山西大学学报(自然科学版)48(2):391—399,2025
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024170

基于图文交互和深层特征融合的多模态讽刺检测方法

王素格[1,2*],李鹏帅[1],李旸[3]

(1.山西大学 计算机与信息技术学院,山西 太原 030006;
2.山西大学 计算智能与中文信息处理教育部重点实验室,山西 太原 030006;
3.山西财经大学 金融学院,山西 太原 030006)

摘 要:多模态讽刺检测是在多模态场景下识别用户的讽刺言语。现有的多模态讽刺识别方法大多对编码之后的特征直接融合,并未关注图像和文本的深层特征以及图像和文本之间的交互。针对上述问题,本文提出基于图文交互和深层特征融合的多模态讽刺检测框架。首先,使用具有长文本能力的对比语言-图像预训练模型(Contrastive Language-Image Pre-training model with Long text capability,Long-CLIP),对图像和文本进行编码,获得图像和文本的特征表示,进一步引入跨模态注意力机制,建立图像和文本特征之间的交互表示;其次,利用多个卷积神经网络从不同角度分别获得图像和文本的深层特征表示,并利用Key_less Attention机制融合图像和文本的深层特征;最后,通过多层感知机实现多模态讽刺检测。在公开的数据集MMSD2.0上进行实验,实验结果表明所提出的方法与现有的最先进的基线模型相比,Acc和*F*1分别提升0.33%和0.15%,表明了图文交互及深层特征可以提升多模态讽刺检测的性能。

关键词:Long-CLIP模型;图文融合;深层特征表示;注意力机制
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2025)02-0391-09

Multimodal Sarcasm Detection Method Based on Image-text Interaction and Deep Feature Fusion

WANG Suge[1,2*], LI Pengshuai[1], LI Yang[3]

(1. *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China;*
2. *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan* 030006*, China;*
3. *School of Finance, Shanxi University of Finance and Economics, Taiyuan* 030006*, China*)

Abstract: Multimodal sarcasm detection is to recognize users' sarcastic speech in multimodal scenarios. Most of the existing multi-modal sarcasm recognition methods directly fuse the features after encoding, and do not pay attention to the deep features of image and text as well as the interaction between image and text. To address the above problems, this paper proposes a multimodal sarcasm detection framework based on image-text interaction and deep feature fusion. Firstly, using the Contrastive Language-Image Pre-training model with Longtext capability (Long-CLIP) model, the image and text are encoded to obtain the feature representations of the image and text, and further the cross-modal attention mechanism is introduced to establish the interactions between the image and text features. Secondly, multiple convolutional neural networks are used to obtain the deep feature representations of the image and text from different perspectives, and the Key_less Attention mechanism is utilized to fuse the deep features of image and text deep features. Finally, multi-layer perceptron is used to realize multi-modal sarcasm detection. The experimental results show that

收稿日期:2024-10-22;**接受日期**:2024-12-23
基金项目:国家自然科学基金(62376143;62076158;62106130);山西省高等学校科技创新项目(2021L284)
* **通信作者**:王素格(1964—),女,河北定州人,博士,教授,研究方向为自然语言处理和情感计算。E-mail:wsg@sxu.edu.cn
引文格式:王素格,李鹏帅,李旸.基于图文交互和深层特征融合的多模态讽刺检测方法[J].山西大学学报(自然科学版),2025,48(2):391-399. DOI:10.13451/j.sxu.ns.2024170.

the proposed method improves Acc and *F*1 by 0.33% and 0.15% respectively, compared with the existing state-of-the-art baseline model, on the publicly available dataset MMSD2.0, which demonstrates that the image-text interaction and deep features can improve the performance of multimodal sarcasm detection.

Key words: Long-CLIP model; image and text fusion; deep feature representation; attention mechanism

0 引言

讽刺是言语反讽的一种语言现象,以其固有的微妙和复杂性,在人类交流中发挥着重要作用,人们经常利用讽刺传达嘲弄或隐藏的意图或情感,对讽刺进行检测,将有助于识别用户隐式的真实情感和意见。随着多媒体技术的发展,用户经常使用文字-图像组合来传达他们的信息,因此,识别多模态场景(如文本和图像模态)中的讽刺越来越受到研究者的关注。如表1所示,有关描述天气的多模态讽刺图-文示例,其中0表示非讽刺,1表示讽刺。

表1 有关描述天气的多模态讽刺图-文示例

Table 1 Example of multimodal sarcasm image-text for describting weather

模态	图-文示例	标签
图像		0
文本	What a wonderful weather!	0
多模态	图像与文本综合	1

对于表1中仅通过文本“What a wonderful weather!”,可以认为是感叹天气很好,不具有讽刺的意味。同样的,文本对应的图像是一张阴雨天的天空照片,仅通过该照片,无法识别出其中的讽刺情感。但是同时关注图像和文本信息时,文本描述的“wonderful weather”与图像中展示的阴雨天形成反差,可以判别其具有讽刺意味。因此,建立图像和文本之间信息的交互显得尤为重要。

现有多模态反讽检测的研究者[1-3]使用图像和文本单独编码,使得模型难以捕捉文本和图像之间复杂的交互信息,尽管Qin等[4]利用对比语言-图像预训练(Contrastive Language-Image Pre-training,CLIP)模型[5]对反讽样本进行编码,并在多模态讽刺检测数据集(Multimodal Sarcasm Detection Dataset 2.0,MMSD2.0)反讽基准上获得了不错的效果,但原始CLIP仅能直接对齐文本和图像,缺乏对于图像和文本深层特征的利用。随着多模态大模型的兴起,也有研究者尝试使用多模态大模型进行多模态讽刺检测,Tang等[6]通过微调生成式的多模态大模型的框架,取得了较好的讽刺检测性能,但是消耗时间和硬件资源较多。

针对上述问题,本文提出一种基于图文交互和深层特征融合的多模态讽刺检测方法,目的是在较低消耗的条件下,充分利用图像和文本包含的信息,提升多模态讽刺检测的性能。首先,对于图文信息交互模块,引进交叉注意力机制,旨在通过图像和文本之间信息的交互,使得模型可以更好地建模图像和文本之间的相互关系。在此基础上,引入卷积神经网络(Convolutional Neural Network,CNN),获取单模态的深层特征表示。然后使用构建的注意力模块,实现特征的充分融合。最后,使用多层感知机(Multilayer Perceptron,MLP),得到更高层次的抽象特征,进一步将文本、图像和融合特征得分相加,实现多模态讽刺检测。本文提出的方法在MMSD2.0反讽数据集上进行了大量的实验,实验结果表明,该模型与目前最先进的方法相比取得了较好的性能。本文的主要贡献如下:

(1)本文提出了一种基于图文交互和深层特征融合的多模态讽刺检测方法,引入跨模态注意力机制,建立图像和文本特征之间的交互表示,并从不同角度分别构建了图像和文本的深层特征表示,以充分利用图像和文本的融合特征以及图像和文本的单模态深层特征,提升多模态讽刺检测的性能。

(2)本文在MMSD2.0数据集上进行了实验,实验结果表明所提出的模型优于现有基线

模型,验证了本文方法的有效性。

1 相关工作

传统的讽刺检测任务旨在识别用户的情绪,并从文本中检测是否存在讽刺[7-9]。随着社交媒体的发展,所产生的多模态数据激增,多模态讽刺检测逐渐受到人们的关注。多模态讽刺检测的方法可以大致分为基于不同模态单独编码的方法、基于多模态Transformers的方法和基于多模态大模型的方法。

基于单模态单独编码的方法:单模态单独编码的方法侧重于为不同的模态使用不同的编码器,再融合多模态进行有效表示,以满足模态内部和模态之间的不一致性。Schifanella等[10]提出的方法将编码器提取的视觉和文本特征进行连接,验证了视觉形态有助于发现社交媒体中的讽刺,但是他们没有考察这种现象的本质,忽略了图像和文本之间存在潜在的联系。Cai等[11]提出了一种分层融合网络,该方法在不同的融合层融合编码后的文本、图像以及图像属性,但是在对上述三种信息编码时仍旧采取独立编码的方式,未能充分利用图像和文本信息之间的联系。Xu等[1]通过构建分解关系网络进一步建模视觉和文本模式之间的共同性和不一致性,该方法通过表示推文的图像和文本的共性和差异,隐式地对跨模态对比信息进行建模。Pan等[2]提出了一个基于Transformer的双向编码器(Bidirectional Encoder Representations from Transformers,BERT)和残差网络(Residual Network,ResNet)的架构,通过在模态内和模态间引入注意力机制解决模态内和模态间的不一致问题。Liang等[3]利用图像属性与文本词之间的相似度,提出了生成交叉模态图方法识别讽刺检测中的重要线索,更好地捕获了不同模态之间的关系。Tian等[12]提出动态路由Transformer网络(Dynamic Routing Transformer Network,DynRT-Net),该网络通过调整动态路径来适应多模态样本条件下图像和文本之间的分层共关注,能够动态捕获跨模态不一致。

基于多模态Transformers的方法:多模态Transformers可以将文本和图像编码到公共特征空间,通过该方法能够有效地识别和理解图像与文本之间的关系。Wang等[13]为了对齐文本特征和图像特征,通过设计桥连接层,将图像特征从ResNet空间映射到BERT空间,通过该方法证明编码到公共特征空间确实对于多模态讽刺检测有帮助。Qin等[4]提出了一个基于CLIP的讽刺检测框架多视角CLIP(Multi-view CLIP),虽然通过CLIP编码的文本特征和图像特征在同一空间中,但该方法未考虑单模态自身深层特征的作用。Chen等[14]引入了交互式CLIP和记忆增强预测器(Interactive CLIP and Memory-Enhanced Predictor,InterCLIP-MEP)框架,通过视觉信息和文本信息的交互,实现更可靠的多模态讽刺检测,但是该方法利用的仍旧只有浅层的特征。

基于多模态大模型的方法:随着大模型的发展,Yang等[15]发布了一个包含多模态讽刺检测任务的基准,他们对各种大模型进行了基准测试,如第四代生成式预训练Transformer模型(Generative Pre-trained Transformer 4,GPT-4),大语言模型Meta AI(Large Language Model Meta AI,LLaMA),引导式语言-图像预训练(Bootstrapping Language-Image Pre-training,BLIP)模型等。Tang等[6]将讽刺检测看成生成任务,尝试使用大模型微调的方式实现多模态讽刺检测,该方法虽然取得了先进的性能,但是在微调大模型的过程中,需要消耗较多的时间和资源。

针对以上工作的不足,本文提出了一种基于图文交互和深层特征融合的多模态讽刺检测方法,该方法利用具备长文本能力的对比语言-图像预训练模型(Contrastive Language-Image Pre-training model with Long-Text capability,Long-CLIP),将文本和图像编码到公共的特征空间中,然后设计一个图像和文本信息之间的交互表示模块,进一步,设计深层特征表示模块,从不同的角度获得图像和文本的深层特征表示。对于得到的图像和文本的深层特征,设计了图像和文本特征的融合模块,将两者充分融合。另外,设计多层感知机模块,获得更高层次的抽象特征,实现多模态场景下的讽刺检测。

2 基于图文交互和深层特征融合的多模态讽刺检测框架

对于多模态讽刺检测,本文主要关注图像

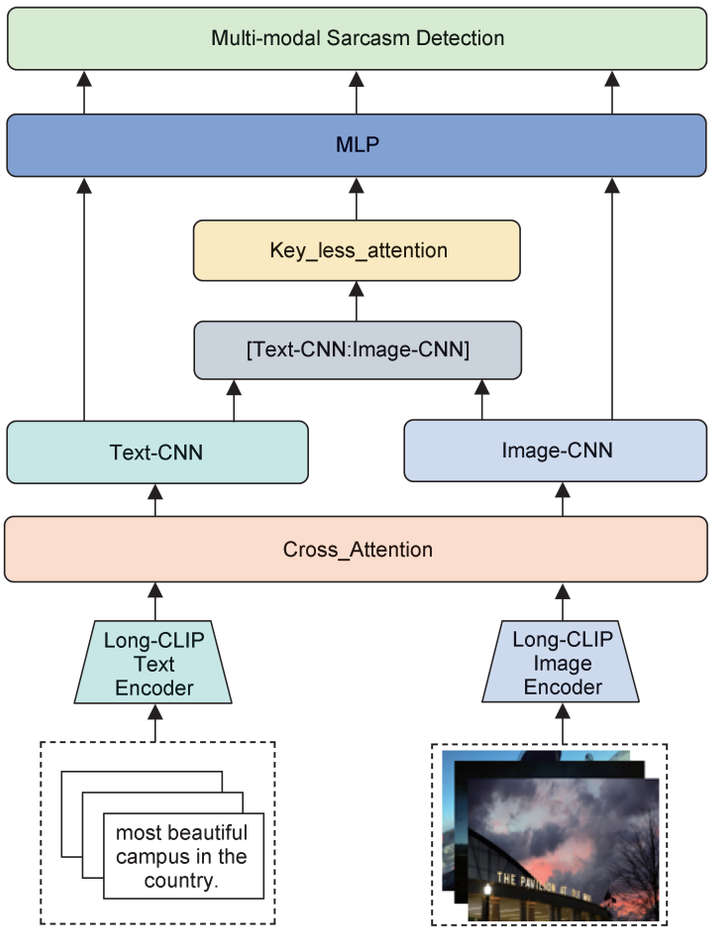

图1 基于图文交互和深层特征融合的多模态讽刺检测框架

Fig. 1 The multimodal sarcasm detection framework based on image-text interaction and deep feature fusion

和文本两种模态信息,提出了基于图文交互和深层特征融合的讽刺检测框架,如图1所示。

该框架的核心在于图像和文本之间的信息交互以及深层特征的获取。框架由以下三部分组成:图文信息交互融合(2.1节)、深层特征表示(2.2节)、图文特征融合(2.3节)。

2.1 图文信息交互融合

图文信息交互融合模块的目的是建立图像和文本之间的信息交互,使得模型能够更有效地建模跨模态数据之间的关系。对于图像和文本信息,本文使用Long-CLIP对图像和文本进行编码,得到图像和文本的浅层特征表示T_S和I_S,并引入图文交互模块,分阶段建立图像和文本之间的信息交互。

第一个阶段,将文本特征(T_S)作为查询(Query),而图像特征(I_S)作为键(Key)和值(Value),通过交叉注意力(Cross_Attention)函数得到的特征标记为Att_{ti}。

$$Att_{ti}=\text{Cross\_Attention}(T_S,I_S,I_S)。\tag{1}$$

同时,为保留原始图像中包含的全局信息,将I_S与Att_{ti}做残差连接,得到新的图像特征I_S,如公式(2)所示。

$$I_S=I_S+Att_{ti}。\tag{2}$$

第二个阶段,类似地,将图像特征(I_S)作为查询(Query),文本特征(T_S)作为键(Key)和值(Value),构建Cross_Attention,得到特征标记为Att_{ti},如公式(3)所示。

$$Att_{ti}=\text{Cross\_Attention}(I_S,T_S,T_S)。\tag{3}$$

类似公式(2),为保留原始文本特征中包含的信息,将T_S与Att_{ti}做残差连接,得到新的文本特征T_S,如公式(4)所示。

$$T_S=T_S+Att_{ti}。\tag{4}$$

2.2 深层特征表示

经过Long-CLIP编码之后的特征,虽然已经具有了丰富的语义信息,但是,编码后的特征向量的表达能力还需要进一步增强。现有的

模型大多未考虑对于编码之后的特征再进行处理,以实现更深层次的特征表示。因此本文通过使用多个不同的卷积神经网络(CNN),进一步改变编码之后特征的分布和组合方式,从多个不同角度对图像和文本的深层特征建模,使模型能够从不同角度获取图像和文本的深层特征,提升模型对讽刺检测的深刻理解。具体地,对于文本和图像特征分别经过卷积核为1、3和5的卷积神经网络,得到不同视角下文本和图像的特征输出。如公式(5)、公式(6)所示。

$$T_{di}=\text{Text\_CNN}_i(T_S) \tag{5}$$

$$I_{di}=\text{Im age\_CNN}_i(I_S) \tag{6}$$

其中T_{di}和I_{di}分别表示使用第i个卷积核得到的文本和图像的深层特征表示,$i\in\{1,2,3\}$。为保证输出特征具有相同的维度,本文对输入特征进行了填充。

进一步,经过CNN得到的多个图像和文本特征,将它们全部执行拼接操作,得到具有深层特征的文本和图像表示。具体实现见公式(7)和(8),T_D和I_D分别表示图像和文本的深层特征表示。

$$T_D=[T_{d1}:T_{d2}:T_{d3}], \tag{7}$$

$$I_D=[I_{d1}:I_{d2}:I_{d3}]。 \tag{8}$$

2.3 图文特征融合

文本和图像模态的特征融合是多模态讽刺检测的关键技术。传统的特征融合方法是将图像和文本直接拼接,其跨模态特征融合效果不理想。为使得融合充分,将图像和文本的深层特征再拼接之后再使用注意力模块。首先将得到的图像和文本深层特征连接,随后,将初步连接的特征F_p传入注意力机制中,得到输出F。由于Key_less_attention是由Long等[16]提出的,它已经在融合多模态特征表示上被证实其有效性,在本文中,通过使用一个Key_less_attention,将图像和文本深层特征充分融合。具体过程如公式(9)—公式(11)所示。

$$F_p=[T_D:I_D], \tag{9}$$

$$p_t,p_v=\text{soft max}(W(T_D,I_D)+b), \tag{10}$$

$$F=p_tT_D+p_vI_D, \tag{11}$$

其中T_D、I_D和F分别为深层特征的文本表示、图像表示以及融合表示。

为了使三种特征实现更加有效地表示,本文设计特征转换模块,该模块基于多层感知机实现。MLP是由多个线性层组成,在每个线性层之后,添加线性整流函数(Rectified Linear Unit,ReLU)函数,其目的是确保网络具有足够的非线性能力捕捉输入数据中的复杂特征,并且在训练过程中有效地传播梯度。对于文本的特征转换的计算过程如公式(12)和(13)所示,类似地,可以得到图像表示I_D和融合表示F。

$$Y_{T_D}=W_{T_D}X_{T_D}+b, \tag{12}$$

$$\text{ReLU}=\max(0,Y_{T_D}), \tag{13}$$

其中X_{T_D}表示经过MLP之前的文本表示T_D,Y_{T_D}是经过MLP之后的文本特征。

最终,利用多层感知机获得具有更高层次的抽象特征,经过得分函数得到最终的得分,用于讽刺检测。

2.4 损失函数

由于多模态讽刺检测是一个分类任务,在训练过程中采用交叉熵损失。在整个训练过程中对文本特征(T_D)、图像特征(I_D)和融合之后的特征(F)分别使用交叉熵损失,并对三者相加进行联合优化,计算过程如公式(14)所示。

$$L=\sum_{i\in\{Y_{T_D},Y_{I_D},Y_F\}}(y^i\log(\hat{y}^i)+(1-y^i)\log(1-(\hat{y}^i))), \tag{14}$$

其中y^i表示真实标签,$\hat{y}^i$表示经过得分函数后,最终的预测概率。

3 实验结果与分析

3.1 数据集

本文使用Qin等[4]提供的MMSD2.0基准数据集,该数据集是在Cai等[11]提出的数据集MMSD基础上,建立的改进版本,具体地,是在MMSD数据集上进一步删除虚假线索,纠正错误标记样本建立的。该数据集包含24 635条来自Twitter的英文数据,其中训练集数据19 816条,验证集数据2 410条,测试集数据2 409条。为直观了解数据集中数据的分布情况,在表2中给出了相关的统计信息。

3.2 评价指标与实验设置

本文参照Qin等[4]的评价指标,采取Acc、P、R、$F1$作为评价指标,其中Acc指Accuracy即准确率,P(Precision)是精确率,R(Recall)是召

表2 MMSD2.0数据分布情况

Table 2 The data distribution for MMSD2.0

Datasets	Split	Positive	Negative	Total	MaxLen	MinLen	AvgLen
MMSD2.0	Train	9 576	10 240	19 816	66	1	13.42
	Valid	1 042	1 368	2 410	55	4	13.64
	Test	1 037	1 372	2 409	52	4	13.52

回率,$F1$(*F*1 Score)是$F1$分数。本文采用Long-CLIP-L作为编码器,对输入的文本和图像进行编码,使用AdamW作为优化器来优化模型中的参数,模型的学习率设置为5×10^{-4},整个训练过程持续30轮,整个模型在单个英伟达NVIDIA 4090 GPU上进行训练,并在单个NVIDIA 4090 GPU上进行测试。

3.3 基线模型

根据之前的工作,本文将提出的方法与MMSD2.0上的讽刺检测的单模态方法和多模态方法进行比较,下面是选取的基线模型。

(1)纯文本的方法

文本卷积神经网络(Text Convolutional Neural Network,TextCNN):该方法是一种基于卷积神经网络的文本分类网络[17]。

双向长短期记忆网络(Bidirectional Long Short-term Memory,BiLSTM):该方法是一种用于文本分类的双向长短期记忆网络[18]。

自匹配讽刺检测模型(Self-matching Sarcasm Detection model,SMSD):该方法是一种用于讽刺检测的低秩双线性池自匹配网络[19]。

稳健优化的BERT预训练方法(Robustly Optimized BERT Pretraining Approach,RoBERTa):该方法是一个鲁棒优化的BERT预训练语言模型[20]。

聊天生成语言模型(Chat Generative Language Model,ChatGLM):该方法是基于通用语言模型框架的开放式大语言模型,具有62亿个参数[21]。

LLaMA2-7B:该方法是一个具有70亿个参数的预训练的基础的大语言模型[22]。

(2)纯图像方法

ResNet:该方法利用池化层产生的图像嵌入来检测讽刺[23]。

视觉转换器(Vision Transformer,ViT):该方法是一种预训练的视觉Transformer模型[24]。

(3)多模态方法

层次融合模型(Hierarchical Fusion Model,HFM):该方法是一个多模态融合的分层网络[11]。

注意力增强的BERT(Attention-Augmented BERT,Att-BERT):该方法采用自注意和共注意机制对模内不协调和模间不协调进行建模[2]。

跨模态图卷积网络(Cross-modal Graph Convolutional Network,CMGCN):该方法是一种细粒度的跨模态图架构来捕获讽刺线索[3]。

分层一致性建模与知识增强(Hierarchical Congruity Modeling with Knowledge Enhancement,HKE):该方法使用基于分层图的框架,并结合图像标题等外部知识进行多模态讽刺检测[25]。

DynRT-Net:该方法是一种动态路由Transformer网络,用来从图像和文本中捕获讽刺线索[12]。

Multi-view CLIP:该方法利用基于CLIP的框架,从图像、文本和图像-文本交互三个视角进行多模态讽刺检测[4]。

基于多模态大模型的方法(Based on Multimodal Large Models,MLLM-Based):该方法是一个基于大语言模型的生成式多模态讽刺模型[6],通过微调的方式来适配多模态讽刺检测任务。

3.4 比较实验结果分析

本文将所提出的模型与基线模型在多模态讽刺数据集MMSD2.0上进行对比实验,其结果如表3所示。其中,最好的结果用粗体突出显示,“*”表示使用BLIP获得图像的标题,将得到的图像标题作为视觉信息输入给大模型。

由表3可以发现:

(1)与现有的基线模型相比,本文提出的基于图文交互和深层特征融合的讽刺检测方法,

表3 本文提出的方法与基线方法的比较实验结果

Table 3 The comparison of experimental results between the proposed method and the baseline method

数据模态	方法	Acc/%	*P*/%	*R*/%	*F*1/%
文本	TextCNN	71.61	64.62	75.22	69.52
	BiLSTM	72.48	68.02	68.08	68.05
	SMSD	73.56	68.45	71.55	69.97
	RoBERTa	79.66	76.74	75.70	76.21
	ChatGLM2-6B	78.41	78.15	78.65	78.23
	ChatGLM2-6B*	80.08	80.52	81.04	80.04
	LLaMA2-7B	82.52	82.15	82.46	82.27
	LLaMA2-7B*	**84.68**	**84.40**	**84.94**	**84.53**
图像	ResNet	65.50	61.17	54.39	57.58
	ViT	72.02	65.26	74.83	69.72
多模态	HFM	70.57	64.84	69.05	66.88
	Att-BERT	80.03	76.28	77.82	77.04
	CMGCN	79.83	75.82	78.01	76.90
	HKE	76.50	73.48	71.07	72.25
	DynRT-Net	71.40	71.80	72.17	71.34
	Multi-view CLIP	85.64	80.33	**88.24**	84.10
	MLLM-Based	86.43	**87.00**	86.30	86.34
	Ours	**86.76**	86.52	86.46	**86.49**

在Acc和*F*1值中取得了领先的性能,相较于目前最好的性能有0.33%和0.15%的提升。这表明,本文所建立的图文信息交互模块以及将深层特征应用于图-文多模态讽刺检测中是有效的。

(2)使用BLIP获得图像标题,并将图像标题作为视觉信息输入给大模型,在纯文本方法中,取得了最好的成绩,这也说明了同时关注文本信息和视觉信息能够更好地捕捉到讽刺信息,也证实了建立不同模态之间信息交互的必要性。

(3)本文提出的模型相较于仅使用文本的方法和仅使用图像的方法,在准确率(Acc)、精确率(*P*)、召回率(*R*)以及*F*1分数(*F*1)均有较大提升,因此认为通过本文设计的框架能够建立不同模态之间的信息交互,模型能够更有效地理解跨模态数据之间的关系。

(4)本文提出的模型相较于目前最先进的多模态方法(MLLM-Based)在准确率(Acc)、召回率(*R*)以及*F*1分数(*F*1)均有一定幅度的提升。因此本文认为在多模态任务中,通过建立图像和文本之间的信息交互,同时利用图像和文本的深层特征,能够更好地建模图像和文本模态之间的信息交互,捕获多模态讽刺信息,提升多模态讽刺检测的性能。

3.5 消融实验

为了探究不同模块的有用性,本文设置了消融实验,分别探究了图文信息交互部分、深层特征表示部分及图文特征融合部分的有效性,设计了在原有的模型上去掉三个模块的方法。-Cross-attention:表示去除图文交互部分;-CNN:表示去除深层特征表示部分;-Fuse"表示去掉图文特征融合部分;

四种方法的比较实验结果如表4所示。

表4 消融实验结果

Table 4 The experimental results of the ablation

Model	Acc/%	*P*/%	*R*/%	*F*1/%
Our	86.76	86.52	86.46	86.49
−Cross-attention	80.87	80.55	81.03	80.67
−CNN	85.97	85.64	85.85	85.73
−Fuse	86.05	85.81	86.43	85.93

通过表4可以发现:

(1)-Cross-attention,与本文方法相比模型效果有较大的下降,原因可能是由于未能充分建模图像和文本之间的信息交互,缺少图像和文本信息的相互补充,使得模型未能充分捕捉文本和图像之间的关联,进而影响最终性能。

(2)-CNN,与本文方法相比模型效果也有一定的下降,原因可能在于模型不能从多个角度表示图像和文本深层特征,仅利用了较为浅层的特征,导致模型效果下降。

(3)-Fuse,与本文方法相比整体效果下降最少,原因可能是由于在经过图文信息交互部分后,模型已经得到图像和文本之间的交互信息,当去除Fuse这部分时,对模型的影响相较于其他模块较小。

4 结论

本文提出了一个基于图文交互和深层特征融合的讽刺检测框架,通过借助交叉注意力机制建立图文交互模块,将图像和文本信息得到充分的交互,然后利用卷积神经网络从不同的角度得到图像和文本的深层特征的表示,同时设计图像和文本的信息融合模块,将图像和文

本特征充分地融合,得到多模态的特征。最后从图像、文本和多模态三种特征入手,分别得到它们更高层次的特征表示,最终得到讽刺检测的结果。实验结果表明,本文提出的讽刺检测框架在部分指标上,达到了最先进的性能。在未来研究中,可以考虑针对讽刺文本所具有的语义特征进行建模,并进一步尝试捕获图像中局部的细节特征进行建模。此外,如何在低时间和资源消耗的情况下使用图像多模态大模型助力于多模态讽刺检测也是一个值得思考的问题。

参考文献:

[1] XU N, ZENG Z X, MAO W J. Reasoning with Multimodal Sarcastic Tweets *via* Modeling Cross-modality Contrast and Semantic Association[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 3777−3786. DOI: 10.18653/v1/2020.acl-main.349.

[2] PAN H L, LIN Z, FU P, *et al*. Modeling Intra and Inter-modality Incongruity for Multi-modal Sarcasm Detection [C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 1383−1392. DOI: 10.18653/v1/2020.findings-emnlp.124.

[3] LIANG B, LOU C W, LI X, *et al*. Multi-modal Sarcasm Detection *via* Cross-modal Graph Convolutional Network[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 1767−1777. DOI: 10.18653/v1/2022.acl-long.124.

[4] QIN L B, HUANG S J, CHEN Q G, *et al*. MMSD2.0: Towards a Reliable Multi-modal Sarcasm Detection System [C]//Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg, PA, USA: Association for Computational Linguistics, 2023: 10834−10845. DOI: 10.18653/v1/2023.findings-acl.689.

[5] RADFORD A, KIM J W, HALLACY C, *et al*. Learning Transferable Visual Models from Natural Language Supervision[EB/OL]. (2021−2−26) [2024−12−27]. https://arxiv.org/abs/2103.00020.

[6] TANG B H, LIN B D, YAN H L, *et al*. Leveraging Generative Large Language Models with Visual Instruction and Demonstration Retrieval for Multimodal Sarcasm Detection[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2024: 1732−1742. DOI: 10.18653/v1/2024.naacl-long.97.

[7] ZHANG M S, ZHANG Y, FU G. Tweet Sarcasm Detection Using Deep Neural Network[C]// Proceedings of the 26th International Conference on Computational Linguistics. Osaka, Japan: ACL, 2016: 2449−2460.

[8] TAY Y, LUU A T, HUI S C, *et al*. Reasoning with Sarcasm by Reading In-between[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 1010−1020. DOI: 10.18653/v1/p18-1093.

[9] BABANEJAD N, DAVOUDI H, AN A J, *et al*. Affective and Contextual Embedding for Sarcasm Detection[C]// Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020: 225−243. DOI: 10.18653/v1/2020.coling-main.20.

[10] SCHIFANELLA R, DE JUAN P, TETREAULT J, *et al*. Detecting Sarcasm in Multimodal Social Platforms[C]// Proceedings of the 24th ACM international conference on Multimedia. New York: ACM, 2016: 10.1145/2964284.2964321. DOI: 10.1145/2964284.2964321.

[11] CAI Y T, CAI H Y, WAN X J. Multi-modal Sarcasm Detection in Twitter with Hierarchical Fusion Model[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 2506−2515. DOI: 10.18653/v1/p19-1239.

[12] TIAN Y, XU N, ZHANG R K, *et al*. Dynamic Routing Transformer Network for Multimodal Sarcasm Detection[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2023: 2468−2480. DOI: 10.18653/v1/2023.acl-long.139.

[13] WANG X Y, SUN X W, YANG T, *et al*. Building a Bridge: A Method for Image-text Sarcasm Detection without Pretraining on Image-text Data[C]//Proceedings of the First International Workshop on Natural Language Processing Beyond Text. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 19−29. DOI: 10.18653/v1/2020.nlpbt-1.3.

[14] CHEN J J, YU H, LIU W D, *et al*. InterCLIP-MEP: Interactive CLIP and Memory-enhanced Predictor for

Multi-modal Sarcasm Detection[EB/OL]. (2024-08-13) [2024-12-27]. https://arxiv.org/abs/2406.16464v4.

[15] YANG X C, WU W F, FENG S, *et al*. MM-BigBench: Evaluating Multimodal Models on Multimodal Content Comprehension Tasks[EB/OL]. (2023-10-13) [2024-12-27]. https://arxiv.org/abs/2310.09036.

[16] LONG X, GAN C, MELO G, *et al*. Multimodal Keyless Attention Fusion for Video Classification[J]. *Proc AAAI Conf Artif Intell*, 2018, **32**(1): 7202-7209. DOI: 10.1609/aaai.v32i1.12319.

[17] KIM Y. Convolutional Neural Networks for Sentence Classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1746-1751. DOI: 10.3115/v1/d14-1181.

[18] ZHOU P, SHI W, TIAN J, *et al*. Attention-based Bidirectional Long Short-term Memory Networks for Relation Classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 207-212. DOI: 10.18653/v1/p16-2034.

[19] XIONG T, ZHANG P R, ZHU H B, *et al*. Sarcasm Detection with Self-matching Networks and Low-rank Bilinear Pooling[C]//The World Wide Web Conference. New York: ACM, 2019. DOI: 10.1145/3308558.3313735.

[20] LIU Y H, OTT M, GOYAL N, *et al*. RoBERTa: A Robustly Optimized BERT Pretraining Approach[EB/OL]. (2019-07-26)[2024-12-27]. https://arxiv.org/abs/1907.11692.

[21] DU Z X, QIAN Y J, LIU X, *et al*. GLM: General Language Model Pretraining with Autoregressive Blank Infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 320-335. DOI: 10.18653/v1/2022.acl-long.26.

[22] TOUVRON H G, MARTIN L, STONE K, *et al*. Llama 2: Open Foundation and Fine-tuned Chat Models[EB/OL]. (2023-07-18) [2024-12-27]. https://arxiv.org/abs/2307.09288.

[23] HE K M, ZHANG X Y, REN S Q, *et al*. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2016: 770-778. DOI: 10.1109/CVPR.2016.90.

[24] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C]// Proceedings of the International Conference on Learning Representations. Vienna, Austria : OpenReview. 2021.

[25] LIU H, WANG W Y, LI H L. Towards Multi-modal Sarcasm Detection *via* Hierarchical Congruity Modeling with Knowledge Enhancement[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 4995-5006. DOI: 10.18653/v1/2022.emnlp-main.333.