http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0365

基于联合交互注意力的图文情感分析方法

胡慧君[1,2]，丁子毅[1,2]，张耀峰[3,*]，刘茂福[1,2]

(1. 武汉科技大学 计算机科学与技术学院，武汉 430065；2. 武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室，武汉 430065；3. 湖北经济学院 湖北数据与分析中心，武汉 430205)

摘　　要：社交媒体中的图文情感对于引导舆论走向具有重要意义，越来越受到自然语言处理（NLP）领域的广泛关注。当前，社交媒体图文情感分析的研究对象主要为单幅图像文本对，针对无时序性及多样性的图集文本对的研究相对较少，为有效挖掘图集中图像与文本之间情感一致性信息，提出基于联合交互注意力的图文情感分析（SA-JIA）方法。该方法使用 RoBERTa 和双向门控循环单元 (Bi-GRU) 来提取文本表达特征，使用 ResNet50 获取图像视觉特征，利用联合注意力来找到图文情感信息表达一致的显著区域，获得新的文本和图像视觉特征，采用交互注意力关注模态间的特征交互，并进行多模态特征融合，进而完成情感分类任务。在 IsTS-CN 数据集和 CCIR20-YQ 数据集上进行了实验验证，结果表明：所提方法能够提升社交媒体图文情感分析的性能。

关 键 词：社交媒体；图文情感分析；联合注意力；交互注意力；多模态融合

中图分类号：TP391

文献标志码：A　　**文章编号**：1001-5965(2025)07-2262-09

随着网络技术的日益发展，越来越多的用户选择通过社交媒体表达观点，社交媒体已成为人们分享和日常交流中必不可少的工具，而用户在表达观点时，不仅会发布文本内容，还会配以图像内容，这些图像信息通常蕴含着用户的情感信息。因此，社交媒体的图文情感分析与理解也逐渐受到广泛关注。

事实上，用户在编辑社交媒体文案时，通常配以多幅图像。本文将包含多幅图像的集合统称为图集。然而，现有社交媒体情感分析多以文本[1]或单幅图文对[2]为主，少见针对图集文本对的情感分析研究。Truong 等[3]开展了面向图集文本对的社交媒体情感分析研究，指出图集中每幅图像对社交媒体整体情感判断的重要性有所不同，因此，以图像视觉特征来对齐文本，针对数据来源单一的餐厅评论图集进行了图文情感分析。而本文聚焦于社交媒体图集，数据来源多样，蕴含情感信息更加丰富，更具有挑战性。随着信息量的不断增加和网络新词的频繁出现，影响文本情感的情感因子信息粒度加大，而情感词典则只是固定长度的情感词汇或短语，虽然准确但难以及时更新，在颗粒度上也很难适应社交媒体上迅速更迭的情感数据。现如今人们常以文本结合图像的方式表达情感，选取能够相互映射的文本和图像来突显自身的情感。但由于图集中图像所表达情感的多样性，从而导致文本与图集之间的情感会表达不同。可以看出，现有方法未能充分考虑图集中部分图像与文本情感不一致的问题。例如：文本“#当代大学生暑假日常#你鲨了我吧，生活压得我喘不过气”，配图为一张生活压垮自己的表情包和一张带有微笑的自拍照。从“鲨了我吧”和配图表情包可以看出，图文的整体情感与文本的情感均表达为负面情感。从配图表

收稿日期：2023-06-15；**录用日期**：2023-10-12；**网络出版时间**：2023-11-23 15:37

网络出版地址：link.cnki.net/urlid/11.2625.V.20231122.1324.001

基金项目：国家自然科学基金 (62271359)；国家社科基金重点项目 (23ATJ005)；“十四五”湖北省优势特色学科（群）项目 (2023D0302)；湖北省教育厅科研重点项目 (20192202)

* **通信作者**. E-mail：yfzhang@hbue.edu.cn

引用格式：胡慧君，丁子毅，张耀峰，等. 基于联合交互注意力的图文情感分析方法 [J]. 北京航空航天大学学报，2025，51（7）：2262-2270.
HU H J，DING Z Y，ZHANG Y F，et al. Analysis of image and text sentiment method based on joint and interactive attention [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2262-2270 （in Chinese）.

情包可以看出, 图像表达为负面情感, 与图文整体情感保持一致, 但是带微笑的自拍照却表达正面情感, 与图文整体情感不一致。这种图集中存在图像与文本表达情感不一致的现象, 会影响图集文本情感分析的效果。

为解决上述问题, 有效挖掘图集中图像和文本情感一致性信息, 本文提出基于联合交互注意力的社交媒体图文情感分析(images-text sentiment analysis in social media based on joint and interactive attention, SA-JIA)方法, 该方法采用联合注意力关注文本与图集之间的情感一致性表达; 将经过联合注意力得到的图文增强特征输入到交互注意力中进行图文的特征融合, 进而完成图文情感分类任务。

本文的主要创新点包括以下 3 个方面:

1) 本文采用联合注意力来关注图文特征的情感显著区域, 对提取的图文特征进行模态间情感信息一致性匹配, 从而增强图文特征表示。

2) 本文提出交互注意力, 对增强的图像特征和文本特征经交互注意力来实行特征信息互补并进行图文特征融合, 进而完成图文情感分析。

3) 为证明 SA-JIA 方法的有效性, 在中文社交媒体图文情感数据集 IsTS-CN 和疫情情感数据集 CCIR20-YQ 上进行了实验验证。与现有主流方法相比, 本文方法在中文社交媒体图集文本对情感分析效果与性能中取得了更好的结果。

1　相关工作

1.1　文本情感分析

早期的文本情感分析主要为基于情感词典的方法。Stone[4] 通过构建情感词典来计算文本的整体情感极性。随着机器学习和深度学习的发展, 对于文本情感分析的方法也逐渐变得复杂多样, Wiebe 等[5] 构建大型语料库, 利用机器学习分类器建立基于情感的预测模型。Basiri 等[6] 利用 2 个独立的双向长短期记忆(bidirectional long short term memory, Bi-LSTM)网络和门控循环单元(gated recurrent unit, GRU)来提取时间流中过去和未来的上下文信息, 极大提高了文本情感分析的准确性。

近年来, 以基于双向变换器的编码器表示(bidirectional encoder representation from transformers, BERT)[7] 为主的预训练语言模型被广泛应用于情感分析领域。Dai 等[8] 基于鲁棒优化的 BERT 预训练(robustly optimized BERT pretraining approach, RoBERTa)[9] 方法生成以情感词为导向的依存句法树, 并将其用于文本方面级情感分析任务。Wu 等[10] 提出一种相对位置编码层, 整合给定方面词的位置信息, 并使用方面注意力机制考虑方面词和上下文词之间的语义关系。

虽然文本能独立地表示一定的情感, 但是人们进行交流总是通过信息的综合表现来进行。因此, 多模态的情感分析更符合人们对情感的感知, 更符合人们表达情感的模式。研究结论也表明, 相比单一文本情感分析, 多模态情感分析效果更好。

1.2　图文情感分析

随着社交媒体迅猛发展, 用户开始以图文方式分享内容, 多模态情感分析吸引了研究者的注意。不同于文本情感分析, 多模态情感分析不仅要学习单模态特征, 并且需要进行模态融合, 捕捉不同模态间的交互信息。Wu 等[11] 提出基于多头自注意力机制的融合网络, 通过合理地分配声学-视觉、声学-文本等特征权重, 获得重要的情感特征。但是该模型无法获得模态间的层次关联信息, 并且会引入大量的噪声, 造成特征冗余。Han 等[12] 提出双向双模态融合网络, 对 2 种模态特征表示进行差异增量操作, 进而提取模态间的相关信息。王靖豪等[13] 提出一种基于多层次特征融合注意力网络, 通过计算“图文”与“文图”特征, 实现多模态情感特征互补。Wen 等[14] 提出跨模态上下文门控卷积网络, 并引入了跨模态上下文门的概念, 使其能有效地捕获模态间的交互信息, 同时减少不相关信息的影响, 从而生成具有情感语义信息的图文特征。并且, 研究者在单模态预训练模型基础上提出包括多模式双向 Transformer (multimodal bitransformers MMBT)[15]、EF-CapTrBERT[16] 在内的多模态预训练模型, 刻画语言和视觉之间的关联信息, 虽能在图像文本检索、视觉问答等诸多下游任务中取得不错效果, 但这些多模态预训练模型多针对单幅图文对, 未能考虑图集与文本之间的内在联系。

1.3　Transformer 网络

Transformer 网络[17] 较早被引入用于神经机器翻译 (neural machine translation, NMT) 任务, 其中, 编码器和解码器端各自利用包含自注意力的 Transformer。在每层自注意力之后, 编码器和解码器通过一个额外的解码器子层连接, 其中, 解码器针对目标文本的每个元素关注源文本的每个元素。Tsai 等[18] 在 Transformer 的基础上设计了一种未对齐多模态序列的多模态 Transformer 来关注跨不同时间步长的多模态序列之间的交互, 并潜在地使一种模态适应另一种模态, 以端到端的形式解决了跨模态之间数据未对齐的问题。本文将 Transformer 应用到多模态, 使用交互注意力替代原有结构, 用以关注图文模态间的特征信息互补, 进而对图文

特征进行融合。

分析上述研究工作发现，已有社交媒体图文情感分析研究存在如下问题：①多模态社交媒体的图像大多不止一幅，而现有研究主要针对单幅图文对进行情感分析，针对图像多样性、无时序性的图集文本对的情感分析研究相对较少；②现有的方法未能充分考虑到图集与文本之间的关系。因此，为有效挖掘图集与文本之间情感一致性的信息，本文提出SA-JIA方法。

2 本文方法

本文SA-JIA方法可以划分为特征提取模块、联合注意力模块、交互注意力融合模块，本文方法框架如图1所示。

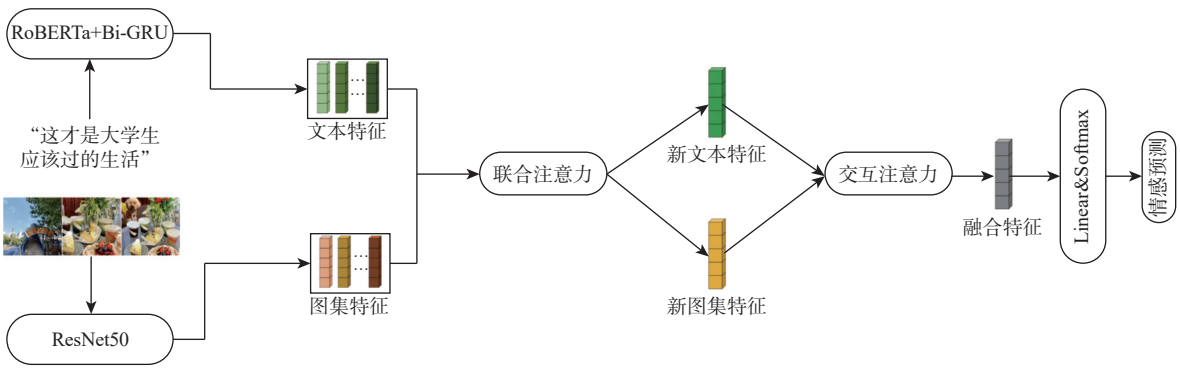

图1 SA-JIA方法框架

Fig. 1 SA-JIA method framework

特征提取模块使用现有的图像和文本特征提取的方法，分别提取图像和文本中表达情感的特征；联合注意力模块使用注意力挖掘图集文本一致性情感信息，以增强图集和文本的特征表达；交互注意力以Transformer为基础，用以关注图文特征之间的互补信息并融合输出，经Linear和Softmax后，对情感进行分类预测。

2.1 特征提取模块

首先，需要单独提取文本和图集的特征。对于文本数据，由于RoBERTa可以学习到更高层次的语义信息，而双向GRU（Bidirectional GRU，Bi-GRU）[19]可以学习到更低层次的语义信息，两者结合可以在编码文本时实现多层次、多粒度的语义信息提取，从而更好地学习文本的语义表示。因此，本文采用RoBERTa和Bi-GRU相结合的方式对文本进行特征提取。使用预训练好的RoBERTa对文本进行字编码，得到文本特征$\boldsymbol{F}_{\boldsymbol{T}-\mathrm{RoBERTa}}$：

$$\boldsymbol{F}_{\boldsymbol{T}-\mathrm{RoBERTa}}=\left[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_m,\cdots,\boldsymbol{f}_N\right]\in\mathbf{R}^{d\times N} \tag{1}$$

式中：下标$\boldsymbol{T}$为文本输入；$\boldsymbol{f}_m$为第m个字的上下文语义特征；N为RoBERTa字编码最大长度；d为RoBERTa输出维度，取值768。

其次，采用Bi-GRU进一步提取文本的上下文信息：

$$\boldsymbol{h}_m=\left[\overrightarrow{\mathrm{GRU}}(\boldsymbol{f}_m)\oplus\overleftarrow{\mathrm{GRU}}(\boldsymbol{f}_m)\right]\in\mathbf{R}^{H} \tag{2}$$

式中：$\overrightarrow{\mathrm{GRU}}$、$\overleftarrow{\mathrm{GRU}}$表示双向门控循环单元；字$\boldsymbol{f}_m$的$\boldsymbol{h}_m$经过Bi-GRU得到；$H=d_h\times2$为Bi-GRU隐藏层维度2倍；⊕表示进行异或运算，d_h为Bi-GRU的隐藏层维度。

最后，可由此得到最终的文本语义特征$\boldsymbol{F}_{\boldsymbol{T}}$为

$$\boldsymbol{F}_{\boldsymbol{T}}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_{\mathrm{N}}]\in\mathbf{R}^{H\times N} \tag{3}$$

对于图像数据，由于ResNet50引入了残差块，减轻了梯度消失的问题。ResNet50可以处理不同大小、角度和光照条件下的图像，而且ResNet50具有更深的网络结构，可以更好地提取图像特征。因此，本文采用ResNet50对图像进行特征提取。首先，将图像设定为K幅图像；然后，采取迁移学习方式，使用在ImageNet数据集[20]上预训练的ResNet50模型[21]提取每幅图像的视觉特征。将每幅图像裁剪成224×224大小作为ResNet50输入$\boldsymbol{I}_i$，对所有图像特征进行拼接得到图集视觉特征$\boldsymbol{F}_{\boldsymbol{I}_r}$：

$$\boldsymbol{F}_{\boldsymbol{I}_r}=\left[\boldsymbol{F}_{\boldsymbol{I}_1},\boldsymbol{F}_{\boldsymbol{I}_2},\cdots,\boldsymbol{F}_{\boldsymbol{I}_K}\right]\in\mathbf{R}^{1\,000\times K} \tag{4}$$

使用单层感知机将视觉特征$\boldsymbol{F}_{\boldsymbol{I}_r}$换成与文本特征具有相同维度的向量，得到最终的图集视觉特征$\boldsymbol{F}_{\boldsymbol{I}}$：

$$\boldsymbol{F}_{\boldsymbol{I}}=\tanh(\boldsymbol{W}_{\boldsymbol{I}}\boldsymbol{F}_{\boldsymbol{I}_r})\in\mathbf{R}^{H\times K} \tag{5}$$

式中：$\boldsymbol{W}_{\boldsymbol{I}}\in\mathbf{R}^{H\times1\,000}$为可训练权重矩阵。

2.2 联合注意力模块

为捕捉图集和文本之间的情感一致性信息，SA-JIA方法采用联合注意力捕捉文本和图集之间的一致性信息，对图文情感表达一致的特征权重进行增强，而图文情感表达不一致的特征权重进行减弱。

联合注意力的结构如图2所示，通过计算得到

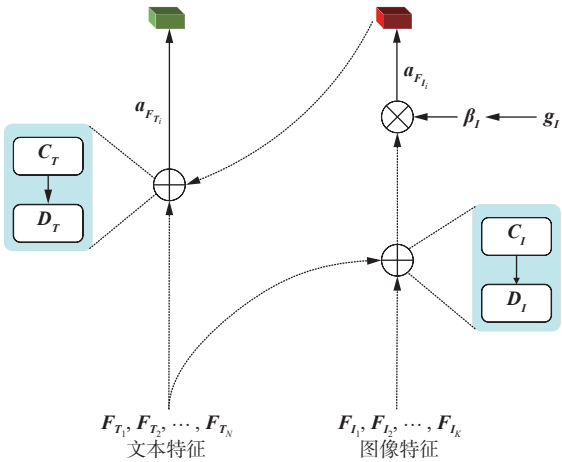

图 2　联合注意力

Fig. 2　Joint attention

图像和文本的交叉矩阵$\boldsymbol{C}_I$:

$$\boldsymbol{C}_I = \tanh(\boldsymbol{F}_I^{\mathrm{T}}\boldsymbol{W}_C\boldsymbol{F}_T) \in \mathbf{R}^{K\times N} \tag{6}$$

式中: 图像特征$\boldsymbol{F}_I \in \mathbf{R}^{H\times K}$, 文本特征$\boldsymbol{F}_T \in \mathbf{R}^{H\times N}$已由 2.1 节得出; $\boldsymbol{W}_C \in \mathbf{R}^{H\times H}$为可训练权重矩阵。

文本引导的图文相似度矩阵为

$$\boldsymbol{D}_I = \tanh\left(\boldsymbol{W}_{F_I}\boldsymbol{F}_I + (\boldsymbol{W}_{F_T}\boldsymbol{F}_T)\boldsymbol{C}_I^{\mathrm{T}}\right) \in \mathbf{R}^{K\times K} \tag{7}$$

式中: $\boldsymbol{W}_{F_I} \in \mathbf{R}^{K\times H}$、$\boldsymbol{W}_{F_T} \in \mathbf{R}^{K\times H}$为可训练权重矩阵。其次, 引入门控机制生成图像情感权重矩阵$\boldsymbol{\beta}_I$和门控机制中对情感一致性信息的表达$\boldsymbol{g}_I$:

$$\boldsymbol{g}_I = \sigma(\boldsymbol{W}_{F_I}\boldsymbol{D}_I) \tag{8}$$

$$\boldsymbol{\beta}_I = \boldsymbol{W}_s^{\mathrm{T}}\tanh(\boldsymbol{g}_I) \tag{9}$$

式中: $\boldsymbol{W}_{F_I} \in \mathbf{R}^{H\times H}$和$\boldsymbol{W}_s \in \mathbf{R}^{K\times H}$为可训练权重矩阵。计算得到视觉注意力权重$\boldsymbol{a}_{F_I}$:

$$\boldsymbol{a}_{F_I} = \mathrm{softmax}(\boldsymbol{\beta}_I) \in \mathbf{R}^{K} \tag{10}$$

对特征进行加权求和获得包含文本一致性信息的图集视觉特征$\tilde{\boldsymbol{F}}_I$为

$$\tilde{\boldsymbol{F}}_I = \sum_{i=1}^{K} \boldsymbol{a}_{F_{I_i}}\boldsymbol{F}_{I_i} \in \mathbf{R}^{H\times K} \tag{11}$$

式中: $\boldsymbol{a}_{F_I} = [\boldsymbol{a}_{F_{I_1}}, \boldsymbol{a}_{F_{I_2}}, \cdots, \boldsymbol{a}_{F_{I_K}}]$; $\boldsymbol{F}_{I_i}$为第 i 幅图像的视觉特征表示。由于获得包含文本一致性情感信息的增强的图集特征, 因此, 可以直接由图集特征引导的文本注意力特征权重$\boldsymbol{a}_{F_T}$, 获得带有文本注意力的文本语义特征$\tilde{\boldsymbol{F}}_T$为

$$\boldsymbol{C}_T = \tanh(\boldsymbol{F}_T^{\mathrm{T}}\boldsymbol{W}_C\tilde{\boldsymbol{F}}_I) \in \mathbf{R}^{K\times N} \tag{12}$$

$$\boldsymbol{D}_T = \tanh\left(\boldsymbol{W}_{F_T}\boldsymbol{F}_T + (\boldsymbol{W}_{\tilde{F}_I}\tilde{\boldsymbol{F}}_I)\boldsymbol{C}_T^{\mathrm{T}}\right) \in \mathbf{R}^{N\times N} \tag{13}$$

$$\boldsymbol{a}_{F_T} = \mathrm{softmax}(\boldsymbol{D}_T) \in \mathbf{R}^{N} \tag{14}$$

$$\tilde{\boldsymbol{F}}_T = \sum_{i=1}^{N} \boldsymbol{a}_{F_{T_i}}\boldsymbol{F}_T \in \mathbf{R}^{H\times N} \tag{15}$$

式中: $\boldsymbol{a}_{F_T} = [\boldsymbol{a}_{F_{T_1}}, \boldsymbol{a}_{F_{T_2}}, \cdots, \boldsymbol{a}_{F_{T_N}}]$; $\boldsymbol{D}_T$为图像引导的图文相似度; $\boldsymbol{W}_{F_T} \in \mathbf{R}^{N\times H}$、$\boldsymbol{W}_{\tilde{F}_I} \in \mathbf{R}^{N\times H}$为可训练权重矩阵。得到的新的文本特征$\tilde{\boldsymbol{F}}_T$和新的图像特征$\tilde{\boldsymbol{F}}_I$即为增强的图文特征。

2.3　交互注意力融合模块

为关注图集和文本之间的情感互补信息, 更好地进行模态的交互融合, SA-JIA 方法在 Transformer[17] 的基础上, 设计了一种融合不同模态信息的交互注意力方法来提供不同模态之间的交互融合。

交互注意力的结构如图 3 所示, 在注意力机制中, 缩放点积注意力问题的输入由 Queries、Keys、Values 组成。对于文本向着图集进行交互融合, 图集特征作为 Queries, 将文本特征作为 Keys 和 Values, 表达 Queries 为$\boldsymbol{Q}_I = \boldsymbol{W}_{Q_I}$, Keys 为$\boldsymbol{K}_T = \boldsymbol{W}_{K_T}$, 并将 Values 表达为$\boldsymbol{V}_I = \boldsymbol{W}_{V_T}$。从文本向着图集进行交互的注意力表示$\boldsymbol{f}_I$为

$$\boldsymbol{f}_I = \boldsymbol{I}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\boldsymbol{Q}_I(\boldsymbol{K}_T)^{\mathrm{T}}/\sqrt{d_K}\right)\boldsymbol{V}_T \tag{16}$$

式中: $\boldsymbol{W}_{Q_I} \in \mathbf{R}^{d_I}$, $\boldsymbol{W}_{K_T} \in \mathbf{R}^{d_T}$, $\boldsymbol{W}_{V_T} \in \mathbf{R}^{d_T}$为可训练权重矩阵; d_T为文本特征维度; d_I为图集特征维度; d_K为键向量的维度。

同理, 从图集向着文本进行交互融合, 将文本特征作为 $\boldsymbol{Q}$, 将图集特征作为 $\boldsymbol{K}$ 和 $\boldsymbol{V}$。从图集向文本进行适应融合的注意力表示$\boldsymbol{f}_T$为

$$\boldsymbol{f}_T = \boldsymbol{T}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\boldsymbol{Q}_T(\boldsymbol{K}_I)^{\mathrm{T}}/\sqrt{d_K}\right)\boldsymbol{V}_I \tag{17}$$

式中: $\boldsymbol{W}_{Q_T} \in \mathbf{R}^{d_T}$, $\boldsymbol{W}_{K_I} \in \mathbf{R}^{d_I}$, $\boldsymbol{W}_{V_I} \in \mathbf{R}^{d_I}$为可训练权重矩阵。

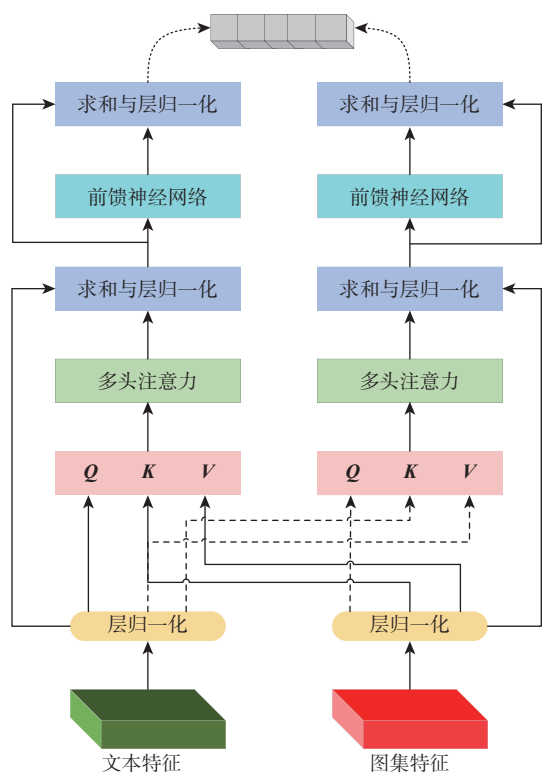

图 3　交互注意力

Fig. 3　Interactive attention

经过前馈神经网络(feedforward neural network, FNN)[22]和层归一化(layer normalization, LN)[23]后得到文本交互特征$\boldsymbol{V}_T$:

$$\boldsymbol{V}_T = \mathrm{LN}(\mathrm{FNN}(\boldsymbol{f}_T(\tilde{\boldsymbol{F}}_T, \tilde{\boldsymbol{F}}_I, \tilde{\boldsymbol{F}}_I))) \tag{18}$$

式中: LN 表示层归一化; FNN 表示前馈神经网络。

同理, 经过交互注意力模块的图集交互特征$\boldsymbol{V}_I$为

$$\boldsymbol{V}_I = \mathrm{LN}(\mathrm{FNN}(\boldsymbol{f}_I(\tilde{\boldsymbol{F}}_I, \tilde{\boldsymbol{F}}_T, \tilde{\boldsymbol{F}}_T))) \tag{19}$$

将文本交互特征$\boldsymbol{V}_T$与图集交互特征$\boldsymbol{V}_I$进行拼接得到所需的图文情感融合特征$\boldsymbol{V}$:

$$\boldsymbol{V} = \boldsymbol{V}_T \oplus \boldsymbol{V}_I \tag{20}$$

将图文特征向量$\boldsymbol{V}$输入全连接层得到结果向量$\boldsymbol{P}_{\mathrm{r}}$, 再经过 softmax 函数得到最终的情感分类结果p:

$$\boldsymbol{P}_{\mathrm{r}} = \mathrm{Linear}(\boldsymbol{W}_{P_{\mathrm{r}}} \boldsymbol{V}) \tag{21}$$

$$p = \mathrm{softmax}(\boldsymbol{W}_p \boldsymbol{P}_{\mathrm{r}}) \tag{22}$$

式中: $\boldsymbol{W}_{P_{\mathrm{r}}}$, $\boldsymbol{W}_p$为可训练权重矩阵。

3 实 验

3.1 数据集

为验证 SA-JIA 方法的有效性, 本文在 IsTS-CN 数据集和 CCIR20-YQ 数据集上进行实验。

IsTS-CN 数据集是由曹梦丽[24]构建的中文社交媒体图文情感数据集。通过设置主题词与“考研”“毕业季”“校招”等与大学生相关的词汇, 进行人工标注后得到的 10 347 条图集文本对数据, 其中, 数据情感表达为正面、中性和负面的分别有 5 449 条、2 191 条和 2 707 条。数据集包含的图像共有 34 807 幅, 平均每条数据包含约有 3 幅图像。

CCIR20-YQ 数据集来源于全国信息检索学术会议 CCIR2020 的疫情期间网民情绪识别评测任务(https://www.datafountain.cn/competitions/423?CCIR2020), 通过设置与“新冠肺炎”相关的 230 个主题词, 进行数据采集并进行人工标注 10 万条数据, 情感标签为正面、中性和负面。本文选取 10 万条中包含图像的 67 556 条图文数据构成 CCIR20-YQ 数据集, 其中, 正面 19 639 条、负面 8 864 条、中性 39 053 条。CCIR20-YQ 数据集共有 138 320 幅图像, 平均每条数据约有 2 幅图像。

在实验过程中, 将 IsTS-CN 和 CCIR20-YQ 数据集按照 8∶1∶1 的比例划分为训练集、验证集、测试集, 数据集详细信息如表 1 所示。

表 1 数据集划分

Table 1 Dataset partitioning

数据集	训练集/条	验证集/条	测试集/条
IsTS-CN	8 277	1 035	1 035
CCIR20-YQ	54 044	6 756	6 756

3.2 实验设置

本文所有模块优化器选择 Adam, Bi-GRU 隐藏层的维度设置为 200, 最优的超参数组合根据在验证集上的实验表现确定, 其中, Dropout 下降率设置为 0.5, 损失函数 Loss 使用交叉熵损失函数 CrossEntropyLoss, 并在测试集上验证实验效果。经过统计, 图集中图像至多为 9 幅, 因此, 图文情感分析图集填充时输入的图片数量统一取值为 9。如表 2 所示。

表 2 SA-JIA 方法参数设置

Table 2 SA-JIA method parameters setting

学习率	句子最大长度	下降率	注意力机制头数	Loss	隐藏层维度
2×10^{-5}	200	0.5	5	CrossEntropyLoss	200

3.3 实验结果与分析

图文情感分析为分类任务, 所以主要使用精确率P、召回率R和F_1值作为评价指标。

3.3.1 对比实验分析

为评估 SA-JIA 方法的性能, 本文从文本和图文两方面进行对比实验。对比方法如下:

Att-Bi-LSTM[25]。基于 Bi-LSTM 与注意力机制的文本情感分析方法。

BERT[7]。基于 BERT 的文本情感分析方法, 通过 BERT 模型上的微调, 有效提取文本语义特征用于微博文本情感分析。

RoBERTa[9]。基于 RoBERTa 的文本情感分析方法。

VistaNet[3]。基于图像与文本对齐的图(集)文情感分析方法。

mBERT[26]。基于 BERT 与图文交互特征的图文情感分析方法。

MMBT[15]。多模态双向 Transformer 方法, 利用卷积神经网络架构增强纯文本表征。

EF-CapTrBERT[16]。利用 BERT 实现基于输入空间转换的多模态目标感知分类方法。

TIBERT[27]。利用特征引导方法获得多模态特征, 然后, 拼接单峰特征并通过注意力模型进行多模态情感分类。

表3 为本文方法与对比方法在IsTS-CN 和CCIR20-YQ 这 2 个数据集上的图文情感分析实验结果。

相较于多模态图文情感分析, 基于文本的情感分析方法性能相对较差, 这说明引入图像视觉特征

表 3　IsTS-CN 和 CCIR20-YQ 数据集上的图文情感分析对比实验结果

Table 3　Comparative experimental results of image and text sentiment analysis on IsTS-CN and CCIR20-YQ datasets

模态	方法	P		R		F_1	
		IsTS-CN	CCIR20-YQ	IsTS-CN	CCIR20-YQ	IsTS-CN	CCIR20-YQ
文本	Att-Bi-LSTM[25]	0.574	0.691	0.601	0.583	0.566	0.607
	BERT[7]	0.617	0.601	0.608	0.634	0.612	0.617
	RoBERTa[9]	0.637	0.682	0.638	0.764	0.637	0.708
图文	VistaNet[3]	0.634	0.627	0.626	0.619	0.628	0.623
	mBERT[26]	0.733	0.709	0.723	0.702	0.727	0.705
	MMBT[15]	0.711	0.732	0.747	0.670	0.722	0.691
	EF-CapTrBERT[16]	0.723	0.717	0.720	0.692	0.721	0.703
	TIBERT[27]	0.716	0.695	0.726	0.687	0.721	0.692
	SA-JIA	0.722	0.741	0.778	0.708	0.734	0.718

能够有效改善单模态文本情感分析方法在社交媒体情感分析任务上的性能表现。而基于 RoBERTa 微调的情感分析方法优于其他方法, 原因在于, 该方法采用 Transformer 捕捉文本语义特征, 能够更加高效地获取文本语义信息, 同时, RoBERTa 在 BERT 上进行优化, 更能充分挖掘文本情感语义信息。

由于现有情感分析方法大多针对单幅图文对进行情感分析, 对于图集中情感信息的捕捉和抽象能力相对较弱, 因此, 图集文本情感分析效果欠佳。VistaNet 虽然针对图集进行情感分析, 但在所有图文情感分析方法中表现最差, 在 IsTS-CN 数据集与 CCIR20-YQ 数据集上较其他方法效果均较低, 可能原因是社交媒体数据表现更复杂, 社交媒体中蕴含着丰富的情感信息, 图集中图像情感表达复杂, 而上述方法未充分考虑到文本与图集之间的一致性表达而导致图文情感分析的效果变差。

由结果看出, SA-JIA 方法的实验结果优于其他主流方法, 表明 SA-JIA 方法能够有效处理社交媒体图文情感分析任务。当前性能表现较好的深度学习方法, 如 EF-CapTrBERT、TIBERT 方法虽然均采用预训练语言模型进行文本特征提取, 但具体的方法不同。EF-CapTrBERT 方法利用图像描述生成方法将图像翻译为辅助句子, 为语言模型提供多模态信息; TIBERT 方法则使用预训练的 ResNet 模型处理输入图像, 将图像与文本特征一起输入到多头注意力模型中, 对文本与图像特征进行融合。虽然以上做法在性能上接近或优于其他图文情感分析方法, 但性能改善并不显著, 这说明在图集文本情感分析任务中有效捕捉图集与文本之间的关联特征和情感信息对结果有重要影响。因此, SA-JIA 方法首先采用联合注意力机制来聚焦图集和文本之间的一致性情感信息, 以此来增强图集和文本的特征表达, 然后, 采用交互注意力来融合图集和文本特征进行情感分类, 最终在社交媒体图集文本情感分析效果上有提升, 在 IsTS-CN 数据集上和在 CCIR20-YQ 数据集上的实验结果可以看出, SA-JIA 方法相较于其他的方法都有较好的效果。

3.3.2　消融实验分析

为评估 SA-JIA 方法中不同模块、不同结构对于情感分析性能的影响程度, 在 IsTS-CN 和 CCIR20-YQ 数据集上对 SA-JIA 方法进行消融实验, 实验结果如表 4 所示, 消融实验所涉方法如下:

JIA-JA。使用联合注意力增强的图文特征融合的情感分析方法。

JIA-IA。使用交互注意力的图文特征融合情感分析方法。

SA-JIA。基于联合交互注意力的图文情感分析方法。

如表 4 所示, JIA-IA 方法在评测指标 F_1 值上要低于 JIA-JA 方法, 主要的原因在于, JIA-IA 方法只是简单地对文本和图集特征直接进行了拼接, 并未进行图文模态的情感一致性匹配增强, 导致图文之

表 4　IsTS-CN 和 CCIR20-YQ 数据集上的消融实验结果

Table 4　Ablation experimental results on IsTS-CN and CCIR20-YQ datasets

方法	P		R		F_1	
	IsTS-CN	CCIR20-YQ	IsTS-CN	CCIR20-YQ	IsTS-CN	CCIR20-YQ
JIA-JA	0.670	0.738	0.724	0.671	0.611	0.661
JIA-IA	0.649	0.718	0.614	0.611	0.602	0.608
SA-JIA	0.722	0.741	0.778	0.708	0.734	0.718

间的特征表示缺少一致性的关联，因此，情感分析效果要低于JIA-JA方法。而SA-JIA方法在评测指标上均高于JIA-IA，主要原因在于，SA-JIA方法首先引入联合注意力，从文本和图像2个方面引导下得到新的图文特征，减少噪声的同时增强了图文模态的情感一致性表达，然后，经过交互注意力处理，使文本和图像进行交互融合，从而提升了情感分析的效果。由实验结果可知，联合注意力和交互注意力的结合优于其他方法，也进一步说明了SA-JIA方法的有效性。

3.3.3 实验实例分析

本文对SA-JIA方法在数据集IsTS-CN上进行实例分析。

对于图4中的正确预测样例，文本标题为“#yqy的大学生活##晚安打工人#”，数据集标注的情感类别为中性，SA-JIA方法的预测情感类别也为中性。从文本信息看，标题“yqy的大学生活”“晚安打工人”是比较普通的一句话，是不带任何情感，从图集来看，每张图像展示的都只是普通城市风景照，也可以不带任何情感，因此，数据集标注情感为中性。SA-JIA方法很好地分析文本语义信息并捕捉到图集特征，所给出的情感预测类别与数据集标注一致，也为中性。

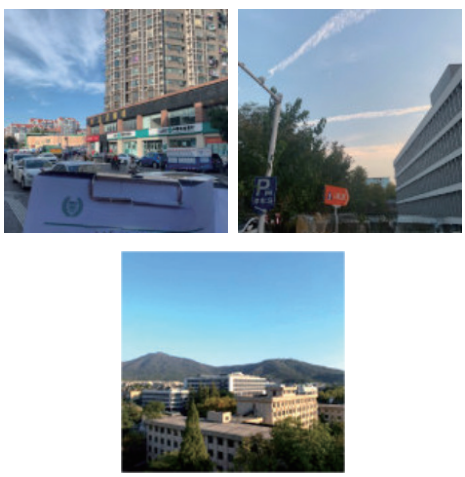

图4 正确预测样例

Fig. 4 Correct prediction example

对于图5中的错误预测样例，文本标题为“#毕业季#认识了814个日夜在第813天拍了毕业照”，数据集标注的整体图文情感类别为正面，但是SA-JIA方法给出的情感类别为中性。单独从文本信息看，“#毕业季#认识了814个日夜在第813天拍了毕业照”表达的情感可以被判断为中性；从图像的角度看，图像内容表现的是毕业的情景，情感表达可以判断为中性，并且，用户在发表图像时，有时候会将大小不一的图像合成一张长图，在处理图像时

图5 错误预测样例

Fig. 5 Incorrect prediction example

由于受神经网络的输入限制，会对图像进行分割，导致丢失部分视觉信息，因此，SA-JIA方法错误地将图文情感判断为中性情感，从而导致预测的情感类别错误。

总结分析结果发现，导致图文情感分析不准确的原因不仅在于文本的情感表达，而且图像处理的方式也会影响图文情感的判断，在一定程度上影响最终图文情感分析效果，因此，对图集的处理方式及图集与文本之间的信息关联极为重要。

4 结 论

1）本文SA-JIA方法解决了主流方法中未充分考虑图集文本对之间情感一致性信息匹配的问题。

2）本文方法通过联合注意力网络，从文本和图集2个方面去关注用户的情感表达，根据相关性来增强文本和图集的情感一致性表达；通过交互注意力对图文特征进行交互融合并进行社交媒体的图文情感分析，从而提升图文情感分析性能。

3）实验结果表明，本文方法与传统单独处理图文信息的方法不同，本文方法以图文信息互补，相互挖掘图文之间的语义关联，提供更全面的情感表达理解，为图集文本对情感分析提供了新的思考方向。

此外，本文方法仍存在局限性，图集中图像视觉信息存在不连续性，无法很好地对其整体处理获取图集视觉特征，在未来的工作中将尝试使用更多

的方法提取图集视觉信息。

参考文献（References）

[1] XIA E, YUE H, LIU H F. Tweet sentiment analysis of the 2020 U. S. presidential election[C]//Proceedings of the 30th Web Conference . New York: ACM, 2021.

[2] BONHEME L, GRZES M. SESAM at SemEval-2020 task 8: investigating the relationship between image and text in sentiment analysis of memes[C]//Proceedings of the Fourteenth Workshop on Semantic Evaluation. Barcelona : International Committee for Computational Linguistics, 2020.

[3] TRUONG Q T, LAUW H W. VistaNet: visual aspect attention network for multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019, 33(1): 305-312.

[4] STONE P J. Thematic text analysis: new agendas for analyzing text content[M]. London: Routledge, 2020: 35-54.

[5] 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. Stroudsburg: Association for Computational Linguistics, 1999.

[6] BASIRI M E, NEMATI S, ABDAR M, et al. ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis[J]. Future Generation Computer Systems, 2021, 115: 279-294.

[7] 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. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[8] 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. Stroudsburg: Association for Computational Linguistics, 2021: 1816-1829.

[9] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. (2019-07-26)[2023-05-20]. https://arxiv.org/abs/1907.11692v1.

[10] WU C, XIONG Q Y, GAO M, et al. A relative position attention network for aspect-based sentiment analysis[J]. Knowledge and Information Systems, 2021, 63(2): 333-347.

[11] WU T, PENG J J, ZHANG W Q, et al. Video sentiment analysis with bimodal information-augmented multi-head attention[J]. Knowledge-Based Systems, 2022, 235: 107676.

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

[13] 王靖豪, 刘箴, 刘婷婷, 等. 基于多层次特征融合注意力网络的多模态情感分析[J]. 中文信息学报, 2022, 36(10): 145-154.
WANG J H, LIU Z, LIU T T, et al. Multimodal sentiment analysis based on multilevel feature fusion attention network[J]. Journal of Chinese Information Processing, 2022, 36(10): 145-154(in Chinese).

[14] WEN H L, YOU S D, FU Y. Cross-modal context-gated convolution for multi-modal sentiment analysis[J]. Pattern Recognition Letters, 2021, 146: 252-259.

[15] KIELA D, BHOOSHAN S, FIROOZ H, et al. Supervised multimodal bitransformers for classifying images and text[EB/OL]. (2020-11-12)[2023-05-20]. https://arxiv.org/abs/1909.02950.

[16] KHAN Z, FU Y. Exploiting BERT for multimodal target sentiment classification through input space translation[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-06-12)[2023-05-20]. https://arxiv.org/abs/1706.03762.

[18] TSAI Y H, BAI S J, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[EB/OL]. (2019-06-01)[2023-05-20]. https://arxiv.org/abs/1906.00295.

[19] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machinetranslation[EB/OL].(2014-06-03)[2023-05-20].https://arxiv.org/abs/1406.1078v3.

[20] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

[21] 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.

[22] BENGIO Y, SCHWENK H, SENÉCAL J S, et al. Neural probabilistic language models[M]. Berlin: Springer, 2006: 137-186.

[23] BA J L, KIROS J R, HINTON G E. Layer normalization[EB/OL]. (2016-07-21)[2023-05-22]. https://arxiv.org/abs/1607.06450v1.

[24] 曹梦丽. 基于辅助信息抽取与融合的社交媒体图文情感分析方法研究[D]. 武汉: 武汉科技大学, 2022.
CAO M L. Research on sentiment analysis method of social media graphics and text based on auxiliary information extraction and fusion[D]. Wuhan: Wuhan University of Science and Technology, 2022(in Chinese).

[25] 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. Stroudsburg: Association for Computational Linguistics, 2016.

[26] YU J F, JIANG J. Adapting BERT for target-oriented multimodal sentiment classification[C]//Proceedings of the 28h International Joint Conference on Artificial Intelligence. Vienna : IJCAI, 2019: 5408-5414.

[27] YU B H, WEI J X, YU B, et al. Feature-guided multimodal sentiment analysis towards industry 4.0[J]. Computers and Electrical Engineering, 2022, 100: 107961.

Analysis of image and text sentiment method based on joint and interactive attention

HU Huijun[1, 2], DING Ziyi[1, 2], ZHANG Yaofeng[3, *], LIU Maofu[1, 2]

(1. School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China;
2. Hubei Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China;
3. Hubei Center for Data and Analysis, Hubei University of Economics, Wuhan 430205, China)

Abstract: The image and text sentiment in social media is an important factor affecting public opinion and is receiving increasing attention in the field of natural language processing (NLP). Currently, the analysis of image and text sentiment in social media has mainly focused on single image and text pairs, while little attention has been given to image and text pairs of atlas that are non-chronological and diverse. To explore the sentiment consistency between images and texts in the atlas, a method for analyzing image and text sentiment in social media based on joint and interactive attention (SA-JIA) was proposed. The method used RoBERTa and bidirectional gated recurrent unit (Bi-GRU) to extract textual expression features and ResNet50 to obtain image visual features. Joint attention was employed to identify salient regions where image and text sentiment align, obtaining new textual and image visual features. Interactive attention was utilized to focus on inter-modal feature interactions and multimodal feature fusion, finally obtaining the sentiment categories. Experimental validation was conducted on the IsTS-CN dataset and the CCIR20-YQ dataset, showing that the proposed method can enhance the performance of analyzing image and text sentiment in social media.

Keywords: social media; image and text sentiment analysis; joint attention; interactive attention; multimodal fusion

Received: 2023-06-15; **Accepted**: 2023-10-12; **Published Online**: 2023-11-23 15:37

URL: link.cnki.net/urlid/11.2625.V.20231122.1324.001

Foundation items: National Natural Science Foundation of China (62271359); Key Projects of the National Social Science Foundation of China (23ATJ005); The "14th Five Year Plan" Hubei Province Advantage Characteristic Discipline (Group) Project (2023D0302); Key Scientific Research Project of Hubei Provincial Department of Education (20192202)

* **Corresponding author**. E-mail: yfzhang@hbue.edu.cn