出一种结合多面图像特征提取与门控融合机制的 MABSA 模型。提供更为全面的视觉情感特征表达，实现与文本模态的信息对齐。

2) 设计门控融合交互模块。通过门控和交互注意力机制，将图文情感信息进行融合交互，减少无关噪声干扰。构建融合图像提示的序列信息，将视觉特征转换为 PLM 的输入空间，弥补图文模态间表示差异。

3) 相比现阶段 MABSA 模型，所提模型在两个数据集 Twitter-2015 和 Twitter-2017 上有明显的性能提升，证明了各模块设计的合理性。

1　相关工作

1.1　方面级情感分析

方面级情感分析 (aspect-based sentiment analysis, ABSA) 作为一种细粒度情感分类任务，旨在识别和分析文本中特定方面的情感极性，所用方法大致可以分为传统机器学习方法、神经网络模型和基于微调的模型。基于传统机器学习方法主要依赖于手工设计的特征和模型以识别文本中方面词的情感极性，但该方法高度依赖于手工设计[10] 的特征，难以捕捉复杂的语言现象。基于神经网络的方法也取得了不错的效果，Liu 等[11] 提出了一种新型的门控交替神经网络 GANN(gated alternate neural network)，学习信息丰富的方面相关情感线索表示；An 等[12] 设计了一个包含 3 种不同节点的异构图神经网络，在当时获得优异表现；段文杰等[13] 利用图卷积网络得到情感与句法的特征表示，取得不错效果；He 等[14] 设计 CABiSTM(multi-channel convolution, multihead self-attention and bidirectional long short-term memory) 模块，在一定程度上提高了方面级情感分析任务准确率。此外，基于注意力[15-17] 的神经网络也被广泛应用于上下文和方面词之间的相关性建模。基于微调的方法利用在大规模语料上预训练的语言模型[18](如 BERT[19]、RoBERTa、GPT 系列等)，通过 Fine-tuning 适应特定的 ABSA 任务，并取得了显著的性能。Liang 等[20] 根据特定的方面来利用句子的情感依赖，提出了一种基于 SenticNet 的图卷积网络；Lee 等[21] 提出了一种通过将低秩自适应 LoRA 引入生成语言模型的微调方法，以提高这些基于生成的 ABSA 模型的性能并实现高效学习。然而，这些方法仅关注文本单一模态，并没有考虑到来自其他模态 (如图像等) 的信息也会为情感分析任务做出一定的贡献。

1.2　多模态方面级情感分析

多模态方面级情感分析 MABSA 属于细粒度

的情感分析任务, 与仅关注纯文本的 ABSA 任务不同, MABSA 侧重于从文本和图像等不同的模态信息中捕获情感特征, 克服了单一模态分析在信息表达和理解上的局限性。Xu 等[3] 基于方面的多模态情感分析, 提出了一种新的多交互网络 MIMN(multi-interactive memory network) 模型, 实现图像和文本的交互; Yu 等 [22] 引入了一个多模态 BERT 框架, 通过目标注意力机制集成图像和文本来增强情感分析效果; Li 等[23] 提出了多级文本−视觉对齐和融合网络 MTVAF(multi-level textual-visual alignment and fusion), 用动态注意力机制生成视觉提示来控制跨模态融合, 取得较好的结果; 杨颖等[24] 提出了一种多粒度视图动态融合模型, 从粗细粒度两个角度, 对图文数据进行向量化编码, 以充分捕捉数据特征, 提升任务表现; Zhao 等[25] 提出了一种 FGSN 网络 (fusion with gcn and se-resnext network) 的融合算法, 有效增强了文本的句法和依赖解析, 结合了高级图像特征提取; 朱超杰等[26] 通过引入目标检测算法提取了原始图像中目标这一细节信息, 提出一种基于目标注意力的方面级多模态情感分析模型。上述研究通过神经网络、注意力和门控机制等, 在一定程度上提高了 MABSA 任务的性能表现。但现阶段学者们仍在努力解决跨模态间不充分不一致的问题, 其关键在于如何去更好地掌握文本和图像之间的交互与协同作用。当前的模型虽在一定程度上能够融合文本和图像信息[25-26], 但在处理复杂场景时, 仍存在诸多挑战。特别是, 模型通常无法检测和捕捉到图像中微妙的情感差异, 而这些差异对于全面理解和分析情感倾向至关重要。为此, 本文提出的模型通过构建多面图像特征提取模块来有效解决这一问题。

2　模型架构

本文所提模型的整体架构如图 1 所示, 该模型共包含 3 个模块: 1) 多面图像特征提取模块, 生成图片场景、人脸、物体和颜色文本描述, 实现跨模态间信息对齐与细节信息提取; 2) 门控融合交互模块, 通过门控机制与交互注意力对特征进行融合交互, 以充分利用各特征信息的优势; 3) 情感预测模块, 构建融合图像提示编码的序列信息, 传入预训练模型 PLM 中进行最终的情感极性预测。

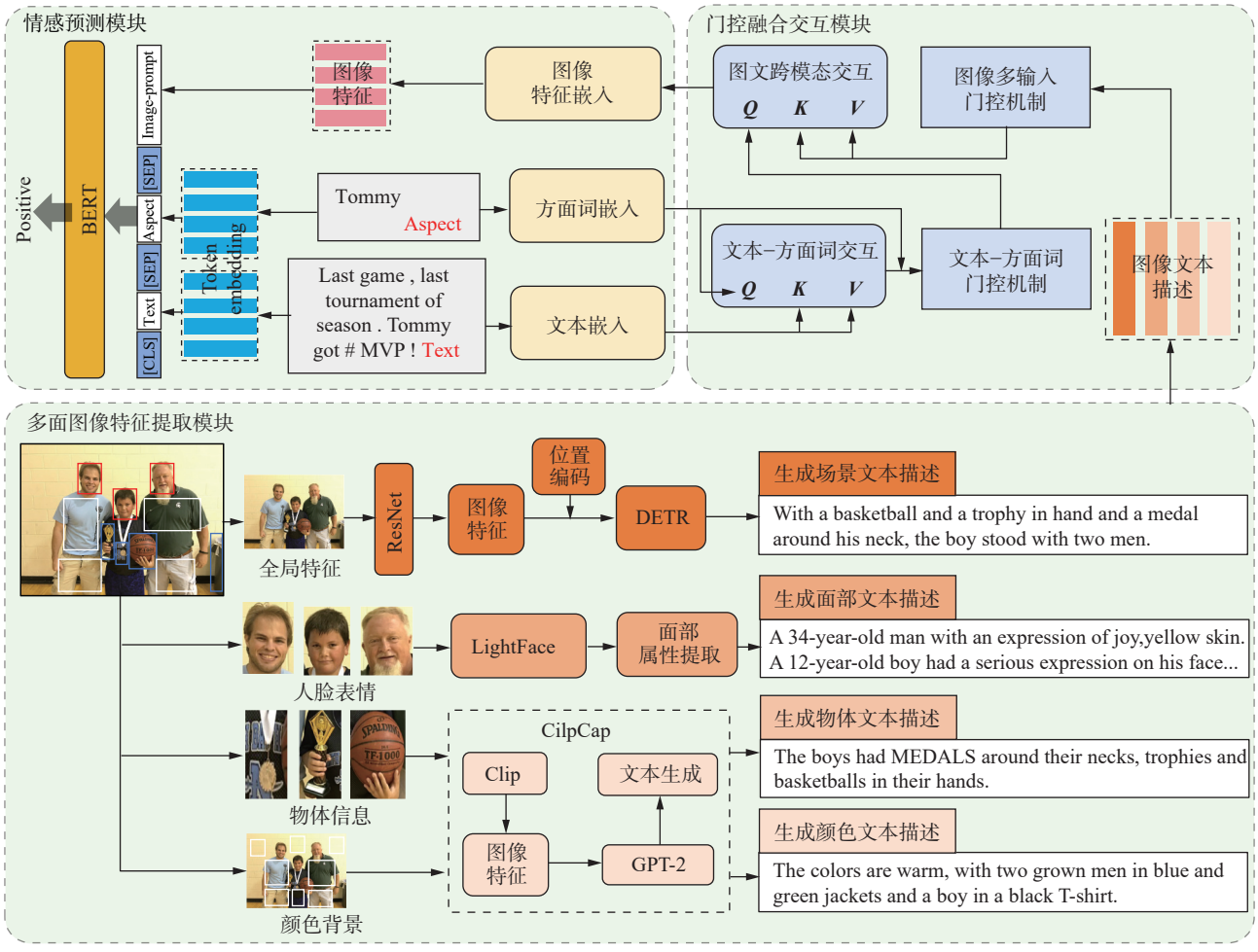


图 1　整体模型结构

Fig. 1　Overall model structure

2.1　任务定义

给定一个多模态方面级数据集，其中的每一个样本包含一个由 n 个单词组成的文本 $\boldsymbol{T}=\{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_n\}$、一幅对应的图像 $\boldsymbol{I}$ 以及一个单词数为 k 的方面词 $\boldsymbol{A}=\{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_k\}$（$\boldsymbol{T}$ 的一个单词子序列），每个方面词 $\boldsymbol{A}$ 都与一个情感标签 $\boldsymbol{y} \in \{\text{positive, negative, neutral}\}$ 相关联。多模态方面级情感分析任务的目标是融合文本 $\boldsymbol{T}$、图像 $\boldsymbol{I}$ 等不同模态的信息，以细粒度地识别和分类特定方面词 $\boldsymbol{A}$ 的情感标签 $\boldsymbol{y}$。

2.2　多面图像特征提取模块

在处理多模态数据推文时，传统方法通常依赖预训练神经网络直接提取图像特征，易忽略视觉情感信息的微妙细节。为此采用跨模态翻译技术，从图像中提取并生成场景、人脸、图中物体和颜色背景的文本描述，以更全面地捕捉视觉中的细节情感信息。

2.2.1　场景描述信息对齐

鉴于图像中的场景往往放映了整体的情感氛围和背景，因此针对全局场景特征信息的提取是任务中不可忽视的一部分。如图 2 所示，为确保场景描述生成的准确性，通过 ResNet101[27] 提取深层全局特征，避免梯度消失；引入正弦−余弦位置编码保留空间位置编码关系；堆叠 DETR(detection Transformer) 并通过注意力建模全局语义、聚焦核心场景，最后结合 BERT 词典概率生成文本描述。

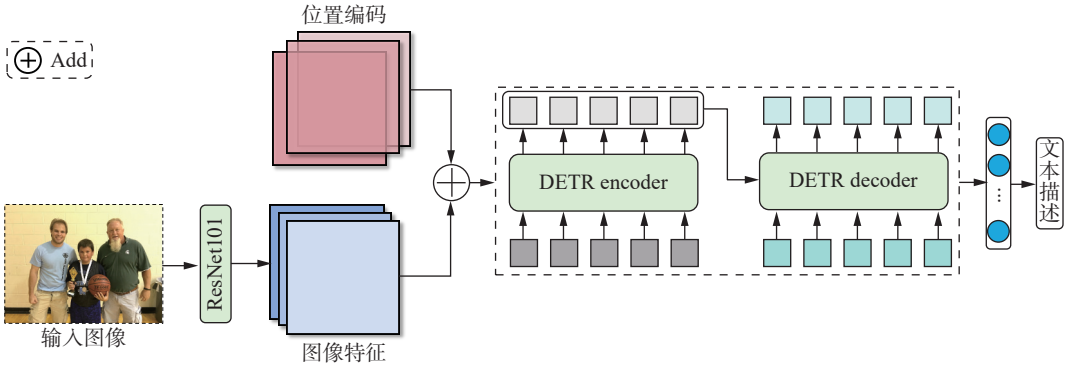


图 2　生成场景描述结构

Fig. 2　Generate the scene description structure

对于输入的图像 $\boldsymbol{I} \in \mathbf{R}^{3\times H\times W}$，采用残差网络 ResNet101 作为图像编码器，相较传统 VGG(visual geometry group) 网络等[28]，ResNet101 在图像识别任务中能够提取更深层次的图像特征信息。具体而言，将输入的图片经过 ResNet 的特征提取之后，得到一个新的特征映射 $\boldsymbol{F}$：

$$\boldsymbol{F} = \text{ResNet}(\boldsymbol{I}) \tag{1}$$

其对应维度为 $K\times H'\times W'$，K 是特征通道数，H'和 W'是特征映射的高度与宽度。

为了更准确地捕获图片全局的结构信息，避免空间信息错乱，增强模型对图像中空间位置关系的理解，在生成的特征映射 $\boldsymbol{F}$ 中增加特定于位置的数据，并相应集成位置编码，减少局部噪声影响。具体来说，针对每个空间位置 $(\boldsymbol{x}, \boldsymbol{y})$，使用正弦和余弦函数对其进行编码。随后，将得到的编码连接起来得出一组向量。因此，每个位置 $(\boldsymbol{x}, \boldsymbol{y})$ 都有了一个对应的位置编码向量 $\boldsymbol{P}(\boldsymbol{x}, \boldsymbol{y})$。随后，将特征映射 $\boldsymbol{F}$ 与位置编码信息 $\boldsymbol{P}$ 进行相加，得到带有位置信息的特征映射 $\boldsymbol{F}'$：

$$\boldsymbol{F}' = \boldsymbol{F} + \boldsymbol{P} \tag{2}$$

为了更好地生成图像的场景描述特征信息，受文献 [29] 启发，采用多个堆叠的基于 Transformer 的 DETR 进行处理，输出特征表示的同时通过注意力过滤无关噪声元素。首先，将处理好的特征映射 $\boldsymbol{F}'$ 转换为一维序列 $\boldsymbol{X}$，再传入 DETR 的编码器和解码器进行处理，生成更丰富的特征表示：

$$\boldsymbol{Z} = \text{DETR\_Encoder}(\boldsymbol{X}) \tag{3}$$

$$\boldsymbol{Y} = \text{DETR\_Decoder}(\boldsymbol{Q}, \boldsymbol{Z}) \tag{4}$$

式中：$\boldsymbol{Q}$ 是可学习的查询向量，$\boldsymbol{Z}$ 是编码器的输出，$\boldsymbol{Y}$ 是解码器的输出。

随后，将输出的特征表示 $\boldsymbol{Y}$ 应用于预测 BERT 词典中的单词概率分布，进而生成全文，对于其中的每个查询点 q，相关的概率预测为

$$\boldsymbol{P}(q) = \text{Softmax}(\boldsymbol{W}_y \boldsymbol{Y}(q) + \boldsymbol{b}_y) \tag{5}$$

式中：$\boldsymbol{W}_y$ 是线性变换矩阵，$\boldsymbol{b}_y$ 是偏置项。

最后，通过概率分布 $\boldsymbol{P}(q)$ 并根据 BERT 词典中选择对应的每个单词，生成一个由概率最高单词组成的序列信息即场景描述 $\boldsymbol{D}_s$：

$$\boldsymbol{D}_s = \text{Scene\_description}(\boldsymbol{P}) \tag{6}$$

式中：Scene_description() 是根据概率 $\boldsymbol{P}(q)$ 生成场景描述序列的函数。

2.2.2　人脸描述信息对齐

在多模态情感分析任务中, 人脸表情作为图像的局部精细信息, 其准确的捕捉与解析至关重要[30], 对整体情感识别的精确度有着决定性影响。人脸属性信息对齐是在全局信息对齐的基础上, 将局部的人脸特征转换到文本空间中, 以获得对齐的面部表情描述。

具体结构如图 3 所示, 针对输入的图像 $\boldsymbol{I}$, 为了更准确地获取面部视觉特征, 首先利用 LightFace[31] 人脸检测器识别出图像中的所有人脸:

$$\boldsymbol{F}_{\text{face}} = \text{LightFace}(\boldsymbol{I}) \tag{7}$$

式中: $\boldsymbol{F}_{\text{face}}$ 为检测到的人脸集合, 其元素为每一张人脸 $\boldsymbol{F}_i$, 同时包括具体的面部区域 $\boldsymbol{f}_i$ 和对应的置信度得分 c_i, 其中 i=1, 2, ⋯, m, m 表示检测到的人脸数量。最终, 检测的人脸集合为 $\boldsymbol{F}_{\text{face}}$={$\boldsymbol{F}_1$, $\boldsymbol{F}_2$, ⋯, $\boldsymbol{F}_m$}。

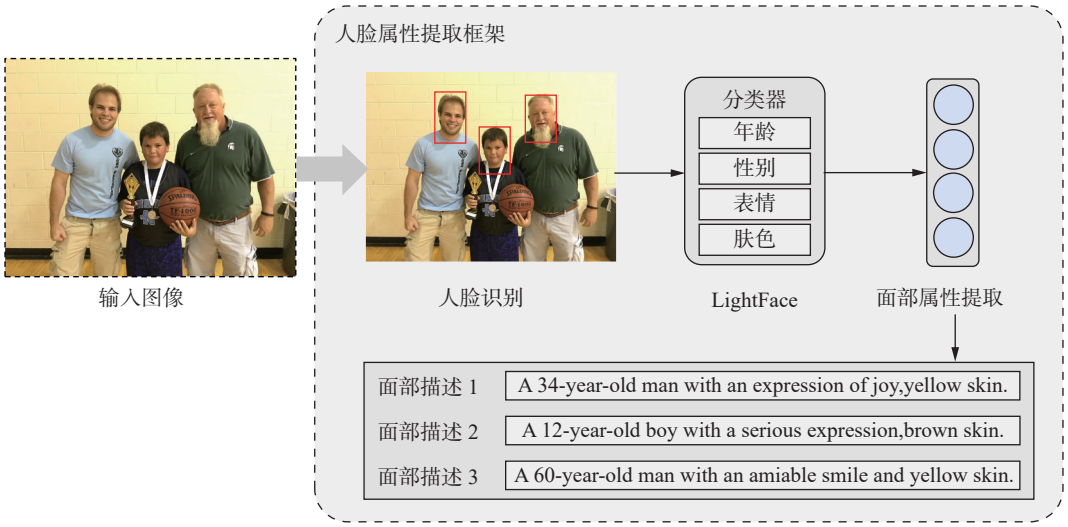


图 3　生成人脸属描述结构

Fig. 3　Generate face genus description structure

对于提取到的面部区域 $\boldsymbol{f}_i$, 提取其面部属性 $\boldsymbol{A}_i$:

$$\boldsymbol{A}_i = \text{Extract\_attributes}(\boldsymbol{f}_i) \tag{8}$$

式中: $\boldsymbol{A}_i$ 是第 i 个人脸的属性集合, 具体包括年龄、性别、表情、肤色等面部特征信息。

随后, 受文献 [6] 启发, 采用其设计提出的人体面部描述模版以生成最终的人脸文本描述 $\boldsymbol{D}_i$:

$$\boldsymbol{D}_i = \text{Face\_Description}(\boldsymbol{A}_i) \tag{9}$$

最终生成的面部描述集合为 $\boldsymbol{D}_{\text{face}}$={$\boldsymbol{D}_1$, $\boldsymbol{D}_2$, ⋯, $\boldsymbol{D}_m$}。

为了更好地关注人脸面部区域的重要相关信息, 根据人脸检测器 LightFace 获取的预测置信度 c_i 进行过滤, 设定置信度阈值 θ=0.5, 过滤掉低于 θ 的面部描述:

$$\boldsymbol{D}'_{\text{face}} = \{\boldsymbol{D}_{\text{face}} | c_i \geqslant \theta\} \tag{10}$$

式中: $\boldsymbol{D}'_{\text{face}}$ 为过滤后保留的面部文本描述集合, 最后根据进行 c_i 降序排序, 得到最终的面部描述集合 $\boldsymbol{D}_f$。

2.2.3　物体描述信息对齐

在情感分析任务中, 图像中的物体检测和描述是关键步骤, 这些物体不仅提供丰富的视觉信息, 还反映场景的情感氛围和背景故事。通过对物体的检测和描述, 可以提取关键的视觉特征, 与文本信息结合, 更全面地理解情感的多维度表达。

对于输入的图像 $\boldsymbol{I}$, 为了更好地实现粒度信息对齐, 本文采用 ClipCap[32] 多模态模型为图像生成高质量且精准的物体信息检测描述 $\boldsymbol{D}_{\text{o}}$:

$$\boldsymbol{D}_{\text{o}} = \text{ClipCap\_Object}(\boldsymbol{I}) \tag{11}$$

式中: $\boldsymbol{D}_{\text{o}}$={$\boldsymbol{D}_1$, $\boldsymbol{D}_2$, ⋯, $\boldsymbol{D}_m$}, m 表示生成物体描述的个数。

2.2.4　物体描述信息对齐

图片的颜色背景承载着丰富的情感信息, 能够直接影响观众的情感感知。通过对背景颜色的精准检测和描述, 能够更深入理解图像所传达的情感倾向, 进而提升情感分析的准确性和全面性。

为了更好地提取图片视觉中的有关色彩背景的特征并生成对应的文本描述, 仍采用 ClipCap 多模态模型为图像生成高质量且精准的色彩背景描述 $\boldsymbol{D}_{\text{c}}$:

$$\boldsymbol{D}_{\text{o}} = \text{ClipCap\_Colour}(\boldsymbol{I}) \tag{12}$$

式中: $\boldsymbol{D}_{\text{c}}$={$\boldsymbol{D}_1$, $\boldsymbol{D}_2$, ⋯, $\boldsymbol{D}_d$}, d 表示生成色彩背景描述的个数。

2.3　门控融合交互模块

2.3.1　图像多输入门控机制

鉴于提取多面视觉特征的多样性，会不可避免地带来相关噪声，进而影响整体的任务性能。为降低噪声带来的负面影响，采用多输入门控机制针对视觉特征进行去噪，如图 4 所示。在本文模型中，针对图像视觉特征的输出，具体包括 4 个方面：场景描述 $\boldsymbol{D}_{\mathrm{s}}$、人脸描述 $\boldsymbol{D}_{\mathrm{f}}$、物体描述 $\boldsymbol{D}_{\mathrm{o}}$ 和颜色描述 $\boldsymbol{D}_{\mathrm{c}}$。首先将 4 个视觉特征输入到一个门控单元，通过学习输入数据的重要性分布，动态地为每个输入分配一个权重分数，该分数决定了每个输入在最终预测任务中的贡献度。

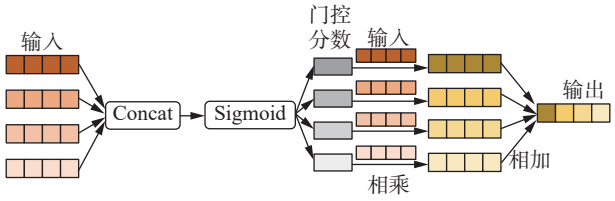


图 4　多输入门控单元结构

Fig. 4　Multi-input gated unit structure

具体而言，首先将 4 个视觉特征进行拼接融合得到整体特征 $\boldsymbol{D}_{\mathrm{total}}$。随后，通过线性变换和 Sigmoid 函数为视觉特征生成门控分数 $\boldsymbol{G}$：

$$\boldsymbol{G}=\sigma(\boldsymbol{W}_{\mathrm{g}}\cdot\boldsymbol{D}_{\mathrm{toal}}+\boldsymbol{b}_{\mathrm{g}}) \tag{13}$$

式中：$\boldsymbol{W}_g$ 是线性层的权重矩阵，$\boldsymbol{b}_g$ 是偏置。

为了使模型能独立地学习每种输入特征对最终输出的重要性，使门控单元可以为每个视觉特征分配不同的权重，以更好利用各特征的综合潜力和作用。为此，本文模型将生成的门控分数 $\boldsymbol{G}$ 在最后一个维度上分成 4 个相等的部分 $\boldsymbol{G}_{\mathrm{s}}$、$\boldsymbol{G}_{\mathrm{f}}$、$\boldsymbol{G}_{\mathrm{o}}$ 和 $\boldsymbol{G}_{\mathrm{c}}$，分别对应每一个视觉特征描述。然后应用门控得分对每种特征进行加权，生成情感增强的视觉描述特征 $\boldsymbol{F}_{\mathrm{total}}$：

$$\boldsymbol{F}_{\mathrm{total}}=\boldsymbol{G}_{\mathrm{s}}\cdot\boldsymbol{D}_{\mathrm{s}}+\boldsymbol{G}_{\mathrm{f}}\cdot\boldsymbol{D}_{\mathrm{f}}+\boldsymbol{G}_{\mathrm{o}}\cdot\boldsymbol{D}_{\mathrm{o}}+\boldsymbol{G}_{\mathrm{c}}\cdot\boldsymbol{D}_{\mathrm{c}} \tag{14}$$

式中：“·”表示逐元素相乘，指两个矩阵或向量中相同位置的元素相乘。

2.3.2　文本−方面词交互

为了更好地实现噪声过滤效果，突出方面词 $\boldsymbol{A}$ 在句子 $\boldsymbol{T}$ 中的重要性。为此，设置交互注意力来学习句子中的上下文与方面词之间的语义权重，最终获得增强后的文本特征：

$$\boldsymbol{E}^{\boldsymbol{A}\to\boldsymbol{T}}=\mathrm{Attention}(\boldsymbol{E}^{\boldsymbol{A}},\boldsymbol{E}^{\boldsymbol{T}/\boldsymbol{A}},\boldsymbol{E}^{\boldsymbol{T}/\boldsymbol{A}}) \tag{15}$$

$$\tilde{\boldsymbol{E}}^{\boldsymbol{A}}=\mathrm{mean}(\boldsymbol{E}^{\boldsymbol{A}},\boldsymbol{E}^{\boldsymbol{A}\to\boldsymbol{T}}) \tag{16}$$

式中：$\boldsymbol{E}^{A}=\{\mathrm{Emb}(\boldsymbol{w}_j)|\boldsymbol{w}_j\in\boldsymbol{A}\}$ 表示方面词 $\boldsymbol{A}$ 的特征嵌入矩阵，$\boldsymbol{E}^{T/A}=\{\mathrm{Emb}(\boldsymbol{w}_j)|\boldsymbol{w}_j\in\boldsymbol{T}\&\boldsymbol{w}_j\notin\boldsymbol{A}\}$ 是句子 $\boldsymbol{T}$ 上下文标记的嵌入矩阵，mean 函数表示元素取均值。

2.3.3　文本−方面词门控机制

为了更好地利用特征之间的优势，针对输出增强后的情感特征进行门控机制处理，通过门控机制动态融合特征，自适应选择更重要的信息，以提升模型的表达能力和性能。

具体地，将输出后的增强文本特征与原始方面词特征矩阵相结合，然后通过线性变换和 Sigmoid 函数，为视觉特征生成门控分数 $\boldsymbol{G}_1$：

$$\boldsymbol{G}_1=\sigma(\boldsymbol{W}_1[\boldsymbol{E}^{\boldsymbol{A}},\tilde{\boldsymbol{E}}^{\boldsymbol{A}}]+\boldsymbol{b}_1) \tag{17}$$

式中：$\boldsymbol{W}_1$ 表示权重矩阵，$\boldsymbol{b}_1$ 表示偏置向量。随后，将门控分数进行动态加权调整，并生成最终过滤后的文本−方面词特征表示 $\boldsymbol{E}_{\mathrm{s}}$：

$$\boldsymbol{E}_{\mathrm{s}}=\boldsymbol{G}_1\cdot\boldsymbol{E}^{\boldsymbol{A}}+(1-\boldsymbol{G}_1)\cdot\tilde{\boldsymbol{E}}^{\boldsymbol{A}} \tag{18}$$

2.3.4　图文跨模态交互

为增强有关文本−方面词和视觉特征之间的情感特征表示，设置交互注意力模块，针对输出的图像特征与文本−方面词信息进行交互，生成最终的图像增强特征表示，将文本−方面词特征表示 $\boldsymbol{E}_{\mathrm{s}}$ 当做 $\boldsymbol{Q}$，视觉描述特征 $\boldsymbol{F}_{\mathrm{total}}$ 当做 $\boldsymbol{K}$ 和 $\boldsymbol{V}$：

$$\boldsymbol{E}^{\boldsymbol{A}\to\boldsymbol{I}}=\mathrm{Attention}(\boldsymbol{E}_{\mathrm{s}},\boldsymbol{F}_{\mathrm{toal}},\boldsymbol{F}_{\mathrm{toal}}) \tag{19}$$

最后，经过输出生成的 $\boldsymbol{E}^{A\to I}$ 即为最终的视觉特征嵌入，并作为最终的图像提示编码输入。

2.4　情感预测模块

考虑到图文模态间存在表示差异，为更好实现模型处理能力，构建融合图像提示的序列信息，将视觉特征转换为 PLM 的输入空间。具体地将图像特征与原始方面词和文本分别嵌入到 CLS 序列中：

$$\boldsymbol{H}^{\mathrm{cls}}=[\mathrm{CLS}]\boldsymbol{E}^{\boldsymbol{T}}[\mathrm{SEP}]\boldsymbol{E}^{\boldsymbol{A}}[\mathrm{SEP}]\boldsymbol{E}^{\boldsymbol{A}\to\boldsymbol{I}} \tag{20}$$

式中：[CLS] 和 [SEP] 分别是用于分类和分离的两个特殊标记，随后将序列 $\boldsymbol{H}^{\mathrm{cls}}$ 输入预训练语言模型中 BERT 中进行处理并通过 Softmax 函数预测出最终的情感极性。

本模型使用最小化交叉熵损失用于优化所提出的方法中的所有参数：

$$\boldsymbol{L}=-\frac{1}{|\boldsymbol{D}|}\sum_{(\boldsymbol{T},\boldsymbol{I},\boldsymbol{A},\boldsymbol{y})\in\boldsymbol{D}}\boldsymbol{y}\cdot\log\boldsymbol{P}(\boldsymbol{y}'|\boldsymbol{T},\boldsymbol{I},\boldsymbol{A}) \tag{21}$$

式中：$\boldsymbol{y}'$表示模型预测出的情感概率分布，$\boldsymbol{y}$ 表示每组数据的真实情感标签，$\boldsymbol{D}$ 表示所有训练样本的集合。

3　实验分析

3.1　实验设置

本文实验选用 Yu 等[22] 提出的基于多模态方面级情感分析的公开数据集 Twitter-2015 和 Twit-

ter-2017 进行实验, 每条数据包含一个句子、一张图片和至少一个方面词以及对应的情感标签, 具体统计信息如表 1 所示 (其中 Pos 表示积极、Neg 表示消极、Neu 表示中性)。

表 1　Twitter 数据集具体信息

Table 1　Twitter datasets specific information

类型	Twitter-2015			Twitter-2017		
	训练集	验证集	测试集	训练集	验证集	测试集
积极	928	303	317	1 508	515	493
消极	368	149	113	416	144	168
中性	1 883	670	607	1 638	517	573
总计	3 179	1 122	1 037	3 562	1 176	1 234

实验中采用预训练的 BERTweet[33] 和 ResNet-101 来初始化所提模型中的 PLM 和图像特征提取器。在交替优化过程中采用 AdamW 作为学习器对参数进行优化。对于模型的超参数设置, 将 Batch-size 设置为 16, 训练 Epoch 设置为 4, 学习率设置为 $1\times e^{-4}$, 交互注意力头数为 8, 随机失活率 dropout 设置为 0.1, 最大文本长度为 128。选用模型 10 次独立训练的最佳结果作为最终的实验结果, 所有实验均基于 PyTorch 以及 NVIDIA RTX 4090 GPU 完成。为了方便与对比模型保持一致, 采用准确率 (accuracy, Acc)、F_1 值 (Macro-F_1) 作为评价指标。

3.2　对比模型

为验证模型在情感分析任务方面的可行性, 选用经典的单模态和具有代表性的多模态模型与所提模型进行对比分析。

1)ResNet-Aspect[34]: 分别使用 ResNet 和 BERT 提取图像和方面词的特征, 采用线性层进行情感分类。

2)AE-LSTM(attention-based long short-term memory)[35]: 通过基于注意力机制的长短期记忆网络关注句子中的不同部分。

3)RAM(recurrent attention network on memory)[36]: 采用多注意力机制捕捉文本中间隔较远的情感特征, 使其对不相关信息具有更强鲁棒性。

4)MGAN(multi-grained attention network)[37]: 利用细粒度和粗粒度的注意力机制来捕捉方面词和上下文之间的交互。

5)BERT[18]: 使用 Transformer 来获取文本与句子之间的交互信息。

6)MIAN(multi-interactive memory network)[3]: 采用两个交互式记忆网络来监督和融合文本、方面词和视觉特征。

7)ESAFN(entity-sensitive attention and fusion network)[38]: 提出一个实体敏感注意力和融合网络来捕获方面词、文本和图像的动态信息。

8)TomBERT: 提出一种多模态 BERT 架构的多模态方面级情感分析模型。

9)CapBERT[5]: 引入双流模型, 将视觉特征转换为文本字幕与文本方面词进行交互处理。

10)KEF-TomBERT(knowledge-enhanced framework)[39]: 在改进多模态 BERT 架构的基础上引入新型知识增强框架 KEF。

11)ITM(image-target matching network)[40]: 提出一种新的粗粒度到细粒度的图像−目标匹配网络。

12)FGSN(fusion with gcn and se-resnext network)[25]: 基于 GCN 与 SE-ResNeXt 网络的融合算法进行方面级情感分析任务。

13)TIGFM(text-image gated gusion mechanism)[41]: 提出一种基于文本和图像门控融合机制的多模态方面级情感分析模型。

14)TISRI(text-image semantic relevance identification)[42]: 提出文本图像语义相关性识别模型来处理情感分析任务。

15)MSPAF(multiscale semantic perception and attention fusion model)[43]: 通过多尺度语义感知和注意力融合实现情感分类。

16)REF(relevance-aware visual entity filter network)[44]: 提出一种面向多模态方面级情感分析的关联感知视觉实体过滤网络。

3.3　实验结果与分析

表 2 给出了本文模型和对比模型在 Twitter-2015 和 Twitter-2017 数据集上的实验结果。从表可见, 本文模型在两个数据集上均取得了较对比模型的最佳结果, 在 Twitter-2015 数据集上的 Acc 提高 0.39%, 在 Twitter-2017 数据集上的 Acc 和 F_1 分别提高 1.46% 和 1.28%, 虽在 Twitter-2015 数据集上 F_1 略低于对比模型的最佳结果, 但整体综合性能仍表现出色。

表 2　实验性能对比

Table 2　Comparison of experimental performance　%

模态	模型	Twitter-2015		Twitter-2017	
		Acc	F_1	Acc	F_1
图像	ResNet-Aspect	59.49	47.79	57.86	53.98
文本	AE-LSTM	70.30	63.43	61.67	57.97
	RAM	70.68	63.05	64.42	61.01
	MGAN	71.17	64.21	64.75	61.46
	BERT	74.25	70.04	68.88	66.12

续表 2

模态	模型	Twitter-2015		Twitter-2017	
		Acc	F_1	Acc	F_1
图像+文本	MIAN	71.84	65.69	65.88	62.99
	ESAFN	73.38	67.37	67.83	64.22
	TomBERT	77.15	71.75	70.34	68.03
	CapBERT	78.01	73.25	69.77	68.42
	KEF-TomBERT	78.68	73.75	72.12	69.96
	ITM	78.27	74.19	72.61	71.97
	TIGFM	78.66	73.89	72.12	70.58
	TISRI	78.50	74.42	72.53	71.40
	MSPAF	78.30	71.75	70.34	69.17
	REF	78.69	75.15	71.88	70.95
	本文模型	**79.08**	**74.39**	**74.07**	**73.25**

注:“_”表示对比模型的最佳结果,加粗表示本文模型实验结果。

具体地,ResNet-Aspect 模型性能低于所有模型结果,这表明视觉特征无法作为独立的模态去主导模型的情感预测;以文本主导的 AE-LSTM、RAM 等模型取得了一定效果,但结合图像和文本的模型整体效果更好,这表明图像信息能够与文本相互补充,进而增强整体任务的性能表现。

MIAN 和 ESAFN 作为最早一批处理 MABSA 任务的模型,由于视觉特征的不充分提取和处理反而干扰整体模型的性能表现。相较而言,TomBERT、KEF-TomBERT 针对任务设计多模态 BERT 架构并引入知识增强网络,在当时取得较好的结果;CapBERT 模型创造性地将视觉特征转换为文本表示,实现了更好的视觉−文本模态对齐,但仍存在不同模态间的融合交互不充分问题,而本文模型通过设计门控融合机制和交互注意力机制相结合,有效地利用了各模态特征间的协同作用与综合潜力。

近几年研究中,ITM、TIGFM、TISRI 等模型在 Twitter-2017 数据集上获得了较大提升,但在视觉特征提取过程中,仍忽略了一些细粒度信息;MSPAF、REF 模型通过多尺度感知、视觉实体过滤等方法,取得了一定效果,但在复杂社交媒体数据中则略显不足。本文模型从图像的多面特征出发,整合场景、人脸、物体和颜色信息,并转换为单模态文本描述,实现了视觉与文本之间的信息对齐;在不同模态信息交互中,结合门控机制和交互注意力进行处理,从而有效地融合和交互情感信息,最终取得的实验结果也较好地证明该模型设计的有效性。

3.4　消融实验

为进一步研究所提模型各部分设计的合理性及其优点,开展消融实验。在 Twitter-2015 和 Twitter-2017 数据集上将所提模型进行分解,分别移除视觉特征中的场景描述、人脸描述、物体描述、颜色描述、图像多输入门控机制、文本−方面词门控机制和交互注意力,以此验证被移除部分对总体模型的有效增益情况,具体实验结果如表 3 所示 (w/o 表示 without)。

表 3　消融实验性能结果

Table 3　Results of ablation experiment　%

模型	Twitter-2015		Twitter-2017	
	Acc	F_1	Acc	F_1
w/o 场景描述	75.37	70.70	71.92	70.23
w/o 人脸描述	77.82	72.52	72.02	70.45
w/o 物体描述	76.51	71.94	72.12	70.89
w/o 颜色描述	77.63	72.40	72.85	70.91
w/o 图像多输入门控机制	78.66	72.81	73.01	71.14
w/o 文本−方面词门控机制	78.54	73.19	73.24	71.32
w/o 交互注意力机制	78.84	73.85	73.36	72.29
本文模型	**79.08**	**74.39**	**74.07**	**73.25**

注:加粗表示最佳结果。

综上所述,在视觉特征方面,移除场景描述后,模型在两个数据集上的 Acc 和 F_1 都呈现大幅下降,可见场景描述对于模型把握整体任务的情感情绪具有重要作用。同样,当去除物体描述时,模型性能出现约 2% 的下降,这表明视觉特征中存在的物体信息在一定程度上能够为方面词和文本提供相应的情感指导。当移除人脸和颜色描述时,仅通过场景描述和物体描述进行视觉特征的表达也能达到不错的效果,但在 Twitter 数据集中,往往存在人体面部表情和相关颜色背景等细粒度信息,而这些信息能够显著提升模型在情感分析任务中的表现,特别是在捕捉用户情感的细微变化和复杂情感时,这些信息起到了关键作用。

在特征融合交互中,由于输出视觉特征的多样性,盲目使用反而会在跨模态间交互中引入不必要噪声,故有必要采用多输入门控机制针对视觉特征进行初步的筛选和过滤;同样文本−方面词门控机制的嵌入,能够进一步挖掘文本句子和方面词之间的协同作用和关键特征,使模型性能提升约 0.5%;最后,当去除注意力机制时,仅单纯通过门控机制也取得了不错的效果,但注意力机制的引入,能够在最终的结果中放大关键信息的重要性,实现更全面的情感特征挖掘和处理。

3.5　参数设置

3.5.1　Batch size 取值

表 4 给出了在 Twitter-2015 和 Twitter-2017 数

据集上, 不同 Batch size 对模型性能的影响。在具体的实验中, 分别设置了 8、16 和 32 共 3 个取值进行分析。

表 4　Batch size 取值对模型性能的影响

Table 4　Impact of Batch size on model performance　%

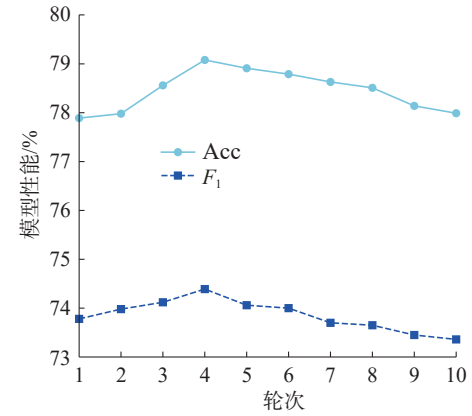
Batch size取值	Twitter-2015		Twitter-2017	
	Acc	F_1	Acc	F_1
8	77.92	73.65	73.34	72.50
16	**79.08**	**74.39**	**74.07**	**73.25**
32	78.47	73.94	73.82	73.01

注: 加粗表示最佳结果。

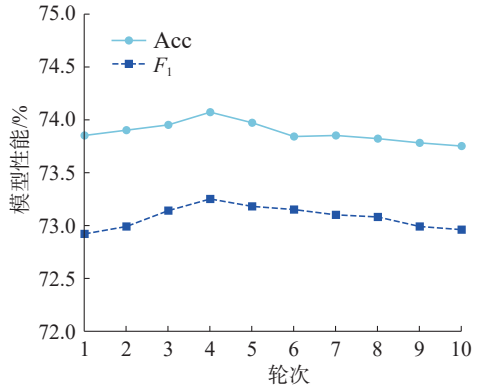
综上所述, 当 Batch size 设置为 16 时, 模型在两个数据集上均取得最佳性能; 当取值为 8 时, 相较两个数据集的样本数量, Batch size 取值偏小易导致模型训练时间过长且难以收敛, 进而引发欠拟合的问题。在一定范围内增加 Batch size 有助于提高模型收敛的稳定性, 当取值为 32 时, 模型容易因为 Batch size 过大而陷入局部极小值, 导致泛化性能下降。因此, 将模型的 Batch size 取值设置为 16。

3.5.2　Epoch 取值

对于本文模型来说, 模型的 Epoch 在很大程度上影响了模型的整体性能, 如图 5 所示。



(a) Twitter-2015



(b) Twitter-2017

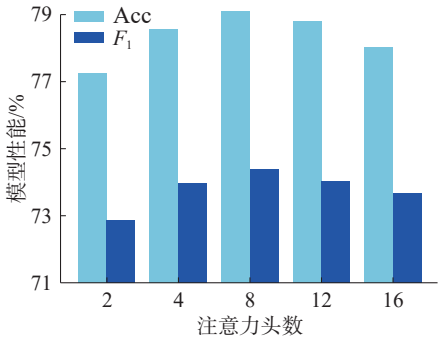
图 5　Epoch 对模型的影响

Fig. 5　Effect of Epoch on the model

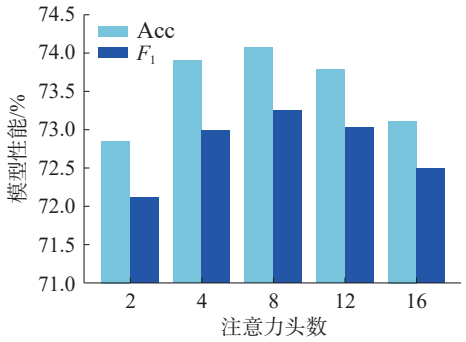
从图 5 可见, 当轮次从 1 不断增加时, 模型性能也在不断上升, 当达到第 4 轮时性能最佳; 若继续增加轮次, 模型的 Acc 和 F_1 值都呈现明显的下降趋势并逐渐趋近于稳定状态, 没有太大变化。这表明, 当轮次较少时, 随着轮次的不断增加, 多模态信息之间的融合也会愈加充分, 进而整体模型性能也越来越好; 反之, 当轮次超过一定的数量时, 模型计算资源的上升, 过于复杂易出现过拟合的现象, 最终影响情感分类的结果。因此, 本文模型在实验过程中轮次确定为 4。

3.5.3　交互注意力头数取值

交互注意力在针对模态内和图文跨模态间特征交互中起着重要作用, 其取值在一定程度上决定了模型处理不同模态数据交互信息的能力和效率。故在具体实验中选取 2、4、8、12、16 进行验证, 得到注意力头数在数据集上的性能表现, 如图 6 所示。



(a) Twitter-2015



(b) Twitter-2017

图 6　注意力头数对模型的影响

Fig. 6　Effect of the number of attention heads on the model

从图 6 可知, 当注意力头数为 8 时, 均取得最好的准确率与 F_1。当取值小于 8 时, 较少的注意力头数对特征信息的捕捉能力存在不足, 在一定范围内增加取值有助于提高模型的准确率。当大于 8 时, 模型可能会过度捕捉训练数据中存在的噪声和细微特征, 进而对情感分类带来干扰。因

此，将注意力头数取值设置为 8。

3.6　案例分析

为了更直观地展示所提模型的有效性，本节选用数据集中 3 个具有代表性的示例并分别与 BERT、CapBERT 模型进行对比分析，示例信息和预测结果如表 5 所示。为更好地解释本文模型，表 5 还特别给出 3 个样本中分别生成的场景描述、人脸描述、物体描述和颜色描述。

表 5　案例分析

Table 5　Case study

类别	示例1	示例2	示例3
图像			
文本	RT @ BadgerMBB: [**Bo**] and [**Butch Ryan**] never missed a Final Four. Touching gesture from Coach [**Williams**] @ UNC Basketball	RT @ RembranceOfPast: [**Lou Reed**], [**Mick Jagger**] and [**David Bowie**] hanging out together at Caf Royale, 1973	RT @ BeschlossDC: [**Coretta Scott King**] with [**Robert**] amp [**Ethel Kennedy**] after husband's assassination, which occurred tonight 1968: # Globe http:
标签	(Bo, Postive) (Butch Ryan, Postive) (Williams, Postive)	(Lou Reed, Neutral) (Mick Jagger, Neutral) (David Bowie, Neutral)	(Coretta Scott King, Neutral) (Robert, Neutral) (Ethel Kennedy, Negative)
场景描述	A man stands on the basketball court with a ticket in his hand.	Three men drinking wine together, each with a different expression.	Two ladies and a man were talking solemnly at the bedroom bedside.
人脸描述	A white-haired 60-year-old man with a happy, friendly, warm smile and white skin	A 20-year-old boy stares into the distance with a color expression, an 18-year-old boy purses his mouth and looks down, and a 22-year-old boy looks surprised	A 35-year-old woman with a yellowish complexion and a serious expression, a 30-year-old man with a serious expression, and a 37-year-old woman with a yellowish expression
物体描述	Man in red jacket, holding a farmed item, large screen, empty seats, basketball court.	Three men, one with a drink, wine and glasses in front of him.	Women sitting on beds, men and blonde women sitting on chairs, murals on the walls, lamps, curtains
颜色描述	The image has a warm, vibrant color tone with reds, oranges, the old man wore a red coat.	The image has a black and white color tone, two men were dressed in white, the other in a black shirt	The color of the picture is yellow, a warm and warm feeling, the women wore long black dresses and long blue dresses, the men wore black suits,
BERT	Neutral×, Neutral×, Positive√	Positive×, Neutral√, Negative×	Negative×, Negative×, Negative√
CapBERT	Neutral×, Neutral×, Positive√	Positive×, Neutral√, Neutral√	Negative×, Neutral√, Negative√
本文模型	Positive√, Positive√, Positive√	Neutral√, Neutral√, Neutral√	Neutral√, Neutral√, Negative√

如表 5 所示，基于文本单模态的 BERT 模型产生了部分准确预测，但由于信息匮乏导致整体任务性能并不理想。CapBERT 通过将图像转换为文本字幕可以提供有关视觉特征信息的补充，但针对方面词呈现中性的情感预测还存在不足，生成的图像描述多数与图像无关，易产生噪声，产生一些预测误差等。

本文模型能够结合图像的场景描述、人脸描述、物体描述、颜色描述等多种信息来丰富视觉特征的输入，并结合门控和注意力机制进行筛选和过滤。如示例 1 所示，生成图像的场景描述、人脸描述都很好地提供了人物积极的情感特征，但文本句子中存在 3 个方面词，而图片中仅存在一个人物，针对其他方面词极性判断带来了一定的挑战性，虽然句子中的相关词语能够带来一些积极信息，但图片的信息干扰，易产生一些噪声，而通过生成物体描述和颜色描述等信息，从侧面补充了积极的情感色彩，进而预测出整体的情感极性。在 MABSA 任务中，针对中性情感的预测是最具挑战性的，如示例 2 所示，3 个方面词标签均为中性，无论是句子还是生成的图像场景描述、物体描述等，都展示一种积极的氛围，这无疑给预测带来了难度，但通过模型生成的人脸描述和颜色描述补充偏中性的氛围，结合门控融合机制，对一些无关特征进行筛选和过滤，最终实现了案例的准确预测。同时如示例 3 所示，本文模

型针对消极情感也取得较好的预测结果。通过3个示例进一步证明了本文模型在全面提取视觉信息方面的有效性, 并验证了门控融合交互模块设计的合理性, 帮助模型降低了无关噪声干扰, 探索了不同信息之间的联系, 最终做出正确的情感预测。

4 结束语

本文提出了一种结合多面图像特征提取和门控融合机制的多模态方面级情感分析模型, 采用跨模态翻译技术生成视觉模态的场景、人脸、物体和颜色的文本描述, 实现图文信息有效对齐; 设计门控融合交互模块, 通过门控和交互注意力机制实现特征之间的融合交互, 能够有效降低无关特征带来的噪声干扰; 考虑到不同模态间的表示差异, 构建融合图像提示的序列信息, 将提取的多面视觉特征与文本−方面词特征相互补充协同, 实现更准确的情感预测。在公开数据集上该模型较现有模型也取得较好的分类效果, 验证了该模型的有效性与合理性。

由于真实数据集中潜在的讽刺现象易被忽略, 限制了复杂场景下的分类效果。下一步工作考虑将讽刺检测机制集成到模型中, 以更全面地捕捉数据中的细节情感信息。

参考文献:

[1] WANG Hongbin, REN Chun, YU Zhengtao. Multimodal sentiment analysis based on multiple attention[J]. Engineering applications of artificial intelligence, 2025, 140: 109731.

[2] 曾子明, 孙守强, 李青青. 基于融合策略的突发公共卫生事件网络舆情多模态负面情感识别[J]. 情报学报, 2023, 42(5): 611−622.
ZENG Ziming, SUN Shouqiang, LI Qingqing. Multimodal negative sentiment recognition in online public opinion during public health emergencies based on fusion strategy[J]. Journal of the China society for scientific and technical information, 2023, 42(5): 611−622.

[3] XU Nan, MAO Wenji, CHEN Guandan. Multi-interactive memory network for aspect based multimodal sentiment analysis[C]//Proceedings of the AAAI conference on artificial intelligence. Honolulu: AAAI Press, 2019: 371−378.

[4] WANG Jiawei, LIU Zhe, SHENG V, et al. SaliencyBERT: recurrent attention network for target-oriented multimodal sentiment classification[M]//Pattern Recognition and Computer Vision. Cham: Springer International Publishing, 2021: 3−15.

[5] KHAN Z, FU Yun. Exploiting BERT for multimodal target sentiment classification through input space translation[C]//Proceedings of the 29th ACM International Conference on Multimedia. [S.l.]: ACM, 2021: 3034−3042.

[6] YANG Hao, ZHAO Yanyan, QIN Bing. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: USAACL, 2022: 3324−3335.

[7] 曾碧卿, 姚勇涛, 谢梁琦, 等. 结合局部感知与多层次注意力的多模态方面级情感分析[J]. 计算机工程, 2025, 51(9): 80−90.
ZENG Biqing, Yao Yongtao, Xie Liangqi, et al. Multimodal aspect level emotion Analysis combining local perception and multi-levelattention[J]. Computer engineering, 2025, 51(9): 80−90.

[8] 余本功, 陈明玥. 基于细粒度图像−方面的情感增强方面级情感分析[J]. 计算机应用研究, 2025, 42(4): 1073−1079.
YU Bengong, CHEN Mingyue. Aspect-oriented affective knowledge enhanced for aspect-based sentiment analysis [J]. Application research of computers, 2025, 42(4): 1073−1079.

[9] 曹银妮, 韩虎, 黄明伟, 等. 基于多视角融合表示的多模态方面级情感分析模型[J/OL]. 数据分析与知识发现, [2024−01−01]. https://link.cnki.net/urlid/10.1478.G2.20250305.1130.002.
CAO Yinni, HAN Hu, HUANG Mingwei, et al. Multimodal aspect level sentiment analysis model based on multi-perspective fusion representation [J/OL]. Data analysis and knowledge discovery, [2024−01−01]. https://link.cnki.net/urlid/10.1478.G2.20250305.1130.002.

[10] ÀLVAREZ-LÓPEZ T, JUNCAL-MARTÍNEZ J, FERNÁNDEZ-GAVILANES M, et al. GTI at SemEval-2016 task 5: SVM and CRF for aspect detection and unsupervised aspect-based sentiment analysis[C]//Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Dieg: USAACL, 2016: 306−311.

[11] LIU Ning, SHEN Bo. Aspect-based sentiment analysis with gated alternate neural network[J]. Knowledge-based systems, 2020, 188: 105010.

[12] AN Wenbin, TIAN Feng, CHEN Ping, et al. Aspect-based sentiment analysis with heterogeneous graph neural network[J]. IEEE transactions on computational social systems, 2023, 10(1): 403−412.

[13] 段文杰, 邓金科, 张顺香, 等. 基于多层次知识增强的方

面级情感分析模型[J]. 智能系统学报, 2024, 19(5): 1287−1297.
DUAN Wenjie, DENG Jinke, ZHANG Shunxiang, et al. Aspect-based sentiment analysis model based on multi-level knowledge enhancement[J]. CAAI transactions on intelligent systems, 2024, 19(5): 1287−1297.

[14] HE Bo, ZHAO Ruoyu, TANG Dali. CABiLSTM-BERT: Aspect-based sentiment analysis model based on deep implicit feature extraction[J]. Knowledge-based systems, 2025, 309: 112782.

[15] LI Xin, BING Lidong, WAI Lam, et al. Transformation networks for target-oriented sentiment classification[C]// Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018: 946−956.

[16] ZHAO Chuanjun, FENG Rong, SUN Xuzhuang, et al. Enhancing aspect-based sentiment analysis with BERT-driven context generation and quality filtering[J]. Natural language processing journal, 2024, 7: 100077.

[17] ALI KANDHRO I, ALI F, UDDIN M, et al. Exploring aspect-based sentiment analysis: an in-depth review of current methods and prospects for advancement[J]. Knowledge and information systems, 2024, 66(7): 3639−3669.

[18] 陈燕, 赖宇斌, 肖澳, 等. 基于 CLIP 和交叉注意力的多模态情感分析模型[J]. 郑州大学学报 (工学版), 2024, 45(2): 42−50.
CHEN Yan, LAI Yubin, XIAO Ao, et al. Multimodal sentiment analysis model based on CLIP and cross-attention [J]. Journal of Zhengzhou University (engineering science), 2024, 45(2): 42−50.

[19] 张铭泉, 周辉, 曹锦纲. 基于注意力机制的双 BERT 有向情感文本分类研究[J]. 智能系统学报, 2022, 17(6): 1220−1227.
ZHANG Mingquan, ZHOU Hui, CAO Jingang. Dual BERT directed sentiment text classification based on attention mechanism[J]. CAAI transactions on intelligent systems, 2022, 17(6): 1220−1227.

[20] LIANG Bin, SU Hang, GUI Lin, et al. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks[J]. Knowledge-based systems, 2022, 235: 107643.

[21] LEE C, LEE Hanyong, KIM K, et al. An efficient fine-tuning of generative language model for aspect-based sentiment analysis[C]//2024 IEEE International Conference on Consumer Electronics. Las Vegas: IEEE, 2024: 1−4.

[22] YU Jianfei, JIANG Jing. Adapting BERT for target-oriented multimodal sentiment classification[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao: International Joint Conferences on Artificial Intelligence Organization, 2019: 5408–5414.

[23] LI You, DING Han, LIN Yuming, et al. Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis[J]. Artificial intelligence review, 2024, 57(4): 78.

[24] 杨颖, 钱馨雨, 王合宁. 结合多粒度视图动态融合的多模态方面级情感分析[J]. 计算机工程与应用, 2024, 60(22): 172−183.
YANG Ying, QIAN Xinyu, WANG Hening. Multimodal aspect-level sentiment analysis based on multi-granularity view dynamic fusion[J]. Computer engineering and applications, 2024, 60(22): 172−183.

[25] ZHAO Jun, YANG Fuping. Fusion with GCN and SE-ResNeXt network for aspect based multimodal sentiment analysis[C]//2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference. Chongqing: IEEE, 2023: 336−340.

[26] 朱超杰, 闫昱名, 初宝昌, 等. 采用目标注意力的方面级多模态情感分析研究[J]. 智能系统学报, 2024, 19(6): 1562−1572.
ZHU Chaojie, YAN Yuming, CHU Baochang, et al. Research on aspect-based multimodal sentiment Analysis using Target Attention[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1562−1572.

[27] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[28] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014−11−18)[2022−01−01]. https://arxiv.org/abs/1409.1556.

[29] XIAO Luwei, WU Xingjiao, YANG Shuwen, et al. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis[J]. Information processing & management, 2023, 60(6): 103508.

[30] FAN Shaojing, SHEN Zhiqi, JIANG Ming, et al. Emotional attention: a study of image sentiment and visual attention[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7521−7531.

[31] SERENGIL S I, OZPINAR A. HyperExtended LightFace: a facial attribute analysis framework[C]//2021 International Conference on Engineering and Emerging Technologies. Istanbul: IEEE, 2021: 1−4.

[32] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervi-

sion[C]//Proceedings of Machine Learning Research. Virtual: PMLR, 2021: 8748−8763.

[33] NGUYEN D Q, VU T, NGUYEN A T. Bertweet: a pre-trained language model for English Tweets[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online: EMNLP, 2020: 9−14.

[34] WANG Qianlong, XU Hongling, WEN Zhiyuan, et al. Image-to-text conversion and aspect-oriented filtration for multimodal aspect-based sentiment analysis[J]. IEEE transactions on affective computing, 2023, 15(3): 1264−1278.

[35] WANG Yequan, HUANG Minlie, ZHU Xiaoyan, et al. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the 2016 Conference on Empirical Methods in NaturalLanguage Processing. Austin: USAACL, 2016: 606−615.

[36] CHEN Peng, SUN Zhongqian, BING Lidong, et al. Recurrent attention network on memory for aspect sentiment analysis[C]//Proceedings of the 2017 Conference on Empirical Methods in NaturalLanguage Processing. Copenhagen: USAACL, 2017: 452−461.

[37] FAN Feifan, FENG Yansong, ZHAO Dongyan. Multi-grained attention network for aspect-level sentiment classification[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: USAACL, 2018: 3433−3442.

[38] YU Jianfei, JIANG Jing, XIA Rui. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification[J]. IEEE/ACM transactions on audio, speech, and language processing, 2019, 28: 429−439.

[39] ZHAO Fei, WU Zhen, LONG Siyu, et al. Learning from adjective-noun pairs: a knowledge-enhanced framework for target-oriented multimodal sentiment classification[C]//Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju: International Committee on Computational Linguistics 2022: 6784−6794.

[40] YU Jjianfei, WANG Jieming, XIA Rui, et al. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. Vienna: IJCAI, 2022: 4482−4488.

[41] 张添植, 周刚, 刘洪波, 等. 基于文本和图像门控融合机制的多模态方面级情感分析[J]. 计算机科学, 2024, 51(9): 242−249.
ZHANG Tianzhi, ZHOU Gang, LIU Hongbo, et al. Text-image gated fusion mechanism for multimodal aspect-based sentiment analysis[J]. Computer science, 2024, 51(9): 242−249.

[42] ZHANG Tianzhi, ZHOU Gang, LU Jicang, et al. Text-image semantic relevance identification for aspect-based multimodal sentiment analysis[J]. PeerJ computer science, 2024, 10: e1904.

[43] 杨丽莎, 马常霞, 仲兆满, 等. 多尺度语义感知和注意力融合的多模态方面级情感分析模型[J]. 南京大学学报(自然科学), 2025, 61(2): 223−236.
YANG Lisha, MA Changxia, ZHONG Zhaoman, et al. Multiscale semantic perception and attention fusion for multimodal aspect-level sentiment analysis model[J]. Journal of Nanjing University (natural science), 2025, 61(2): 223−236.

[44] CHEN Yifan, XIONG Haoliang, LI Kuntao, et al. Relevance-aware visual entity filter network for multimodal aspect-based sentiment analysis[J]. International journal of machine learning and cybernetics, 2025, 16(2): 1389−1402.

作者简介:

赵雪峰, 副教授, 博士, 江苏省科技副总, 主要研究方向为多模态情感分析、数字图像处理与无损检测。作为主要成员主持、参与完成省市级项目 4 项, 发表学术论文 20 余篇。E-mail: zhaoxf@jou.edu.cn。

狄恒西, 硕士研究生, 主要研究方向为多模态情感分析、自然语言处理。E-mail: dihx@jou.edu.cn。

柏长泽, 硕士研究生, 主要研究方向为多模态情感分析、自然语言处理。E-mail: 2023220901@jou.edu。