

# 基于 GCN 和目标视觉特征增强的多模态方面级情感分析

赵雪峰, 柏长泽, 狄恒西, 仲兆满, 仲晓敏

(江苏海洋大学计算机工程学院, 连云港 222005)

**摘要:** 多模态方面级情感分析旨在整合图文模态数据, 以精准预测方面词的情感极性。现有方法在精确定位文本相关的图像区域特征及有效处理模态间信息交互方面仍存在显著局限, 同时模态内的上下文信息理解存在偏差, 导致产生额外的噪声。为了解决上述问题, 本文提出一种基于图卷积神经网络和目标视觉特征增强(Graph convolutional network and target visual feature enhancement, GCN-TVFE)的多模态方面级情感分析模型。首先, 本文采用 CLIP(Contrastive language-image pre-training)模型处理文本、方面词和图像数据, 通过计算文本与图像之间的相似度以及方面词与图像之间的相似度, 并结合这两项相似度, 实现对文本与图像、方面词与图像匹配程度的量化评估。再通过 Faster R-CNN 模型去快速且精确地识别并定位图像中的目标区域, 进一步增强模型提取与文本相关的图像特征能力。其次, 通过图文 GCN 网络, 利用文本之间的依存句法关系构建文本图结构, 同时借助 K 近邻(K-nearest neighbor, KNN)算法生成图像图结构, 从而深入挖掘模态内的特征信息。最后, 采用多模态交互注意力机制, 有效捕捉方面词与文本之间、目标视觉特征与图像生成文本描述特征之间的关联信息, 显著减少噪声干扰, 增强模态间的特征交互。实验结果表明, 本文提出的模型在公共数据集 Twitter 2015 和 Twitter 2017 上的综合性能优越, 验证了该模型在多模态情感分析领域的有效性。

**关键词:** 多模态方面级情感分析; 目标视觉特征; 依存句法关系; KNN 算法; 多模态交互注意力机制  
**中图分类号:** TP391 **文献标志码:** A

## Multimodal Aspect-Level Sentiment Analysis Based on GCN and Target Visual Feature Enhancement

ZHAO Xuefeng, BAI Changze, DI Hengxi, ZHONG Zhaoman, ZHONG Xiaomin

(School of Computer Engineering, Jiangsu Ocean University, Lianyungang 222005, China)

**Abstract:** Multimodal aspect-level sentiment analysis aims to integrate graphic modal data to accurately predict the emotional polarity of aspect words. However, the existing methods still have significant limitations in accurately locating text-related image region features and effectively processing the information interaction between modalities. At the same time, the understanding of context information within modalities is biased, which leads to additional noise. In order to solve the above problems, a multimodal aspect-level sentiment analysis model based on graph convolutional network and target visual feature enhancement (GCN-TVFE) is proposed. First of all, this paper uses the contrastive language-image pre-

**基金项目:** 国家自然科学基金(72174079); 江苏省“青蓝工程”优秀教学团队项目(2022-29)。

**收稿日期:** 2024-12-21; **修订日期:** 2025-02-17

training (CLIP) model to process text, aspect words, and image data. By calculating the similarity between text and image and the similarity between aspect words and image, and then combining these two similarities, the quantitative evaluation of the matching degree between text and image and the matching degree of aspect words and image is realized. Then, the Faster R-CNN model is used to quickly and accurately identify and locate the target region in the image, which further enhances the ability of the model to extract image features related to text. Secondly, through the GCN network, the text graph structure is constructed by using the dependency syntactic relationship between texts, and the image graph structure is generated by the K-nearest neighbor (KNN) algorithm, to dig the feature information in the mode deeply. Finally, the multi-layer and multi-modal interactive attention mechanism is used to effectively capture the correlation information between aspect words and text, and between target visual features and image-generated text description features, which significantly reduces noise interference and enhances feature interaction between modes. Experimental results show that the model proposed in this paper has superior comprehensive performance on the public datasets Twitter-2015 and Twitter-2017, which verifies the effectiveness of the model in the field of multimodal sentiment analysis.

**Key words:** multimodal aspect-level sentiment analysis; target visual features; dependency syntactic relation; KNN algorithm; multimodal interactive attention mechanism

## 引言

随着社交媒体与电子商务平台的盛行,用户在发布文本评论时往往会选择附加一些相关的图片,这使得多模态数据在情感分析中的应用日益重要<sup>[1]</sup>。研究者通过深入挖掘多模态数据中的情感信息,并从中分析引发情感分歧的具体事件原因,不仅可以帮助企业 and 组织更精准地理解用户需求和反馈,还能够让其在潜在危机爆发前及时采取预防措施,有效遏制舆情带来的负面影响。

近年来,多模态方面级情感分析作为情感分析领域中的一项重要且细致的任务,已经成为学术界广泛关注的热点研究领域,吸引了众多研究者的积极参与。它的核心目标在于综合分析图像与文本数据,精准预测文本中具体方面词的情感极性。Xu等<sup>[2]</sup>在早期的研究中提出了一种新的交互记忆网络模型,通过两个交互记忆网络监督文本和视觉信息,主要利用多个门控循环单元和注意力机制对文本、方面词和图像进行交互处理,以预测方面词的情感极性。Khan等<sup>[3]</sup>提出了EF-CaTrBERT模型,通过基于转换器的翻译将图像转换为文本表示,并将辅助句输入集成到BERT编码器中,采用多模态进行融合。尽管上述方法在情感分析任务中取得了一定的成效,但由于文本中的方面词与情感因子之间的关联性不够直接,并且多模态信息内容极其丰富且复杂,其中包含大量与情感无关的噪声信息,这使得模型在判断方面词的情感极性时面临显著挑战,难以保持清晰和准确。例如,在图1(a)所示的案例中,文本中的方面词“Harry Gulliver”在缺乏明确情感指示的情况下,情感极性无法通过文本内容直接判定。同时,图像模态中的信息复杂且多样化,难以直接定位出与“Harry Gulliver”相关的特征,这进一步增加了情

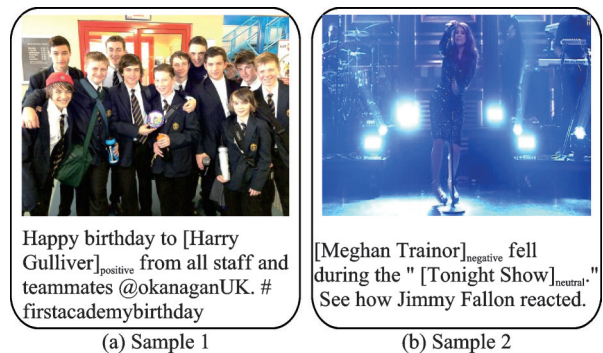


图1 数据集中的示例样本

Fig.1 Example samples in the dataset

感极性判断的难度。另外,在图1(b)中,方面词“Meghan Trainor”和“Tonight Show”表达的情感极性不同,并且与方面词相关的图像中这些信息的位置区域并未体现出直接的关联。另外,文本中未提供具体的情感信息,这会导致模型在分析时未能充分理解模态内的上下文信息,也忽略了多模态间细粒度相关联的内在联系。

针对上述问题,本文提出了一种基于图卷积神经网络和目标视觉特征增强(Graph convolutional network and target visual feature enhancement,GCN-TVFE)的多模态方面级情感分析模型。首先,为了深入挖掘图像与文本模态内部的深层关联特征信息,本文利用依存句法关系来确定文本中词与词之间的关联性,并通过GCN捕捉与方面词相关的特征信息。在图像处理方面,图像会先分割成相同大小的图块,并采用K近邻(K-nearest neighbor, KNN)算法挖掘图块之间的关联性。随后,通过GCN进一步提取细粒度的视觉特征。其有效地整合了图像与文本模态内部的特征信息,提升方面级情感分析的精确度和深度。其次,为了精准定位与方面词相关的图像区域,并降低无关噪声信息对模型判断情感的影响,本文设计了一种由方面词和文本引导的目标视觉特征提取模块,该模块主要通过CLIP(Contrastive language-image pre-training)模型将文本、方面词与图像进行联合建模,映射到一个共享的嵌入空间。在此过程中,分别提取各个模态的特征信息,并通过计算余弦相似度来量化它们之间的关联性,再利用Faster R-CNN进一步匹配出与方面词相关的最佳图像区域。最后,为了捕捉各个模态间的关联特征,本文首先使用BLIP(Bootstrapping language-image pre-training)模型提取图像文本描述的特征信息作为查询,然后将上述与方面词相关的最佳图像区域特征信息作为键和值,传入多头交叉注意力机制中,以提取与方面词相关联的图像特征信息。同时,将方面词特征作为查询,文本特征作为键和值,通过多头交叉注意力机制进一步提取与方面词相关联的文本特征信息,增强了模型对关联特征的捕捉能力。在两个基准数据集上的实验结果显示,该模型通过其总体结构建模,显著提升了多模态方面级情感分析任务的准确性。

本文主要的贡献归纳如下:(1)针对各模态内特征信息挖掘不充分及关联性不足的问题,通过结合句法依存关系和KNN算法,分别强化模态内部的信息关联,随后利用图文GCN进一步挖掘各模态内的深层关联特征信息,以提升特征的丰富性和关联性;(2)提出由方面词和文本引导的目标视觉特征提取模块,定位方面词相关的最佳图像区域;减少噪声信息对模型的判断,并进一步使用多头交叉注意力机制捕捉各模态间的关联信息;(3)在两个公开的基准数据集上进行验证,结果表明了本文模型在方面级情感分析任务中的有效性。

## 1 相关工作

### 1.1 方面级情感分析

方面级情感分析主要依赖文本信息来判断方面词情感倾向。近年来,随着深度学习技术的发展,注意力机制、卷积神经网络(Convolutional neural network, CNN)和长短期记忆网络(Long short-term memory, LSTM)等方法逐渐成为方面级情感分析任务的主流方法。因此,Wang等<sup>[4]</sup>通过引入注意力机制的长短期记忆网络,深入挖掘方面词与其观点表达之间的内在联系。Nguyen等<sup>[5]</sup>提出的交互式注意机制模型通过LSTM在词级上获取目标及其上下文的隐藏状态,利用注意力机制交互学习目标和上下文的注意,进一步增强方面词和观点之间的关联关系。Fan等<sup>[6]</sup>提出了多粒度注意力网络,通过整合细粒度和粗粒度注意机制,并引入方面对齐损失函数以捕捉方面级交互信息。

尽管基于LSTM和注意力机制的方法在利用上下文结构有效提取方面词与观点词的相关性方面表现优异,然而面对句子中可能存在的多个情感极性各异的方面词,这些方法在细致区分每个方面与

相关观点词之间的具体联系上,仍凸显其局限性。因此,研究者们将研究焦点逐渐转向利用句子的句法结构。Aziz等<sup>[7]</sup>提出一种基于CoreNLP依赖句法分析的数据处理技术模型,用于识别方面词与情感词之间的关键联系,从而有效减少了语法在特定方面情感识别中噪声影响。Gao等<sup>[8]</sup>提出双通道相对位置引导注意网络,该网络综合了目标意见的语义和句法表示,以实现情感动态融合和预测。Liu等<sup>[9]</sup>引入一种基于动态情感知识和静态外部知识图的新型图增强网络,以深入挖掘方面词与其相关情感词之间的关系。谢珺等<sup>[10]</sup>提出一种基于知识增强的双Transformer网络方面级情感分析模型,利用情感常识知识库SenticNet7中情感得分改进句法依赖图并考虑对多种句法依赖关系类型分类降噪,然后使用双Transformer网络增强处理长距离词的性能。尽管这些基于依存树的模型通过学习文本的语法依赖关系,但忽视了文本中词语间依赖关系的细微差别及其不同的重要性。此外,上述方法仅聚焦于文本模态数据的研究,未能充分挖掘用户评论中图像模态数据的情感特征。

## 1.2 多模态方面级情感分析

社交媒体完善的评论功能已成为用户各类平台上交流观点与见解的核心载体。这些评论内容不仅涵盖了丰富的文本信息,还包含了多样化的图像信息。因此,如何在多模态数据融合的基础上精准判断方面词的情感极性,已成为多模态方面级情感分析领域的关键挑战之一。

近年来,Yu等<sup>[11]</sup>提出一种实体敏感的注意力融合网络,通过门控机制消除视觉上下文中的噪声,并利用双线性交互融合文本和视觉表示,从而有效捕捉模态间的动态关系。Zhou等<sup>[12]</sup>设计一种多模态交互模型,该模型通过3种交互机制学习各模态之间的关系,并采用对抗性策略将文本和图像特征对齐到一个公共空间,从而显著提升模型的目标情感预测性能。Song等<sup>[13]</sup>采用文本中的目标导向主题,构建多头注意网络以学习文本、视觉和主题信息之间的多模态交互关系。尽管上述方法有效解决了模态间的交互问题,但忽略了图像中方面词的语义信息,并未能充分弥合文本与图像表示之间的语义差距。因此,Yu等<sup>[14]</sup>提出一种通用的分层交互多模态变压器模型,该模型通过目标检测方法从图像中提取目标实体相关的显著特征,并基于自编码器设计了一个辅助重构模块,用以融合各模态间的语义信息。Wang等<sup>[15]</sup>提出一种双视角融合网络,该网络构建图结构来深入探索文本和图像中的细粒度信息,从而挖掘并充分利用与方面词密切相关的细粒度多模态信息。Yang等<sup>[16]</sup>设计了一种新的多粒度自蒸馏融合网络,用以精细提取多粒度语义信息。该网络通过结合相似度计算动态过滤图像中的潜在噪声,并引入自蒸馏机制,实现硬标签与软标签知识的转换,从而提升多模态表示的质量,进而实现更为精准的情感分析。Wan等<sup>[17]</sup>提出了一种知识增强异构图卷积网络,通过使用动态知识选择算法,高效获取最相关的外部知识,从而显著提升对文本中隐含情感表达的理解能力。

## 2 多模态方面级情感分析模型

本文提出的GCN-TVFE模型架构如图2所示。整体结构包括图文特征提取、目标视觉特征提取、图文图神经网络、多模态交互注意力、多模态特征融合以及情感预测模块。本节首先明确定义多模态方面级情感分析的研究任务,然后对GCN-TVFE模型中各核心模块进行详细阐述。

### 2.1 多模态方面级情感分析任务定义

本任务基于给定的1组图文多模态数据集 $D$ ,其中每个样本 $d \in D$ 包含1个文本评论 $T = \{t_1, t_2, \dots, t_n\}$ 、1幅图像 $I$ 以及1个方面序列 $A = \{a_1, a_2, \dots, a_r\}$ 。其中,方面序列 $A$ 是文本评论 $T$ 的子序列。任务的核心目标是,综合运用图文信息 $(T, I)$ ,针对方面序列 $A$ 进行的情感极性 $y \in Y$ 的预测。具体来说, $n$ 代表文本评论的长度, $r$ 则是方面词的长度, $Y$ 共有积极、中性和消极3种情感类别。

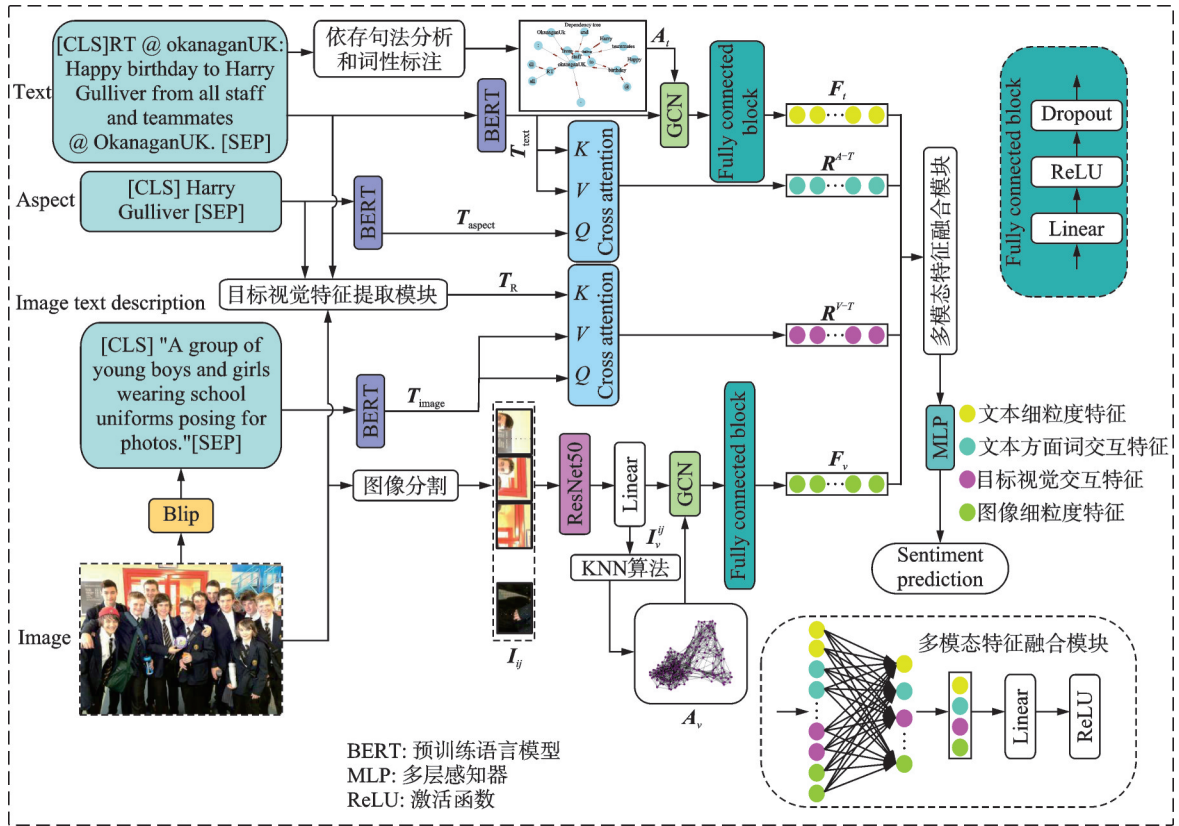


图2 GCN-TVFE模型的整体结构

Fig.2 Overall structure of GCN-TVFE model

### 2.2 图文特征提取模块

本文采用了BERT预训练语言模型<sup>[18]</sup>作为文本 $T$ 和方面词 $A$ 的特征提取编码器。BERT通过双向Transformer架构,能够同时捕获单词的前后上下文信息,从而在理解词义、处理多义词和上下文依赖问题时展现出更强的语境理解能力。由于BERT生成的词嵌入是上下文相关的,即同一个词在不同语境中会有不同的表示,这使模型能够更加精细地捕捉文本中的语义差异。这种特性使得BERT在特征提取上具有显著的优势,能够更准确地表达文本的深层含义。具体的计算过程如下

$$T_{\text{text}} = \text{BERT}(T) \tag{1}$$

$$T_{\text{aspect}} = \text{BERT}(A) \tag{2}$$

式中: $T_{\text{aspect}} \in \mathbb{R}^{q \times d}$ ,  $T_{\text{text}} \in \mathbb{R}^{s \times d}$ ,  $s$ 和 $q$ 分别表示编码后的文本和方面词序列的长度,  $d$ 表示每个单词向量的隐藏维度。

ResNet50<sup>[19]</sup>作为图像 $I$ 的预训练模型,其深度网络结构能捕捉图像中的丰富抽象特征。通过残差学习机制,残差块有效地解决了深层网络中的梯度消失问题,提高了训练的稳定性和有效性。在图像预处理阶段,首先将输入图像分割成 $14 \times 14$ 个大小的图像块 $I_{ij}$ ,并调整至 $224 \text{像素} \times 224 \text{像素}$ 的标准尺寸,以满足模型输入要求。随后,图像被转换为张量形式,并进行图像归一化和标准化处理,确保图像数据与预训练网络高度一致。具体的计算过程如下

$$N_h = \left\lceil \frac{H}{14} \right\rceil, N_w = \left\lceil \frac{W}{14} \right\rceil \tag{3}$$

$$I_{ij} = I \lfloor i \times N_h : (i+1) \times N_h, j \times N_w : (j+1) \times N_w \rfloor \quad (4)$$

$$I_{\text{image}}^{ij} = \text{ResNet50}(I_{ij}) \quad (5)$$

式中： $\lfloor \cdot \rfloor$ 表示向下取整运算； $I_{\text{image}}^{ij} \in \mathbf{R}^{196 \times 2048}$ ，196代表图像经过ResNet50处理后输出的特征图的空间维度大小（即 $14 \times 14$ ），而2048则表示该特征图所拥有的卷积神经网络输出通道数； $H$ 和 $W$ 分别为原图像的高度和宽度； $N_h, N_w$ 表示图像块的像素； $I_{ij}$ 表示分割的图像块。

为了确保图文模态在同一维度空间中的对齐，定义一个线性变换层，将图像模态的特征维度调整至与文本模态相同。通过这种维度对齐，模型能够更有效地学习不同模态数据之间的内在关系，从而显著提升多模态任务的整体性能。具体过程如下

$$I_v^{ij} = W_0 I_{\text{image}}^{ij} + b_0 \quad (6)$$

式中： $I_v^{ij} \in \mathbf{R}^{196 \times d}$ ； $W_0$ 和 $b_0$ 为线性变换层的可学习参数。

### 2.3 图文图神经网络模块

#### 2.3.1 文本图卷积网络层

图3所示结构图可以有效捕捉句子中单词之间的长距离依赖关系，并深入理解句子的语法结构。首先对输入文本 $T$ 进行分词。使用NLTK工具将句子分解为独立的单词，加载预训练的Biaffine依存句法分析器来解析句子结构。Biaffine模型通过双向LSTM和双仿射变换来建模词语之间的依赖关系，能够捕捉到复杂的句法结构，显著提高解析的准确性。将各个单词之间的依存关系生成图的邻接矩阵 $A_l \in \mathbf{R}^{s \times s}$ ，这些依存关系为GCN提供了丰富的语义信息。最后，将邻接矩阵 $A_l$ 和节点特征矩阵 $T_{\text{text}}$ 作为图的边和节点特征，输入到GCN中，以清晰地展示单词之间的连接和关联。具体的计算公式如下

$$H_i^{l+1} = \text{ReLU} \left( \tilde{D}^{-\frac{1}{2}} A_l \tilde{D}^{-\frac{1}{2}} H_i^l W^l \right) \quad (7)$$

$$F_i = H_i^{l+1} \quad (8)$$

式中： $H_i^l$ 表示第 $l$ 层的GCN的输入特征； $H_i^{l+1}$ 表示第 $l+1$ 层GCN的输出特征； $\tilde{D} = \sum_j A_{ij}$ 为邻接矩阵 $A_l$ 的度矩阵； $W^l$ 表示可训练参数； $F_i$ 表示文本细粒度特征向量。

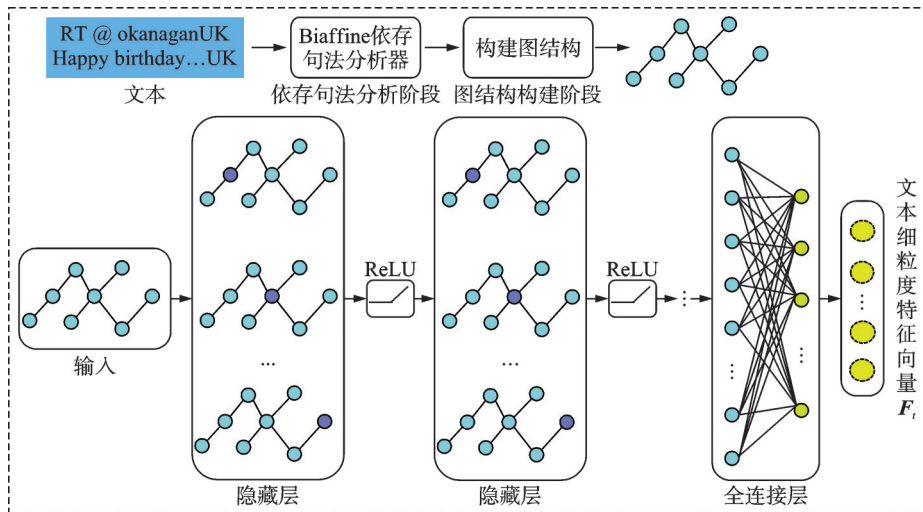


图3 文本细粒度特征提取结构图

Fig.3 Text fine-grained feature extraction structure

### 2.3.2 图像图卷积网络层

如图4所示,为了捕捉图像块特征向量之间的局部结构和依赖关系,本文采用KNN算法为每个图像块 $I_v^{ij}$ 特征向量寻找 $K$ 个最近邻居,并记录这些邻居的索引。通过基于距离的搜索,KNN算法能够有效捕捉特征向量之间的局部相似性和关联性,从而为GCN提供高质量的输入特征信息。这种基于邻居的构建方式使得GCN能够更好地捕捉特征向量之间的语义关系,增强其表征能力。基于这些邻居信息构建图的邻接矩阵 $A_v \in \mathbf{R}^{m \times m}$ ,将邻接矩阵 $A_v$ 和图像块 $I_v^{ij}$ 特征向量作为图的边和节点特征,输入到GCN中,具体的计算公式如下

$$H_v^{s+1} = \text{ReLU} \left( \tilde{D}^{-\frac{1}{2}} A_v \tilde{D}^{-\frac{1}{2}} H_v^s W^s \right) \quad (9)$$

$$F_v = H_v^{s+1} \quad (10)$$

式中: $H_v^s$ 表示第 $s$ 层的GCN的输入特征, $H_v^{s+1}$ 表示第 $s+1$ 层GCN的输出特征, $\tilde{D} = \sum_j A_{vij}$ 为邻接矩阵 $A_v$ 的度矩阵, $W^s$ 表示可训练参数, $F_v$ 表示图像细粒度特征向量。最近邻居数量 $K$ 默认设置为8。

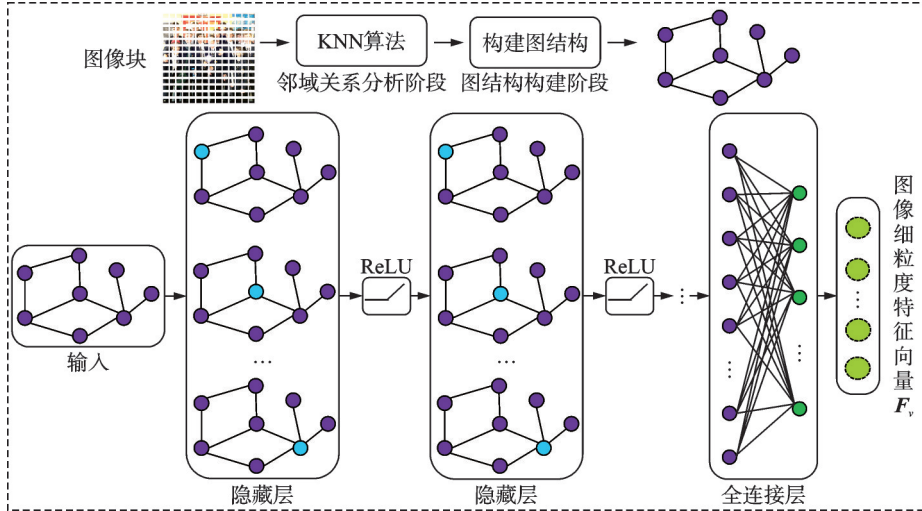


图4 图像细粒度特征提取结构图

Fig.4 Image fine-grained feature extraction structure

## 2.4 目标视觉特征提取模块

为了深入理解图像和文本之间所蕴含的深层语义关系,本文设计了一种目标视觉特征提取模块,如图5所示。该模块采用了CLIP模型<sup>[20]</sup>同时处理文本和图像信息,从而有效捕捉二者之间的语义联系。CLIP是OpenAI开发的一种多模态学习模型,旨在通过对比学习的方法,将文本和图像映射到同一语义空间。通过训练,使得图像和相应文本之间的表示尽可能接近,而与不相关文本或图像则表示保持较大的距离。然后,通过计算相似度,能够量化地评估文本与图像之间的匹配程度。这一过程不仅有助于识别图像中的主题内容,还能深入分析图像中的相关特征,进而提取出支持文本意图的关键信息。具体的计算公式如下

$$F_{\text{image}}, F_{\text{text}}, F_{\text{aspect}} = \text{CLIPModel}(I, T, A) \quad (11)$$

$$\text{sim}_{\text{Text}, I} = \frac{F_{\text{text}} \cdot F_{\text{image}}}{\|F_{\text{text}}\| \cdot \|F_{\text{image}}\|} \quad (12)$$

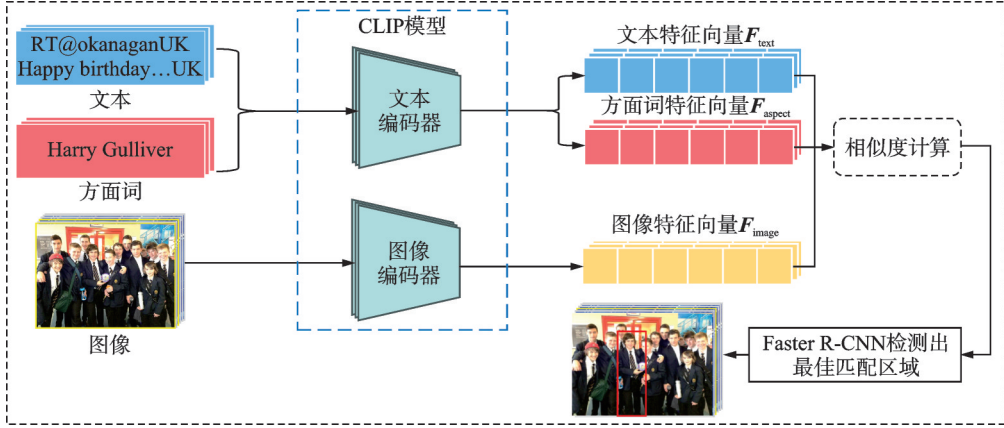


图5 目标视觉特征提取模块

Fig.5 Extraction module for target visual feature

$$\text{sim}_{\text{Aspect},I} = \frac{F_{\text{Aspect}} \cdot F_{\text{image}}}{\|F_{\text{Aspect}}\| \cdot \|F_{\text{image}}\|} \quad (13)$$

$$\text{sim}_{\text{combined}} = \frac{(1 - \mu)\text{sim}_{\text{Text},I} + \mu\text{sim}_{\text{Aspect},I}}{2} \quad (14)$$

式中： $F_{\text{text}}$ 、 $F_{\text{aspect}}$ 、 $F_{\text{image}}$  分别表示文本、方面词、图像的特征向量； $\text{sim}_{\text{Text},I}$ 、 $\text{sim}_{\text{Aspect},I}$  分别表示文本和图像、方面词和图像的余弦相似度； $\text{sim}_{\text{combined}}$  表示文本和方面词与图像的综合余弦相似度，其值域在  $[-1, 1]$ ，值越接近 1 表示向量越相似； $\mu$  表示为相似度权重参数比例值。

为了提升模块的整体性能，本文还结合了 Faster R-CNN 模型<sup>[29]</sup>来生成候选区域。Faster R-CNN 是一种高效的目标检测框架，基于区域提议网络的工作原理，对图像中的潜在目标进行快速识别和定位。通过将 Faster R-CNN 与 CLIP 结合，能够在图像中精准地生成候选区域，确保只关注与文本描述相关的部分。这种整合不仅提高了目标检测的精确度，也使得后续的语义分析与特征提取更加高效。具体的计算公式如下

$$D = \text{Faster R-CNN}(I) \quad (15)$$

$$F_R = \underset{R_i \in D}{\text{argmax}} \text{sim}_{\text{combined}} \quad (16)$$

式中： $D$  包含物体的边界框和对应的置信分数； $F_R$  为最佳目标视觉特征向量信息。

## 2.5 多模态交互注意力模块

为了精确提取与方面词高度相关的特征信息并有效减少噪声干扰，本文引入了多头交叉注意力机制，将方面词  $T_{\text{aspect}}$  设为查询，文本  $T_{\text{text}}$  则作为键和值，从而实现了对文本信息的精细化筛选。该机制通过多个注意力头并行计算，从多维度全面捕捉输入序列中的特征，不仅显著减少了冗余信息的干扰，还极大提升了对关键信息的识别能力。通过这种处理方式，模型能够动态地调整注意力分布，使得与方面词高度相关的文本信息得到更多关注，而与其无关的内容则被有效抑制。具体的计算公式如下

$$Q_i = T_{\text{aspect}} W_i^Q, K_i = T_{\text{text}} W_i^K, V_i = T_{\text{text}} W_i^V \quad (17)$$

$$\text{Attention}_i(Q_i, K_i, V_i) = \text{soft max} \left( \frac{Q_i K_i^T}{\sqrt{d_k}} \right) V_i \quad (18)$$

$$\text{head}_i = \text{Attention}_i(Q_i, K_i, V_i) \quad (19)$$



$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) \mathbf{W}^O \quad (20)$$

式中:  $\mathbf{W}_i^Q$ 、 $\mathbf{W}_i^K$ 、 $\mathbf{W}_i^V$  为权重矩阵;  $h$  表示注意力头的数量。令  $\mathbf{R}^{A-T} = \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$  表示由方面词引导的文本特征信息, 其中  $\mathbf{R}^{A-T} \in \mathbf{R}^{g \times d}$ 。

为突出图像核心信息, 确保模型聚焦于关键视觉内容, 本文采用最佳目标视觉的特征作为查询, 并利用 BLIP 模型<sup>[22]</sup>生成文本描述,  $T_{\text{image}}$  作为键和值, 得到目标视觉特征引导的文本特征信息  $\mathbf{R}^{V-T} \in \mathbf{R}^{196 \times d}$ 。该方法有效增强了图像与文本的语义关联, 确保生成描述的精准性和相关性, 并且借助多头交叉注意力机制, 模型得以深度融合跨模态信息, 多维度捕捉图像与文本的丰富上下文, 增强了模型的鲁棒性和泛化能力, 使其在复杂场景下表现更为卓越。

## 2.6 多模态特征融合模块

本文通过拼接多个模块输出的特征向量, 实现了对不同模态信息的全面整合, 有效降低了模型的时间复杂度和计算负担, 进一步提升了模型的性能。具体的计算公式如下

$$\mathbf{E}_{\text{ATV}} = \text{concat}(\mathbf{F}_i; \mathbf{F}_v; \mathbf{R}^{A-T}; \mathbf{R}^{V-T}) \quad (21)$$

$$\mathbf{Z}_{\text{ATV}} = \text{ReLU}(\mathbf{W}_{\text{ATV}} \mathbf{E}_{\text{ATV}} + \mathbf{b}_{\text{ATV}}) \quad (22)$$

式中  $\mathbf{W}_{\text{ATV}}$ 、 $\mathbf{b}_{\text{ATV}}$  为可训练权重参数。

## 2.7 情感预测模块

将特征向量  $\mathbf{Z}_{\text{ATV}}$  通过 MLP 再进行深层次的处理与整理, 以进一步捕捉特征间的复杂关系并提升模型的表达能力。最后使用 softmax 得到一个情感极性的预测概率分布值, 并通过标准交叉熵损失函数加上  $L_2$  正则项作为损失函数对模型进行标准梯度下降训练, 表达式为

$$y = \text{softmax}(\mathbf{W}_s \text{MLP}(\mathbf{Z}_{\text{ATV}}) + \mathbf{b}_s) \quad (23)$$

$$\text{loss} = - \sum_i^l y_i \log \hat{y}_i + \lambda \|\theta\|_2 \quad (24)$$

式中: softmax 为激活函数;  $\mathbf{W}_s$  和  $\mathbf{b}_s$  为可训练权重矩阵;  $i$  表示数据集中的样本;  $l$  表示包含所有样本的集合;  $y_i$  为样本标签真实值;  $\hat{y}_i$  为样本的预测标签值;  $\lambda$  为正则化系数;  $\theta$  为所有可训练参数。

# 3 实验分析

## 3.1 数据集

为了验证本文模型的有效性, 本文采用 Yu 等<sup>[11]</sup>提出的基于多模态方面级的数据集 Twitter 2015 和 Twitter 2017 进行实验。这两个数据集分别收集了 2014—2015 年及 2016—2017 年期间 Twitter 平台用户发布的推文。数据集包含图像、文本和方面词, 每个样本中包含 1 个或多个方面词, 每个方面词均标注有明确的情感标签, 情感标签分别为消极、中性和积极。数据集按照 3:1:1 的比例划分为训练集、验证集和测试集。表 1 展示了这两个数据集的详细信息统计。

## 3.2 实验设置及评估指标

实验基于深度学习框架 PyTorch 1.2.0 实现, 使用 Python 3.9 作为开发编程语言, 在 AutoDL 算力云平台上进行训练和测试。实验环境配置包括: NVIDIA RTX 4090 显卡 (24 GB 显存), 16 核 Intel(R) Xe-

表 1 Twitter 2015 和 Twitter 2017 数据集统计  
Table 1 Statistics of Twitter 2015 and Twitter 2017 datasets

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

on(R) Platinum 8352V 处理器(主频 2.10 GHz), 120 GB 内存, 以及 Cuda 版本为 11.8。为获取模型参数的最优组合, 采用 Adam 优化器<sup>[23]</sup>对模型参数进行调整更新, 经多次实验确定了最佳参数设置。具体实验参数如表 2 所示。

为了衡量多模态方面级情感分析任务的模型性能, 实验采用准确率(Acc)、Macro- $F_1(F_1)$ 、精确率( $P$ )和召回率( $R$ )作为模型的最终评估指标。具体公式如下

$$P = \frac{TP}{TP + FP} \quad (22)$$

$$R = \frac{TP}{TP + FN} \quad (23)$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (24)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (25)$$

式中: TP 表示实际为正类且被模型预测为正类的样本数; TN 表示实际为负类且被模型预测为负类的样本数; FP 表示实际为负类但被模型预测为正类的样本数; FN 表示实际为正类但被模型预测为负类的样本数。

### 3.3 对比实验

#### 3.3.1 基线模型

为了验证模型的性能, 本文模型在 Twitter 2015 和 Twitter 2017 数据集上与以下具有代表性的多模态方面级情感分析基线模型进行了比较。

(1) EF-Net<sup>[24]</sup>: 该模型采用多头注意力网络和 ResNet-152 残差网络分别处理文本和图像, 通过整合多头注意力和胶囊网络, 旨在捕捉多模态输入之间的相互作用。

(2) JML<sup>[25]</sup>: 该模型通过创建一种辅助跨模态关系检测的多模态联合学习方法, 采用层次化框架连接 MATE 和 MASC, 并对每个子模块分别进行视觉引导, 联合提取特定方面词的情感极性。

(3) SMP<sup>[26]</sup>: 该模型设计了一个基于视觉和文本信息交互的跨模态对比学习模块, 并引入了额外的情感感知预训练目标, 从情感丰富的数据集中捕获细粒度的情感信息; 通过掩码语义建模器和掩码自动编码器, 从文本和图像中提取语义信息, 从而实现更精确的情感分析和跨模态理解。

(4) VLP-MABSA<sup>[27]</sup>: 该模型采用了一种新颖的多头交叉注意力机制来捕捉文本和视觉特征, 并通过对比学习的无监督联合训练方法, 显著提高了模型的有效性。

(5) ITM<sup>[28]</sup>: 该模型利用图像目标关联和对象目标对齐两个辅助任务来捕捉目标匹配关系。

(6) KEF-TomBERT<sup>[29]</sup>: 该模型利用从图像中提取的形容词-名词对来对齐文本和图像, 并设计了一种新的知识增强框架, 包含视觉增强器以提高视觉注意的有效性, 以及情感预测增强器降低情感预测的难度。

(7) M2DF<sup>[30]</sup>: 该模型定义了两个噪声度量: 粗粒度噪声度量和细粒度噪声度量, 用于衡量每个训练实例中噪声图像的程度, 并设计了一种单一的去噪课程和多重去噪课程, 旨在减少噪声图像对模型学习的负面影响。

(8) HIMT<sup>[14]</sup>: 该模型通过对象检测方法从图像中提取具有语义概念的显著特征, 并采用方面-文本和方面-图像交互建模及辅助重建模块, 实现模态交互并消除文本和图像之间的语义差距。

表 2 GCN-TVFE 模型参数

Table 2 Parameters of GCN-TVFE model

参数名	参数值	
	Tiwtter 2015	Twitter 2017
训练轮数	25	30
批次大小	32	32
学习率	1e-4	1e-4
文本嵌入维度	768	768
图像嵌入维度	2 048	2 048
优化器	Adam	Adam
GCN 层数	2	2
交叉注意力头数	8	8
图像块大小/个	14×14	14×14
暂退概率	0.5	0.5

(9) MGFN-SD<sup>[16]</sup>:该网络通过单模态特征提取、多粒度表示学习和基于自蒸馏的情感预测,实现了情感预测的提升。在多粒度表示学习模块中,通过计算方面图像相关性和相似度,探索细粒度和粗粒度交互,并动态过滤图像噪声。情感预测模块引入自蒸馏机制,从硬标签和软标签中迁移知识,以提高预测准确性。

### 3.3.2 对比实验结果分析

表3列出了本文模型与各主流基线模型在Twitter 2015和Twitter 2017数据集上的性能对比。结果表明,本文GCN-TVFE模型在Acc和 $F_1$ 值上均优于绝大多数基线模型。

表3 不同模型对比实验结果  
Table 3 Results of comparative experiments of different models %

模型	Twitter 2015		Twitter 2017	
	Acc	$F_1$	Acc	$F_1$
EF-Net	74.13	68.24	67.85	64.20
JML	77.05	71.88	70.02	67.67
SMP	77.53	72.24	71.15	69.47
HIMT	78.10	73.70	71.10	69.20
VLP-MABSA	78.60	73.80	—	71.80
ITM	78.30	74.20	72.60	72.00
KEF-TomBERT	78.68	73.75	72.12	69.96
M2DF	78.90	74.80	—	—
MGFN-SD	79.36	74.81	72.77	72.07
GCN-TVFE(本文)	79.16	75.51	72.81	72.21

具体来看,EF-Net和VLP-MABSA模型通过引入注意力机制捕捉模态间的关联信息,从而提升模型性能,但它们忽略了各模态内部的细粒度特征信息。针对这一问题,GCN-TVFE模型利用文本中存在的句法依赖关系和KNN算法,将各模态内部的细粒度特征信息关联起来,从而显著增强了模型的性能。与VLP-MABSA相比,GCN-TVFE在Twitter 2015数据集上的Acc和 $F_1$ 值分别提高了0.56%和1.71%。其次,HIMT、ITM、KEF-TomBERT、M2DF和MGFN-SD模型通过方面词和文本模态挖掘图像中的深层语义特征,减少了视觉噪声对模型的影响。相比之下,GCN-TVFE模型设计了目标视觉提取模块,以准确地定位与目标方面词相关联的视觉特征,进一步增强视觉特征信息并降低噪声信息的影响。与MGFN-SD相比,GCN-TVFE在Twitter 2017数据集上的Acc和 $F_1$ 值分别提高了0.04%和0.14%。尽管本文GCN-TVFE模型在Twitter 2015数据集上的Acc相较于MGFN-SD模型略有下降,但其在综合性能上仍表现出色,这充分证明了GCN-TVFE在细粒度特征提取和目标视觉特征定位方面具有显著优势。

### 3.3.3 GCN层数分析

为了探究GCN层数对模型性能的影响,本文将文本和图像模态中的GCN层数同时设置为1~6层,并在两个数据集上评估了不同层数设置下的模型性能。实验结果如图6和图7所示。从实验结果可以看出,当GCN层数设置为2层时,模型在Twitter 2015和Twitter 2017数据集上的性能达到最优。然而,随着GCN层数的增加,模型性能逐渐下降。这一现象可以归因于以下几点:(1)增加GCN层数会提高模型的复杂度,从而可能导致过拟合;(2)GCN层数增多导致梯度消失或爆炸的问题,特别是在

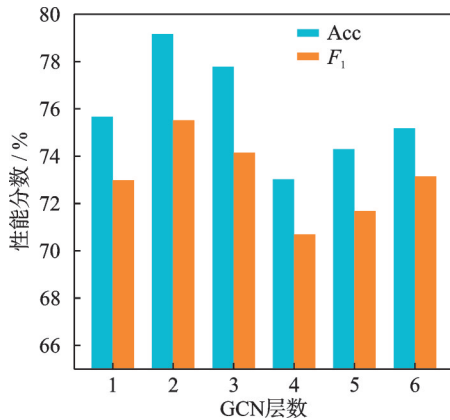


图6 GCN层数在Twitter 2015数据集上的性能

Fig.6 Performance of the number of GCN layers on the Twitter 2015 dataset

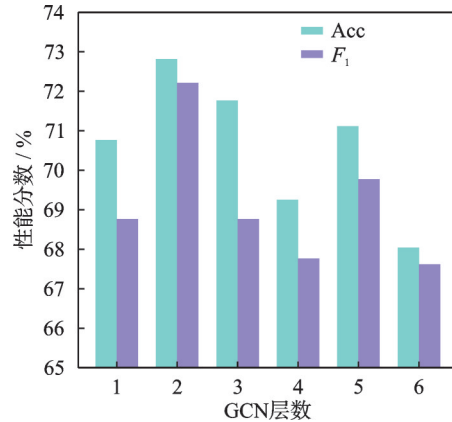


图7 GCN层数在Twitter 2017数据集上的性能

Fig.7 Performance of the number of GCN layers on the Twitter 2017 dataset

训练深度网络时,这会干扰模型的训练过程,影响其收敛性和性能;(3) 随着GCN层数的增加,计算资源的消耗也会显著上升;(4) 过度的层数设置可能导致信息传递的稀释,在图卷积网络中,每一层都会聚合邻居节点的信息,层数过多时,信息传递过远,远距离节点的信息被过度平滑,从而导致局部特征细节的丢失。实验分析表明,本文模型设置GCN层数为2是最优选择。这种设置方案让模型发挥了最佳性能,保证了计算效率并提升了模型的泛化能力。

### 3.3.4 相似度权重参数 $\mu$ 值分析

为了分析文本与图像以及方面词与图像相似度的权重比例,本文评估了不同相似度权重参数值对模型性能的影响,实验结果如图8和图9所示。实验结果显示,当 $\mu$ 值设置为0.4时,模型的性能在两个数据集上达到最优。具体而言,方面词与图像相似度权重比例占0.4,文本与图像相似度比例占0.6。这表明在定位最佳目标视觉区域时,文本信息发挥了更为突出的作用,并且文本信息能够提供直接和明确的语义信息,有助于更准确地理解图像内容并与方面词建立关联。然而,随着相似度权重参数 $\mu$ 值的逐渐增加,模型性能呈现出不稳定性并最终下降,这说明过高比例的关联权重可能导致模型过度依赖于多模态中某种单一信息源,从而忽视了其他重要信息。因此,最佳的权重分配需要平衡多种信息源的贡献,以确保模型的稳定性和性能的优化。

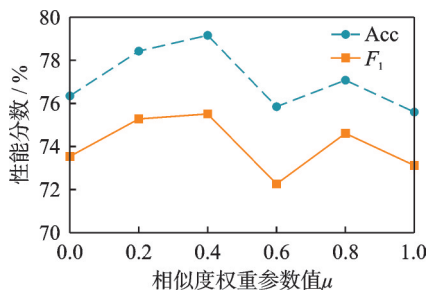
图8  $\mu$ 值在Twitter 2015数据集上的性能

Fig.8 Performance of the  $\mu$  value on the Twitter 2015 dataset

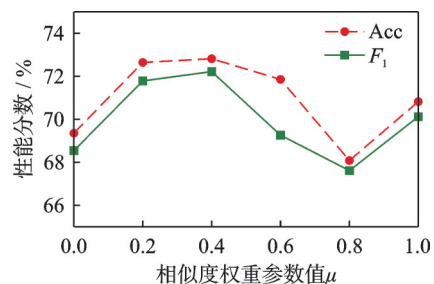
图9  $\mu$ 值在Twitter 2017数据集上的性能

Fig.9 Performance of the  $\mu$  value on the Twitter 2017 dataset

### 3.4 消融实验

为了研究GCN-TVFE模型中各个组成部分对模型性能的影响,本文分别对文本图卷积网络层(TextGCN)、图像图卷积网络层(ImageGCN)、多模态交互注意力模块(Multi-head cross-attention)和目标视觉特征提取模块依次进行消融实验。表4展示了在Twitter 2015和Twitter 2017数据集上的消融实验结果。

表4 消融实验结果  
Table 4 Results of ablation experiments

模型	Twitter 2015		Twitter 2017		%
	Acc	$F_1$	Acc	$F_1$	
不包含TextGCN	74.13	71.24	68.35	67.20	
不包含ImageGCN	73.96	71.81	69.74	68.81	
不包含Multi-head cross-attention	77.16	74.42	71.36	70.93	
不包含目标视觉特征提取模块	75.19	73.21	71.86	71.27	
GCN-TVFE(本文)	79.16	75.51	72.81	72.21	

实验结果表明,去除文本图卷积网络层和图像图卷积网络层会导致模型性能显著下降。这说明TextGCN和ImageGCN对于充分挖掘模态内部深层的细粒度特征信息以及实现模态内信息的准确关联至关重要。此外,去除多模态交互注意力模块会导致文本、方面词和图像模态之间的特征信息无法充分交互和关联,从而使得模型在捕捉特征信息时容易忽略重要的关联信息,影响其整体性能。这凸显了多模态交互注意力模块在强化模态间交互和全面理解模态之间复杂信息方面的重要作用。


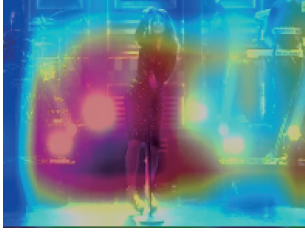

最后,去除目标视觉特征提取模块会使模型在两个数据集上的Acc和 $F_1$ 值分别下降3.97%和2.3%,以及0.95%和0.94%,这表明目标视觉模块在帮助模型精确定位与方面词相关联的图像区域方面具有显著作用。通过这一模块,模型不仅能够增强对图像特征信息的提取能力,还能够有效降低图像噪声信息的干扰,从而提升整体性能。因此,目标视觉特征提取模块在优化模型对视觉信息的利用和提高模型鲁棒性方面发挥着关键作用。

### 3.5 案例分析

为了验证本文模型的性能优势,本文从两个数据集中选取了3个具有代表性的数据样本进行实验。表5展示了本文模型GCN-TVFE与基线模型HIMT在这些样本中的预测结果对比情况。首先,在第1个样本中,方面词“Harry Gulliver”的情感极性为积极情感,GCN-TVFE和HIMT都能正确判断出结果,说明两个模型对于样本中表达情感的词“Happy”都能充分提取。其次,在第2个样本中,由于文本中没有直接给出富有情感的词汇,导致HIMT在预测方面词“Meghan Trainor”时出现错误,这说明分析文本的情感不仅需要捕捉显著的情感词汇,还需进一步对文本内部的语义结构进行关联。在这一方面,GCN-TVFE结合句法依赖关系增强了文本细粒度特征信息的提取,从而在预测方面词“Meghan Trainor”时得到了正确的情感预测结果。最后,在第3个样本中,HIMT在预测方面词“nba”时判断错误,这是因为它没有充分结合图像信息提取出与方面词相关的信息。相较之下,GCN-TVFE设计了一个目标视觉特征提取模块来关注与方面词相关的图像区域,这使得GCN-TVFE能够正确预测出方面词“nba”的情感极性。

为了进一步验证模型中目标视觉特征提取模块的效果,本文对3个样例图像进行了深入的可视化分析。实验结果表明,当样例图像中的方面词分别为“Harry Gulliver”“Meghan Trainor”和“Kyrie Irving”

表5 案例研究  
Table 5 Case study

Image			
Text	Happy birthday to <b>Harry Gulliver</b> from all staff and teammates @ okanaganUK .	<b>Meghan Trainor</b> fell during the <b>Tonight Show</b> . See how Jimmy Fallon reacted .	<b>Kyrie Irving</b> ' s handshakes are in playoff form # <b>nba</b>
HIMT	[ <b>Harry Gulliver</b> ] positive (✓)	[ <b>Meghan Trainor</b> ] negative (×) [ <b>Tonight Show</b> ] neutral (✓)	[ <b>Kyrie Irving</b> ] positive (✓) [ <b>nba</b> ] neutral (×)
GCN-TVFE	[ <b>Harry Gulliver</b> ] positive (✓)	[ <b>Meghan Trainor</b> ] negative (✓) [ <b>Tonight Show</b> ] neutral (✓)	[ <b>Kyrie Irving</b> ] positive (✓) [ <b>nba</b> ] neutral (✓)

时,GCN-TVFE能够精准地检测出图像中与这些方面词相关的部分,说明目标视觉特征提取模块不仅显著提升了模型的性能,还有效降低了图像中的噪声干扰。

#### 4 结束语

本文提出了一种基于GCN和目标视觉特征增强的多模态方面级情感分析模型。该模型通过句法依存关系和KNN算法充分挖掘模态内部信息,有效解决了细粒度特征信息的全面提取问题。在处理图像信息时,模型设计了目标视觉特征提取模块,结合CLIP模型与相似度计算来度量图像中与方面词最相关的区域,并利用Faster R-CNN精确定位出最佳目标视觉特征。为进一步增强各模态之间的关联,模型采用交互注意力机制来深入挖掘模态间的深层关系。实验结果表明,GCN-TVFE模型在两个公开数据集上表现出色,验证了其有效性。未来工作将致力于优化模型的性能,设计更加轻量化的结构以减少复杂度并节省计算资源。同时,引入更多模态信息辅助文本判断方面词的情感极性,改进目标视觉模块以并行提取多个方面词的图像区域。

#### 参考文献:

- [1] 刘强,朱金森,赵龙龙,等.基于字句动态特征和自注意力的情感分析方法[J].数据采集与处理,2024,39(1):193-203.  
LIU Qiang, ZHU Jinsen, ZHAO Longlong, et al. Emotional analysis approach based on dynamic word-sentence features and self-attention[J]. *Journal of Data Acquisition and Processing*, 2024, 39(1): 193-203.
- [2] XU N, MAO W, CHEN G, et al. Multi-interactive memory network for aspect based multimodal sentiment analysis[C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. [S.l.]: ACM, 2019: 371-378.
- [3] KHAN Z, FU Y. 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.
- [4] WANG Y, HUANG M, ZHU X, et al. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA: ACL, 2016: 606-615.
- [5] NGUYEN H T, NGUYEN M. Effective attention networks for aspect-level sentiment classification[C]//Proceedings of 2018

- 10th International Conference on Knowledge and Systems Engineering (KSE). [S.l.]: IEEE, 2024.
- [6] FAN F, FENG Y, ZHAO D. Multi-grained attention network for aspect-level sentiment classification[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2018: 3433-3442.
- [7] AZIZ M M, ABU BAKAR A, YAAKUB M R. Core NLP dependency parsing and pattern identification for enhanced opinion mining in aspect-based sentiment analysis[J]. Journal of King Saud University: Computer and Information Sciences, 2024, 36(4): 102035.
- [8] GAO X, LIU F, ZHUANG X, et al. Dual-channel relative position guided attention networks for aspect-based sentiment analysis[J]. Expert Systems with Applications, 2024, 253: 124271.
- [9] LIU H, LI X, LU W, et al. Graph augmentation networks based on dynamic sentiment knowledge and static external knowledge graphs for aspect-based sentiment analysis[J]. Expert Systems with Applications, 2024, 251: 123981.
- [10] 谢珺, 高婧, 续欣莹, 等. 基于知识增强的双Transformer网络的方面级情感分析模型[J]. 数据分析与知识发现, 2024, 8(11): 47-58.
- XIE Jun, GAO Jing, XU Xinying, et al. Aspect-based sentiment analysis model of dual-transformer network based on knowledge enhancement[J]. Data Analysis and Knowledge Discovery, 2024, 8(11): 47-58.
- [11] YU J, JIANG J, XIA R. 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.
- [12] ZHOU J, ZHAO J, HUANG J X, et al. MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis[J]. Neurocomputing, 2021, 455: 47-58.
- [13] SONG Z, XUE Y, GU D, et al. Target-oriented multimodal sentiment classification by using topic model and gating mechanism[J]. International Journal of Machine Learning and Cybernetics, 2023, 14(7): 2289-2299.
- [14] YU J, CHEN K, XIA R. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis[J]. IEEE Transactions on Affective Computing, 2022, 14(3): 1966-1978.
- [15] WANG D, TIAN C, LIANG X, et al. Dual-perspective fusion network for aspect-based multimodal sentiment analysis[J]. IEEE Transactions on Multimedia, 2023, 26: 4028-4038.
- [16] YANG J, XIAO Y, DU X. Multi-grained fusion network with self-distillation for aspect-based multimodal sentiment analysis[J]. Knowledge-Based Systems, 2024, 293: 111724.
- [17] WAN Y, CHEN Y, LIN J, et al. A knowledge-augmented heterogeneous graph convolutional network for aspect-level multimodal sentiment analysis[J]. Computer Speech & Language, 2024, 85: 101587.
- [18] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of NAACL-HLT. Minneapolis, MN, USA: Association for Computational Linguistics, 2019: 2.
- [19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778.
- [20] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2021: 8748-8763.
- [21] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.
- [22] LI J, LI D, XIONG C, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2024.
- [23] KINGMA D P. Adam: A method for stochastic optimization[EB/OL]. (2014-12-22). <https://arxiv.org/pdf/1412.6980v6>.
- [24] GU D, WANG J, CAI S, et al. Targeted aspect-based multimodal sentiment analysis: An attention capsule extraction and multi-head fusion network[J]. IEEE Access, 2021, 9: 157329-157336.
- [25] JU X, ZHANG D, XIAO R, et al. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2021: 4395-4405.
- [26] YE J, ZHOU J, TIAN J, et al. Sentiment-aware multimodal pre-training for multimodal sentiment analysis[J]. Knowledge-

Based Systems, 2022, 258: 110021.

- [27] LING Y, YU J, XIA R. Vision-language pre-training for multimodal aspect-based sentiment analysis[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics. [S.l.]: Association for Computational Linguistics, 2022.
- [28] YU J, WANG J, XIA R, 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, Austria: Artificial Intelligence Organization, 2022: 4482-4488.
- [29] ZHAO F, WU Z, LONG S, 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, Korea: [s.n.], 2022: 6784-6794.
- [30] ZHAO F, LI C, WU Z, et al. M2DF: Multi-grained multi-curriculum denoising framework for multimodal aspect-based sentiment analysis[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023: 9057-9070.

#### 作者简介:



赵雪峰(1976-),通信作者,男,博士,副教授,硕士生导师,研究方向:自然语言处理和机器视觉,E-mail: zhaoxf@jou.edu.cn。



柏长泽(1999-),男,硕士研究生,研究方向:自然语言处理和多模态情感分析,E-mail: 2023220901@jou.edu.cn。



狄恒西(2000-),男,硕士研究生,研究方向:自然语言处理和多模态情感分析,E-mail: dihx@jou.edu.cn。



仲兆满(1977-),男,博士,教授,硕士生导师,研究方向:人工智能、自然语言处理和社交网络分析,E-mail: zhongzhaoman@163.com。



仲晓敏(1977-),女,硕士,讲师,研究方向:软件缺陷检测,E-mail: 2001000049@jou.edu.cn。

(编辑:张黄群)