山西大学学报(自然科学版)49(1):1—14,2026
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2025091

基于全局局部交互模型的多模态情感分析

李梦晗[1],仲兆满[1,2*],徐俊康[1],陈柯含[1]
(1.江苏海洋大学 计算机工程学院,江苏 连云港 222005;
2.江苏省海洋资源开发研究院,江苏 连云港 222005)

摘 要:多模态方面级情感分析(Multimodal Aspect-Based Sentiment Analysis, MABSA)作为情感计算领域的关键研究方向,致力于融合文本、图像、音频等多种模态信息,用以实现对特定方面情感的精细化分析。在当前的多模态方面级情感分析研究中,存在着图像噪声干扰以及过度依赖局部特征等问题,进而影响了分析的准确性和全面性。针对这些局限,本文提出了一种创新的全局局部交互情感分析模型(Global-Local Interactive Emotion Analysis Model, GLIEAM)。一方面,采用视觉Transformer(Vision Transformer, ViT)模型与生成式预训练Transformer(Generative Pre-trained Transformer, GPT)模型串联生成图像描述并将其与原始文本特征进行拼接,有效强化了信息融合效果,从而更全面地捕捉多模态数据中的情感线索。另一方面,为解决图像噪声问题,结合小波变换和非局部均值方法对图像进行去噪处理;同时,利用卷积神经网络(Convolutional Neural Network, CNN)和T2T视觉Transformer(Tokens-to-Token Vision Transformer, T2T-ViT)分别提取局部和全局图像特征,避免了对局部特征的过度依赖,实现了对图像特征的全面、均衡提取。通过在基准数据集上进行实验,结果表明该方法显著优于现有方法,在Twitter-15上准确率达到78.46%,Twitter-17数据集上准确率达到75.21%,尤其在低资源场景下展现出卓越的性能。

关键词:多模态方面级情感分析;视觉Transformer;特征融合;图像去噪;注意力机制
中图分类号:TP39 **文献标志码**:A **文章编号**:0253-2395(2026)01-0001-14

Multimodal Sentiment Analysis Based on Global-local Interaction Model

LI Menghan[1], ZHONG Zhaoman[1,2*], XU Junkang[1], CHEN Kehan[1]
(1. *School of Computer Engineering, Jiangsu Ocean University, Lianyungang* 222005*, China;*
2. *Jiangsu Institute of Marine Resources Development, Lianyungang* 222005*, China*)

Abstract: Multimodal aspect-based sentiment analysis (MABSA) is a critical research direction in the field of affective computing, aiming to integrate multimodal information, such as text, images, and audio to achieve fine-grained analysis of sentiment toward specific aspects. Current research in MABSA faces challenges such as image noise interference and excessive reliance on local features, which compromise the accuracy and comprehensiveness of the analysis. To address these limitations, this paper proposes an innovative global-local interactive emotion analysis model (GLIEAM). On the one hand, the model employs a tandem architecture of vision transformer (Vision Transformer, ViT) and generative pre-trained transformer (GPT) to generate image descriptions, which are then concatenated with original text features, significantly enhancing information fusion and enabling a more comprehensive capture

收稿日期:2025-06-05;**修回日期**:2025-09-10
基金项目:国家自然科学基金(72174079);江苏省"青蓝工程"大数据优秀教学团队(2022-29);连云港市重点研发(产业前瞻与关键核心技术)项目(CG2323)
作者简介:李梦晗(2001—),女,内蒙古赤峰人,硕士研究生,研究方向为多态方面级情感分析。E-mail:18347338521@163.com
* **通信作者**:仲兆满(ZHONG Zhaoman),E-mail:zhongzhaoman@163.com
引文格式:李梦晗,仲兆满,徐俊康,等.基于全局局部交互模型的多模态情感分析[J].山西大学学报(自然科学版),2026,49(1):1-14. DOI:10.13451/j.sxu.ns.2025091.

of emotional cues in multimodal data. On the other hand, to mitigate image noise, a hybrid approach combining wavelet transform and non-local means is applied for image denoising. Additionally, convolutional neural networks (CNN) and tokens-to-token vision transformer (T2T-ViT) are utilized to extract local and global image features, respectively, avoiding over-reliance on local features and achieving balanced and holistic image feature extraction. Experimental results on benchmark datasets demonstrate that the proposed method outperforms existing approaches, the accuracy reached 78.46% on the Twitter-15 dataset and 75.21% on the Twitter-17 dataset, particularly exhibiting superior performance in low-resource scenarios.

Key words: multimodal aspect-based sentiment analysis; vision Transformer; feature fusion; image denoising; attention mechanism

0 引言

在数字化信息呈指数级增长的当下,社交媒体、在线视频平台以及各类智能交互设备等,成为人们日常交流与表达的重要阵地。在这些平台上,信息以丰富多样的形式呈现,文本、图像、音频等多种模态相互交织,共同承载着人们的情感与态度。多模态方面级情感分析,正是在这样的背景下应运而生,成为自然语言处理、计算机视觉以及人工智能等领域的研究热点,其旨在综合分析多种模态的数据,精准识别人们针对特定方面的情感倾向,这一研究方向对于深入理解人类情感、优化人机交互以及挖掘用户潜在需求,都有着不可忽视的重要意义。

回顾多模态方面级情感分析的发展历程,其起源于早期较为单一的文本情感分析。起初,研究人员主要借助词袋模型、词频-逆文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)等传统方法[1],对文本中的情感词汇进行提取与分析。这类方法虽能初步捕捉文本中的情感信息,但存在诸多局限性,例如对词序和上下文关系的忽视,导致其难以准确把握复杂语境下的情感语义。随着深度学习技术的兴起,循环神经网络(Recurrent Neural Network, RNN)及其变体,如长短期记忆网络(Long Short-Term Memory, LSTM)、门控循环单元(Gated Recurrent Unit, GRU)等[2],凭借对序列数据的出色处理能力,在文本情感分析中崭露头角,大幅提升了分析的准确性。然而,仅依赖文本模态的情感分析,无法全面涵盖人类情感表达的丰富性。于是,多模态情感分析开始进入研究视野,早期的多模态情感分析,只是简单地将不同模态的特征进行拼接,便用于后续分析,这种方式未能充分挖掘不同模态之间的内在联系与协同作用。

近年来,随着研究的深入,学者们不断探索更为复杂、有效的模型与方法。华南理工大学研究团队提出的半监督模态内交互学习网络(Semi-IIN)[3],创新性地结合半监督学习与动态选择机制,有效攻克了多模态情感分析中高标注成本与标签歧义的难题,同时在多模态交互信息的选择与利用方面取得了显著突破。而多模态方面级情感分析在传统多模态情感分析的基础上,实现了显著进化。细粒度分析的实现[4]使得情感识别更加精准。相比于全局情感分析,方面级情感分析专注于特定方面(如产品特性或服务维度),能够深入挖掘用户反馈中的细微情感差异。此外,通过结合文本、图像和音频等多种模态,能够更全面地理解用户情感,例如通过分析评价文本与相关图像的结合,从而提升情感分析的整体准确性。其次,复杂的特征融合策略得以应用,使得不同模态的信息能够互补。通过引入注意力机制和特征选择,模型能够构建更为复杂的特征表示,从而捕捉多模态数据中的上下文信息和情感表达的多层次细节。

尽管多模态方面级情感分析已取得一定成果,但在实际发展过程中,仍面临着诸多严峻挑战。首先,信息融合的复杂性使得不同模态(如文本、图像、音频)之间的语义差异和表达方式不一致,可能导致情感倾向的误判[5]。其次,图像质量和噪声问题影响分析准确性,尽管有去噪技术,但关键特征可能在去噪过程中丢失[6]。此外,许多模型过度依赖局部特征,忽视全局信息,导致情感态度的捕捉不够全面。文本语义理解方面,处理复杂句子(如隐喻或情感倾向模糊的句子)仍具挑战,对于深层次的语义理解难以实现。同时,在多模态分析场景下,模型的可解释性也至关重要[7]。因为用户往往希望深入了解情感判断背后的依据,从而更好地利用分析结果。此外,数据的

不平衡性可能导致某些情感类别表现不佳,需要有效平衡不同情感类别以提高模型泛化能力[8]。在某些应用场景中,实时处理能力也是一大挑战,要求模型在速度和准确性之间取得平衡。最后,情感分析的结果依赖于标注数据的质量。例如,一张看似“中性”的图片可能因特定上下文(如伴随文本)隐含积极情感,但现有数据集缺乏细粒度对齐标注,导致模型训练偏差。进而影响模型的训练效果。这些挑战亟需进一步研究与创新,以提升多模态方面级情感分析的准确性和鲁棒性。

本文的主要贡献涵盖以下几个方面:

(1)采用视觉 Transformer(Vision Transformer, ViT)与生成式预训练 Transformer(Generative Pre-trained Transformer, GPT)的串联,构建一个生成图像描述的转换器。该模型将图像转换为文本描述特征,并利用注意力机制将这些特征与原始文本特征进行拼接,以增强信息融合的效果。该方法结合了全局和局部信息,解决了当下方面级情感分析中会忽略整体情感倾向的问题。

(2)结合小波变换和非局部均值方法对图像进行去噪处理,以提高图像质量。随后,通过卷积神经网络(Convolutional Neural Networks, CNN)提取局部图像特征,并利用 T2T 视觉 Transformer(Tokens-to-Token Vision Transformer, T2T-ViT)获取全局图像特征,从而实现更全面的图像理解。解决当下多模态方面级情感分析中图像噪声以及许多模型过度依赖局部特征而忽略全局特征等问题

(3)应用依存句法解析技术对文本进行分析,以加强语义理解。再利用面向任务的多模态 BERT(Task-Oriented Multimodal BERT, TomBERT)模型提取文本特征,增强了文本信息的深层次表达。

1 相关工作

1.1 文本特征提取

在自然语言处理领域,文本特征提取技术的发展是一个从简单到复杂、从静态到动态的进阶过程。早期,词袋模型(Bag of Words, BoW)和词频-逆文档频率方法[1]通过统计词频来对文本进行表征。这两种方法实现起来简单便捷,然而,它们存在明显的局限性,无法有效捕捉文本中的词序和语义信息。这就好比拼图时,只关注每一块拼图的数量,却忽略了它们之间的拼接顺序和整体图案的含义。为了突破这一局限,词嵌入技术横空出世,像 Word2Vec、GloVe 和 FastText[9],它们将单词映射到低维向量空间,显著提升了语义表示能力。这就像是给每个单词赋予了一个独特的“语义指纹”,让模型能更好地理解单词的含义。不过,这些方法生成的是静态词向量,在面对上下文变化时,表现得有些力不从心。为了进一步捕捉上下文信息,上下文词嵌入方法应运而生,其中代表性的有 ELMo、双向编码器表示 Transformer(Bidirectional Encoder Representations from Transformers, BERT)和 GPT[10]。这些模型利用预训练语言模型,生成能够根据上下文动态变化的词向量。这一突破,极大地提升了模型对复杂语言现象的理解能力,就像为模型戴上了一副“上下文理解眼镜”,让它能更准确地把握文本的内涵。此外,主题模型(如隐含狄利克雷分布(Latent Dirichlet Allocation, LDA))、句法特征、字符级特征以及图嵌入等方法,也在特定的自然语言处理任务中展现出独特的优势,为解决不同类型的文本分析问题提供了多样化的思路。近年来,预训练语言模型(如 BERT、GPT 和 T5)[11]成为了自然语言处理领域的主流。这些模型通过在大规模语料上进行预训练,然后针对不同的自然语言处理(Natural Language Processing, NLP)任务进行微调,在多种任务上都取得了突破性的进展,引领了自然语言处理技术的新变革。在句法解析领域,传统规则方法逐渐被神经架构替代。Wang 等[12]在 BERT 中引入图传播层,显式建模依存树结构,在 PTB(Penn Tree Bank)数据集无标记依存正确率(Unlabeled Attachment Score, UAS)中高达 96.20%;Xu 等[13]提出跨语言依存投影框架,利用多语言 BERT 对齐语法结构,在 50 种语言上平均带标签依存关系得分(Labeled Attachment Score, LAS)达 78.30%。Zhang 等[14]通过异构图网络实现依存关系与语义角色的联合推理。

1.2 图像特征提取

图像特征提取技术经历了从传统手工设计到深度学习驱动下多模态融合的深刻变革,实现了显著的技术飞跃。早期,图像特征提取依赖于手工设计的特征,如尺度不变特征变换(Scale-invariant Feature Transform,SIFT)、方向梯度直方图(Histogram of Oriented Gradient,HOG)和颜色直方图[15],这些方法在尺度变换和光照变化下展现了一定的鲁棒性,但难以捕捉图像的深层语义信息,限制了其在复杂场景中的应用。随着深度学习的兴起,以AlexNet卷积神经网络、视觉几何组(Visual Geometry Group,VGG)和残差网络(Residual Network,ResNet)[16]为代表的预训练卷积神经网络通过端到端学习,自动提取具有丰富语义的图像特征,使得图像在视觉表达上具备更强的辨识度与区分度,进而显著提升了图像在各类视觉分析任务(如目标识别、图像分类、语义分割等)中的性能,为后续任务奠定了坚实基础。在多模态任务不断涌现的背景下,研究者开始探索图像与文本的联合建模方法[17]。注意力机制(如交叉注意力)和联合嵌入空间(如对比语言-图像预训练(Contrastive Language-Image Pre-training,CLIP))的引入,实现了跨模态信息的有效交互和深度融合。基于Transformer的多模态架构(如从Transformer中学习跨模态编码器表示(Learning Cross-Modality Encoder Representations from Transformers,LXMERT))进一步推动了跨模态特征的动态对齐与协同工作,显著提升了多模态任务的性能。

近年来,ViT和自监督预训练模型(如简单的对比学习表示(Simple Contrastive Learning of Representations,SimCLR)、掩码自编码器(Masked Autoencoders,MAE))取得了突破性进展[18]。ViT通过引入Transformer架构,解决了图像特征的长距离依赖建模问题,而自监督学习减少了对标注数据的依赖,降低了数据成本。图神经网络(Graph Neural Network,GNN)[19]的引入则为图像中复杂物体关系的建模提供了新维度。图像去噪技术在计算机视觉领域取得了显著进展,小波变换和非局部均值方法作为两种经典方法,各自经历了重要的技术演进。在小波变换方面,研究者们主要从三个方向进行了改进。首先,针对传统固定小波基函数的局限性,Zhang等[20]提出了基于卷积神经网络的动态小波基学习方法,显著提升了基函数的自适应能力。其次,为了克服离散小波变换可能导致的伪吉布斯效应,Chen等[21]开发了融合小波域与空间域特征的创新方法。此外,Wang等[22]将小波变换与视觉Transformer相结合,通过引入小波引导的注意力机制,实现了更有效的特征提取。在非局部均值方法方面,研究进展主要体现在两个方面。Liu等[23]采用深度学习方法改进了传统的相似性度量方式,使用CNN提取的高层语义特征替代了原有的像素块匹配方法。

1.3 图像特征文本化相关工作

在多模态方面级情感分析领域,图像特征文本化是实现图像与文本信息深度融合的核心技术,旨在将图像中的视觉信息转化为文本形式,使其能够在统一的分析框架下与文本模态数据协同处理,从而提升情感分析的准确性和全面性。图像字幕生成技术是图像特征文本化的重要路径之一,它结合了计算机视觉与自然语言处理技术。以经典的CNN-RNN架构为例[24],首先通过CNN提取图像的语义特征(如物体形状、颜色、空间关系等),并将其编码为高维特征向量;随后,利用循环神经网络(Recurrent Neural Network,RNN)或其变体(如LSTM)生成与图像内容契合的自然语言描述。例如,对于一张儿童在公园放风筝的图片,模型可以生成“阳光明媚的日子里,一个孩子在绿草如茵的公园里欢快地放风筝”这样的生动描述。这种方法在社交媒体平台中具有实际应用价值,能够自动生成图片描述,提升内容的可检索性和用户体验。然而,图像字幕生成技术可能面临描述不准确或语义单一的问题,尤其在复杂场景中容易遗漏关键细节。

另一种图像特征文本化的方法是直接将图像特征映射到文本特征空间,利用预训练的多模态模型(如基于Transformer的架构[25])和注意力机制构建图像与文本特征之间的关联。这种方法在图像检索任务中表现出色,能够高效匹配文本查询与图像内容。例如,用户输入“白色高帮篮球鞋,带有蓝色装饰条纹”时,模

型可以快速检索出符合条件的商品图像。尽管这种方法在检索效率和准确性上具有优势,但其生成的文本特征缺乏明确的语义解释,难以直接转化为人类可理解的自然语言描述。

1.4 特征融合

在多模态方面级情感分析领域,特征融合是提升模型性能的关键环节,其重要性不言而喻。随着研究的深入,特征融合方法的探究对于挖掘多模态数据间的复杂关联、提升情感分析的精准度起着决定性作用。早期融合(Early Fusion)在数据预处理阶段将不同模态的原始数据直接整合,便于模型捕捉数据间的内在规律,但易受噪声和冗余信息影响,对数据质量和一致性要求极高。中期融合(Mid-level Fusion)[26]在特征提取阶段融合不同模态的特征,避免了早期融合的噪声问题,但如何科学选择融合方式和确定特征权重仍是难题。晚期融合(Late Fusion)在各模态独立处理后进行结果融合,灵活性高但可能忽视模态间的交互信息。深度学习框架下的端到端融合[27](End-to-end Fusion)通过深层神经网络自动学习模态间的内在联系,适应性强但对数据量和计算资源要求高,且可解释性较差。综上所述,特征融合方法的选择需根据具体任务和数据特点,权衡各方法的优劣,以实现最优的情感分析效果。

2 基于全局局部交互模型的多模态情感分析

2.1 任务定义

我们给定一组多峰样本M,对于每个样本$m\in M$,其包含一个样本$S=(w_1,w_2,\cdots,w_N)$、与句子相关联的伴随图像I以及作为句S的子序列的体项T。体项通常是句子中提到的实体,其与情感标签y相关联,情感标签y可以是积极的(positive)、消极的(negative)、中性的(neutral)。多模态方面级情感分析的目标是学习一个面向目标的情感分类器,以便它可以正确地预测未见过样本中的意见目标的情感标签,以便它可以正确地预测未见过样本中的意见目标的情感标签。例如,通过输入诸如“这家甜品店的环境很好,但甜品味道很糟糕”的文本以及图片,模型可以预测目标“甜品店的环境”是积极的,而目标“甜品”是消极的。

2.2 系统框架图

本文的全局局部交互情感分析模型(Global-Local Tnteractive Emotion Analysis Model, GLIEAM)如图1所示。该模型包括三个主要部分:全局特征增强模块、图像特征增

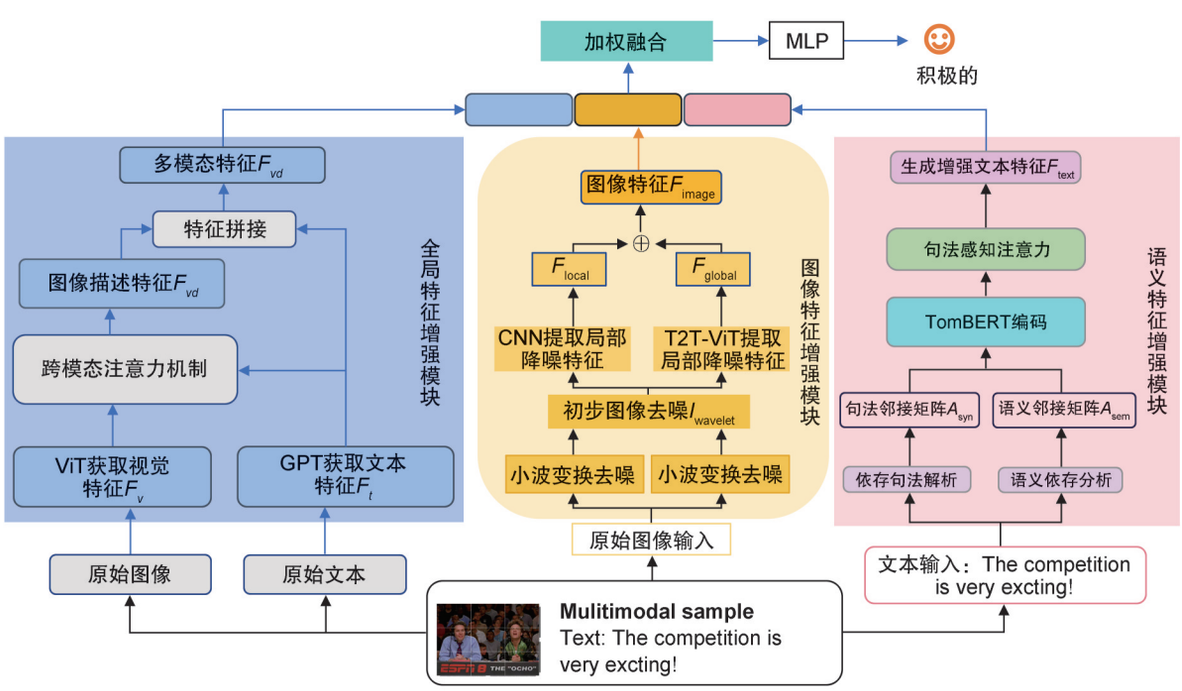

图1 模型框架结构图

Fig. 1 Structure diagram of model framework

强模块和语义特征增强模块。每个模块的具体功能和作用如下:

在全局特征增强模块中,首先通过ViT[28]将图像转换为文本描述特征,以弥合视觉与语言之间的鸿沟。再将这些图像描述特征与原始文本特征通过注意力机制进行拼接,作为预训练的GPT模型[29]的输入,以提取全局语义信息。该方法结合了全局和局部信息,解决了当下方面级情感分析中会忽略整体情感倾向的问题。第2.2.1节介绍了全局特征增强模块的详细信息。

在图像特征增强模块中,首先结合小波变换和非局部均值方法对图像进行去噪处理,以提高图像质量。随后,通过CNN提取局部图像特征,并利用T2T-ViT获取全局图像特征,从而实现更全面的图像理解。该方法解决了当下多模态方面级情感分析中图像噪声以及许多模型过度依赖局部特征忽略全局特征等问题。第2.2.2节介绍了图像特征增强模块的详细信息。

在语义特征增强模块中,首先应用依存句法解析技术对文本进行分析,以加强语义理解。再利用TomBERT[30]模型提取文本特征,增强了文本信息的深层次表达。第2.2.3节介绍了语义特征增强模块的详细信息。

2.2.1 全局特征增强模块

首先,给定输入图像$I\in R^{H\times W\times C}$和文本序列$T=\{t_1,t_2,\cdots,t_n\}$,其中$H,W,C$分别表示图像的高度、宽度和通道数,$n$为文本序列长度。在视觉特征提取阶段,采用ViT对图像进行编码:将图像I划分为m个不重叠的patch $P=\{p_1,p_2,\cdots,p_m\}$,每个patch通过线性投影映射为d维向量,得到patch嵌入$E_p\in\mathbb{R}^{m\times d}$。随后,ViT通过多头自注意力机制(Multi-Head Self-Attention,MSA)和前馈网络(Feedforward Neural Network,FFN)[31]对patch嵌入进行编码,生成视觉特征$F_v\in\mathbb{R}^{m\times d}$,用于与文本特征融合,以解决方面级情感分析中忽略整体情感倾向的问题,实现更准确的情感分类:

$$F_v=\mathrm{ViT}(E_p)=\mathrm{FFN}(\mathrm{MSA}(E_p))。\tag{1}$$

在文本特征提取阶段,使用GPT对输入文本T进行编码,生成文本特征$F_t\in\mathbb{R}^{n\times d}$:

$$F_t=\mathrm{GPT}(T)。\tag{2}$$

为了生成图像描述特征,设计一个跨模态注意力机制(Cross-modal Attention),将视觉特征F_v作为键值对,文本特征F_t作为查询,计算注意力权重并生成图像描述特征$F_{vd}\in\mathbb{R}^{n\times d}$:

$$F_{vd}=\mathrm{CrossAtt}(F_t,F_v,F_v)=\mathrm{Softmax}\left(\frac{F_t\boldsymbol{W}_q(F_v\boldsymbol{W}_k)^T}{\sqrt{d}}\right)F_v\boldsymbol{W}_v,\tag{3}$$

其中$\boldsymbol{W}_q,\boldsymbol{W}_k,\boldsymbol{W}_v\in\mathbb{R}^{d\times d}$为可学习的投影矩阵。为了增强信息融合效果,将图像描述特征F_{vd}与原始文本特征F_t进行拼接,并通过一个全连接层进行融合,得到多模态特征$F_{\mathrm{fused}}\in\mathbb{R}^{n\times d}$:

$$F_{\mathrm{fused}}=\mathrm{ReLU}([F_t;F_{vd}]W_f+b_f),\tag{4}$$

其中$W_f\in\mathbb{R}^{2d\times d}$和$b_f\in\mathbb{R}^d$为可学习参数。最后,将多模态特征$F_{\mathrm{fused}}$输入到一个分类器中,以生成$x$—方面级情感分布$y\in\mathbb{R}^C$,$x$代表输入文本中的特定方面(如“甜品店的环境”),该分布用于细化情感分析,帮助模型识别用户对不同方面的情感倾向,其中C为情感类别数:

$$y=\mathrm{Softmax}(F_{\mathrm{fused}}W_c+b_c),\tag{5}$$

其中$W_c\in R^{d\times C}$和$b_c\in\mathbb{R}^C$为分类器的可学习参数。此处的分类器可作为中间监督,支持多任务学习框架,以提升模型的整体性能。通过上述流程,模型能够有效结合全局视觉信息和局部文本信息,解决方面级情感分析中忽略整体情感倾向的问题,从而实现更准确的情感分类。

算法1 ViT-GPT跨模态特征融合

输入:图像I,文本序列S

输出:融合特征$Z_{\mathbf{fused}}$

① // ViT图像编码(对应式1)

② 将I分割为16×16块→$X_{\mathrm{patches}}\in\mathbb{R}$

③ $E_{\mathrm{patch}}\leftarrow\mathrm{Linear}(X_{\mathrm{patches}})+\mathrm{PosEmbed}$

④ for $i\leftarrow$ 1 to 12 do

⑤ $Z_{\mathrm{vit}}\leftarrow\mathrm{MultiHeadAttention}(E_{\mathrm{patch}},$ num_heads为8, head_dim为64, dropout为0.1)

⑥ $Z_{\mathrm{vit}}\leftarrow\mathrm{LayerNorm}(E_{\mathrm{patch}},Z_{\mathrm{vit}})$

⑦ end for

⑧ //GPT文本编码(对应式2)

⑨ $E_{\mathrm{patch}}\leftarrow\mathrm{Embedding}(S)\leftarrow\mathrm{PosEmded}$

⑩ for $j\leftarrow$ 1 to 6 do

⑪ $Z_{\text{text}} \leftarrow$ MaskedMultiHeadAttention(E_{text}, num_heads为8, dropout为0.1)

⑫ $Z_{\text{cross}} \leftarrow$ MultiHeadAttention($Z_{\text{text}}, Z_{\text{vit}}$, num_heads为8, head_dim为64)//式(3)

⑬ end for

⑭ // 特征融合(对应式(4)—式(5))

⑮ $Z_{\text{fused}} \leftarrow \text{ReLU}(\text{Concat}[Z_{\text{cross}}, Z_{\text{text}}] \cdot W_f)$

⑯ return Z_{fused}

其中多头注意力:头数(num_heads)设为8,每头的维度 d_k(head_dim)为64,dropout率为0.1。

2.2.2 图像特征增强模块

给定输入图像 $I \in \mathbb{R}^{H \times W \times C}$,首先通过离散平稳小波变换将图像分解为低频子带 c_A 和高频子带 $\{c_H, c_V, c_D\}$,并对高频子带应用自适应阈值收缩:

$$\hat{c}_X(j,k) = \text{sign}\left(c_X(j,k)\right) \cdot \max\left(\left|c_X(j,k)\right| - \tau_j, 0\right), \quad (6)$$

其中 $\tau_j = \frac{\sigma_j^z}{\sigma_x}$ 为自适应阈值,σ_j 为第 j 层子带噪声方差,σ_x 为信号标准差。随后,通过逆小波变换重构去噪图像 I_{wavelet}。为进一步抑制残留噪声,采用非局部均值方法计算像素相似性权重:

$$w(p,q) = \exp\left(-\frac{/\!/ N(p) - N(q) /\!/ _{2,a}^2}{h^2}\right), \quad (7)$$

并加权聚合邻域像素值得到最终去噪图像,解决多模态方面级情感分析中图像噪声问题:

$$I_{\text{denoised}}(p) = \frac{1}{Z(p)} \sum_{q \in \Omega(p)} \mathcal{W}(p,q) \cdot I_{\text{wavelet}}(q), \quad (8)$$

其中 $Z(p) = \sum_q \mathcal{W}(p,q)$ 为归一化因子。接下来,通过CNN提取局部特征 F_{local},同时利用T2T-ViT逐步重组图像Token并编码全局特征 $F_{\text{global}} = \text{ViT}(T_L)$,其中 T_L 为最终Token序列。设计门控融合机制,动态整合局部与全局特征:

$$\begin{aligned} F_{\text{fused}} = {} & \sigma\left(W_g\left[F_{\text{local}}; F_{\text{global}}\right]\right) \odot F_{\text{local}} + \\ & \left(1 - \sigma\left(W_g\left[F_{\text{local}}; F_{\text{global}}\right]\right)\right) \odot F_{\text{global}}, \end{aligned} \quad (9)$$

其中 $W_g \in \mathbb{R}^{(d_{\text{local}} + d_{\text{global}}) \times 1}$ 为可学习参数,σ 为Sigmoid函数。最后,通过交叉注意力对齐文本特征 F_t 与图像特征 F_{fused},生成对齐特征:

$$F_{\text{align}} = \text{Softmax}\left(\frac{\left(F_t \boldsymbol{W}_q\right)\left(F_{\text{fused}} \boldsymbol{W}_k\right)^T}{\sqrt{d}}\right) \cdot \left(F_{\text{fused}} \boldsymbol{W}_v\right), \quad (10)$$

并将其输入分类器预测情感分布:

$$y = \text{Softmax}\left(\text{MLP}\left(\left[F_t; F_{\text{align}}\right]\right)\right)。 \quad (11)$$

分类器作为中间监督,增强特征学习的有效性。该方法通过联合去噪、多尺度特征提取和跨模态对齐,有效解决了图像噪声和局部-全局特征失衡问题,显著提升了多模态情感分析的性能。

该方法通过联合去噪、多尺度特征提取和跨模态对齐,有效解决了图像噪声和局部-全局特征失衡问题,显著提升了多模态情感分析的性能。

2.2.3 语义特征增强模块

给定输入文本序列 $T = \{t_1, t_2, \cdots, t_n\}$,首先通过依存句法解析技术(如Stanford Parser)对文本进行分析,生成依存句法树。依存句法树以有向图的形式表示词与词之间的语法关系,其中每个节点代表一个词,边表示词与词之间的依存关系(如主谓关系、动宾关系等)。基于依存句法树,构建句法邻接矩阵 $\boldsymbol{A}_{\text{syn}} \in \{0,1\}^{n \times n}$,其中 $\boldsymbol{A}_{\text{syn}}(i,j) = 1$ 表示词 t_i 与 t_j 存在依存关系,否则为0。为了进一步增强句法信息的表达能力,引入语义依存分析(Semantic Dependency Parsing),生成语义邻接矩阵 $\boldsymbol{A}_{\text{sem}} \in \mathbb{R}^{n \times n}$,其中 $\boldsymbol{A}_{\text{sem}}(i,j)$ 表示词 t_i 与 t_j 之间的语义关联强度。接下来,利用TomBERT模型对文本进行编码,生成词级特征 $F_t = \{f_1, f_2, \cdots, f_n\} \in \mathbb{R}^{n \times d}$,其中 d 为特征维度。TomBERT通过多层Transformer编码器捕捉文本的上下文信息,其自注意力机制如下:

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (12)$$

其中 $Q = F_t \boldsymbol{W}_q$,$K = F_t \boldsymbol{W}_k$,$V = F_t \boldsymbol{W}_v$ 分别为查询、键和值,$\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d \times d}$ 为可学习参数。为了将句法信息融入TomBERT的语义表征中,设计句法感知注意力机制(Syntax-aware Attention)。该机制在标准自注意力的基础上,引入句法邻接矩阵 $\boldsymbol{A}_{\text{syn}}$ 和语义邻接矩阵 $\boldsymbol{A}_{\text{sem}}$ 作为先验知识,增强模型对语法结构的建模能力。

句法感知注意力如下:

$$\text{Attention}_{\text{syn}}(Q,K,V)=\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}+\beta\times A\right)V, \tag{13}$$

其中λ_1和λ_2为可学习的句法和语义权重系数,用于动态调节句法和语义信息对注意力分布的贡献。通过多任务学习框架,联合优化句法解析任务和语义表征任务。句法解析任务的损失函数为交叉熵损失:

$$\mathcal{L}_{\text{syn}}=-\sum_{i,j}\boldsymbol{A}_{\text{syn}}(i,j)\log\hat{\boldsymbol{A}}_{\text{syn}}(i,j), \tag{14}$$

其中$\boldsymbol{A}_{\text{syn}}(i,j)$为模型预测的句法邻接矩阵。语义表征任务的损失函数为情感分类的交叉熵损失:

$$\mathcal{L}_{\text{sem}}=-\sum_{c}y_c\log\hat{y}_c, \tag{15}$$

其中y_c为真实情感标签,$\hat{y}_c$为模型预测的情感分布。最终,总损失函数如下:

$$\mathcal{L}=\mathcal{L}_{\text{syn}}+\alpha\mathcal{L}_{\text{sem}}, \tag{16}$$

其中α为任务权重系数。将句法增强的文本特征$F_{\text{syn}}\in R^{n\times d}$输入下游任务(如情感分类器),预测情感分布:

$$y=\sqrt{1}\,\text{Softmax}\big(\text{MLP}(\text{Fsyn})\big)。 \tag{17}$$

分类器作为中间监督,整合多任务学习的策略,进一步优化效果。该方法通过结合依存句法解析与TomBERT模型,显著提升了文本语义理解的准确性和鲁棒性,为多模态情感分析提供了更丰富的文本特征表示。

2.2.4 特征融合模块

在特征融合阶段,本模型采用了一种基于加权融合的多模态特征集成策略,通过自适应权重分配机制将来自不同模态的特征进行深度融合。具体而言,设ViT-GPT模块提取的图像描述特征为$F_{\text{text}}\in\mathbb{R}^{d_1}$,该特征通过ViT编码器获取图像的全局语义信息,并经由GPT解码器生成具有上下文感知能力的文本描述特征;小波变换与T2T-ViT模块提取的局部和全局图像特征为$F_{\text{image}}\in\mathbb{R}^{d_2}$,其中局部特征通过CNN提取,全局特征则由T2T-ViT捕获二者通过特征拼接实现互补;依存句法解析与TomBERT模块提取的文本特征为$F_{\text{syntax}}\in\mathbb{R}^{d_3}$,该特征通过依存句法解析树构建文本的语法结构信息,并结合TomBERT模型增强语义表示的深度。最终的特征融合通过以下公式实现:

$$\begin{aligned}F_{\text{final}}=&\alpha\cdot\text{LayerNorm}(F_{\text{text}})+\\&\beta\cdot\text{LayerNorm}\big(F_{\text{image}}\big)+\\&\gamma\cdot\text{LayerNorm}\big(F_{\text{syntax}}\big),\end{aligned} \tag{18}$$

其中α、β、γ为可学习的权重参数,通过反向传播优化,以确保各模态特征在融合过程中的贡献度能够根据任务需求动态调整。LayerNorm(层归一化)用于对特征进行标准化处理,避免特征尺度不一致导致的融合偏差。此外,为了进一步增强特征间的交互,融合后的特征F_{final}会经过一个多头自注意力机制(Multi-Head SeIf-Attention, MHSA)模块以捕获跨模态特征之间的高阶关联:

$$F'_{\text{final}}=\text{MHSA}(F_{\text{final}})。 \tag{19}$$

最终,F'_{final}将被输入到任务特定的输出层(如全连接层或Softmax层),用于生成图像描述、情感分类或其他下游任务的预测结果。这种融合策略不仅解决了传统多模态模型中局部特征与全局特征不平衡的问题,还通过语法结构信息的引入提升了语义理解的准确性,从而显著提高了模型的整体性能。

3 实验

3.1 数据集

为了评估全局局部交互情感分析模型的效果,使用Zhang等[32]和Wang等[33](2018)注释的两个基准数据集进行面向目标的多模态情感分类。即Twitter-2015和Twitter-2017。在多模态情感分析领域,Twitter-15和Twitter-17数据集作为广泛使用的基准数据集,为研究者提供了丰富的多模态数据资源。Twitter-15数据集包含约15 000条推文,涵盖了积极、消极和中性三类情感标签。每条推文不仅包含文本内容,还附带了相关的图像或视频信息,这种多模态特性为研究文本与视觉信息之间的情感关联提供了重要基础。Twitter-17数据集作为Twitter-15的扩展版本,进一步扩充了样本规模至约17 000条推文,并对情感类别进行了更细粒度的划分,包括愤怒、快乐、悲伤等类别。这种细粒度的标注支持了更复杂的情感分析任务,例如方面级情感分类和情感强度预测。此外,Twitter-17保留了图像和视频的多模态特

性,为模型提供了更丰富的上下文信息,有助于提升情感识别的鲁棒性和准确性。

这些数据集的研究不仅推动了多模态情感分析技术的发展,还为理解文本与视觉信息之间的交互机制提供了重要的实证基础。例如,通过分析模型在Twitter-15和Twitter-17上的表现,研究者能够揭示多模态融合策略的有效性,并进一步探索模态间的情感一致性、互补性以及噪声干扰等问题。由于篇幅限制,本文仅在表1中展示了Twitter-15和Twitter-17数据集的基本统计数据,包括样本数量、情感类别分布以及多模态数据的比例等信息。

3.2 实验参数

为了获取模型的最优参数集合,在进行模型的优化时,需要调节的超参数以及获得最优模型时的超参数如表2所示。

表1 数据集统计表

Table 1 Statistics of dataset

数据	Twitter-15							Twitter-17						
划分	POS	NEG	Neutral	Total	AvgTargets	Words	Length	POS	NEG	Neutral	Total	AvgTargets	Words	Length
Train	928	368	1 883	3 179	1.348	9 023	16.72	1 508	416	1 638	3 562	1.410	6 027	16.21
Dev	303	149	670	1 122	1.336	4 238	16.74	515	144	517	1 176	1.439	2 922	16.37
Test	317	113	607	1 037	1.354	3 919	17.05	493	168	573	1 234	1.450	3 013	16.38

表2 实验参数设置

Table 2 Experimental parameter settings

类别	参数名称	参数值	说明
硬件环境	GPU	NVIDIA RTX 3090	使用单卡训练
	CUDA版本	11.8	与pyTorch2.0.1兼容
软件环境	pyTorch版本	3.9.21	开发语言版本
	pyTorch版本	2.01+cu118	深度学习框架版本
模型配置	文本编辑器	TomBERT	基于BERT的句法增强文本编码器
	图像编码器	ViT-B/6	VisinTransformer模型,输入图像尺寸为224×224
	多模态融合模块	FusinGCN	基于图卷积网络的多模态特征

GLIEAM时间复杂度:$O(N^2d+T^2d+TNd+(T+N)d)$,其中N为图像被划分为N个patch,d为嵌入维度,T为输入文本的长度。

3.3 基准模型

随着互联网和社交媒体的广泛普及,多模态数据的情感分析也得到了显著发展。在多模态情感分析领域,出现了众多具有优异表现的模型。本文实验选择了近年来表现突出的几种模型进行对比分析。将提出的模型与现有的几种竞争性模型进行比较,这些模型依据所使用的模态可分为视觉基、文本基及文本-视觉融合基的模型。

视觉基模型:

Res-Target:一种基于残差网络(ResNet)的目标检测模型,结合了残差学习和目标检测任务的特点。

文本基模型:

AE-LSTM:自编码器(Autoencoder)与LSTM的结合,用于文本序列的编码-解码,支持文本生成或降噪。

MGAN(Multi-Generator Adversarial Network):多任务生成对抗网络(Multi-task GAN),通过多任务学习生成多样化文本,提升生成质量。

BERT:预训练双向Transformer模型,通过掩码语言建模捕捉上下文语义,广泛应用于文本分类、问答等任务。

BERT+BL:BERT与双向LSTM(BiLSTM)的混合架构,增强序列建模能力,适用于需要长距离依赖的任务。

BERT-Pair-QA:基于BERT的问答模型,处理成对输入(如问题-答案对),优化语义匹配与推理能力。

文本-视觉融合基模型:

MIMN(Multi-Interactive Memory Network):MIMN是一种多模态交互记忆网络,通过记忆

网络存储和更新视觉与文本特征,实现模态间的深度融合。其核心在于多层交互机制,该机制能够动态捕捉并增强视觉与文本之间的关联性。

ESAFN(Entity-sensitive Attention and Fusion Network):一种情感语义感知融合网络,专注于动态融合视觉与文本特征以提升情感分析性能。它通过引入情感语义注意力模块,有效捕捉与情感相关的特征,并结合自适应融合策略平衡不同模态的贡献。

TomBERT:扩展的BERT架构,专门用于融合文本与视觉对象特征。它利用预训练的BERT提取文本特征,并结合CNN提取的视觉对象特征,通过注意力机制实现模态间的有效融合。

SalienCyBERT(Recurrent Attention Network for Target-Oriented Multimodal Sentiment Classification):一种显著性感知的跨模态BERT模型,通过视觉显著性检测模块突出图像中的重要区域,并结合跨模态注意力机制优化特征对齐。

Res-MGAN(Combining Residual Networks and Multi-Generator Adversarial Networks):结合残差网络与多任务GAN的跨模态模型,支持图像到文本的生成与对齐。

Res-BERT(A Model that Integrates ResNet Visual Features with BERT Text Features):融合ResNet视觉特征与BERT文本特征的模型,用于视觉-语言联合任务(如视觉问答)。

Res-BERT+BL(A Model that Integrates ResNet Visual Features, BERT Text Features, and Bidirectional LSTM):在Res-BERT基础上引入双向LSTM,增强对时序多模态数据的建模能力。

3.4 实验结果对比分析

实验结果如表3所示,在Twitter-15和Twitter-17两个广泛使用的社交媒体情感分析数据集上,GLIEAM模型通过多模态特征融合与技术创新,在性能上显著超越了现有的基线模型,展现了其在复杂多模态情感分析任务中的优越性。具体而言,相较于纯视觉模型Res-Target,GLIEAM的核心突破在于其创新性地引入了视觉Transformer(ViT)与生成式预训练模型(GPT)的串联架构,通过ViT将输入图像编码为视觉特征,并利用GPT生成具有上下文感知能力的图像描述文本。随后,通过多头注意力机制将生成的文本描述特征与原始视觉特征进行深度拼接,从而实现了视觉与文本模态的高效融合。这一设计有效解决了传统视觉模型因缺乏文本信息而难以捕捉整体情感倾向的问题,尤其是在方面级情感分析任务中,能够同时兼顾局部细节与全局情感语义。

与纯文本模型(如BERT系列)相比,GLIEAM不仅通过依存句法解析技术增强了文本的语法结构理解能力,还结合了TomBERT模型进一步提升了文本特征的深层次语义表达能力。更重要的是,GLIEAM充分利用了多模态数据的互补性,通过视觉与文本特征的协同作用,弥补了纯文本模型对图像上下文信息依赖不足的问题。这种多模态融合策略使得模型在处理包含丰富视觉信息的社交媒体数据时,能够更准确地捕捉用户的情感倾向。

在与其他多模态模型的对比中,GLIEAM的优势尤为显著。例如,与依赖传统卷积神经网络(CNN)提取视觉特征的TomBERT模型相比,GLIEAM采用了T2T-ViT架构,能够同时提取图像的局部细节特征与全局语义信息,并结合小波变换与非局部均值去噪技术,显著提升了图像信息的质量与鲁棒性。此外,与SalienCyBERT模型相比,GLIEAM通过创新的图像描述生成策略与噪声处理机制,在复杂场景下表现出更强的鲁棒性,尤其是在噪声干扰较大的数据环境中。

与生成式对抗网络模型MGAN相比,GLIEAM摒弃了传统的LSTM架构,转而采用基于Transformer的跨模态对齐机制,实现了视觉与文本特征的高效交互,避免了LSTM在长距离依赖建模中的表达局限性。尽管在Twitter-15数据集的Macro-$F1$指标上,GLIEAM略低于SalienCyBERT,这一差异可能源于特征融合权重的初始化或优化不足,但GLIEAM在噪声更多、模态交互更复杂的Twitter-17数据集中表现尤为突出,充分验证了其全局-局部特征融合策略、深度文本解析技术以及图像去噪机制的综合效能。

通过对GLIEAM模型在不同情感类别上的

表3 实验结果

Table 3 Experimental result

Model	Twitter-15		Twitter-17		Params(M)	FLOPs(G)
	Acc/%	Macro-*F*1/%	Acc/%	Macro-*F*1/%		
Visual						
Res-Target	59.90	46.60	58.60	54.00	23.50	3.80
Text						
AE-LSTM	70.30	63.43	61.67	59.57	12.80	2.10
MGAN	71.17	64.21	64.75	61.46	15.60	2.60
BERT	74.30	70.00	68.90	66.10	110.00	17.50
BERT+BL	74.30	70.00	68.90	66.10	115.00	19.20
BERT-Pair-QA	74.40	67.70	63.10	59.70	112.00	18.00
Visual+Text						
MIMN	71.80	65.70	65.90	63.00	85.60	18.30
ESAFN	73.40	67.40	67.80	64.20	92.40	19.80
TomBERT	77.20	71.80	70.50	68.00	120.70	25.60
SalienCyBERT	77.03	72.36	69.69	67.19	125.30	26.90
Res-MGAN	71.70	63.90	66.40	63.00	88.20	18.90
Res-BERT	73.58	68.74	67.18	64.29	125.80	24.30
Res-BERT+BL	75.00	69.20	69.20	66.50	138.00	24.90
GLIEAM(ours)	78.46	72.05	75.21	76.75	130.2	22.30

性能分析(表4),发现其表现出明显的类别差异性:积极情感识别效果最佳(Acc 82.10%,*F*1 80.30%),这主要得益于积极文本中包含更多明确的情感词汇(如“优秀”“满意”),这些词汇具有更强的语义区分度;消极情感次之(Acc 73.50%,*F*1 72.80%),其性能下降主要源于负面表达常采用隐晦措辞以及伴随图像的噪声干扰;中性情感表现相对较弱(Acc 70.20%,*F*1 69.10%),这与中性情感在语义和视觉特征上都缺乏显著判别性有关。跨模态注意力机制分析表明,积极情感判断中文本模态贡献度达65.00%,而消融实验证实图像去噪模块对消极情感识别尤为关键(移除后*F*1下降5.10%)。

表4 GLIEAM在不同情感类别上的性能表现

Table 4 Performance of GLIEAM in different emotion categories

情感类别	Acc/%	*F*1/%
积极	82.10	80.30
消极	73.50	72.80
中性	70.20	69.10

3.5 消融实验

为了验证全局局部交互情感分析模型(GLIEAM)中各个核心模块的有效性,我们设计了系统的消融实验,在Twitter-2015和Twitter-2017数据集上评估性能变化。实验目的是验证以下问题:全局特征增强模块(ViT-GPT图像描述生成与融合)对捕捉整体情感倾向的作用,图像特征增强模块(小波去噪、CNN与T2T-ViT特征提取)对噪声抑制和多尺度特征提取的贡献,语义特征增强模块(依存句法解析与TomBERT)对文本深层语义理解的影响,特征融合模块(加权融合与多头自注意力)对跨模态交互的重要性。实验结果如表5所示。

表5 消融实验结果

Table 5 Results of ablation experiment

模块	Twitter-2015		Twitter-2017	
	Acc/%	*F*1/%	Acc/%	*F*1/%
w/o 标注预处理	73.21	67.05	70.34	72.18
w/o 依存句法解析	75.82	69.45	72.63	73.92
w/o ViT-GPT融合	74.95	68.70	71.88	73.45
w/o Global	74.10	68.20	71.50	73.10
w/o Image Denoise	75.30	69.80	72.80	74.20
w/o T2T-ViT	76.20	70.10	73.40	75.00
w/ LSEM(BiLSTM)	75.60	69.00	72.10	73.50
GLIEAM(ours)	78.46	72.05	75.21	76.75

消融实验结果说明,移除标注优化(w/o标注预处理)导致Twitter-15和Twitter-17的Acc分别下降5.25%和4.87%,其中高熵样本

($H(y_i)>0.5$)的$F1$降幅达7.30%,显著高于其他模块消融的影响;相比之下,移除T2T-ViT或图像去噪主要影响Twitter-17的图像相关指标($F1$降1.75%),这验证了标注优化对文本模态稳定性的关键作用以及图像模块的模态特异性,而完整GLIEAM模型在所有测试场景下均展现出最优性能。移除依存句法解析(w/o 依存句法解析)导致Twitter-15和Twitter-17的F1分别下降2.60%和2.83%,其中复杂句式样本(如含否定或被动语态)的$F1$降幅达5.10%,凸显了语法结构对深层语义理解的重要性;而移除ViT-GPT融合(w/o ViT-GPT融合)引发更广泛的性能衰减(Acc下降3.51%/3.33%),视觉相关样本的$F1$骤降6.20%,远高于纯文本样本的1.80%降幅,证实了跨模态特征生成在多模态分析中的核心地位。全局与局部特征的互补性:移除全局模块(w/o Global)对Twitter-2015影响更大(Acc下降4.36%),因其数据更依赖整体场景语义;而移除局部模块(w/o T2T-ViT)对Twitter-2017影响更显著(Macro-$F1$下降1.75%),因其需细粒度分析多目标交互。这种互补性同样体现在模态交互层面——ViT-GPT融合模块通过全局视觉语义转换弥补了局部语法解析(w/o 依存句法解析)的不足,而依存句法解析则为跨模态对齐提供了文本侧的结构化约束。图像去噪的必要性:在低质量图像占比高的Twitter-2017中,未去噪(w/o ImageDenoise)导致Macro-$F1$下降2.55%,验证了小波变换与非局部均值对噪声抑制的有效性。消融实验证实了GLIEAM各模块的协同作用:全局特征增强模块为情感分析提供整体语境,图像特征增强模块通过去噪与多尺度提取提升鲁棒性,语法增强模块通过依存解析深化文本语义理解,跨模态生成模块(ViT-GPT)构建视觉与语言的语义桥梁,语法增强模块深化文本语义理解,动态融合模块实现跨模态精准对齐。未来工作将探索模块间的动态权重分配,以进一步优化模型效率与可解释性。

4 结论

针对多模态方面级情感分析领域中,现有方法普遍面临的模态融合不充分、图像噪声干扰显著、局部特征过度依赖等问题,本研究设计并实现了全局局部交互情感分析模型(GLobal-Local Interactive Emotion Analysis Model,GLIEAM)。通过ViT-GPT串联架构生成的图像描述文本,与原始文本模态形成语义互文矩阵,采用交叉注意力机制建立跨模态语义关联。该方法有效弥补了传统方法中视觉特征与文本特征简单拼接导致的语义鸿沟问题,创新性地引入小波变换与非局部均值(Non-Local Means)的混合去噪算法,提出基于T2T-ViT与CNN的协同特征提取框架。实验结果表明,该研究在两个公开数据集上均取得了显著的性能提升,验证了其有效性和合理性。然而,本研究仍存在一定局限性,未完全覆盖语音等动态时序模态。未来工作将拓展至视频流数据分析领域,研发时空协同的多模态情感计算框架,以提升模型在复杂场景下的泛化能力与实用性。另外,未来研究将针对标注数据质量问题,探索半监督学习与跨模态一致性校验的结合,例如通过注意力机制自动检测并修正模态间冲突的标签,同时引入领域自适应技术以降低对高质量标注数据的依赖。

参考文献:

[1] TALIB R, KASHIF M, AYESHA S, *et al*. Text Mining: Techniques, Applications and Issues[J]. *Int J Adv Comput Sci Appl*, 2016, **7**(11): 414-418. DOI: 10.14569/ijacsa.2016.071153.

[2] XU P, ZHU X T, CLIFTON D A. Multimodal Learning with Transformers: A Survey[J]. *IEEE Trans Pattern Anal Mach Intell*, 2023, **45**(10): 12113-12132.

[3] LIN J, WANG Y, XU Y, *et al*. Semi-IIN: Semi-Supervised Intra-Inter Modal Interaction Learning Network for Multimodal Sentiment Analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia: AAAI Press, 2025: 1411-1419. DOI: 10.1609/aaai.v39i2.32131.

[4] GUO X T, YU W, WANG X D. An Overview on Fine-grained Text Sentiment Analysis: Survey and Challenges [J]. *J Phys: Conf Ser*, 2021, **1757**(1): 012038. DOI: 10.1088/1742-6596/1757/1/012038.

[5] 罗渊贻, 吴锐, 刘家锋, 等. 面向情感语义不一致的多模态情感分析方法[J]. 计算机研究与发展, 2025, **62**(2): 374-382. DOI: 10.7544/issn1000-1239.202330199.
LUO Y Y, WU R, LIU J F, *et al*. Multimodal Sentiment Analysis Method for Sentimental Semantic Inconsistency

[J]. *J Comput Res Dev*, 2025, **62**(2): 374−382. DOI: 10.7544/issn1000-1239.202330199.

[6] 刘佳, 宋泓, 陈大鹏, 等. 非语言信息增强和对比学习的多模态情感分析模型[J]. 电子与信息学报, 2024, **46**(8): 3372−3381. DOI: 10.11999/JEIT231274.
LIU J, SONG H, CHEN D P, *et al*. A Multimodal Sentiment Analysis Model Enhanced with Non-verbal Information and Contrastive Learning[J]. *J Electron Inf Technol*, 2024, **46**(8): 3372−3381. DOI: 10.11999/JEIT231274.

[7] 张淼萱, 张洪刚. 人脸表情识别可解释性研究综述[J]. 计算机学报, 2024, **47**(12): 2819−2851. DOI: 10.11897/SP.J.1016.2024.02819.
ZHANG M X, ZHANG H G. A Survey on Interpretability of Facial Expression Recognition[J]. *Chin J Comput*, 2024, **47**(12): 2819−2851. DOI: 10.11897/SP.J.1016.2024.02819.

[8] 王宁, 武芳宇, 赵宇轩, 等. 基于集成学习与多模态大语言模型的图文情感分析方法[J/OL]. 计算机工程与应用, 2025: 1−11. (2025−06−05). https://kns.cnki.net/kcms/detail/11.2127.tp.20250604.1653.002.html.
WANG N, WU F Y, ZHAO Y X, *et al*. Image-text Sentiment Analysis Method Based on Ensemble Learning and Multimodal Large Language Model[J/OL]. *Comput Eng Appl*, 2025: 1−11. (2025−06−05). https://kns.cnki.net/kcms/detail/11.2127.tp.20250604.1653.002.html.

[9] MIKOLOV T, CHEN K, CORRADO G, *et al*. Efficient Estimation of Word Representations in Vector Space[C]// Proceedings of the 1st International Conference on Learning Representations (ICLR 2013). Scottsdale, AZ, USA: ICLR, 2013.

[10] 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. Minneapolis: Association for Computational Linguistics, 2019: 4171−4186. DOI: 10.18653/v1/N19−1423.

[11] RAFFEL C, SHAZEER N, ROBERTS A, *et al*. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer[J]. *J Mach Learn Res*, 2020, **21**(140): 1−67.

[12] WANG Y, LIU J, CHEN Z, *et al*. Syntax-aware BERT with Graph Propagation for Dependency Parsing[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: ACL, 2023: 123−135.

[13] XU H, KOEHN P. Zero−Shot Cross−Lingual Dependency Parsing through Contextual Embedding Transformation[C]//Proceedings of the Second Workshop on Domain Adaptation for NLP. Kyiv: Association for Computational Linguistics, 2021: 204−213. DOI: 10.18653/v1/2021.adaptnlp−1.21.

[14] ZHANG R, WANG L, SUN H, *et al*. Dependency-driven Semantic Role Labeling with Heterogeneous Graph Networks[C]//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023: 4567−4575. DOI: 10.1609/aaai.v37i11.26520.

[15] PLESTED J, GEDEON T. Deep Transfer Learning for Image Classification: A Survey[EB/OL]. (2022−05−20) [2025−08−27]. https://doi.org/10.48550/arXiv.2205.09904.

[16] LOWE D G. SIFT: A Retrospective on the Scale-Invariant Feature Transform[J]. *J Comput Vis Image Process*, 2022, **15**(3): 45−60.

[17] YANG D, LI X H, LI Z, *et al*. Prompt Fusion Interaction Transformer for Aspect−based Multimodal Sentiment Analysis[C]//2024 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2024: 1−6. DOI: 10.1109/ICME57554.2024.10687885.

[18] CALDERON-RAMIREZ S, YANG S X, ELIZONDO D. Semisupervised Deep Learning for Image Classification with Distribution Mismatch: a Survey[J]. *IEEE Trans Artif Intell*, 2022, **3**(6): 1015−1029. DOI: 10.1109/TAI.2022.3196326.

[19] JIE Z ,GANQU C ,SHENGDING H, *et al*. Graph Neural Networks: A Review of Methods and Applications[J].*AI Open*, 2020, 157−81. DOI: 10.1016/J.AIOPEN.2021.01.01.

[20] ZHANG Y, LI X, WANG Q. Dynamic Wavelet Learning for Image Denoising Based on Convolutional Neural Networks[J]. *IEEE Trans Image Process*, 2022, **31**: 4567−4580. DOI: 10.1109/TIP.2022.3185597.

[21] CHEN H, LIU Z, SUN T. Wavelet-Spatial Domain Feature Fusion for Image Restoration Using Residual Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 12345−12354. DOI: 10.1109/CVPR52729.2023.01191.

[22] WANG L, ZHOU Y, ZHANG R. Wavelet-Guided Attention Mechanism in Vision Transformers for Image Enhancement[J]. *IEEE Trans Multimedia*, 2023, **25**(3): 102−115. DOI: 10.1109/TMM.2022.3205024.

[23] LIU Z, WANG Q, ZHANG Y. Deep Feature Similarity for Non-Local Means Image Denoising[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 12345−12354.

[24] GEETHA R, THILAGAM T, PADMAVATHY T. Retraction Note: Effective Offline Handwritten Text Recognition Model Based on a Sequence-to-sequence Approach with CNN-RNN Networks[J]. *Neural Comput Appl*, 2024, **36**(24): 15227. DOI: 10.1007/s00521-024-10104-6.

[25] BALTRUSAITIS T, AHUJA C, MORENCY L P. Multi-

modal Machine Learning: A Survey and Taxonomy[J]. *IEEE Trans Pattern Anal Mach Intell*, 2019, **41**(2): 423-443. DOI: 10.1109/tpami.2018.2798607.

[26] CAI Y J, LI X G, ZHANG Y Y, *et al*. Multimodal Sentiment Analysis Based on Multi-layer Feature Fusion and Multi-task Learning[J]. *Sci Rep*, 2025, **15**(1): 2126. DOI: 10.1038/s41598-025-85859-6.

[27] XU M Q, MA Q T, ZHANG H J, *et al*. MEF-UNet: an End-to-end Ultrasound Image Segmentation Algorithm Based on Multi-scale Feature Extraction and Fusion[J]. *Comput Med Imag Graph*, 2024, **114**: 102370. DOI: 10.1016/j.compmedimag.2024.102370.

[28] WANG A, CHEN H, LIN Z J, *et al*. Rep ViT: Revisiting Mobile CNN from ViT Perspective[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2024: 15909-15920. DOI: 10.1109/CVPR52733.2024.01506.

[29] MIAH M S U, KABIR M M, SARWAR T B, *et al*. A Multimodal Approach to Cross-lingual Sentiment Analysis with Ensemble of Transformer and LLM[J]. *Sci Rep*, 2024, **14**(1): 9603. DOI: 10.1038/s41598-024-60210-7.

[30] YU J F, JIANG J. Adapting Bert for Target-oriented Multimodal Sentiment Classification[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019. Macao, China: Ijcai. 2019: 5408-5414.

[31] BOLYA D, FU C Y, DAI X L, *et al*. *Hydra* Attention: Efficient Attention with Many Heads[M]//Computer Vision-ECCV 2022 Workshops. Cham: Springer Nature Switzerland, 2023: 35-49. DOI: 10.1007/978-3-031-25082-8_3.

[32] ZHANG Q, FU J, LIU X, *et al*. Adaptive Co-Attention Network for Named Entity Recognition in Tweets[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI Press, 2018: 5674-5681. DOI: 10.1609/aaai.v32i1.11962.

[33] WANG B, LU W. Learning Latent Opinions for Aspect-Level Sentiment Classification[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI Press, 2018: 5537-5544.