计算机科学
COMPUTER SCIENCE

跨模态不一致感知下双视角交互融合的多模态情感分析

卜韵阳, 齐彬廷, 卜凡亮

引用本文

卜韵阳, 齐彬廷, 卜凡亮. 跨模态不一致感知下双视角交互融合的多模态情感分析[J]. 计算机科学, 2026, 53(1): 187-194.

BU Yunyang, QI Binting, BU Fanliang. Multimodal Sentiment Analysis for Interactive Fusion of Dual Perspectives Under Cross-modalInconsistent Perception [J]. Computer Science, 2026, 53(1): 187-194.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241100029

跨模态不一致感知下双视角交互融合的多模态情感分析

卜韵阳　齐彬廷　卜凡亮
中国人民公安大学信息网络安全学院　北京 100038
(1252300321@qq.com)

摘　要　在社交媒体上,人们的评论通常会描述对应图像中的某一情感区域,图像和文本之间是具有对应信息的。以往的大多数多模态情感分析方法只是从单一视角探索图像和文本的相互影响,捕获图像区域和文本单词的对应关系,导致结果不是最优的。此外,社交媒体上的数据具有强烈的个人主观性,数据中的情感是多维和复杂的,导致出现了图像和文本情感一致性弱的数据。针对上述问题,提出了一种跨模态不一致感知下双视角交互融合的多模态情感分析模型。一方面,从全局和局部两种视角对图文特征进行跨模态交互,提供更全面、准确的情感分析,从而提升模型的表现和应用效果。另一方面,计算图文特征的不一致分数,用于代表图文不一致程度,以此来动态调控单模态表示和多模态表示的最终情感特征的权重,从而提高模型的鲁棒性。在 MVSA-Single 和 MVSA-Multiple 两个公共数据集上进行广泛实验,结果证明所提出的多模态情感分析模型与现有基线模型相比 F1 值分别提高 0.59 个百分点和 0.39 个百分点,具有有效性和优越性。

关键词:多模态情感分析;跨模态不一致感知;双视角交互融合;动态调控;跨模态交互
中图分类号　TP391.41;TP391.1

Multimodal Sentiment Analysis for Interactive Fusion of Dual Perspectives Under Cross-modal Inconsistent Perception

BU Yunyang, QI Binting and BU Fanliang
College of Information Network Security, People's Public Security University of China, Beijing 100038, China

Abstract　In social media, people's comments usually describe a certain sentiment region in the corresponding image, and there is correspondence information between image and text. Most previous multimodal sentiment analysis methods only explore the interactions between images and text from a single perspective, capturing the correspondence between image regions and text words, leading to results that are not optimal. In addition, data on social media is strongly personal and subjective, and the sentiment in the data is multidimensional and complex, which leads to the emergence of data with weak image and text sentiment consistency. To address the above two problems, a multimodal sentiment analysis model with interactive fusion of two perspectives under cross-modal inconsistency perception is proposed. On the one hand, cross-modal interaction of graphic and textual features from both global and local perspectives provides a more comprehensive and accurate sentiment analysis, which improves the performance and application of the model. On the other hand, the inconsistency scores of the graphical features are calculated to represent the degree of graphical inconsistency, as a way to dynamically regulate the weights of the unimodal and multimodal representations in the final sentiment features, thus improving the robustness of the model. Extensive experiments are conducted on two public datasets, MVSA-Single and MVSA-Multiple, and the results demonstrate the validity and superiority of the proposed multimodal sentiment analysis model compared to the existing baseline models, with F1 values increasing by 0.59 persentage points and 0.39 persentage points, respectively.

Keywords　Multimodal sentiment analysis, Cross-modal inconsistent perception, Dual-view interactive fusion, Dynamic regulation, Cross-modal interaction

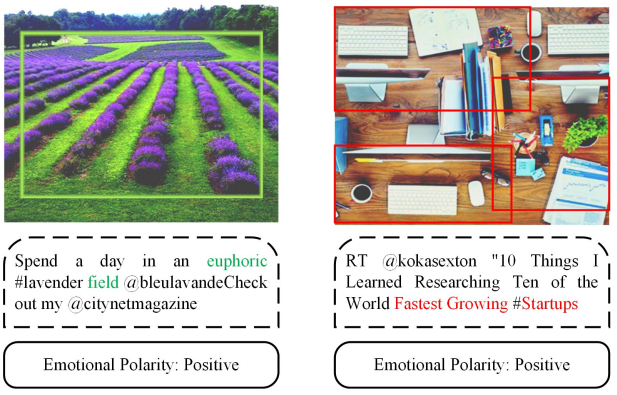
1　引言

随着 Facebook, Twitter, Tumblr 等社交媒体蓬勃发展,用户可以不用拘泥于文本的方式来表达自己的情感,各种形式和主题的文章(如图像、文本和视频)出现在社交平台上,社交媒体的情感分析随之变得更加复杂。社交媒体中多模态情

到稿日期:2024-11-05　返修日期:2025-02-12
基金项目:中国人民公安大学安全防范工程双一流专项(2023SYL08)
This work was supported by the Double First-Class Innovation Research Project for People's Public Security University of China(2023SYL08).
通信作者:卜凡亮(bufanliang@sina.com)

感分析引起了学术界的广泛关注[1]，具有广阔的应用前景，包括跨模态情感检索[2]、舆情监测、医疗病情诊断、产品用户体验分析等。

单模态情感分析仅使用一种模态（如文本、音频、图像）进行情感分析，但是仅用单模态分析可能会成为预测正确情感的阻碍。与单模态情感分析相比，多模态情感分析可以更加全面地捕捉人类情感的复杂性。早期的多模态工作中，模态特征通常是基于手工设计提取的[3]。然而，手工设计的特征通常依赖于领域知识和经验，容易受到主观因素的影响。随着深度学习的发展，目前的研究更倾向于使用神经网络自动学习多模态数据的表示，从而避免手工设计特征的问题。虽然之前的研究取得了令人瞩目的成功，但是它们通常忽视了一些问题。一方面，大多数多模态融合策略的重心在局部或全局，仅分析局部情感可能忽略整体情感趋势，而仅分析全局情感则可能错过细节情感变化，忽略全局和局部情感之间的互动，无法充分挖掘和利用多模态数据的潜在信息，使得模型性能无法达到最佳，结合两者才能更好地理解和预测情感；另一方面，图像的情感是由图像中的情感区域体现的，这些情感区域通常在对应的文本评论中的某些单词上[3]，但是数据集中存在图像情感区域和文本情感单词之间相关性弱和相关性强的数据，如果忽略这种情感差异，可能会导致实验结果不是最优。以图 1(a)为例，可以看到绿色方框内的一片田野，在相关的文本中可以找到被标绿的“field”一词，在图文数据语义对齐的基础上，文本中“field”前的“euphoric”信息和图像中的多彩信息都表达了积极的情感，因此仅通过单模态信息就可以判断情感，此时对图文特征进行跨模态交互反而会引入多余信息和噪声，进而加大分析难度；图 1(b)中，图像中红色方框内的情感区域与对应文本中被标红的情感词相关性较弱，此时对图文特征进行跨模态交互能捕捉到这种情感差异，有助于模型正确分类情感。



(a)情感一致性强 (b)情感一致性弱

图 1 多模态情感分析的两个例子(电子版为彩图)

Fig. 1 Two examples of multimodal sentiment analysis

基于以上问题，本文提出了一种跨模态不一致感知下双视角交互融合的多模态情感分析方法(Multimodal Sentiment Analysis for Interactive Fusion of Dual Perspectives Under Cross-Modal Inconsistent Perception，CIPIFDP)，目的是结合全局和局部两个视角的跨模态交互方法，提供更全面、准确的情感分析，并通过灵活调整单模态表示和多模态表示之间的权重，使得情感分析在面对复杂和多维的数据时更为有效和精准。具体地，在特征提取阶段，为了从两个视角挖掘图像和文本之间的相关度，提取图像和文本的全局特征和局部特征；在特征交互阶段，分别对全局和局部图文特征进行跨模态交互，使用张量融合整合全局图文特征的整体信息，利用注意力机制捕捉局部图文特征中细微的变化和细节。此外，计算全局(局部)图文特征之间的 Kullback-Leibler(KL)散度来表示全局(局部)图文特征之间的不一致分数，动态调整单模态和多模态表示的权重；在情感预测阶段，连接经过动态调整的全局图文表示和局部图文表示，以提供最终的情感预测。

本文的主要贡献可概括如下：

1)提出了一种跨模态不一致感知下双视角交互融合的多模态情感分析方法，从全局和局部两个视角处理情感分析不全面和图文数据情感不一致的问题，从而更全面准确地理解和预测情感，提升模型性能。

2)设计了两种跨模态交互方法，提取了图文数据对的全局特征和局部特征，利用张量融合和注意力机制分别对全局图文特征和局部图文特征进行跨模态交互融合，实现图像和文本之间的多层次交互；引入了一种衡量跨模态不一致的方法，通过计算全局(局部)图文特征之间的不一致分数，动态调控单模态表示和多模态表示在最终情感特征中的权重。

3)在两个真实的社交媒体数据集上进行广泛的实验和比较，实验结果证明了所提方法的有效性和优越性。

2 相关工作

2.1 文本情感分析

近年来，文本情感分析方法大致可分为两类：基于词典的方法和基于机器学习的方法[4]。基于词典的方法利用预先构建好的词汇资源，通过统计词频来判断情感属性。Taboada 等[5]提出了 1 种带有语义取向注释的单词词典(SO-CAL)，用 SO-CAL 来为文本分配正面或负面标签。Rao 等[6]提出了 1 种算法和 3 种剪枝策略来构建用于社交情感检测的词级情感词典。Hamouda 等[7]使用 SentiWordNet 词典为 WordNet 的每个同义词集分配积极性分数、消极性分数和中性分数。基于词典的方法虽然简单易实现，但不能识别语境、语义对情感的影响。机器学习技术的发展使得研究者开始将机器学习技术应用于文本情感分类。Pang 等[8]使用朴素贝叶斯、最大熵和支持向量机等机器学习算法对文本进行情感分类，取得了比词典方法更高的准确率。近年来，深度神经网络的发展使得模型可以处理更加复杂的文本结构和语言现象。Kim[9]利用 CNN 进行句子级分类任务。Socher 等[10]使用递归神经网络对电影评论中的短语和句子进行五分类。Akhtar 等[11]提出了一个多任务学习框架，采用 BiLSTM 和自注意力机制来识别给定句子中的方面术语，然后利用 CNN 框架来预测所识别方面术语的情绪。

2.2 图像情感分析

图像可以传递丰富的情感，如何使计算机能够检测和识别这些信息是一项具有挑战性的任务。用于图像情感分析的视觉特征一般可以分为 3 类：低级特征、中级特征和高级特征。Machajdik 等[12]基于心理学和艺术理论中的一些理论和

经验，提取出了图像的低级特征来进行情感分析。Siersdorfer 等[13]利用 SentiWordNet 同义词库从随附的文本元数据中提取其情感的数值，然后基于信息论方法进行判别性特征分析，并使用机器学习技术来预测图像的情感。Borth 等[14]提出了视觉概念检测器 SentiBank，用于检测图像中是否存在 1 200 个形容词名词对（ANP）。Yuan 等[15]提出了一种图像情感预测框架 Sentribute，它引入了面部表情检测作为附加的中级特征，然后利用图像的中级特征预测其情感。深度学习技术作为一种强大的机器学习技术，可以有效地解决图像情感分析中的一些难题。You 等[16]设计了一个适合图像情感分析的 CNN 框架，在使用基线情感算法标记的 Flickr 图像上采用渐进策略来微调网络。Yang 等[17]通过联合优化分类和分布预测开发了一个多任务深度框架，并进一步利用两种先验知识来生成每个类别的情感分布。Li 等[18]利用由深度残差网络和长短期记忆网络组成的基于深度学习的图像描述框架来生成图像的初始文本描述，通过 SentiBank 引入图像的形容词-名词对描述，将响应值最大的 4 组形容词-名词对嵌入到图像字幕模型获得的文本描述中。

2.3 多模态情感分析

如今各类社交平台上多模态数据量快速增长，多模态情感分析受到广泛关注。Wang 等[19]提出了一种用于微博情感分析的新型跨媒体词袋模型（CBM），其将微博推文的文本和图像表示为统一的词袋表示，并使用 Logistic 回归对微博情感进行分类，结果表明该模型比基于文本的方法表现更好。You 等[20]利用最先进的视觉和文本情感分析技术进行联合视觉文本情感分析，首先微调用于图像情感分析的卷积神经网络，并训练用于文本情感分析的段落向量模型，结果表明联合视觉文本特征比单独的文本和视觉情感分析性能更好。Li 等[21]提出了一种多窗口卷积变换器，能够有效捕获重要的局部 n-gram 特征，同时提出了一种情感感知注意力机制，利用外部知识库 sentiWordNet 来合并每个单词的情感强度信息。各类数据（如文字、图片等）表达情感的方式不同，但背后的意图和目标是一致的。通过分析这些一致性，可以更准确地判断情感。例如，He 等[22]提出了一种动态不变特定表示融合网络，通过改进的联合域分离网络获得所有模式的联合域分离表示，有效利用融合信息。Liu 等[23]提出了一种基于跨模态一致性建模的知识蒸馏框架，通过设计混合课程学习策略来衡量多模态数据的语义一致性。Xu 等[24]提出了一种名为跨模态联合表示变换器的多模态情感分析框架，通过层次化交互将双模态的联合表示转移到单模态，以获得模态间的一致性和互补性。Chen 等[25]使用视觉情感概念分类器提取的中层表示来确定信息相关性，并将其与其他特征结合，包括关注的文本和视觉特征，然后应用网格搜索来调整决策融合方案的权重系数，采用多模态自适应方法进行基于图像文本相关性的联合情感分析。

上述一致性多模态情感分析方法虽然强调了不同模态之间通用信息的重要性，但是分析维度单一。受此启发，本文从全局和局部两个维度衡量图文模态的不一致性，动态调整单模态和多模态表示的权重。

3 CIPIFDP 模型

本文提出的 CIPIFDP 模型的架构如图 2 所示，主要由 4 个模块组成，包括多维度特征提取模块、全局语义交互模块、局部语义交互模块以及跨模态动态调控模块。多维度特征提取模块提取文本和图像的全局特征和局部特征来实现多尺度信息联合。全局语义交互模块通过外积来学习全局文本和图像特征之间的交互。局部语义交互模块使用跨模态注意力来充分发挥局部图文特征之间的互相关联优势。跨模态动态调控模块通过学习全局和局部图文模态的情感不一致来动态调控多模态和单模态表示在最终情感特征中所占的权重。

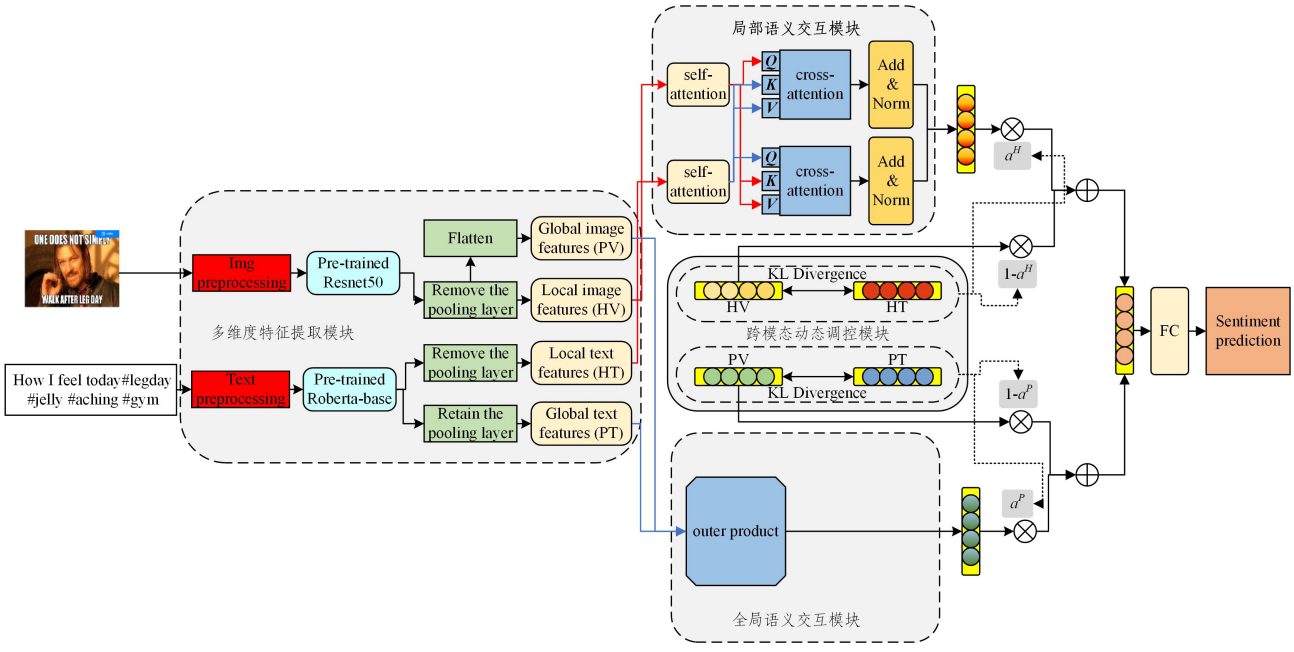


图 2 CIPIFDP 模型的架构图

Fig. 2 Overall architecture of CIPIFDP model

3.1 任务定义

给定一组三元图文数据集 $D=\{(T_1,V_1,L_1),\cdots,(T_N,V_N,L_N)\}$，其中 T_i 代表文本模态，V_i 代表图像模态，L_i 代表第 i 个图文数据对的标签信息，N 代表数据集中图文对的数量。多模态情感分析的目标是将每个图文数据对 (T_i,V_i) 分类为预定义好的标签信息 L_i，其中 L 包括积极、中性和消极 3 种情感类别。

3.2 多维度特征提取模块

3.2.1 图像特征提取

对于图像模态，使用在 ImageNet-1k 数据集上预训练过的 ResNet[26] 来提取图像的全局特征和局部特征。对原始图像进行预处理后得到图像 V，对于每一组图像 V_i，利用 ResNet 的残差卷积层来提取用于感知图像细节信息的中低级特征 $\boldsymbol{R}_i$：

$$\boldsymbol{R}_i=f_v(V_i,\boldsymbol{\theta}_v),\boldsymbol{R}_i\in R^{M*N}$$

其中，θ_v 表示残差卷积层的参数，M 表示图像的区域量，N 表示每个图像区域的特征维度。在获得高分辨率但较小视野的图像局部特征后，为了使该局部特征更加适合后面的任务与模块，将特征图压缩成固定维度的特征向量。

$$\boldsymbol{\alpha}_k=Flatten(Conv2D(\boldsymbol{R}_i))$$

$$\boldsymbol{H}_v=Relu(\boldsymbol{W}_v\boldsymbol{\alpha}_k+\boldsymbol{b}_v)$$

其中，$Conv2D$ 将 ResNet 生成的动态特征维度映射成固定值，函数 $Flatten$ 将批量轴和通道轴展平合并，$\boldsymbol{W}_v$ 和 $\boldsymbol{b}_v$ 为需要学习的参数，$\boldsymbol{H}_v$ 为局部图像特征。为了后面识别图像主体概念，进一步利用之前得到的中低级特征 $\boldsymbol{R}_i$ 提取高级图像特征。

$$\boldsymbol{Y}_i=Flatten(\boldsymbol{Avg}(\boldsymbol{R}_i))$$

$$\boldsymbol{P}_v=Relu(\boldsymbol{W}_y\boldsymbol{Y}_i+\boldsymbol{b}_y)$$

其中，$\boldsymbol{Avg}$ 提取特征图每个通道的平均值，$\boldsymbol{P}_v$ 为全局图像特征。

3.2.2 文本特征提取

对于文本模态，使用经过预训练的 RoBERTa[27] 来提取文本的全局特征和局部特征。给定文本 $T=\{x_1,x_2,\cdots,x_m,\cdots,x_S\}$，其中 x 代表文本单词，S 表示文本的单词数量。对预训练的 RoBERTa 模型进行微调，保留池化层和移除池化层以获得 768 维的全局文本序列特征向量和 768 维的局部文本特征向量，然后经过一层全连接层得到 $\boldsymbol{P}_T$ 和 $\boldsymbol{H}_T$。以上过程可以表示为：

$$\boldsymbol{P}_T,\boldsymbol{H}_T=FC(Roberta(\boldsymbol{T}_i))$$

其中，$\boldsymbol{FC}$ 表示全连接层，$\boldsymbol{P}_T$ 表示全局文本特征向量，$\boldsymbol{H}_T$ 表示局部文本特征向量。

3.3 全局语义交互模块

在得到图像和文本的整体情感特征后，为了捕捉图像和文本之间的复杂关系，使用张量融合来处理图像和文本模态数据。张量融合不仅操作简单，而且可以利用两种模态之间的互补信息，提高模型的预测能力和准确性。具体地，对两种模态的特征矩阵进行外积运算得到最终输出：

$$\boldsymbol{P}_{\text{fusion}}=\boldsymbol{P}_{\boldsymbol{P}_T}\times\boldsymbol{D}_{\boldsymbol{P}_V},\boldsymbol{P}_{\text{fusion}}\in R^{d_{TV}}$$

3.4 局部语义交互模块

图文全局特征描述了图文数据的整体属性，例如图像的颜色分布和纹理结构、文本的长度和整体语言风格等。图文局部特征是从图文数据的局部区域提取的特征，如图像的某个区域纹理和物体的边缘特征、文本中的关键词等。相比图文全局特征的交互，局部特征交互能够捕捉到图像和文本中细微的变化和细节，可以提供更加丰富的语义信息，从而提高模型的识别精度。在跨模态交互之前，为了促进模态内部不同区域之间的交互，先将图像局部区域视觉特征信息与文本局部上下文语义信息输入自注意力网络[28]，以学习丰富的内部表征。具体来说，先计算每个单词向量 $\boldsymbol{\alpha}_i$ 与所有单词的相似度。

$$e_{i,t}=(W_q\boldsymbol{\alpha}_i*W_k\boldsymbol{\alpha}_t),t\in(1,S)$$

为了更直观地观察哪些单词向量与 $\boldsymbol{\alpha}_i$ 最相关，首先将相似度归一化得到 $\beta_{i,t}$，$\beta_{i,t}$ 表示单词向量 $\boldsymbol{\alpha}_i$ 与 $\boldsymbol{\alpha}_t$ 的相关程度。

$$\beta_{i,t}=\frac{\exp(e_{i,t})}{\sum_{j=1}^{S}\exp(e_{i,j})},t\in(1,S)$$

然后计算每个单词的注意力分数 $\boldsymbol{b}_i$：

$$\boldsymbol{b}_i=\sum_{t=1}^{S}W_v\alpha_t\beta_{i,t}$$

其中，$\boldsymbol{W}_v$，$\boldsymbol{W}_q$ 和 $\boldsymbol{W}_k$ 为需要学习的参数。用这种方式，模型能够自动识别输入序列中元素之间的依赖性，关注序列中的特定部分，从而提高后面模态交互的效率。图像局部区域特征也是同样的处理方式。将上面的过程表示为：

$$\boldsymbol{G}_T,\boldsymbol{G}_v=f_a(\boldsymbol{H}_T,\boldsymbol{H}_v)$$

经过自注意力网络得到的输出分别为 $\boldsymbol{G}_T$ 和 $\boldsymbol{G}_v$。为了学习文本和图像模态的交互信息，将 $\boldsymbol{G}_T$ 和 $\boldsymbol{G}_v$ 经过线性变换转化成 query(Q)，key(K)，value(V)，当信息从图像模态输入文本模态时，query 为 $\boldsymbol{Q}_T$，key 为 $\boldsymbol{K}_v$，value 为 $\boldsymbol{V}_v$。利用以下计算式计算注意力分数：

$$Att_{v\to t}(\boldsymbol{G}_T,\boldsymbol{G}_v)=\text{softmax}\left(\frac{\boldsymbol{Q}_T\boldsymbol{K}_v^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_v$$

当信息从文本模态输入图像模态时，query 为 Q_v，key 为 K_T，value 为 V_T。利用同样的计算式计算注意力分数：

$$Att_{T\to v}(\boldsymbol{G}_T,\boldsymbol{G}_v)=\text{softmax}\left(\frac{\boldsymbol{Q}_v\boldsymbol{K}_T^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_T$$

以上只是一个注意力头的 Q,K,V。为了提高跨模态注意力网络的性能，首先使用 h 个注意力头的 Q,K,V 得到输出 $H_{v\to T}(\boldsymbol{G}_T,\boldsymbol{G}_v)$ 和 $H_{T\to v}(\boldsymbol{G}_T,\boldsymbol{G}_v)$。然后在网络中加入残差层和归一化层：

$$\boldsymbol{CH}_{v\to T}=LayerNorm(\widetilde{H}_A+H_{v\to T}(\boldsymbol{G}_T,\boldsymbol{G}_v))$$

$$\boldsymbol{CH}_{T\to v}=LayerNorm(\widetilde{H}_A+H_{T\to v}(\boldsymbol{G}_T,\boldsymbol{G}_v))$$

其中，$\boldsymbol{CH}_{v\to T}$ 和 $\boldsymbol{CH}_{T\to v}$ 是图像和文本模态交互的输出特征表示。最后，直接连接经过交互的图文局部特征得到最终输出。

$$\boldsymbol{H}_{\text{fusion}}=Concat(\boldsymbol{CH}_{v\to T},\boldsymbol{CH}_{T\to v})$$

3.5 跨模态动态调控模块

原始数据中存在图像和文本情感一致性弱的数据，也存在图像和文本情感一致性强的数据。为了检测图像和文本情感表达的差异来调整不同情感信息在情感预测中的重要性，进而提高最终的预测准确性，引入 Wang 等[29] 的做法，计算全局(局部)图文特征之间的不一致分数。当图文模态间的情感表现不一致(即高评分)时，情感预测更依赖跨模态表示，本

文的跨模态表示为全局图文跨模态表示和局部图文跨模态表示；而在图文情感表现一致（即低评分）时，则更多依赖单模态表示。在这种方法中，通过计算 KL 散度来得到不一致分数。具体来说，对于每个局部文本特征和全局文本特征 $\boldsymbol{H}_T^i$ 和 $\boldsymbol{P}_T^i$，通过计算文本模态的变分后验来近似于真实的概率分布。

$$q_T(z_O^i|O_T^i)=\partial(\boldsymbol{C}\boldsymbol{z}_O^i|\mu_{O_T^i},\sigma_{O_T^i})$$

其中，$O\in\{H,P\}$，$\mu_{O_T^i}$ 表示均值，$\sigma_{O_T^i}$ 表示方差。类似地，对于每个局部图像特征和全局图像特征 $\boldsymbol{H}_v^i$ 和 $\boldsymbol{P}_v^i$，图像模态的变分后验可以定义为：

$$q_v(z_O^i|O_v^i)=\partial(z_O^i|\mu_{O_v^i},\sigma_{O_v^i})$$

其中，$\mu_{O_v^i}$ 表示均值，$\sigma_{O_v^i}$ 表示方差。然后，在整个数据集上计算平均变分后验：

$$q(z_O^T)=\frac{1}{N}\sum_{i=1}^{N}q_T(z_O^i|O_T^i)$$

$$q(z_O^v)=\frac{1}{N}\sum_{i=1}^{N}q_v(z_O^i|O_v^i)$$

KL 散度用于衡量两个概率分布之间的差异。因此，可以通过计算文本模态概率分布和图像模态概率分布之间的 KL 散度来衡量这两种数据类型在情感表达上的差异程度。两个概率分布之间的 KL 散度的定义如下：

$$KL(P(x)\parallel Q(x))=\sum_x P(x)\log\frac{P(x)}{Q(x)}$$

根据 KL 散度的定义，通过计算全局文本和全局图像、局部文本和局部图像之间的平均散度，得到全局文本和全局图像、局部文本和局部图像之间的不一致分数。

$$\beta_i^{O(T\to v)}=\left(\frac{KL(q_T(z_O^i|O_T^i)\parallel q_v(z_O^i|O_v^i))}{KL(q(z_O^T)\parallel q(z_O^v))}\right)$$

$$\gamma_i^{O(v\to T)}=\left(\frac{KL(q_v(z_O^i|O_v^i)\parallel q_T(z_O^i|O_T^i))}{KL(q(z_O^v)\parallel q(z_O^T))}\right)$$

$$\alpha_i^O=sigmoid\left(\frac{1}{2}(\beta_i^{O(T\to v)}+\gamma_i^{O(v\to T)})\right)$$

其中，$\beta_i^O(T\to v)$ 衡量的是使用图像分布来近似文本分布的信息损失，$\gamma_i^O(v\to T)$ 衡量的是使用文本分布来近似图像分布的信息损失，$sigmoid$ 将输入值映射到 0～1 之间，α_i^O 表示全局和局部图像文本之间的不一致分数。

3.6 情感预测

文字的精确性和表达力使得复杂的想法和情感能够被准确理解，这些在非语言形式中可能难以表达的元素，通过文字得以清晰传递[30]。因此，本文的单模态表示为全局文本特征和局部文本特征。通过 α_i^O，$O\in\{H,P\}$ 控制文本特征和跨模态表示之间的权重。

$$\boldsymbol{F}^H=(\alpha_i^H\otimes\boldsymbol{H}_{\text{fusion}})\oplus((1-\alpha_i^H)\otimes\boldsymbol{H}_T)$$

$$\boldsymbol{F}^P=(\alpha_i^P\otimes\boldsymbol{P}_{\text{fusion}})\oplus((1-\alpha_i^P)\otimes\boldsymbol{P}_T)$$

其中，$\oplus$ 表示连接操作，$\otimes$ 表示乘积，$\boldsymbol{F}^H$ 和 $\boldsymbol{F}^P$ 表示文本和跨模态表示之间的动态输出。然后，连接 $\boldsymbol{F}^H$ 和 $\boldsymbol{F}^P$ 输入线性层预测最终的情感标签。

$$\bar{y}=\text{softmax}(\boldsymbol{F}\boldsymbol{C}(\boldsymbol{F}^H\oplus\boldsymbol{F}^P))$$

最后使用交叉熵函数计算真实标签 y 和预测标签 $\bar{y}$ 之间的损失，计算式如下：

$$L=-\sum_i y_i\log(\bar{y}_i)$$

其中，y_i 是真实标签。

4 实验

4.1 实验数据集

在两个公开数据集 MVSA-Singe 和 MVSA-Multiple[31] 上验证所提模型的有效性。MVSA 中的所有图文对是从推特社交平台收集的，其中 MVSA-Singe 包括 5129 个图像文本对，MVSA-Multiple 包括 19600 个图像文本对。数据集中的每个图像和文本被标记成积极、中性和消极 3 个标签中的 1 种。为了公平比较，采用与 Xu 等[32] 一致的方法来对 MVSA 数据集进行预处理，得到新的 MVSA 数据集，如表 1 所列。

表 1 经过预处理的 MVSA 数据集

Table 1 Preprocessed MVSA dataset

Dataset	Positive	Neutral	Negative	Total
MVSA-S	2683	470	1358	4511
MVSA-M	11318	4408	1298	17024

4.2 实验设置与评价标准

使用基于 Python3.8 的 PyTorch 深度学习框架和 HuggingFace Transformers[33] 来验证基线模型和本文方法，利用 RTX 3090（显存 24 GB）显卡训练模型。使用 HuggingFace Transformers 提供的 Roberta-base 提取文本特征，用预训练过的 resnet50 提取图像特征。在实验中，将数据集按照 8∶1∶1 的比例随机划分为训练集、验证集和测试集，如表 2 所列。使用 AdamW[34] 作为优化器，权重衰减为 0.01，初始学习率为 1×10^{-4}，使用 ReduceLROnPlateau 学习率衰减策略来自动调整学习率。考虑到 MVSA-Singe 数据集和 MVSA-Multiple 数据集的数据量不一样，将 MVSA-Singe 和 MVSA-Multiple 的批量大小分别设置为 16 和 64。实验使用的评价指标是准确率（ACC）和 F1 值（F1-Score）。

表 2 数据集的统计

Table 2 Statistics of the datasets

Dataset	Train	Val	Test	Total
MVSA-S	3611	450	450	4511
MVSA-M	13624	1700	1700	17024

4.3 基线模型与比较方法

为了验证所提模型的有效性，将其与单模态基线模型与多模态基线模型进行比较，对比模型如下：

1）仅仅只有文本的单模态基线模型。CNN[9] 在预训练的词向量上进行句子级分类任务。Bi-LSTM[35] 捕获了句子中重要的语义信息。预训练的 Bert[36] 可以仅通过添加额外的输出层来微调。

2）仅仅只有图像的单模态基线模型。InceptionV3[37] 是一种对图像进行情感分类的模型。OSDA[38] 根据不同视角的图像特征来进行情感分类。

3）多模态基线模型。HSAN[39] 提出了一种基于图像标题的分层注意力网络。MultiSentiNet[32] 从图像中提取目标特征和场景特征，并使用一个视觉特征引导的注意力机制。CNN-Multi[40] 将两个单独的 CNN 用来学习文本特征和图像特征。CoMN-Hop6[41] 提出了共同记忆网络来对文本和视觉内容进行交互，性能超过了传统的动态记忆网络和其他基准模型。MGNNS[42] 提出了一种新的图卷积操作，可以有效地

融合多模态的特征，并且可以适应不同的图结构。MVAN[38]提出了一种多视图注意网络，利用记忆网络来获取深层语义特征，在多模态情感分类任务上取得了很好的结果。SMP[43]设计了一个跨模态对比学习模块，并引入了额外的情感感知预训练目标捕获细粒度情感信息。MVCN[44]提出了一种文本引导融合模块，设计了基于情感的一致性约束任务和自适应损失校准策略，以解决忽视模态异构性的问题。

4.4 实验结果与分析

所提出的 CIPIFDP 模型与基线模型的性能比较如表 3 所列。通过实验结果可以观察到以下几点：

1)模型在两个数据集上对图像的情感分类能力较差。用户上传到社交媒体的图像种类繁多，其中包含大量语义特征模糊的图像，这些图像有很多噪声和干扰因素，且图像特征过于抽象。提取这些图像的特征是一项非常困难的任务，需要更多地分析才能获取情感信息。因此，图像情感分类的结果通常低于文本情感分类的结果。

2)在两个数据集上，大部分多模态基线模型的表现优于单模态基线模型。CNN-Multi 模型的效果比 Bert 模型和 BiLSTM 模型的效果差，这可能是因为 CNN-Multi 是基于卷积神经网络的模型，BERT 和 BiLSTM 是专为处理文本而设计的，擅长捕捉文本的深层次语义信息。相比之下，CNN 虽然在图像处理中表现出色，但在捕捉文本的细腻语义方面可能稍逊一筹。这种差异或许是导致 CNN-Multi 在性能上不如 BERT 和 BiLSTM 的原因之一。

表 3 两个数据集上的准确率和 F1 分数指标

Table 3 Metrics of accuracy and F1-score on two datasets

Modality	Model	MVSA-Single		MVSA-Multiple	
		Accuracy	*F1*	*Accuracy*	*F1*
Text	CNN	0.6819	0.5590	0.6564	0.5766
	BiLSTM	0.7012	0.6506	0.6790	0.6790
	Bert	0.7111	0.6970	0.6759	0.6624
Image	InceptionV3	0.6362	0.6304	0.6341	0.6207
	OSDA	0.6675	0.6651	0.6662	0.6623
Text-Image	CNN-Multi	0.6120	0.5837	0.6639	0.6419
	HSAN	0.6988	0.6690	0.6796	0.6776
	MultiSentiNet	0.6984	0.6984	0.6886	0.6811
	CoMN-Hop6	0.7051	0.7001	0.6892	0.6883
	MVAN	0.7298	0.7139	0.7183	0.7038
	MGNNS	0.7377	0.7270	0.7249	0.6934
	SMP	—	—	0.7289	0.6696
	MVCN	**0.7606**	0.7455	0.7207	0.7001
	CIPIFDP(Ours)	0.7556	**0.7514**	**0.7382**	**0.7040**

3)所提模型优于大多数强基线模型，原因在于：首先，局部特征能捕捉到特定细节和瞬时情感，而全局特征能识别出宏观情感模式，这种结合能够更全面地识别和分析情感，让情感分析模型在不同层次上进行校验和补充，增强模型的稳健性和准确性；其次，所提模型通过跨模态动态调控模块捕获图像和文本之间的情感差异信息，并动态调控单模态表示和跨模态表示在最终情感特征中的权重，从而减小了图文信息情感不一致对情感预测的影响。

4.5 消融实验

为了验证所提模型每个模块的重要性，在两个数据集上对模型进行了消融实验。

w/o PI：去除模型中的全局语义交互模块，结合局部语义交互模块和跨模态动态调控模块得到最终的情感特征。

w/o HI：去除模型中的局部语义交互模块，结合全局语义交互模块和跨模态动态调控模块得到最终的情感特征。

w/o CDR：去除模型中的跨模态动态调控模块，直接连接跨模态表示与单模态表示。

Only PI：模型仅有全局语义交互模块，利用全局语义交互模块的输出特征进行情感分类。

Only HI：模型仅有局部语义交互模块，利用局部语义交互模块的输出特征进行情感分类。

Only PT：模型仅提取图文全局特征，然后连接起来进行情感分类。

Only HT：模型仅提取图文局部特征，然后连接起来进行情感分类。

消融实验结果如表 4 所列。从实验结果可以观察到：1)完整的模型在两个数据集上取得了最好的性能，表明了每个模块都有助于提高性能；2)Only PI 和 Only HI 在两个数据集上的性能比 Only PT 和 Only HT 好，验证了全局语义交互模块和局部语义交互模块中跨模态注意力机制和张量融合能够充分利用多模态数据的丰富信息，显著提升了情感分析的精度和鲁棒性；3)w/o PI，w/o HI 和 w/o CDR 的性能都优于 Only PI 和 Only HI，说明当其中一个模块被移除时，另外两个模块可以部分弥补缺失的信息，性能没有显著下降，但当两个模块被移除，只剩下一个模块时，模型失去了大部分关键信息和功能，性能明显下降；4)在 w/o PI，w/o HI 和 w/o CDR 这 3 个模型中，w/o CDR 的性能最差，表明与直接连接跨模态表示与单模态表示相比，动态调节跨模态表示和单模态表示有助于提高情感分析能力；5)相比完整模型，w/o PI 和w/o HI 的性能略有下降，说明结合全局特征和局部特征可以实现对数据的多层次和多角度理解，从而更好地刻画情感的复杂性。

表 4 消融实验结果

Table 4 Ablation experiment results

Model	MVSA-Single		MVSA-Multiple	
	Accuracy	*F1*	*Accuracy*	*F1*
Only PT	0.7011	0.6818	0.6904	0.6760
Only HT	0.7033	0.6873	0.6915	0.6735
Only HI	0.7156	0.7046	0.7065	0.6889
Only PI	0.7176	0.7059	0.7041	0.6864
w/o CDR	0.7208	0.7102	0.7119	0.6876
w/o HI	0.7396	0.7389	0.7229	0.6984
w/o PI	0.7424	0.7401	0.7276	0.7006
CIPIFDP	**0.7556**	**0.7514**	**0.7382**	**0.7040**

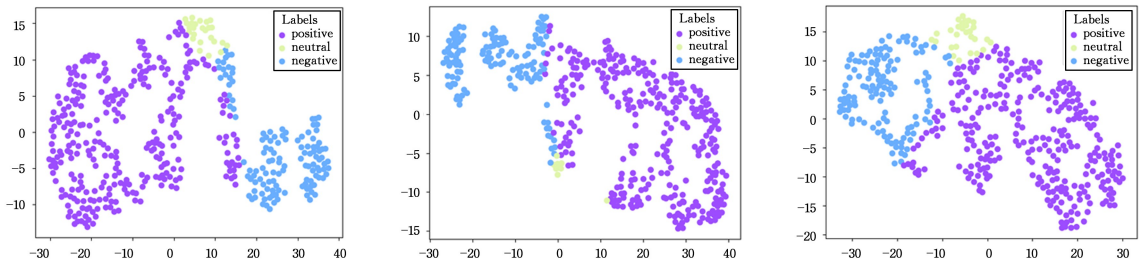
总之，实验结果证明了模型中的 3 个子模块是不可或缺的，它们相互作用，对最后的情感分析做出了贡献。

4.6 可视化分析

为了更直观地展示所提模型的优越性，使用 t-SNE[45]方法可视化了经过动态调节的全局语义交互模块、局部语义交互模块和整个模型的情感特征。

图 3 展示了 MVSA-Single 数据集上 3 个实验对象情感特征的可视化结果。从可视化结果可以观察到，所提模型在数据集上的 3 个情感标签的分布更具区分性，且相同的标签点在降维后仍然相对集中。这是因为在模型中，各个模块的

情感特征相互作用和影响，这些相互作用有助于定义更清晰的数据结构和边界，而单个模块可能无法捕捉到这些复杂的关系。在情感标签分布图中，经过动态调节后，全局语义交互模块和局部语义交互模块的情感标签分布中 3 种情感互相混淆，不容易区分，每种情感标签相对分散。相比之下，完整模型的情感标签分布中，3 种情绪数据点的分界明显，各种标签数据点分布集中。这一发现表明，所提模型将 3 个模块综合在一起产生了更好的聚类效果。



(a) 经过动态调节的全局语义交互模块　(b)经过动态调节的局部语义交互模块　(c)CIPIFDP 模型

图 3　情感特征可视化

Fig. 3　Emotion feature visualization

结束语　本文提出了一种跨模态不一致感知下双视角交互融合的多模态情感分析方法，旨在从全局和局部双视角分析多模态情感，提高情感预测精度。本文模型主要包括全局语义交互模块、局部语义交互模块和跨模态动态调控模块。全局语义交互模块使用张量融合建模图文模态之间的复杂关系，使得图文模态的数据能够在同一空间中比较和分析，从而更好地捕捉交互信息。局部语义交互模块利用注意力机制关注图像和文本中突出的关键部分，在图像和文本之间建立更深层次的语义连接，提高图文融合结果，使得多模态分析更加高效。此外，为了解决社交媒体图文数据中图像和文本情感一致性弱的问题，提出了跨模态动态调控模块，通过计算图文情感特征的不一致分数，动态调控单模态表示和多模态表示在最终情感特征中的权重。在两个公开的实验数据集上进行实验，结果表明 CIPIFDP 模型的各项评价指标都优于具有强大竞争力的基准模型，从而验证了模型的有效性。然而，CIPIFDP 模型针对的数据是情感标签一致的数据，对于情感标签相反的数据分析能力相对较弱。在未来的研究中，将使用对抗训练增加模型的鲁棒性，以帮助模型更好地处理标签相反的数据，并进一步对图文特征进行更加细化的视角分析，如话题层级、语法结构、细粒度区域分析等，提高多模态情感分析的深度和精度，从而提供更有价值的情感洞察。

参 考 文 献

[1] ZHANG L, WANG S, LIU B. Deep learning for sentiment analysis: A survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4): e1253.

[2] PANG L, ZHU S, NGO C W. Deep multimodal learning for affective analysis and retrieval[J]. IEEE Transactions on Multimedia, 2015, 17(11): 2008-2020.

[3] ZHU T, LI L, YANG J, et al. Multimodal sentiment analysis with image-text interaction network[J]. IEEE Transactions on Multimedia, 2022, 25: 3375-3385.

[4] XU J, HUANG F, ZHANG X, et al. Visual-textual sentiment classification with bi-directional multi-level attention networks [J]. Knowledge-Based Systems, 2019, 178: 61-73.

[5] TABOADA M, BROOKE J, TOFILOSKI M, et al. Lexicon-based methods for sentiment analysis[J]. Computational linguistics, 2011, 37(2): 267-307.

[6] RAO Y, LEI J, LIU W, et al. Building emotional dictionary for sentiment analysis of online news[J]. World Wide Web, 2014, 17: 723-742.

[7] HAMOUDA A, ROHAIM M. Reviews classification using sentiwordnet lexicon[C]// World Congress on Computer Science and Information Technology. 2011: 104-105.

[8] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques[C]// Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. 2002: 79-86.

[9] KIM Y. Convolutional Neural Networks for Sentence Classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014: 1746-1751.

[10] SOCHER R, PERELYGIN A, WU J, et al. Recursive deep models for semantic compositionality over a sentiment treebank [C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1631-1642.

[11] AKHTAR M S, GARG T, EKBAL A. Multi-task learning for aspect term extraction and aspect sentiment classification[J]. Neurocomputing, 2020, 398: 247-256.

[12] MACHAJDIK J, HANBURY A. Affective image classification using features inspired by psychology and art theory[C]// Proceedings of the 18th ACM International Conference on Multimedia. 2010: 83-92.

[13] SIERSDORFER S, MINACK E, DENG F, et al. Analyzing and predicting sentiment of images on the social web[C]// Proceedings of the 18th ACM International Conference on Multimedia. 2010: 715-718.

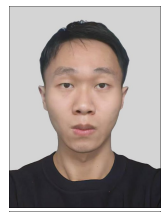
[14] BORTH D, JI R, CHEN T, et al. Large-scale visual sentiment ontology and detectors using adjective noun pairs[C]// Proceedings of the 21st ACM International Conference on Multimedia. 2013: 223-232.

[15] YUAN J, MCDONOUGH S, YOU Q, et al. Sentribute: image sentiment analysis from a mid-level perspective[C]// Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. 2013: 1-8.

[16] YOU Q, LUO J, JIN H, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2015.

[17] YANG J, SHE D, SUN M. Joint Image Emotion Classification

and Distribution Learning via Deep Convolutional Neural Network[C]// IJCAI. 2017:3266-3272.
[18] LI Z, SUN Q, GUO Q, et al. Visual sentiment analysis based on image caption and adjective-noun-pair description[J]. Soft Computing, 2021:1-13.
[19] WANG M, CAO D, LI L, et al. Microblog sentiment analysis based on cross-media bag-of-words model[C]// Proceedings of International Conference on Internet Multimedia Computing and Service. 2014:76-80.
[20] YOU Q, LUO J, JIN H, et al. Joint visual-textual sentiment analysis with deep neural networks[C]// Proceedings of the 23rd ACM International Conference on Multimedia. 2015:1071-1074.
[21] LI P, ZHONG P, ZHANG J, et al. Convolutional transformer with sentiment-aware attention for sentiment analysis[C]// 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020:1-8.
[22] HE J, YANGA H, ZHANG C, et al. Dynamic Invariant-Specific Representation Fusion Network for Multimodal Sentiment Analysis[J]. Computational Intelligence and Neuroscience, 2022, 2022(1):2105593.
[23] LIU H, LI K, FAN J, et al. Social Image-Text Sentiment Classification With Cross-Modal Consistency and Knowledge Distillation[J]. IEEE Transactions on Affective Computing, 2022, 14(4):3332-3344.
[24] XU M, LIANG F, SU X, et al. Cmjrt: Cross-modal joint representation transformer for multimodal sentiment analysis[J]. IEEE Access, 2022, 10:131671-131679.
[25] CHEN D, SU W, WU P, et al. Joint multimodal sentiment analysis based on information relevance[J]. Information Processing & Management, 2023, 60(2):103193.
[26] 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.
[27] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv:1907, 11692, 2019.
[28] 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: 6000-6010. .
[29] WANG J, YANG Y, LIU K, et al. CiteNet: Cross-modal incongruity perception network for multimodal sentiment prediction [J]. Knowledge-Based Systems, 2024, 295:111848.
[30] ZHAN F, YU Y, WU R, et al. Multimodal image synthesis and editing: A survey and taxonomy[J]. arXiv:2112. 13592, 2023.
[31] NIU T, ZHU S, PANG L, et al. Sentiment analysis on multiview social data[C]// MultiMedia Modeling: 22nd International Conference (MMM 2016). Miami, FL, USA, Part II 22. 2016:15-27.
[32] 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.
[33] WOLF T, DEBUT L, SANH V, et al. Transformers: State-of-the-art natural language processing[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020:38-45.
[34] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[J]. arXiv:1711, 05101, 2017.
[35] 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. 2016:207-212.
[36] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810, 04805, 2018.
[37] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:2818-2826.
[38] YANG X, FENG S, WANG D, et al. Image-text multimodal emotion classification via multi-view attentional network[J]. IEEE Transactions on Multimedia, 2020, 23:4014-4026.
[39] XU N. Analyzing multimodal public sentiment based on hierarchical semantic attentional network[C]// 2017 IEEE International Conference on Intelligence and Security Informatics (ISI). 2017, IEEE:152-154.
[40] CAI G, XIA B. Convolutional neural networks for multimedia sentiment analysis[C]// 4th CCF Conference Natural Language Processing and Chinese Computing (NLPCC 2015). 2015:159-167.
[41] XU N, MAO W, CHEN G. A co-memory network for multimodal sentiment analysis[C]// The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018:929-932.
[42] YANG X, FENG S, ZHANG Y, 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 (Volume 1: Long Papers). 2021:328-339.
[43] YE J, ZHOU J, TIAN J, et al. Sentiment-aware multimodal pretraining for multimodal sentiment analysis[J]. Knowledge-Based Systems, 2022, 258:110021.
[44] WEI Y, YUAN S, YANG R, 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. 2023:5240-5252.
[45] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(86): 2579-2605.



BU Yunyang, born in 2000, postgraduate. His main research interest is multimodal sentiment analysis.

BU Fanliang, born in 1965, Ph. D, professor, Ph. D supervisor. His main research interests include computer control and information processing.

（责任编辑：何杨）