基于对偶注意力机制的图文情感分析

李劲哲,吴宇贤,蔡珺恺,王成济,蒋兴鹏

(华中师范大学 计算机学院,湖北 武汉 430079)

摘 要: 传统情感分析方法无法有效处理社交平台中的大量多模态图文数据,暴露出多模态特征融合效果不佳的问题。为此,结合注意力机制与前馈神经网络建立基于对偶注意力机制融合多模态的情感分析模型。该模型利用预训练模型提取文本与图像特征,采用跨模态特征融合模块强化属于多个模态的公有特征,采用单模态自注意力模块提取单个模态私有特征中的有效信息,最终拼接融合多模态特征,实现对多模态数据的高效表征。在推特图文数据集上进行验证实验,通过与多种方法进行比较以及对内部各模态进行消融实验,证实所提模型具有较好的情感分类效果。

关键词: 多模态情感分析;多头注意力;特征融合;对偶融合

DOI: 10. 11907/rjdk. 231294 **开放科学(资源服务)标识码(OSID):**

中图分类号: TP391 **文献标识码:** A **文章编号:** 1672-7800(2024)004-0178-08

Graphic Sentiment Analysis Based on Pairwise Attention Mechanisms

LI Jinzhe, WU Yuxian, CAI Junkai, WANG Chengji, JIANG Xingpeng

(*School of Computer Science*, *Central China Normal University*, *Wuhan* 430079, *China*)

Abstract: Traditional sentiment analysis methods are unable to effectively handle a large amount of multimodal graphic and textual data on social platforms, exposing the problem of poor performance in multimodal feature fusion. To this end, a multimodal sentiment analysis model based on dual attention mechanism fusion is established by combining attention mechanism and feedforward neural network. This model utilizes pre trained models to extract text and image features, strengthens public features belonging to multiple modalities using a cross modal feature fusion module, extracts effective information from private features belonging to a single modality using a single modal self attention module, and finally concatenates and fuses multimodal features to achieve efficient representation of multimodal data. Validation experiments were conducted on the Twitter image and text dataset, comparing with various methods and conducting ablation experiments on internal modalities, confirming that the proposed model has good sentiment classification performance.

Key Words: multimodal sentiment analysis; multi-head attention; feature fusion; dual fusion

0 引言

情感分析又称为意见挖掘。自互联网兴起,各类型网络产品成为用户分享生活经历的重要平台,如传统网络论坛、在线社区等。这类平台中带有情感信息的数据在企业管理、商品推荐、舆情分析与情绪干预等方面都有着巨大的应用价值[1]。以往情感分析数据以文本为主,然而随着带有高质量摄像头的手机不断普及,以图片和视频为代表的视觉数据推动多模态数据呈现出爆炸式增长的趋势[2],以推特、微博等为代表的社交平台以及以亚马逊、淘宝等为代表的网络购物平台出现大量视觉与文本互补的数据,对经济社会发展具有深刻且重要的意义[3]。因此,多模态情感分析成为一项研究热点。

收稿日期: 2023-03-22

基金项目: 国家语委“十四五”科研规划研究基地项目(重点项目)(ZDI145-56);中央高校基本科研业务费资助项目(CCNU23XJ001);中国博士后科学基金面上项目(2023M741305)

作者简介: 李劲哲(2002-),男,华中师范大学计算机学院学生,研究方向为多模态情感分析;吴宇贤(2003-),男,华中师范大学计算机学院学生,研究方向为自然语言处理;蔡珺恺(1998-),男,华中师范大学计算机学院硕士研究生,研究方向为生物信息学;王成济(1993-),男,CCF专业会员,华中师范大学计算机学院讲师,研究方向为计算机视觉、行人重识别、多模态表示学习;蒋兴鹏(1979-),男,CCF专业会员,华中师范大学计算机学院教授、博士生导师,研究方向为医学人工智能、生物信息学。本文通讯作者:蒋兴鹏、王成济。

1　相关研究

情感分析旨在从文本、图像与音频等信息中提取情感信息，帮助了解人类的情感、态度与意见。根据所含内容种类，可将信息分为单模态和多模态两种。

1.1　单模态情感分析

单模态情感分析主要针对文本信息，大致可分为基于词典的方式与基于机器学习的方式两种[4]。基于词典的方式是通过构建情感词典，并对其进行极性标注，从而完成文本情感分析任务。例如，徐琳宏等[5]采用手工分类与自动获取相结合的方法构建情感词汇本体；管雨翔等[6]基于微博平台数据构建针对网络舆情领域的情感词典。然而基于词典的文本情感分析往往需要投入大量人力物力，且适用领域较为单一，对长文本分析效果不佳。因此，利用机器学习方法完成文本情感分析任务逐渐流行。基于机器学习的文本情感分析将已经完成情感类别标注的文本作为训练数据集，对训练数据进行特征提取与特征选择，然后利用机器学习模型进行训练，最终完成文本情感分析任务。例如，Kiritchenko 等[7]使用支持向量机（Support Vector Machine，SVM）进行文本情感分析。然而，该类方法的效果很大程度上依赖于特征工程的完成质量。

1.2　多模态情感分析

随着社交平台上多模态数据爆炸式增长，单模态情感分析模型已不能满足实际需求。多模态情感分析问题逐渐受到国内外研究人员的高度重视。图 1 为情感分析常用的 MVSA 数据集[8]图文示例，其中图 1(a)的图片与文本均表达积极情绪，最终表示的情绪为积极；图 1(b)的图片表达积极情绪，文本却存在争议，有的人觉得是积极，有的人则觉得是中立，数据集标注的是积极；图 1(c)的图片表达消极情绪，文本表达中立情绪，最终表示的情绪为消极；图 1(d)的图文情绪表达均为中立，因此最终结果也是中立。可以看出，多模态图文数据相辅相成，包含更加丰富的信息，能够更客观地体现用户情感。

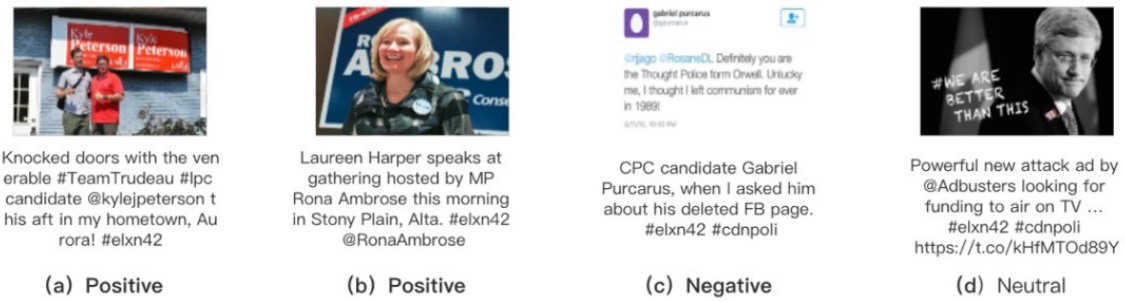

Fig. 1　Graphic example of MVSA dataset

图 1　MVSA 数据集图文示例

多模态情感分析的目的是利用更多信息来源实现效果更优的决策。研究人员一直在积极探索将深度学习方法应用于情感分析中，并设计融合多模态特征的分析模型。该类模型主要包括特征提取与多模态特征融合两部分，首先构建模态相关的特征提取模块，将不同模态的特征送入跨模态特征融合模块中，然后将融合特征送入决策模块。研究人员认为在以上过程中特征融合起到了关键作用，因此目前多模态情感分析的研究重心主要集中在跨模态特征融合模块上。例如，Gandihi 等[9]将多模态特征融合方式分为张量融合[10]、层次融合[11]、词级融合[12]与基于注意力机制融合[13]等几种类型，以上方法旨在通过融合多模态特征挖掘不同模态的公有特征，从而提取多模态数据中隐含的情感信息；Poria 等[14]提出一种先通过卷积神经网络（Convolutional Neural Network，CNN）提取文本特征，再采用多核学习（Multiple Kernel Learning，MKL）方式进行特征融合的方法；而后 Poria 等[15]又提出一种基于长短期记忆（Long Short-Term Memory，LSTM）的视频多模态情感分析模型，然而不同模态之间的特征信息不一定存在顺序关系，仅考虑时间关系使用 LSTM 进行多模态特征融合的方法无法适用于全部情况；Yu 等[16]设计了一个基于自监督学习实现标签自动生成的模型，可联合训练单模态与多模态任务，从而学习到一致性和差异性，然而该模型训练的时间与人工成本较高；Tang 等[17]通过耦合学习建模模态间的相互作用，以保证对缺失模态特征学习的鲁棒性。

然而，上述方法忽略了专属于一个模态的私有特征。很多研究认为模态的私有特征属于干扰信息，致使该类特征一直被忽略。事实上，私有特征同样编码了丰富的情感信息，这些信息对于情感分析具有至关重要的作用[18]。为此，部分研究人员选择使用注意力机制解决这一问题，其能够专注于每个模态中最重要的特征，并生成更有效的表示。例如，Truong 等[19]提出视觉注意网络 VistaNet，在特征融合中没有使用视觉信息作为特征，而是利用注意力机制将图像与文本中的重要内容对齐，从而提高特征信息的有效性，该方法能有效区分公有特征与私有特征。

为充分挖掘公有特征与私有特征中的情感信息，本文借鉴 Transformer[20]思想，基于多头注意力模型（Multi-Head Attention，MHA）设计一个对偶注意力机制融合模型（Pairwise Attention Mechanisms，PAM），通过调整 MHA 的输入提出跨模态特征融合模块和单模态注意力模块，分别处理公有特征和私有特征，同时弱化私有特征中的干扰信

息。与以往研究相比，本文研究主要有以下贡献：①针对多模态特征信息分为公有特征与私有特征，而私有特征往往被忽略的问题，提出PAM，其可以有效提取公有特征与私有特征，提高多模态表征的准确性；②提出跨模态特征融合模块与单模态注意力模块，实现公有特征与私有特征的区分，并弱化干扰信息的影响；③在多模态数据集MVSA-Single与MVSA-Multi上进行实验，证实本文方法相较以往研究在准确性方面有较大提升。

2 模型建立

2.1 MHA

MHA是自然语言处理领域广泛应用的神经网络结构，其在机器翻译、问答系统、语言生成等任务中具有较好效果。MHA基于自注意力(Self-Attention)模型建立，旨在建模长序列内部不同词(token)之间的依赖关系，其有查询Q(Query)、键K(Key)和值V(Value)3个输入。自注意力模型首先计算查询Q与键K的相关性，根据相关性对值V进行加权组合，进而建模长距离依赖关系。计算公式为：

$$Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d}}V\right) \tag{1}$$

如图2(彩图扫OSID码可见，下同)所示，MHA采用不同的映射矩阵将输入序列分别映射到不同的线性空间中，然后利用自注意力模型对不同空间中的特征进行加权组合。每个空间会学习到一组特定特征，不同空间中特征的加权组合可得到多个输出，也被称为头(Head)。对不同头进行拼接得到最终输出，表示为：

$$head_i=Attention\left(QW_i^Q,KW_i^k,VW_i^V\right) \tag{2}$$

式中：$\{W_i^Q,W_i^K,W_i^V\}\in R^{d\times\frac{d}{h}}$分别为第$i$个嵌入空间的参数矩阵；$d$为输入特征维度；$h$为头的个数。

最终模型可表示为：

$$\mathrm{MHA}(Q,K,V)=\mathrm{Concat}\left(head_1,\cdots,head_h\right)\mathrm{W}^o \tag{3}$$

式中：$W^o\in R^{d\times d}$为线性变换参数矩阵。

一般来说，自注意力模型和MHA的输入Q、K和V可以相同也可以不同。本文基于MHA设计了两个注意力模块，分别为输入来自于同一模态的单模态多头注意力模块(Single-modal Multi-Head Attention，S-MHA)和输入来自于不同模态的跨模态多头注意力模块(Cross-modal Multi-Head Attention，C-MHA)。

2.2 PAM

为充分融合多模态特征完成情感分析任务，本文从模型输入方面改进MHA，提出基于对偶注意力机制的图文情感分析模型PAM，结构如图3所示。该模型分别挖掘公有特征和私有特征，并将它们融合在一起，解决特征融合只关注多个模态共有特征而忽略专属于单个模态私有特征的问题。具体来说，PAM分为以下4个部分：①模态特征提取。利用不同神经网络对文本、图像特征进行提取，然后进行线性变化等操作为跨模态特征融合作准备；②跨模态特征融合。改进MHA，构建文本—图像特征融合模块(Text-Image)与图像—文本特征融合模块(Image-Text)，得到多模态融合表征；③单模态自注意力。基于MHA获得单模态私有特征；④情感分类。基于多模态融合表征，使用情感分类器获得最终预测结果。

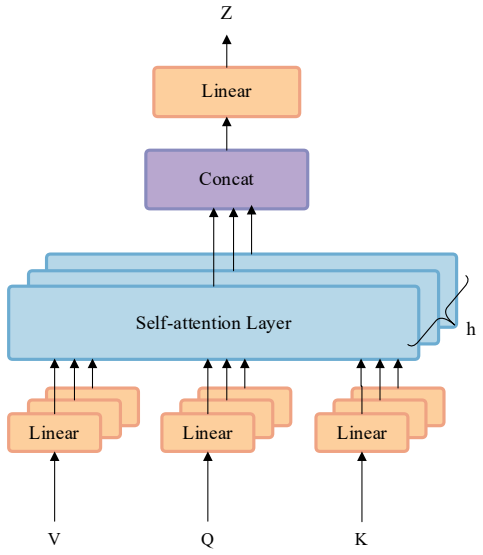

Fig. 2 Structure of multi-head attention mechanism

图2 多头注意力机制结构

2.2.1 模态特征提取

(1)文本特征。给定文本T，使用基于注意力机制的BERT预训练模型[21]提取文本特征。将文本特征输入到BiLSTM(Bi-directional Long Short-Term Memory)中建模词的上下文信息，然后使用全连接层(又称线性层，Linear，LN)调整特征维度，最终得到$X_t\in R^{n\times d}$的文本特征序列。计算公式为：

$$X=BERT(T) \tag{4}$$

$$X'=BiLSTM(X) \tag{5}$$

$$X_t=LN(X') \tag{6}$$

(2)图像特征。给定图像V，使用ResNet152的前4层网络对图像特征进行提取，获得$X_v'\in R^{2048\times 7\times 7}$[22]。将特征$X_v'$的第2和第3个维度展开，使用平均池化调整token个数，得到与文本特征维度大小相同的特征$X_v\in R^{n\times d}$。计算公式为：

$$X_v'=ResNet152_{layer4}(V) \tag{7}$$

$$X_v=average\left(flatten(X_v')\right) \tag{8}$$

2.2.2 跨模态特征融合

为充分融合图像与文本特征，本文改进MHA，使用一个模态的特征作为查询，引导并聚合另一个模态的特征，从而挖掘多模态数据的公有特征。为此，设计两个跨模态特征融合模块：①文本—图像特征融合模块。使用图像特征作为查询聚合文本特征；②图像—文本特征融合模块。

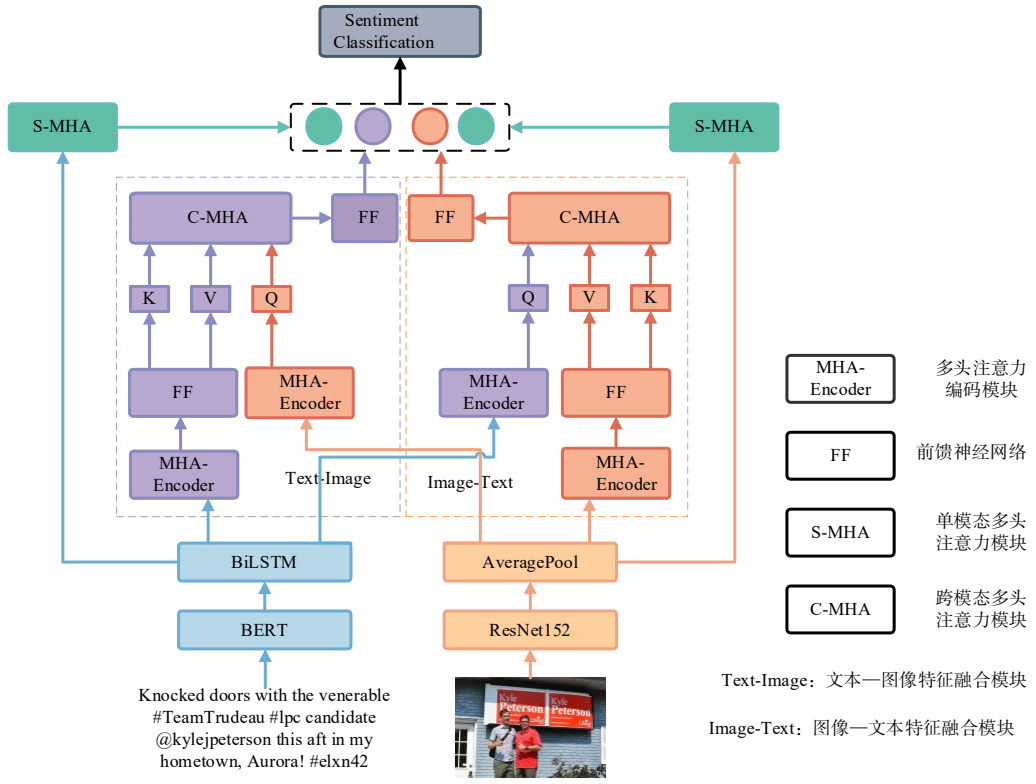

Fig. 3　Model structure

图 3　PAM 结构

使用文本特征作为查询聚合图像特征。

在文本—图像特征融合模块中，首先将文本特征 X_t 与图像特征 X_v 分别输入到两个 MHA 模块中，还要将文本特征 X_t 输入到一个两层前馈神经网络（Feed Forward Neural Network，FF）中；然后将图像特征 X_v 作为查询 Q，文本特征 X_t 作为键 K 和值 V，将其输入到 MHA 模块中进行跨模态特征融合，获得跨模态特征 $z_{t\to v}\in R^{n\times d}$。详细计算过程为：

$$Q_{t\to v}=MHA(X_v,X_v,X_v) \tag{9}$$

$$K_{t\to v},V_{t\to v}=MHA(X_t,X_t,X_t) \tag{10}$$

$$z_{t\to v}=FF(MHA(Q_{t\to v},K_{t\to v},V_{t\to v})) \tag{11}$$

在图像—文本特征融合模块中，首先将图像特征 X_v 与文本特征 X_t 分别输入到两个 MHA 模块中，还要将图像特征 X_v 输入到一个两层 FF 中；然后将文本特征 X_t 作为查询 Q，图像特征 X_v 作为键 K 和值 V，将其输入到 MHA 模块中进行跨模态特征融合，获得跨模态特征 $z_{v\to t}\in R^{n\times d}$。

2.2.3　单模态自注意力

跨模态特征融合模块侧重于对公有特征进行强化。为此，本文对单模态特征使用 MHA 机制，从而获取私有特征中的有效信息。将文本特征 X_t 与图像特征 X_v 分别输入到单模态多头注意力模块中，获得 $X_t^{new}\in R^{n\times d}$ 与 $X_v^{new}\in R^{n\times d}$。计算过程为：

$$X_t^{new}=MHA(X_t,X_t,X_t) \tag{12}$$

$$X_v^{new}=MHA(X_v,X_v,X_v) \tag{13}$$

2.2.4　情感分类

首先将 $z_{t\to v}$、$z_{v\to t}$、X_t^{new} 和 X_v^{new} 拼接得到最终的特征 $X_{all}\in R^{1\times d}$。表示为：

$$X_{all}=W^TConcat(z_{t\to v},z_{v\to t},X_t^{new},X_v^{new}) \tag{14}$$

式中：$W\in R^{4n\times 1}$。

然后去除 X_{all} 中多余的维度，将其输入到包含两个全连接层的网络进行分类，得到 y'。将 y' 送入 softmax 层进行归一化，得到不同情感类别的得分 y。表示为：

$$y=softmax(y') \tag{15}$$

使用交叉熵作为损失函数训练模型。计算公式为：

$$Loss=-\sum_{i=1}^{N}y_i logy_i' \tag{16}$$

式中：N 为 batch_size 大小，y_i 为模型预测概率，y_i' 为真实结果。测试时，取得分最高的类别作为预测结果。

3　实验方法与结果分析

3.1　实验环境

使用 Pytorch 框架。在文本特征提取模块，BiLSTM 的隐藏层特征维度设置为 2 048。在图像特征提取模块对所有图像进行裁剪缩放，使其大小统一变为 224×224。学习率设置为 1×10^{-5}。采用基于随机梯度下降的 Adam 优化器对模型进行训练[23]。模型训练与测试均使用一张 NVIDIA 4090 显卡，内存为 24 G。

3.2 数据集

使用推特图文情感分析数据集MVSA，其包含MVSA-Single和MVSA-Multi两个子任务数据集。MVSA-Single数据集包含5 219条数据，每个模态进行一次单独标注；MVSA-Multi数据集包含19 600条数据，表1给出其多模态数据标签确定示例。

Table 1 Multimodal data label determination example

表1 多模态数据标签确定示例

图像	文本	图像标签	文本标签	是否有效	最终标签
	Today is#InternationalDayofDemocracy? Oh，great day to remind @pmharper that he is Good to Go! #elxn42 #cdnpoli	消极	积极	无效	-
	Great AM with riding neighbors@Carolyn_Bennett &amp@marcomendicino sharing our#LPC plan for #RealChange! #DVW#elxn42	积极	中立	有效	积极

表2为MVSA-Single和MVSA-Multi数据集处理后的数据分布情况。

Table 2 Data distribution of two datasets

表2 两个数据集数据分布情况

数据集	数据数量	有效数据	积极数据	中立数据	消极数据
MVSA-Single	5 129	4 511	2 683	470	1 358
MVSA-Multi	19 600	17 024	11 318	4 408	1 298

3.3 评价指标

使用准确率(Accuracy)与F1值(F1-score)对模型性能进行评估。计算公式为：

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{17}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{18}$$

式中：TP为真正例，指正样本且预测为正的结果个数；FN为假负例，指正样本却预测为负的结果个数；FP为假正例，指负样本却预测为正的结果个数；TN为真负例，指负样本且预测为负的结果个数。

3.4 比较实验

为验证本文模型的有效性，将其与文本单模态、图像单模态和多模态3类方法进行比较。具体对照模型包括：①单模态文本情感分类模型BERT-LSTM。其使用BERT预训练模型对文本信息进行编码，将得到的文本特征信息输入到LSTM网络结构中，在多个全连接层构成的多层感知机中完成情感分类；②单模态图像情感分类模型。ResNet152[21]使用预训练的ResNet152图像分类模型的前4层进行特征提取，然后输入到全连接网络结构中进行情感分类；VGG16[24]使用预训练的VGG16图像分类模型微调最终输出进行情感分类；③多模态情感分类模型。CNN-Concat使用CNN获取两个模态的特征信息，并直接拼接输入到多层感知机中进行分类；MultiSentiNet[25]是一种首先提取图像深层语义特征，然后结合注意力机制强化图文关联的多模态情感分析模型；AMABF[26]是一种基于双线性融合的特征融合模型，其在文本特征模块使用BERT-GRU，在图像特征模块使用ResNet。

表3为各模型在MVSA两个子任务数据集上的表现。可以看出，本文模型除在MVSA-Multi数据集上的F1值低于AMABF模型外，其余各项指标均为最优。

Table 3 Performance of each model on the two subtask datasets of MVSA

表3 各模型在MVSA两个子任务数据集上的表现 (%)

模态	模型	MVSA-Single		MVSA-Multi	
		Acc	F1	Acc	F1
Text	BERT-LSTM	69.91	70.33	67.90	65.73
Image	ResNet152	69.24	68.93	62.69	60.21
	VGG16	62.73	59.77	61.15	59.17
Text + Image	CNN-Concat	62.05	60.44	60.75	59.13
	MultiSentiNet	69.84	69.63	68.86	68.11
	AMABF	72.47	71.69	72.15	**71.47**
	本文模型	**75.11**	**73.64**	**72.68**	71.40

图4为MVSA数据集的两个预测实例，其实际标签均为积极。可以看出，本文提出的PAM模型预测结果为积极的得分远超其他标签，而CNN-Concat模型则预测错误；BERT-LSTM模型虽然预测正确，但积极标签的得分远小于PAM模型。

3.5 消融实验

为进一步分析PAM中跨模态特征融合与单模态自注意力机制对模型性能的贡献，设计两组消融实验，在MVSA-Single与MVSA-Multi数据集上的消融实验结果如图5、图6所示。可以看出，舍弃PAM模型中的任一模块都会导致模型性能下降。由图5可知，舍弃跨模态特征融合模块后，模型在MVSA-Single数据集上的准确率下降4.98%，F1值下降3.78%；在MVSA-Multi数据集的准确率下降

8.15%，F1 值下降 7.23%。

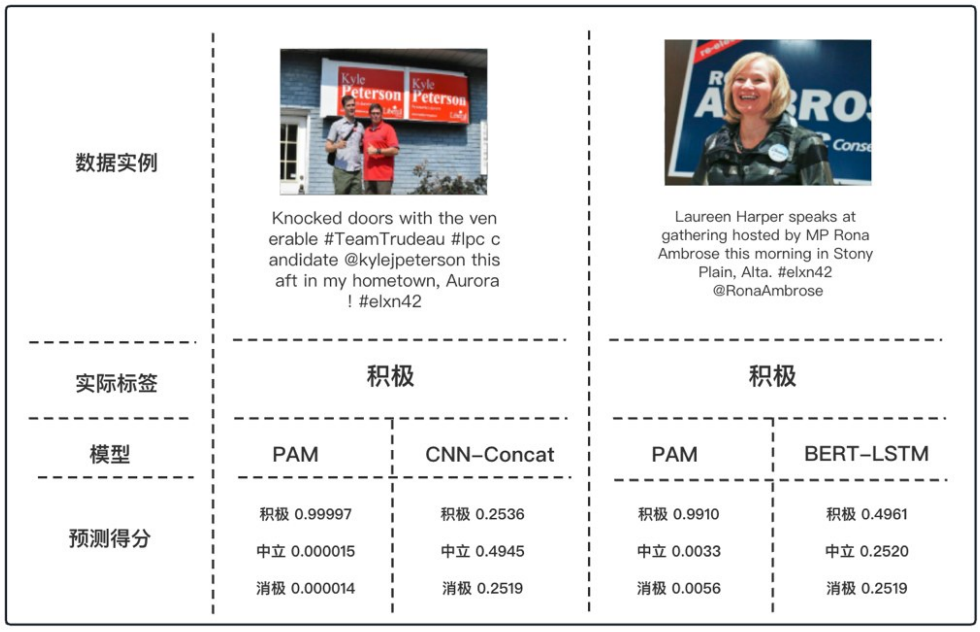

Fig. 4　MVSA dataset prediction example

图 4　MVSA 数据集预测实例

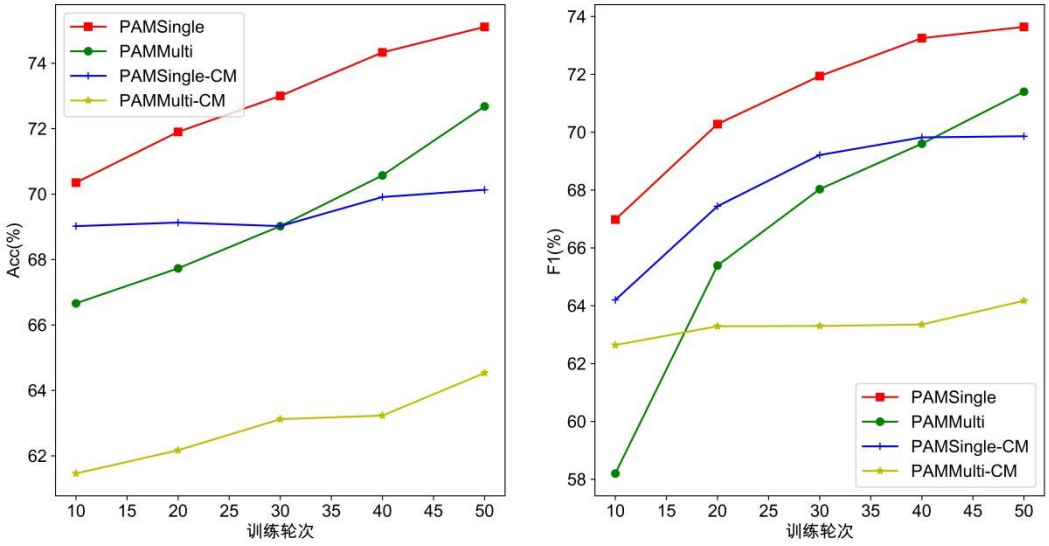

Fig. 5　Ablation experimental results of removal of cross-modal fusion features fusion

图 5　去跨模态融合特征融合消融实验结果

由图 6 可知，舍弃单模态自注意力机制模块后，模型在 MVSA-Single 数据集上的准确率下降 2.55%，F1 值下降 2.85%；在 MVSA-Multi 数据集上的准确率下降 5.78%，F1 值下降 7.21%。

为研究 PAM 层数对模型性能的影响，分别对 1、2、3 层 PAM 模型性能进行比较，结果见图 7。可以看出，虽然多层 PAM 在 MVSA-Single 数据集上的表现有所提升，但在 MVSA-Multi 数据集上却出现过拟合现象，在实验中表现为最终预测结果全部为积极。因此，多层 PAM 的效果并不会优于单层 PAM。

4　结语

本文针对多模态情感分析任务提出一种基于对偶注意力机制的多模态特征融合模型。该模型分为模态特征提取、跨模态特征融合、单模态自注意力与情感分类 4 个模块：在跨模态特征融合模块通过使用来自另一个模态的特征引导融合该模态的特征强化公有特征；单模态自注意力模块旨在强化私有特征，通过融合公有特征和私有特征获取高效多模态情感表征，进而通过情感分类模块完成多模态情感分析任务。实验结果表明，所提模型具有较好的

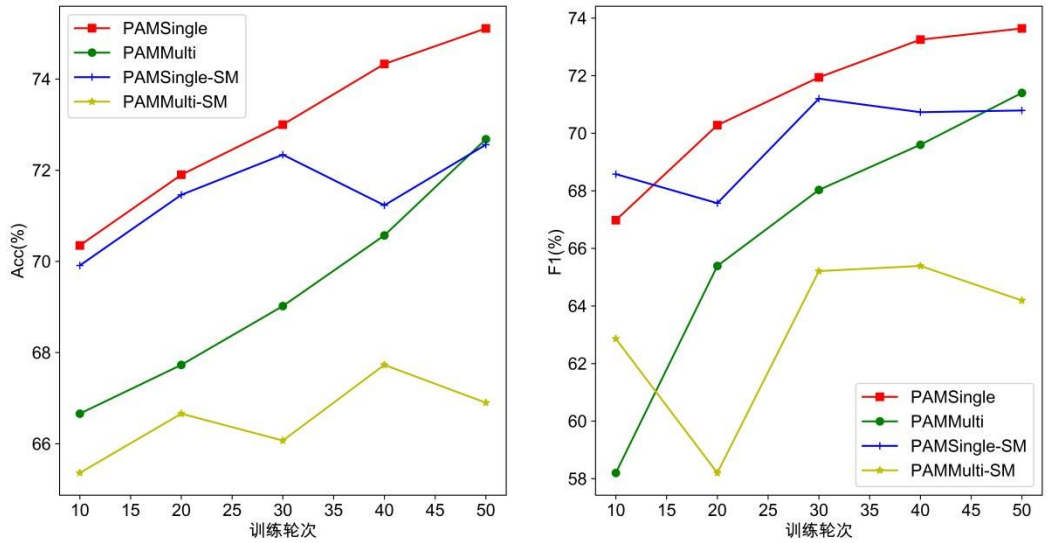

Fig. 6 Ablation experiments results of removal of unimodal self-attentive mechanisms

图 6 去单模态自注意力机制消融实验结果

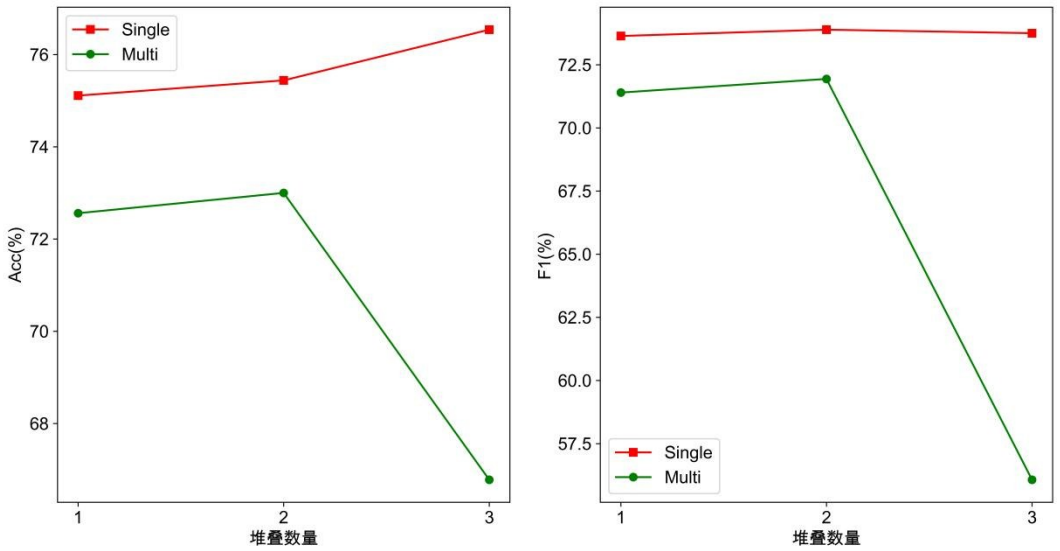

Fig. 7 The impact of PAM layers on model performance

图 7 PAM 层数对模型性能的影响

情感分类效果，优于单模态模型；同时证明了在多模态情感分析任务中，图文之间具有信息互补作用。然而，本文研究仍存在一些不足之处，尽管模型通过单模态自注意力模块强化了私有特征，但在处理高度复杂或非常细微的情感表达时，仍可能无法完全捕捉到所有关键信息。为此，未来研究可以考虑引入更先进的自然语言处理技术或更深层次的图像处理技术，以更精确地识别与提取模态之间的微妙差异。

参考文献：

[1] YANG C W, LI X R, ZHU C L. A sentiment analysis method based on hybrid word feature representation for online reviews and its application[J]. Software Guide, 2023, 22(4): 48-53.
杨成伟，李希茹，祝翠玲. 一种基于混合词特征表示的在线评论情感分析方法及应用[J]. 软件导刊，2023，22(4)：48-53.

[2] HAN W, CHEN H, GELBUKH A, et al. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis[C]//Proceedings of the 2021 International Conference on Multimodal Interaction, 2021: 6-15.

[3] ZHANG Y, SONG D, ZHANG P, et al. A quantum-inspired multimodal sentiment analysis framework[J]. Theoretical Computer Science, 2018, 752: 21-40.

[4] LIU S, ZHAO J X, YANG H Y, et al. A review of text sentiment analysis[J]. Software Guide, 2018, 17(6): 1-4.
刘爽，赵景秀，杨红亚，等. 文本情感分析综述[J]. 软件导刊，2018，17(6)：1-4.

[5] XU L H, LIN H F, PAN Y, et al. The construction of emotion vocabulary ontology[J]. Journal of Intelligence, 2008, 27(2): 180-185.
徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 情报学报，2008，27(2)：180-185.

[6] GUAN Y X, WANG J, LIU J, et al. Construction of an emotion lexicon in the field of sudden-onset online public opinion[J]. Intelligence Inquiry, 2023(2): 1-8.
管雨翔，王娟，刘静，等. 突发事件网络舆情领域情感词典构建[J]. 情

报探索,2023(2):1-8.

[7] KIRITCHENKO S, ZHU X, CHERRY C, et al. NRC-Canada-2014: detecting aspects and sentiment in customer reviews[C]//Proceedings of the 8th International Workshop on Semantic Evaluation,2014:437-442.

[8] NIU T, ZHU S, PANG L, et al. Sentiment analysis on multi-view social data[C]// 22nd International Conference on MultiMedia Modeling, 2016: 15-27.

[9] GANDHI A, ADHVARYU K, PORIA S, et al. Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions[J]. Information Fusion, 2022,91: 424-444.

[10] LIN M H, MENG Z Q. Multimodal sentiment analysis based on attention neural network [J]. Computer Science, 2020, 47(S2): 518-524, 558.
林敏鸿,蒙祖强. 基于注意力神经网络的多模态情感分析[J]. 计算机科学,2020,47(S2):508-514,548.

[11] MAJUMDER N, HAZARIKA D, GELBUKH A, et al. Multimodal sentiment analysis using hierarchical fusion with context modeling[J]. Knowledge-Based Systems, 2018, 161: 124-133.

[12] CHEN M, WANG S, LIANG P P, et al. Multimodal sentiment analysis with word-level fusion and reinforcement learning[C]//Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017: 163-171.

[13] YADAV A, VISHWAKARMA D K. A deep multi-level attentive network for multimodal sentiment analysis[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(1): 1-19.

[14] PORIA S, CAMBRIA E, GELBUKH A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,2015:2539-2544.

[15] PORIA S, CAMBRIA E, HAZARIKA D, et al. Context-dependent sentiment analysis in user-generated videos[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 873-883.

[16] YU W, XU H, YUAN Z, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis[DB/OL]. https://arxiv. org/abs/2102. 04830.

[17] TANG J, LI K, JIN X, et al. CTFN: hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing ,2021: 5301-5311.

[18] WU Y, LIN Z, ZHAO Y, et al. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis[C]// Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics, 2021: 4730-4738.

[19] TRUONG Q T, LAUW H W. Vistanet: visual aspect attention network for multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2019:305-312.

[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems,2017:5998-6008.

[21] DEVLIN J , CHANG M W , LEE K , et al. BERT: pre-training of deep bidirectional transformers for language understanding[DB/OL]. https://arxiv. org/abs/1810. 04805,2018.

[22] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:770-778.

[23] DIEDERIK P K, JIMMY B. Adam: a method for stochastic optimization [DB/OL]. https://arxiv. org/abs/1412. 6980,2014.

[24] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [DB/OL]. https://arxiv. org/abs/1409. 1556.

[25] XU N, MAO W. Multisentinet: a deep semantic network for multimodal sentiment analysis[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management,2017: 2399-2402.

[26] XIA H, HU Y. Graphic-text sentiment analysis based on attention and bilinear fusion[J]. Information Technology and Informatization, 2022(5): 38-41.
夏欢,胡勇. 基于注意力和双线性融合的图文情感分析[J]. 信息技术与信息化,2022(5):38-41.

(责任编辑:尹晨茹)