http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2024.0545

基于情感表征校准的图文情感分析模型

张顺香[1,2,3,*]，刘佳佳[1,2]，焦熠璇[1,2]，张玉雷[1,2]，李自亮[1,2]

(1. 安徽理工大学 计算机科学与工程学院，淮南 232001；2. 合肥综合性国家科学中心人工智能研究院，合肥 230000；
3. 淮南师范学院 计算机学院，淮南 232038)

摘　　要：随着社交媒体用户更多地使用图文表达观点，挖掘多模态数据关联性以辅助分析用户情感成为挑战。目前的研究方法对模态间信息的交互不充分，且跨模态交互易引入噪声，导致学习到的图文表示出现情感表征偏移。为此，提出一种基于情感表征校准的图文情感分析模型。通过特征提取模块获取图文特征；在跨模态交互模块中，利用跨模态的注意力进行“图-文”和“文-图”2个方向上的双向交互，以充分利用模态间的互补信息，再通过门控校准模块过滤交互后特征中的噪声信息，来校准交互后图文表示的情感表征偏移；通过多头注意力融合图文特征进行情感识别。在2个公开的多模态数据集上进行实验，结果表明：所提方法优于CLMLF、MVCN等先进方法，能提高情感识别的准确性。

关 键 词：多模态；图文情感分析；跨模态交互；情感表征偏移；注意力机制
中图分类号：TP391.41
文献标志码：A　　**文章编号**：1001-5965(2026)11-3724-11

随着社交网络媒体的发展，人们拥有更多的方式去表达自己的情感，如同时发布图像和文字表达自己的情绪或观点[1]。多模态情感分析的目的是从多种模态的数据中分析出其中的情感，相较基于单一模态的情感分析，能够更准确地理解人们的情感表达，因此，受到越来越多的关注和研究[2]，并在情感跨模态检索[3]、观点挖掘[4]、图文匹配[5]等领域逐渐得到广泛应用。

本文的研究重点是识别图文评论中的情感。该任务需要根据图像和文本的内容，判断出该言论表达的情感。与单模态情感分析相比，基于图文的情感分析面临的主要挑战在于如何交互不同模态间的信息，以挖掘图文间的语义关联性。通常来说，图像和文本具有互补性，在用户表达情感时，只有同时结合图像和文本才能更好地理解其中的情感。

现有的图文情感分析方法中，许多研究将图像作为文本的附加信息，通过跨模态交互，利用图像帮助模型定位文本中的情感信息以用于情感识别，但这种单向的交互方式没有充分利用图文间的互补关系。同时，跨模态交互过程中，容易引入其他模态的冗余信息，导致交互结果含噪声，无法准确表征真实情感。这种情感表征的偏移，从空间上看，会导致模型学习到的图文特征向量与其他情感类别数据的特征向量混杂在一起，区分度不足，降低情感识别的效果。

为解决上述问题，本文提出一种基于情感表征校准的图文情感分析模型(sentiment representation calibration-based model for image-text sentiment analysis, SRC-Model)。首先，将图文对输入特征提取模块，获得重要的图像和文本表示，用于跨模态交互。然后，使用一种跨模态的注意力交互机制，在“图-文”和“文-图”2个方向上进行模态间的双向

收稿日期：2024-07-18；**录用日期**：2024-09-23；**网络出版时间**：2024-11-19 15:27
网络出版地址：link.cnki.net/urlid/11.2625.V.20241119.1312.002
基金项目：国家自然科学基金面上项目(62076006)；认知智能全国重点实验室开放课题(COGOS-2023HE02)；安徽高校协同创新项目(GXXT-2021-008)
* **通信作者**. E-mail：sxzhang@aust.edu.cn

引用格式：*张顺香，刘佳佳，焦熠璇，等. 基于情感表征校准的图文情感分析模型 [J]. 北京航空航天大学学报，2026，52（11）：3724-3734.*
Zhang S X，Liu J J，Jiao Y X，et al. Sentiment representation calibration-based model for image-text sentiment analysis[J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52（11）：3724-3734（in Chinese）.

交互。该注意力机制结合图文模态的信息, 调整图文 2 个模态的注意力权重矩阵, 从而突出图像和文本中与情感表达有关的重要特征信息。针对交互结果可能出现的情感表征偏移, 进一步设计出一种简单有效的门控校准模块, 该模块基于门控机制过滤交互所得特征中的噪声, 提高交互后特征的情感表征能力。最后, 将图文特征输入多头注意力情感识别模块中, 进行特征融合和情感识别。在 2 个公开的基准数据集上进行的实验表明, 相较于先进的图文情感分析模型, SRC-Model 具有更加优越的性能表现。

综上, 本文的主要贡献如下:

1) 提出 SRC-Model, 通过跨模态的注意力机制进行图文信息间的双向交互, 从而利用二者的语义互补关系, 突出重要的情感特征信息, 增强图文表征能力。

2) 针对跨模态交互后的图文表示可能存在情感表征偏移的问题, 设计门控校准机制, 校准情感表征偏移, 提升模型情感识别的准确性。

1　相关工作

1.1　文本情感分析

早期的文本情感分析主要通过提取文本中包含的情感词来建立情感词典用于情感分析。Wiebe 等[6] 构建了大型语料库, 利用机器学习分类器建立基于情感的预测模型。但是, 这种依靠构建词典的情感分析方法, 其准确率受限于词典和规则集的完备性。

因此, 随着机器学习的发展, 情感分析方法也越来越多样化。Zhang 等[7] 根据情感强度将数据集分为强情感和弱情感 2 个子集, 使虚假评论失去隐蔽性, 再通过半监督的 PU 算法进行检测。此外, 也有工作结合无监督技术和情感词典, 研究评论中的情感特征及其相关规范[8]。

受深度学习模型在自然语言处理上的卓越性能的影响, Kim[9] 使用多通道卷积神经网络提取文本的 N-gram 特征, 并形成最终的句子表示, 以便进行分类。Wang 等[10] 引入树状长短期记忆模型表示语法属性的树结构网络拓扑, 从而深入挖掘文本语义特征。此外, 由于 Transformer[11] 结构的出现, 预训练模型在提取文本特征方面展现出了卓越的性能。以 BERT[12] 为代表的文本预训练模型在情感分析领域得到广泛应用。Dai 等[13] 基于 RoBERTa[14] 生成以情感词为导向的依存句法树, 应用在文本方面级情感分析中, 提升了情感分析的准确性。

1.2　图文情感分析

由于图文模态异质性带来的挑战, 如何对图文进行交互, 有效挖掘出其中的情感语义联系是迫待解决的研究难题。

传统多模态研究主要根据不同模态的分析结果得出最终决策。Liu 等[15] 通过独立为图像和文本建立情感分类器, 并根据各自的预测得分分配权重后相加, 来判断最终的情感倾向。这种方法忽略了图文之间的情感语义关联, 因此, Xu[16] 通过图像描述提取视觉语义特征, 作为文本的辅助信息分析情感表达。为了更充分地利用图像中的深层语义特征, Xu 等[17] 又提出了 MultiSentiNet, 通过识别图像中的对象和场景等显著特征, 辅助提取关键情感词汇, 并与视觉语义特征、对象和场景信息相结合进行情感识别。Huang 等[18] 提出以文本为中心的跨模态对齐方法, 得到文本中与情感分析任务相关的特征信息。

这些方法虽初步考虑了图文交互, 但忽略了图文模态彼此间的相互作用, 对图文间语义关系的利用不够充分。Yang 等[19] 通过注意力记忆网络模块对文本和图像特征进行交互式学习, 实现对图像和文本之间跨模态依赖关系的建模。Zhu 等[20] 使用跨模态注意力对齐图像的情感区域和文本单词, 充分探索图文间的交互作用。这种使用注意力的方式能够有效地进行图文交互。为了进一步探索图文间的关联性, Chen 等[21] 引入图像-文本相关性分类器, 解决图像和文本描述存在矛盾的情况。吴思思等[22] 通过自注意力融合多个模态的特征信息, 从而综合分析多模态上下文信息, 解决模态间的一致性和差异性问题。Li 等[23] 利用多层融合模块对齐和融合文本与图像特征, 学习多模态情感相关特征, 同时引入对比学习, 但该方法粒度较粗, 未充分考虑图像中大量的冗余视觉信息的存在。因此, Wei 等[24] 考虑视觉冗余信息的影响, 利用标签信息估计情感质心, 通过限制样本与质心间的距离来校准特征偏移, 但其简单地以文本为中心, 未考虑图像在情感表达中的重要作用。

以上工作未充分利用图像和文本之间的互补关系, 或忽视了跨模态交互过程会引入冗余信息, 这 2 种原因都会导致模态交互不够充分, 交互所得的图文特征对真实情感的表征不准确。因此, 本文在跨模态交互阶段利用跨模态的注意力交互机制进行图文间的双向交互, 该注意力在每个方向上都通过引入另一模态的信息调整当前模态的自注意力权重, 从而精细化地为图文中的情感特征信息加权, 获取在情感表达中具有一致性的信息。在注意力交互后, 通过门控校准模块, 过滤交互后特征中的噪声, 以提升其情感表征能力。

2 本文方法

如图1所示，本文提出的SRC-Model包括3个部分：特征提取、跨模态交互和情感识别。首先，在特征提取模块，通过预训练的BERT[12]和ResNet50[25]分别提取文本和图像的特征信息；然后，输入跨模态交互模块中，在“图-文”和“文-图”2个方向上，通过跨模态的注意力交互机制实现图文间的双向交互，并利用门控校准模块校准交互后可能出现的情感表征偏移；最后，将校准所得的图文表示输入情感识别模块，利用多头注意力进行特征融合，再实现情感识别。

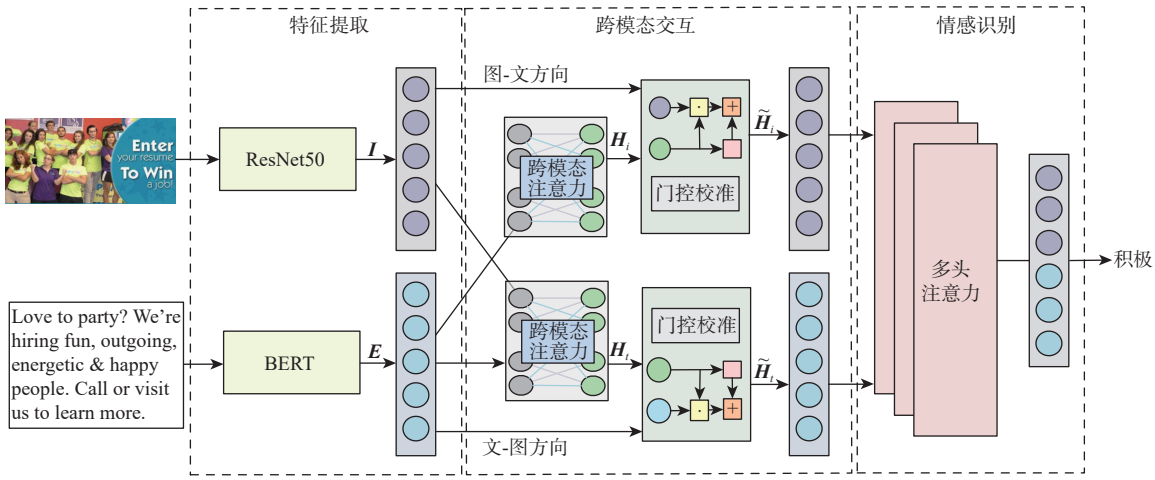


图1 SRC-Model结构

Fig. 1 Structure of SRC-Model

2.1 问题定义

图文情感分析的目的是根据用户发布的图像和文本预测其情感倾向。给定一个图文对(T,M)，图文对中句子$T=\{t_1,t_2,\cdots,t_n\}$，t_i为句子T的第i个词，M为文本T对应的一张图像。模型根据输入的图文对预测出该图文对表达的情感极性。

2.2 特征提取

为获得图文交互所需的图像和文本特征，本文使用2种不同的单模态编码器来提取其特征。

1）文本特征提取。BERT是一种基于Transformer的预训练语言模型，能够理解文本中的复杂语义关系。因此，本文使用预训练的BERT模型作为文本模态的编码器。

对于给定的图文对中的文本，将其最大长度设置为n，超过长度的部分截断，不足的补零。输入BERT模型中，编码为向量化的文本表示$\boldsymbol{E}$：

$$\boldsymbol{E}=\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n\}=\mathrm{BERT}(T)\in\mathbf{R}^{n\times d_t} \tag{1}$$

式中：n为文本的最大长度；d_t为BERT输出的词向量维度，d_t=768；$\boldsymbol{e}_i\in\mathbf{R}^{d_t}$为第$i$个单词经BERT编码后的词向量。

2）图像特征提取。ResNet是一种深度卷积神经网络，通过残差连接解决梯度消失问题，从而构建更深层的网络结构，捕获视觉特征。因此，本文使用预训练的ResNet50作为图像模态的编码器。

首先，将图像转换成224像素×224像素的尺寸，输入预训练的ReNet50模型中，将其最后一个卷积层得到的特征图$\boldsymbol{I}'$作为输出：

$$\boldsymbol{I}'=\mathrm{ResNet50}(M)\in\mathbf{R}^{p\times p\times 2\,048} \tag{2}$$

式中：p为特征图的高度和宽度，p=7；2 048为特征图的维度。该特征图$\boldsymbol{I}'$具有$p\times p$个视觉块，每个视觉块由一个2 048维的向量表示。

为将三维特征图表示矩阵转换为与文本一致的二维表示矩阵，以用于注意力交互，先通过全连接层降低特征图中每个视觉块的维度，避免过高的维度带来的计算负担，再通过Trans( )操作得到用于注意力交互的图像表示。Trans( )操作包括：先重塑特征图$\boldsymbol{I}'$，参考Li等的工作[23]，将形状为$p\times p\times 2\,048$的特征图的前2个维度展平成1个维度，得到由p^2个视觉块向量组成的图像表示$\tilde{\boldsymbol{I}}\in\mathbf{R}^{p^2\times d_i}$；再通过拼接零向量，使视觉块的数目与文本的最大长度相同，均为n，以用在后续的注意力交互中。表示如下：

$$\boldsymbol{I}=\mathrm{Trans}(\boldsymbol{I}'\boldsymbol{W}_{\boldsymbol{I}'}+\boldsymbol{b}_{\boldsymbol{I}'})=\{\boldsymbol{i}_1,\boldsymbol{i}_2,\cdots,\boldsymbol{i}_n\}\in\mathbf{R}^{n\times d_i} \tag{3}$$

式中：$\boldsymbol{W}_{\boldsymbol{I}'}$和$\boldsymbol{b}_{\boldsymbol{I}'}$为全连接层的可训练参数；$d_i=40$。

2.3 跨模态交互

2.3.1 跨模态注意力机制

不同模态的数据具有独特的情感信息，表达情感时又能互为补充，相互促进。用户在表达情感时，可能情感色彩更加偏向于图像，也可能更加偏

向于文本, 因此, 设计有效的方式去探索二者相互间的互补关系能够捕获更多、更准确的情感信息。本文通过一种跨模态注意力实现“图-文”和“文-图”2 个方向上的交互。该注意力机制引入另一方向上模态的信息, 来调整当前方向上模态的自注意力矩阵, 以利用模态间的互补关系关注在情感上具有联系的情感特征信息。经过注意力交互后的图像和文本表示, 既通过自注意力关注模态内部的特征信息, 又能捕获图文模态间的共性。

跨模态注意力应用在文本引导的“文-图”和图像引导的“图-文”2 个方向上。以“文-图”方向为例, 如图 2 所示。

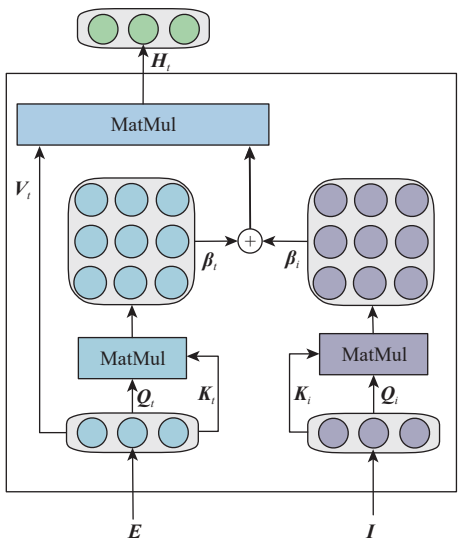


图 2　文-图方向上的跨模态注意力

Fig. 2　Cross-modal attention in text-to-image direction

对于文本模态, 定义注意力机制的 query 和 key 为文本的特征提取结果: $\boldsymbol{Q}_t = \boldsymbol{K}_t = \boldsymbol{E}$。对于图像模态, 定义注意力机制的 query 和 key 为图像的特征提取结果: $\boldsymbol{Q}_i = \boldsymbol{K}_i = \boldsymbol{I}$。分别计算文本和图像的自注意力分数矩阵, 文本自注意力分数矩阵$\boldsymbol{\beta}_t$和图像自注意力分数矩阵$\boldsymbol{\beta}_i$的计算过程如下:

$$\boldsymbol{\beta}_t = \mathrm{ReLU}\left(\boldsymbol{Q}_t \boldsymbol{K}_t^{\mathrm{T}}\right) \tag{4}$$

$$\boldsymbol{\beta}_i = \mathrm{ReLU}\left(\boldsymbol{Q}_i \boldsymbol{K}_i^{\mathrm{T}}\right) \tag{5}$$

为通过图文间的交互动态地调整文本中不同特征信息的权重, 受到 Yang 等[26] 在语音文本工作上的启发, 本文对文本自注意力分数矩阵$\boldsymbol{\beta}_t$和图像自注意力分数矩阵$\boldsymbol{\beta}_i$进行求和, 得到加权融合后的文本注意力分数矩阵$\boldsymbol{A}_t$。

$$\boldsymbol{A}_t = \boldsymbol{\beta}_t \boldsymbol{W}_{tt} + \boldsymbol{\beta}_i \boldsymbol{W}_{ti} + \boldsymbol{b}_{ti} \tag{6}$$

式中: $\boldsymbol{W}_{tt}$、$\boldsymbol{W}_{ti}$和$\boldsymbol{b}_{ti}$为加权融合计算的可训练参数。

将加权融合后得到的文本自注意力分数矩阵$\boldsymbol{A}_t$经过 Softmax 函数得到注意力权重矩阵, 为文本特征表示$\boldsymbol{E}$的不同特征赋予权重, 突出重要的情感信息, 计算出“文-图”方向上经过跨模态注意力交互后的输出$\boldsymbol{H}_t$, 如下:

$$\boldsymbol{H}_t = \mathrm{Softmax}\left(\boldsymbol{A}_t\right)\boldsymbol{E} \tag{7}$$

同理, 经过在“图-文”方向上的跨模态注意力, 得到图像在交互后的输出$\boldsymbol{H}_i$。

2.3.2　门控校准

跨模态交互易引入不同模态的冗余信息, 导致输出结果中含噪声。由于受到噪声信息的掩盖, 跨模态注意力的输出$\boldsymbol{H}_t$和$\boldsymbol{H}_i$中, 情感特征信息不能很好地表达, 因而产生情感表征偏移。为解决该问题, 一种可行的方式是从未经交互的原始图文表示($\boldsymbol{E}$和$\boldsymbol{I}$)中获取信息来增强情感特征的表达。因此, 本文利用门控机制去校准跨模态注意力的输出可能存在的情感表征偏移。

以图 3 中“文-图”方向上的门控校准模块的结构为例:

$$\boldsymbol{g}_t = \sigma\left(\boldsymbol{H}_t \boldsymbol{W}_{\mathrm{g}}\right) \tag{8}$$

$$\boldsymbol{H}'_t = \boldsymbol{g}_t \odot \boldsymbol{H}_t + (\boldsymbol{O} - \boldsymbol{g}_t) \odot \boldsymbol{E} \tag{9}$$

式中: σ表示 Sigmoid 函数; $\boldsymbol{W}_{\mathrm{g}}$为可训练参数; $\odot$表示矩阵间的逐元素乘法; $\boldsymbol{g}_t$为门控值; $\boldsymbol{O}$为全 1 矩阵。当$\boldsymbol{g}_t$值较高时, 说明$\boldsymbol{H}_t$中含有较少的噪声, 门控会偏向于来自跨模态注意力交互后的文本信息, 因此给出更高的门控值, 当$\boldsymbol{g}_t$值较低时, 说明$\boldsymbol{H}_t$中含有较多的噪声, 在一定程度上掩盖了情感特征的表达, 门控会偏向于来自跨模态注意力交互前的文本信息, 因此给出较低的门控值。根据门控值的动态调控, 模型可以在跨模态注意力的输出$\boldsymbol{H}_t$存在较多噪声时, 更多地选择$\boldsymbol{E}$中的低噪信息融合进来, 从而过滤噪声, 加强$\boldsymbol{H}_t$中情感特征的表达, 以校准情感表征偏移。经过该计算, 得到校准了情感表征偏移的文本特征表示$\boldsymbol{H}'_t$。

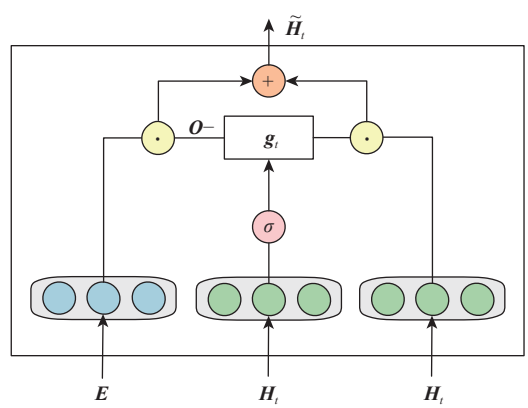


图 3　文-图方向上的门控校准模块

Fig. 3　Gated calibration module in text-to-image direction

进一步，为保证值的区间稳定，将$\boldsymbol{H}_t'$归一化，如下：

$$\tilde{\boldsymbol{H}}_t = \mathrm{ReLU}\left(\boldsymbol{H}_t'\boldsymbol{W}_c + \boldsymbol{b}_c\right) \tag{10}$$

式中：$\boldsymbol{W}_c$、$\boldsymbol{b}_c$为可训练参数。

经过式(8)～式(10)，得到跨模态交互模块最终输出的文本特征表示$\tilde{\boldsymbol{H}}_t$。

与上述在“文-图”方向上计算原理一致，可以获得在“图-文”方向上经过门控校准模块的图像特征表示$\tilde{\boldsymbol{H}}_i$。

2.4 情感识别

为使用图文中的情感特征信息进行情感识别，将跨模态交互输出的图像和文本特征拼接在一起，作为情感识别阶段的输入。

$$\boldsymbol{F} = \left[\tilde{\boldsymbol{H}}_t;\tilde{\boldsymbol{H}}_i\right] \tag{11}$$

式中：[;]表示拼接操作；$\boldsymbol{F}$为包含全部情感信息的图文情感特征表示。

由于图像和文本中的不同情感特征信息在表达情感时的贡献不同，不适合直接用来进行情感识别，因此，参考Du等的工作[27]，本文使用多头自注意力对图文特征进行融合，来分析不同特征之间的关系用于情感识别。$\boldsymbol{F}_x$是$\boldsymbol{F}$在多头注意力中每个头的输入，公式如下：

$$\boldsymbol{Q}_j = \boldsymbol{F}_x\boldsymbol{W}_{\boldsymbol{Q}_j} + \boldsymbol{b}_{\boldsymbol{Q}_j} \tag{12}$$

$$\boldsymbol{K}_j = \boldsymbol{F}_x\boldsymbol{W}_{\boldsymbol{K}_j} + \boldsymbol{b}_{\boldsymbol{K}_j} \tag{13}$$

$$\boldsymbol{V}_j = \boldsymbol{F}_x\boldsymbol{W}_{\boldsymbol{V}_j} + \boldsymbol{b}_{\boldsymbol{V}_j} \tag{14}$$

式中：$\boldsymbol{W}_{\boldsymbol{Q}_j}$、$\boldsymbol{W}_{\boldsymbol{K}_j}$、$\boldsymbol{W}_{\boldsymbol{V}_j}$为全连接层的权重矩阵，$\boldsymbol{b}_{\boldsymbol{Q}_j}$、$\boldsymbol{b}_{\boldsymbol{K}_j}$、$\boldsymbol{b}_{\boldsymbol{V}_j}$为偏置项，均为可训练参数，$j$为多头注意力的第$j$个头，本文将多头个数设置为8。

多头注意力每个头都进行相同的注意力计算，其中，注意力的计算公式如下：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{15}$$

通过上述多头注意力计算公式，可以计算出每个头的注意力结果，如下：

$$\boldsymbol{u}_j = \mathrm{Attention}\left(\boldsymbol{Q}_j,\boldsymbol{K}_j,\boldsymbol{V}_j\right) \tag{16}$$

得到每个头的注意力计算结果$\boldsymbol{u}_j$后，将其拼接起来，得到图文联合特征$\boldsymbol{U}'$，如下：

$$\boldsymbol{U}' = \mathrm{concat}(\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_{\mathrm{num}}) \tag{17}$$

为充分利用图文联合特征$\boldsymbol{U}'$的所有特征信息去计算不同类别下的情感预测概率，将$\boldsymbol{U}'$展开得到一个高维向量，再通过全连接层映射到不同情感类别的目标空间，得到预测情感向量，公式如下：

$$\boldsymbol{U} = \mathrm{flatten}(\boldsymbol{U}') \tag{18}$$

$$\boldsymbol{Y} = \mathrm{Tanh}\left(\boldsymbol{U}\boldsymbol{W}_f + \boldsymbol{b}_f\right) \tag{19}$$

式中：flatten( )表示展平操作；$\boldsymbol{W}_f$和$\boldsymbol{b}_f$为全连接层的可训练参数；Tanh为激活函数；s为标签数目；$\boldsymbol{Y} \in \mathbf{R}^s$为映射到$s$个情感类别空间的预测情感向量。

使用Softmax计算不同情感类别的预测概率$\boldsymbol{\delta}$，具体公式如下：

$$\boldsymbol{\delta} = \mathrm{Softmax}(\boldsymbol{Y}) \in \mathbf{R}^s \tag{20}$$

取$\boldsymbol{\delta}$中最高预测概率值的索引为预测标签，如下：

$$p_t = \mathrm{argmax}(\boldsymbol{\delta}) \tag{21}$$

由于数据的类别不平衡会导致模型无法充分学习少类别样本的信息，损失函数使用Focal Loss[28]，使模型更加关注难分类的少类别样本。Focal Loss公式如下：

$$L = -\alpha_t(1-p_t)^{\gamma}\ln p_t \tag{22}$$

式中：p_t为模型预测的结果；α_t为类别权重系数，是可训练参数；γ为可调节因子，默认为2；L为计算到的损失值。

3 实验结果与分析

3.1 数据集

本文在2个公开的多模态数据集[29]上进行实验：MVSA-Single和MVSA-Multiple。2个数据集的样本均是从Twitter上爬取到的图文评论。MVSA-Single包括5 129个图文对，每个图文对由单个注释者分别标注为积极、中性或消极情感标签。而MVSA-Multiple则包括19 600个图文对，每个图文对由3个注释者独立标注情感标签。

由于图文标签存在不一致，本文采取与文献[24]相同的预处理策略。对于MVSA-Single数据集，若图像和文本的标签感情极性相反，则舍弃该条数据；若其中一方为中性，则采用另一方标签作为该图文对的最终情感标签。对于MVSA-Multiple数据集，则选择出现次数最多的注释标签作为图文对的最终情感标签。处理好的数据集中每个标签数据个数的统计信息如表1所示。

表1 MVSA数据集数据

Table 1 Statistics of MVSA datasets

数据集	积极	中性	消极	总计
MVSA-Single	2 683	470	1 358	4 511
MVSA-Multiple	11 318	4 407	1 299	17 024

3.2 实验细节

3.2.1 参数设置

所有代码均在pytorch框架下完成，模型的训

练在一块 3080Ti GPU 上进行。对数据集按照 8∶1∶1 的比例拆分为训练集、验证集和测试集。

BERT 的词嵌入维度为 768。由于数据集文本属于短评, 大部分文本长度处于 50 以内, 因此, 最大文本长度设置为 50。所有图像均转换为 224 像素×224 像素的三通道 RGB 图像, 输入 ResNet50 中。模型使用 Adam 优化器更新参数。多头注意力的头数为 8。其余有关参数设置如表 2 所示。

表 2　参数设置

Table 2　Parameter settings

数据集	批大小	学习率	权重衰减系数	训练轮数	γ
MVSA-Single	16	1×10^{-5}	1×10^{-4}	10	2
MVSA-Multiple	32	8×10^{-5}	1×10^{-4}	10	2

3.2.2　评价指标

对模型的评价指标, 同 Wei 等[24] 的工作一致, 使用准确率和 Weighted-F1 作为评价指标, 其中, 准确率和 F_1 定义如下:

$$A=\frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{23}$$

$$P=\frac{T_P}{T_P+F_P} \tag{24}$$

$$R=\frac{T_P}{T_P+F_N} \tag{25}$$

$$F_1=\frac{2PR}{P+R} \tag{26}$$

式中: T_P、T_N、F_P、F_N分别表示真正例的数量、真负例的数量、假正例的数量、假负例的数量。Weighted-F1 即对每个类别的 F_1 值进行加权平均, 每个类别的权重为该类别样本占全部样本数量的比重。

3.3　对比实验

3.3.1　对比模型

为对比模型在图文情感分析上的表现, 本文选择以下具有代表性的对比模型在 MVSA-Single 和 MVSA-Multiple 数据集上进行对比。

1) BERT[12]: 一种基于 Transformer 架构的预训练语言模型, 经过微调用于文本情感分析。

2) ResNet50[25]: 一种深度残差网络结构, 经过微调进行基于图像的情感分析。

3) Vision Transformer (ViT)[30]: 一种基于 Transformer 的视觉感知模型, 在图像分类等任务上具有出色的能力, 因此用于基于图像的情感分析。

4) HSAN[16]: 一种基于图像描述的分层语义注意力网络。对于输入的图像-文本对, 使用图像描述提取视觉语义特征, 作为文本的附加信息。

5) MultiSentiNet[17]: 从图像中提取物体和场景作为视觉语义信息, 通过注意力引导提取文本重要单词, 将所有的视觉和语义特征聚合在一起进行预测。

6) Co-MN-Hop6[31]: 迭代地对图文交互进行建模, 即利用文本查询图像中的关键信息, 并利用图像信息来定位文本重要词语。

7) MGNNS[32]: 用于图像文本情感识别的多通道图神经网络, 通过多通道图神经网络学习多模态数据集的全局特征。

8) CLMLF[23]: 通过基于标签和数据的对比学习对齐和融合文本和图像模态, 提高模型提取和融合多模态数据特征的能力。

9) MVCN[24]: 文本引导的融合模块, 基于情感的一致性约束和用于多模态情感识别的自适应损失校准策略, 以系统地解决来自不同视图的模态异构问题。

3.3.2　对比实验结果及分析

将本文模型与 3.3.1 节对比模型在 2 个基准数据集上进行对比, 结果如表 3 所示。可知, 在 MVSA-Single 数据集上, SRC-Model 性能优于其他的对比模型, 准确率为 78.32%, Weighted-F1 值为 76.79%, 较 SOTA 模型 MVCN 分别提高了 2.26% 和 2.24%; 在 MVSA-Multiple 数据集上, SRC-Model 的准确率为 72.81%, Weighted-F1 值为 70.13%, 较先进模型 MVCN 分别提高了 0.74% 和 0.12%。这主要是因为: MVCN 将图像作为文本的附加信息, 利用图像模态增强文本模态的表示。这种方式对图文间的互补关系利用不足, 因此交互不够充分。而本文通过跨模态的注意力进行图文间的双向交互, 并利用门控校准机制对交互结果中的噪声进行过滤, 能够充分利用模态间的互补关系增强每个模态的情感表征能力。

与单模态模型 BERT、ResNet50 和 ViT 相比, SRC-Model 在 2 个基准数据集上的对比指标均提高了约 10%。仅仅基于单个模态(如文本或图像), 难以有效针对用户的多元化表达方式分析其情感倾向, 因此, 结合图文的多模态模型整体上的实验结果均优于单模态模型。同时也可以发现, 基于文本的模型在情感识别时的效果优于基于图像的模型, 这种实验结果在一定程度上可以表明, 文本中的情感信息比图像更为丰富, 反映真实情感的能力更强。

与 MultiSentiNet 和 HSAN 模型相比, SRC-Model 在 MVSA-Single 数据集上的准确率和 Weighted-F1 值均提升了近 10%, 在 MVSA-Multiple 数据集上的准确率和 Weighted-F1 值均提升了约 5%。以上 2 个模型均采用了图文间的单向交互机制, 即将图像作为文本的附加信息, 利用图像增强文本模态的

表3 不同模型在MVSA-Single和MVSA-Multiple数据集上的对比实验结果

Table 3 Comparative experimental results of different models on MVSA-Single and MVSA-Multiple datasets

模态	模型	准确率/%		Weighted-F1/%	
		MVSA-Single	MVSA-Multiple	MVSA-Single	MVSA-Multiple
文本	BERT[12]	72.35	67.70	70.81	67.10
图像	ResNet50[25]	67.04	62.83	64.41	60.94
	ViT[30]	67.92	65.12	67.88	62.99
多模态	HSAN[16]	69.88	67.96	66.90	67.76
	MultiSentiNet[17]	69.84	68.86	69.63	68.11
	Co-MN-Hop6[31]	70.51	68.92	70.01	68.83
	MGNNS[32]	73.77	72.49	72.70	69.34
	CLMLF[23]	75.33	72.00	73.46	69.83
	MVCN[24]	76.06	72.07	74.55	70.01
	SRC-Model(本文)	78.32	72.81	76.79	70.13

情感表达信息,这种方法忽视了图文在情感表达时具有相互补充的关系,因此对图像中表达的情感信息利用不足。

与Co-MN-Hop6、MGNNS和CLMLF模型相比,SRC-Model在2个数据集上均具有优势。这3种模型直接对图文数据进行交互和融合,在结合图文进行挖掘情感信息的同时,也会引入更多的冗余信息,这使得模型较难准确聚焦于重点的信息上。而本文通过跨模态注意力和门控校准机制建立起图文间的双向交互,不仅能够有效地利用图文间的互补关系挖掘情感信息,并且过滤了交互结果的噪声,准确获取图文中的重要情感信息,为情感分析提供有力支撑。

3.4 消融实验

本文在MVSA-Single和MVSA-Multiple数据集上对模型进行了消融实验,进一步评估了模型的不同组成模块对情感识别的贡献程度。具体实施方案如下:

1) BERT: 仅基于文本模态,使用BERT模型进行情感识别。

2) ResNet50: 仅基于图像模态,使用ResNet50进行情感识别。

3) SRC+M: 分别独立地使用BERT和ResNet50提取文本和图像的特征,利用多头注意力融合图文特征,再预测情感标签。

4) SRC+M+A: 在SRC+M的基础上添加跨模态注意力交互模块,进行图像和文本间的双向交互。

5) SRC-TI: 在SRC+M+A的基础上添加门控校准模块,并且仅实现“文-图”方向的交互。

6) SRC-Model(本文模型): 在SRC+M+A的基础上添加门控校准模块,进行图文间双向交互。

由表4、表5可以看出:

1) SRC+M在2个数据集上的准确率和Weighted-F1值都高于独立处理文本和图像的BERT和ResNet50模型,说明加入的多头注意力能够有效整合图像和文本中的情感信息。

2) SRC+M+A相比SRC+M,在2个数据集上的指标均有所提升,主要是因为:跨模态注意力能够捕捉图文模态间的情感联系,从而准确地发现图文中的情感特征信息。

3) SRC-Model相比SRC+M+A有进一步的提升,这表明加入门控校准模块能够提升跨模态交互的效果。主要原因在于:门控校准模块可以通过门控机制自适应地增强跨模态注意力输出中的情感特征信息,从而过滤掉噪声的影响,校准情感表征偏移。

表4 MVSA-Single数据集上的消融实验结果

Table 4 Ablation results on MVSA-Single dataset

模型	准确率/%	Weighted-F1/%
BERT	72.35	70.81
ResNet50	67.04	64.41
SRC+M	74.56	74.54
SRC+M+A	76.33	76.07
SRC-TI	77.21	76.72
SRC-Model	78.32	76.79

表5 MVSA-Multiple数据集上的消融实验结果

Table 5 Ablation results on MVSA-Multiple dataset

模型	准确率/%	Weighted-F1/%
BERT	67.70	67.10
ResNet50	62.83	60.94
SRC+M	69.41	67.47
SRC+M+A	71.99	67.89
SRC-TI	72.58	68.72
SRC-Model	72.81	70.13

4) SRC-Model 的 2 个评测指标值要优于 SRC-TI, 这是因为 SRC-TI 仅采取了“文-图”方向上的单向交互, 这种利用图像增强文本的方法并不能很好地利用图文间的互补性, 因此性能不足。

总体而言, SRC-Model 通过跨模态注意力机制进行图文间的双向交互, 可以引入模态间的互补关系, 从而发现图像和文本中与情感相关的特征信息。此外, 门控校准机制的引入解决了注意力交互结果可能含噪声产生的情感表征偏移问题。因此, SRC-Molde 充分探索了模态间的交互作用, 实验结果也验证了本文方法在情感识别任务上的有效性。

3.5　注意力头数分析

为探究不同多头注意力头数对实验的影响, 根据以往工作[11]将头数设置为 1、4、8、12、16。在 MVSA-Single 数据集上进行实验, 结果如图 4 所示。

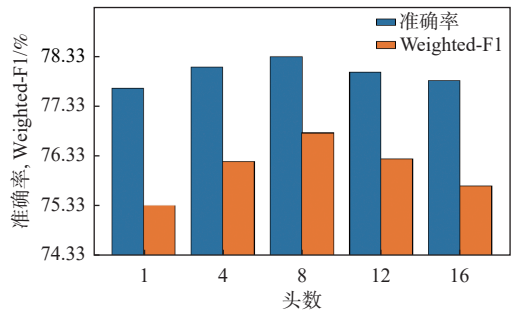


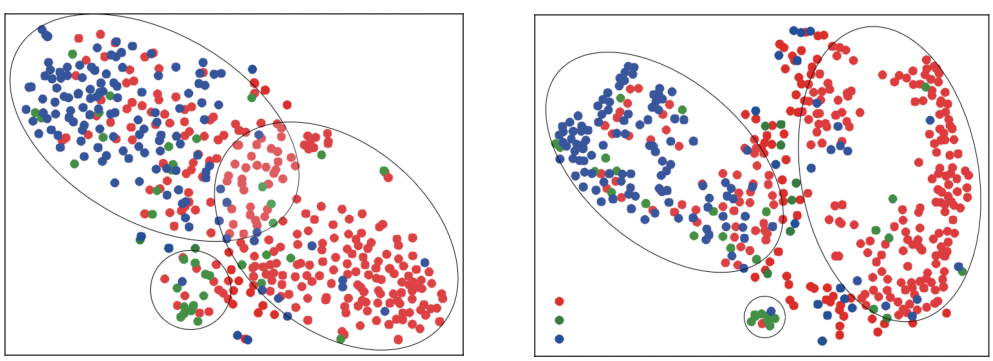
图 4　注意力头数对准确率和 Weighted-F1 值的影响

Fig. 4　Effect of attention heads on accuracy and Weighted-F1

通过观察可知, 注意力头数为 8 时, 模型的性能最优。注意力头数大于 8 时, 效果逐渐下降。说明增加注意力头数可以较好地融合图像和文本特征, 这是因为头数增加, 模型可以从更多的子语义空间关注图文内容与情感具有联系的相关特征信息。但注意力头数过多时, 也会导致语义空间过于碎片化, 使模型更容易记住一些不重要的信息, 造成注意力分散, 导致每个头无法捕捉到有价值的信息片段, 使得模型整合来自每个注意力头的信息时效果不理想。因此, 在利用多头注意力融合图像和文本信息时, 需要合适的头数设置, 模型才能更好地学习图文模态间的复杂关系, 从而提升对情感的识别能力。

3.6　特征分布可视化

为验证门控校准模块是否校准了图文特征的情感表征偏移, 本文设计了直观的可视化实验观察这一现象。在 MVSA-Single 数据集上, 运用 TSNE 降维算法对跨模态交互模块输出的特征向量进行了可视化分析。图 5(a)为去除门控校准模块后的可视化结果, 图 5(b)为添加门控校准模块后的可视化结果。



(a) 去除门控校准模块后的可视化结果　　(b) 添加门控校准模块后的可视化结果

图 5　特征分布可视化

Fig. 5　Feature distribution visualization

通过对比可以发现, 模型加入门控校准模块后, 跨模态交互得到的图文特征表示向量有明显的聚合趋势, 情感表征能力得到提高。具体分析如下: 图像和文本中客观存在大量与情感识别无关的特征, 被引入到交互过程, 导致交互到的图文特征向量中存在噪声, 无法准确地表征情感。因此, 相同情感类别数据的特征向量在空间中分布更为分散, 不同情感类别数据的特征向量混淆程度较高。然而, 在引入门控校准模块后, 由于对交互后结果中的噪声进行了过滤, 其中的情感信息得到增强, 各类别样本数据的特征向量区分度明显提高, 数据的聚合效果得到增强, 因此, 校准情感表征偏移能够优化不同情感类别数据的特征向量的情感表征能力, 有效提高图文情感分析的准确性。

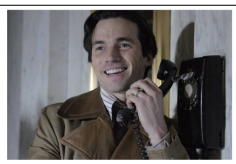
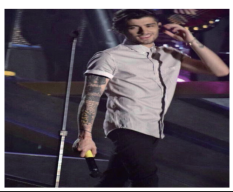
3.7　案例分析

为进一步证明本文模型的有效性, 比较了基于 BERT、ResNet50 和 SRC-Model 的分类结果, 如表 6 所示。

通过观察分类结果可以看出, 在单模态内容没有传达出明确的情感时, 单模态模型容易判断错

表6 不同模型的分类结果示例

Table 6 Examples classified by different models

图像	文本	真实情感	BERT	ResNet50	SRC-Model
	RT@dynamitefilm: A light hearted moment on set with our actor@IANMHARDING	积极	积极(√)	积极(√)	积极(√)
	Martha said for Valentine's Day she wanted a heart shaped pancake for lunch. Being the doting husband I am...	积极	积极(√)	中性(×)	积极(√)
	I HAVE NO WORDS IM SPEECHLESS	积极	消极(×)	积极(√)	积极(√)

误,而多模态模型能够较好地结合多模态内容给出正确的判断。例如,第1行的图像和文本分别通过笑脸和“light hearted moment”等内容表达了明确的积极情绪,因此模型均判断正确。第2行的图像中无法查询与情绪相关的线索,第3行中,仅根据文本内容难以确定其表达了积极还是消极的情绪,因此单模态模型均出现判断错误。

总体而言,图像和文本模态在表达情感上具有相互补充的作用,只有结合多模态内容,才能更全面准确地理解用户表达的情感。

4 结 论

1)本文针对社交媒体上的图文情感分析任务,提出一种基于情感表征校准的图文情感分析模型,在基准数据集MVSA上进行了大量实验,验证了模型的有效性。

2)引入门控校准机制,来校准特征空间中情感表征向量的情感表征偏移问题,消融实验和可视化实验的结果表明了该机制的有效性。

本文工作假设图文具有相互促进的作用,但部分情况下,图像和文本表达的语义并非一致,如有关讽刺的表达。图文语义的不一致为模态交互带来了挑战。未来将从模态间语义的一致性和不一致性出发,更加全面地分析多模态数据中的情感。

参考文献(References)

[1] 张亚洲, 戎璐, 宋大为, 等. 多模态情感分析研究综述[J]. 模式识别与人工智能, 2020, 33(5): 426-438.
Zhang Y Z, Rong L, Song D W, et al. A survey on multimodal sentiment analysis[J]. Pattern Recognition and Artificial Intelligence, 2020, 33(5): 426-438(in Chinese).

[2] Das R, Singh T D. Multimodal sentiment analysis: a survey of methods, trends, and challenges[J]. ACM Computing Surveys, 2023, 55(13s): 1-38.

[3] Kaur P, Pannu H S, Malhi A K. Comparative analysis on cross-modal information retrieval: a review[J]. Computer Science Review, 2021, 39: 100336.

[4] Serrano-Guerrero J, Romero F P, Olivas J A. Fuzzy logic applied to opinion mining: a review[J]. Knowledge-Based Systems, 2021, 222: 107018.

[5] 郑祎豪, 郭奕君, 毋立芳, 等. 知识驱动的多模态语义理解研究综述[J]. 模式识别与人工智能, 2023, 36(12): 1127-1138.
Zheng Y H, Guo Y J, Wu L F, et al. A survey on knowledge-driven multimodal semantic understanding[J]. Pattern Recognition and Artificial Intelligence, 2023, 36(12): 1127-1138(in Chinese).

[6] Wiebe J M, Bruce R F, O'Hara T P. Development and use of a gold-standard data set for subjectivity classifications[C]//Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. New York: ACM, 1999: 246-253.

[7] Zhang S X, Zhu A Q, Zhu G L, et al. Building fake review detection model based on sentiment intensity and PU learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(10): 6926-6939.

[8] Pashchenko Y, Rahman M F, Hossain M S, et al. Emotional and the normative aspects of customers' reviews[J]. Journal of Retailing and Consumer Services, 2022, 68: 103011.

[9] Kim Y. Convolutional neural networks for sentence classification [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Kerrville: ACL, 2014: 1746-1751.

[10] Wang L, Cao H, Yuan L. Gated tree-structured RecurNN for detecting biomedical event trigger[J]. Applied Soft Computing, 2022, 126: 109251.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural In-

formation Processing Systems. New York: ACM, 2017: 6000-6010.

[12] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Kerrville: ACL, 2019: 4171-4186.

[13] Dai J Q, Yan H, Sun T X, et al. Does syntax matter? a strong baseline for aspect-based sentiment analysis with RoBERTa[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Kerrville: ACL, 2021: 1816-1829.

[14] Liu Y H, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach[PP/OL]. arXiv (2019-07-26)[2024-07-14]. https://arxiv.org/abs/1907.11692.

[15] Liu N N, Dellandréa E, Chen L M, et al. Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme[J]. Computer Vision and Image Understanding, 2013, 117(5): 493-512.

[16] Xu N. Analyzing multimodal public sentiment based on hierarchical semantic attentional network[C]//Proceedings of the IEEE International Conference on Intelligence and Security Informatics. Piscataway: IEEE Press, 2017: 152-154.

[17] Xu N, Mao W J. MultiSentiNet: a deep semantic network for multimodal sentiment analysis[C]//Proceedings of the ACM on Conference on Information and Knowledge Management. New York: ACM, 2017: 2399-2402.

[18] Huang C Q, Zhang J L, Wu X M, et al. TeFNA: text-centered fusion network with crossmodal attention for multimodal sentiment analysis[J]. Knowledge-Based Systems, 2023, 269: 110502.

[19] Yang X C, Feng S, Wang D L, et al. Image-text multimodal emotion classification via multi-view attentional network[J]. IEEE Transactions on Multimedia, 2021, 23: 4014-4026.

[20] Zhu T, Li L D, Yang J F, et al. Multimodal sentiment analysis with image-text interaction network[J]. IEEE Transactions on Multimedia, 2023, 25: 3375-3385.

[21] Chen D L, Su W, Wu P, et al. Joint multimodal sentiment analysis based on information relevance[J]. Information Processing & Management, 2023, 60(2): 103193.

[22] 吴思思, 马静. 基于感知融合的多任务多模态情感分析模型[J]. 数据分析与知识发现, 2023, 7(10): 74-84.
Wu S S, Ma J. Multi-task & multi-modal sentiment analysis model based on aware fusion[J]. Data Analysis and Knowledge Discovery, 2023, 7(10): 74-84(in Chinese).

[23] Li Z, Xu B, Zhu C H, et al. CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection[C]// Proceedings of the Findings of the Association for Computational Linguistics. Kerrville: ACL, 2022: 2282-2294.

[24] Wei Y W, Yuan S Z, Yang R S, et al. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Kerrville: ACL, 2023: 5240-5252.

[25] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[26] Yang K C, Xu H, Gao K. CM-BERT: cross-modal BERT for text-audio sentiment analysis[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 521-528.

[27] Du Y P, Liu Y, Peng Z, et al. Gated attention fusion network for multimodal sentiment classification[J]. Knowledge-Based Systems, 2022, 240: 108107.

[28] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[PP/OL]. arXiv (2018-02-07)[2024-07-14]. https://doi.org/10.48550/arxiv.1708.02002.

[29] Niu T, Zhu S A, Pang L, et al. Sentiment analysis on multi-view social data[C]//Proceedings of the MultiMedia Modeling. Berlin: Springer, 2016: 15-27.

[30] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[PP/OL]. arXiv (2021-06-03)[2024-07-14]. https://doi.org/10.48550/arxiv.2010.11929.

[31] Xu N, Mao W J, Chen G D. A co-memory network for multimodal sentiment analysis[C]//Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York: ACM, 2018: 929-932.

[32] Yang X C, Feng S, Zhang Y F, et al. Multimodal sentiment detection based on multi-channel graph neural networks[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Kerrville: ACL, 2021: 328-339.

Sentiment representation calibration-based model for image-text sentiment analysis

ZHANG Shunxiang[1, 2, 3, *], LIU Jiajia[1, 2], JIAO Yixuan[1, 2], ZHANG Yulei[1, 2], LI Ziliang[1, 2]

(1. School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China;
2. Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230000, China;
3. School of Computer Science, Huainan Normal University, Huainan 232038, China)

Abstract: Exploring the association of multi-modal data for sentiment analysis is made more difficult by the growing trend of social media users expressing their opinions through both text and visuals. Existing research methods lack sufficient interaction between modalities and cross-modal interactions often introduce noise, which leads to sentiment representation bias of learned image and text representations. To address this issue, we introduce a sentiment representation calibration-based model for image-text sentiment analysis. Firstly, we encode the image and text through a feature extraction module to capture image and text features. In order to fully utilize the inter-modal complementary information, the cross-modal attention is then employed in the cross-modal interaction module to carry out bidirectional interaction in the "image-to-text" and "text-to-image" directions. Next, noise is filtered out from the interaction-enhanced features through a gated calibration module to correct the sentiment representation bias in the interaction-enhanced features. Finally, in the sentiment detection module, feature fusion is achieved by using multi-head attention to integrate the calibrated feature representations for sentiment detection. Experimental results on two public benchmark datasets demonstrate that the proposed model outperforms state-of-the-art image-text sentiment analysis methods, such as CLMLF and MVCN, thereby improving the accuracy of sentiment recognition.

Keywords: multimodal; image-text sentiment analysis; cross-modal interaction; sentimental representation bias; attention mechanism

Received: 2024-07-18; **Accepted**: 2024-09-23; **Published Online**: 2024-11-19 15:27

URL: link.cnki.net/urlid/11.2625.V.20241119.1312.002

Foundation items: General Program of National Natural Science Foundation of China (62076006); Opening Foundation of State Key Laboratory of Cognitive Intelligence (COGOS-2023HE02); University Synergy Innovation Program of Anhui Province (GXXT-2021-008)

* **Corresponding author**. E-mail: sxzhang@aust.edu.cn