智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于卷积交叉注意力与跨模态动态门控的多模态情感分析模型

仲兆满, 樊继冬, 张渝, 王晨, 吕慧慧, 张丽玲

引用本文:
仲兆满, 樊继冬, 张渝, 等. 基于卷积交叉注意力与跨模态动态门控的多模态情感分析模型[J]. 智能系统学报, 2025, 20(4): 999-1009.
ZHONG Zhaoman, FAN Jidong, ZHANG Yu, et al. Multimodal sentiment analysis model with convolutional cross-attention and cross-modal dynamic gating[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(4): 999-1009.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202409012

DOI: 10.11992/tis.202409012
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250221.0909.004

基于卷积交叉注意力与跨模态动态门控的多模态情感分析模型

仲兆满[1,2], 樊继冬[1], 张渝[1], 王晨[1], 吕慧慧[1], 张丽玲[1]
(1. 江苏海洋大学 计算机工程学院, 江苏 连云港 222005; 2. 江苏省海洋资源开发研究院, 江苏 连云港 222005)

摘 要: 在多模态情感分析任务中, 现有方法由于忽视了图像与文本之间的情感关联性, 导致融合特征存在大量冗余特征。为此, 提出了一种基于卷积交叉注意力与跨模态动态门控的多模态情感分析模型 (convolutional cross-attention and cross-modal dynamic gating, CCA-CDG)。CCA-CDG 通过引入卷积交叉注意力模块 (convolutional cross-attention module, CCAM) 来捕捉图像与文本间的一致性表达, 获取图文之间的对齐特征; 同时利用跨模态动态门控模块 (cross-modal dynamic gating module, CDGM), 根据图文之间的情感关联性动态调节情感特征的融合。此外, 考虑到图文上下文信息对于理解情感的重要性, 还设计了一个全局特征联合模块, 将图文交互特征与全局特征权重融合, 实现更可靠的情感预测。在 MVSA-Single 和 MVSA-Multi 数据集上进行实验验证, 所提出的 CCA-CDG 能够有效改善多模态情感分析的效果。

关键词: 多模态融合; 情感分析; 情感关联性; 注意力机制; 卷积交叉注意力; 跨模态动态门控; 全局特征联合; 权重融合
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2025)04−0999−11

中文引用格式: 仲兆满, 樊继冬, 张渝, 等. 基于卷积交叉注意力与跨模态动态门控的多模态情感分析模型 [J]. 智能系统学报, 2025, 20(4): 999–1009.
英文引用格式: ZHONG Zhaoman, FAN Jidong, ZHANG Yu, et al. Multimodal sentiment analysis model with convolutional cross-attention and cross-modal dynamic gating[J]. CAAI transactions on intelligent systems, 2025, 20(4): 999–1009.

Multimodal sentiment analysis model with convolutional cross-attention and cross-modal dynamic gating

ZHONG Zhaoman[1,2], FAN Jidong[1], ZHANG Yu[1], WANG Chen[1], LYU Huihui[1], ZHANG Liling[1]
(1. School of Computer Engineering, Jiangsu Ocean University, Lianyungang 222005, China; 2. Jiangsu Institute of Marine Resources Development, Lianyungang 222005, China)

Abstract: In multimodal sentiment analysis tasks, ignoring the emotional correlation between images and text leads to a large amount of redundant features in the fused representation. To mitigate this challenge, this paper introduces a multimodal sentiment analysis model grounded in convolutional cross-attention and cross-modal dynamic gating (CCA-CDG). The CCA-CDG model incorporates a convolutional cross-attention module to capture consistent expressions between images and text effectively, thereby obtaining aligned features. Furthermore, the model employs a cross-modal dynamic gating module to modulate the fusion of emotional features dynamically based on their interrelations across modalities. Additionally, recognizing the importance of contextual information from images and text for accurate sentiment interpretation, this paper devises a global feature fusion module that integrates interaction features with global feature weights, which leads to more reliable sentiment predictions. Experiments conducted on the MVSA-Single and MVSA-Multi datasets validate that the proposed CCA-CDG model remarkably enhances performance in multimodal sentiment analysis.

Keywords: multimodal fusion; emotion analysis; emotional relevance; attention mechanism; convolutional cross attention; cross-modal dynamic gating; global feature association; weight fusion

随着移动设备和社交网络的发展, 人们更倾向借助图像和文字来传递情感和观点, 这推动了多模态情感分析领域的兴起。多模态情感分析的

收稿日期: 2024−09−06. **网络出版日期**: 2025−02−21.
基金项目: 国家自然科学基金项目(72174079); 江苏省“青蓝工程”大数据优秀教学团队项目(2022-29).
通信作者: 仲兆满. E-mail: zhongzhaoman@163.com.

概念最早是在 20 世纪 90 年代末提出的，研究者主要基于文本和图像的双模态情感识别进行研究，分析两者对最终情感识别贡献度的差异性。如今，多模态情感分析[1]已经成为情感计算领域的一个重要研究方向，并广泛应用于个性化广告、情感跨模态检索、意见挖掘和决策支持等领域。

相较于单模态情感分析，多模态情感分析能更好地整合文本和图像的优势：文本揭示语言层面的情感观点，图像则展现视觉情感。通过结合两者，可以更准确地理解用户的情感状态。尤其是在社交媒体中，用户常使用图文来表达自己情感。这些图像信息往往蕴含着丰富的情感，使得图文情感分析变得尤为重要。然而，当前的研究主要关注文本与图像的特征融合，对图像与文本之间的情感关联性尚待深入探索。图像中富含的情感表达，与伴随的文本在情感传达上可能存在显著差异。例如，在某些场景中，视觉内容与文字描述可能协同表达一致的情感倾向；而在另一些情况下，图像传达的情感基调可能与文本描述形成鲜明对比——如消极的视觉内容与积极的文字表述并存。对于这类图文情感表达不一致的情况，若直接将图像情感特征与文本情感特征融合，可能对后续情感分析结果造成干扰，甚至降低准确性。

因此，为了有效探究图文之间的情感关联性，解决图文未对齐融合导致的特征冗余问题，本文提出了基于卷积交叉注意力与跨模态动态门控的多模态情感分析模型 (convolutional cross-attention and cross-modal dynamic gating, CCA-CDG)。模型通过卷积交叉注意力模块获得图文情感关联特征；通过跨模态动态门控模块控制图文特征融合；最后，联系全局特征作为补充实现更可靠的情感预测。本文的主要贡献包括以下 3 个方面：

1) 本文提出了一种基于卷积交叉注意力与跨模态动态门控的多模态情感分析模型 (CCA-CDG)。CCA-CDG 关注图文数据中情感对齐融合，结合全局特征实现更可靠的情感预测。

2) 本文所提出的卷积交叉注意力模块关注图文情感关联特征，增强图文情感对齐特征表示，降低图文未对齐带来冗余特征。跨模态动态门控模块通过动态权重分配机制，细粒度地调节图文情感特征的融合过程，确保情感一致性的图文对能够得到有效融合。而对于未有效对齐的图文对，则保留原始模态特征，从而避免了不匹配特征的负面干扰。

3) 为了验证 CCA-CDG 的有效性，在 MVSA-single 和 MVSA-multi 数据集进行实验验证。与基线模型和主流模型对比，CCA-CDG 在情感分析任务中效果更优。此外，在 MVSA-single 数据集上进行消融实验，验证了所提出的 CCA-CDG 合理性。

1　相关工作

1.1　单模态情感分析

单模态情感分析是指仅使用一种模态来进行情感分析的方法，主要集中在文本和图像领域。

构建情感词典是早期文本情感分析的主流方法。Taboada 等[2]提出了语义方向计算器 (semantic orientation calculator, SO-CAL)，使用带有情感极性标注的词典，从文本中提取情感特征。吴杰胜等[3]通过改进和扩充情感词典，添加程度副词、否定词等，实现了微博网民评论数据的情感分析。随着机器学习与深度学习的不断发展，文本分析方法也变得更加复杂和准确。李洋等[4]通过卷积神经网络 (convolutional neural network, CNN) 提取局部特征与 BILSTM(bidirectional long short-term memory) 提取全局特征互补，提升了文本分类准确性。刘继等[5]将 BERT(bidirectional encoder representations from Transformers) 与 BILSTM 结合，将 BERT 隐藏层序列转成向量，得到语义特征输入到 BILSTM，该模型在新冠数据集中实验效果较好。

早期图像情感分析的方法主要基于低端视觉特征的图像情感分类方法，使用基础的人工特征来对图像进行情感分类。Datta 等[6]通过艺术理论的图像特征来进行情感图像分类的方法，通过使用基于亮度、色彩饱和度、色调等视觉特征，并结合支持向量回归的方法，研究了图像特征与情感之间的联系。Machajdik 等[7]运用心理学和艺术理论中的相关理论和经验概念，提取了专门针对具有情感表达的艺术品领域的图像特征，在国际情感图片系统 (international affective picture system, IAPS) 上进行了测试，改进了分类结果。当今对于图像情感分类的主流模型主要基于深度学习模型，尤其是基于 CNN 的预训练模型 (如 VGG19、Resnet18、Inception) 来提取图像特征，并结合全连接层进行情感分类。Zhou 等[8]基于 Resnet18 网络模型，将平均池化层改成双卷积的平均池化，提升了分类准确率。Meena 等[9]基于 VGG19 迁移学习方法进行图像情感分析，实现对不同情绪的情绪检测和分类。尽管单模态情感分析在某些情况下可以取得较好的效果，但由于

情感表达的复杂性和多样性，仅使用单一模态往往难以全面、准确地识别情感。因此，多模态情感分析逐渐成为研究的热点，即结合多种模态的信息来进行情感分析，从而提高情感识别的准确性。

1.2　多模态情感分析

图像和文本是人们在社交媒体和其他在线平台表达情感时常用的两种形式。多模态情感分析的大多数方法使用图像和文本两种模态来获得情感信息。Yang 等[10] 针对图像与文本信息，提出了多视图注意网络情感分析模型，通过不断更新记忆网络来获取深层语义特征。张继东等[11] 在多模态情感模型基础上，引入注意力机制捕获文本中情感信息，提升了情感分析的准确率。杨力等[12] 通过门控循环单元和多头注意力提取模态特征，提高内部重要特征权重。在 CMU-MOSI (multimodal opinion sentiment intensity) 和 CMU-MOSEI(multimodal opinion sentiment and emotion intensity) 数据集进行实验，评价指标均有所提高。Zadeh 等[13] 关注模态内和模态间的动态建模，提出了张量融合网络。Han 等[14] 利用独立性和相关性的动态模式达到最高绩效的限制，提出了双向双峰融合网络，对两两模态表示进行融合和分离，利用 Transformer 中门控机制改进最终输出，模型分类性能明显优于其他基线模型。曾子明等[15] 针对新冠肺炎的相关图文数据，构建了混合融合策略多模态细粒度负面情感识别模型，提升了在负面情感识别方面的精度。杨颖等[16] 针对情感特征提取不充分问题，提出了一种多粒度视觉动态融合模型，对粗粒度与细粒度捕获情感特征，进行两阶段动态融合，在 Twitter-2015 和 Twitter-2017 数据集上，相比其他基线模型分类效果有所提升。Gan 等[17] 提出了一个具有多头注意机制的多模态融合网络，通过神经网络和注意力机制将不同模态之间的噪声干扰降到最低，从而获得独立的视觉和文本特征。

尽管先前的研究在多模态情感分析领域取得了显著成就，但这些研究主要集中在图像和文本的特征提取以及多模态特征融合策略上，往往忽视了图像与文本之间的内在联系和相互影响的复杂性。具体而言，“内在联系”指的是图像与文本之间的情感关联性，即它们在表达情感时的一致性；而“相互影响的复杂性”则涵盖了当图像与文本情感对齐或不对齐时，这种对齐状态如何影响多模态特征的融合。为解决这一问题，本文提出了一种基于卷积交叉注意力与跨模态动态门控的多模态情感分析模型（CCA-CDG）。CCA-CDG 通过引入卷积交叉注意力模块，获取图像与文本之间的关联特征。同时，跨模态动态门控模块的设计确保了图文之间的对齐融合，有效抑制了情感不一致的干扰，保留了原始特征的完整性。最后，结合全局特征作为补充，CCA-CDG 能够实现更为可靠和精确的情感预测。

2　CCA-CDG

CCA-CDG 分为图文特征提取模块、卷积交叉注意力模块、跨模态动态门控模块、全局特征联合模块，模型结构如图 1 所示。

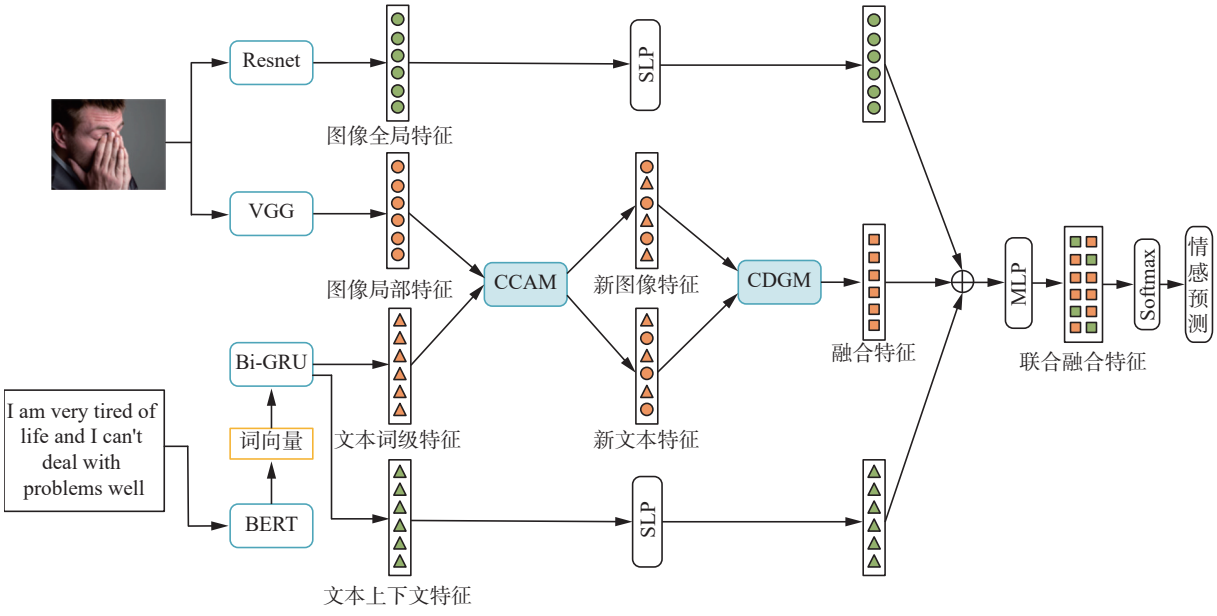


图 1　CCA-CDG 结构

Fig. 1　CCA-CDG structure

与现有研究方法相比，CCA-CDG 关注图像与文本之间的关联性影响，通过卷积交叉注意力获得图像与文本的交互特征。跨模态动态门控模块关注图像与文本的对齐，促进正对融合，抑制负对融合，从而减弱图文不匹配对情感分析结果的负面影响。最终联系全局特征作为补充，实现更全面的情感预测。

2.1　图文特征提取模块

BERT 模型在处理文本数据时，采用一种特定的格式作为输入，即在句子的开头添加 [CLS] 标记，随后是句子中的每个单词，最后以 [SEP] 标记结束。这种输入方式使得 BERT 有效地处理单个句子或句子对。双向门控循环单元 (bidirectional gated recurrent unit, Bi-GRU) 是循环神经网络 (recurrent neural network, RNN) 的一种扩展，它改进了传统的单向 GRU 模型。Bi-GRU 旨在更好地捕捉输入序列中的过去和未来信息，从而提供更全面的上下文理解。对于给定包含 n 个单词的句子，使用预训练的 BERT-base[18] 将每个单词嵌入到 768 维度向量空间中，得到嵌入向量特征 $\boldsymbol{F}_T$：

$$\boldsymbol{F}_T = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n]$$

式中 $\boldsymbol{x}_n$ 表示第 n 个单词的特征向量。

为了进一步捕捉文本的情感信息，本文采用了一个双向 GRU[19] 模型。这个模型能够综合考虑单词的前后文信息，从而增强对句子整体意义的理解。

$$\overrightarrow{\boldsymbol{h}_i} = \mathrm{GRU}(\boldsymbol{x}_i, \overrightarrow{\boldsymbol{h}_{i-1}}),\ i \in [1,n]$$

$$\overleftarrow{\boldsymbol{h}_i} = \mathrm{GRU}(\boldsymbol{x}_i, \overleftarrow{\boldsymbol{h}_{i+1}}),\ i \in [1,n]$$

$$\boldsymbol{t}_i = \frac{1}{2}(\overrightarrow{\boldsymbol{h}_i} + \overleftarrow{\boldsymbol{h}_i}),\ i \in [1,n]$$

式中：$\overrightarrow{\boldsymbol{h}_i}$ 表示正向状态，$\overleftarrow{\boldsymbol{h}_i}$ 表示反向状态，$\boldsymbol{t}_i$ 表示第 i 个单词特征。最后将单词特征拼接得到文本的词级特征 $\boldsymbol{T}$：

$$\boldsymbol{T} = [\boldsymbol{t}_1\ \boldsymbol{t}_2\ \cdots\ \boldsymbol{t}_n]$$

对于给定的输入图像 $\boldsymbol{R}$，将图像裁剪成 224 像素×224 像素大小。对图像的 m 个区域，使用在 ImageNet[20] 数据集上预训练的 VGG16[21] 前卷积部分对图像局部特征提取，通过单层感知机 (single layer perceptron, SLP)，将图像特征向量通过线性变换转换成与文本特征向量相同维度，得到图像局部特征 $\boldsymbol{W}$：

$$\boldsymbol{W} = \mathrm{SLP}([\boldsymbol{w}_1\ \boldsymbol{w}_2\ \cdots\ \boldsymbol{w}_m])$$

2.2　卷积交叉注意力模块

为了捕捉图像与文本之间的关联性特征，CCA-CDG 依据 Transformer 结构，在多头注意力机制基础上加入多核卷积神经网络，旨在有效地提取并整合图像与文本之间的相关特征。这种融合设计使得模型能够更全面地理解图像和文本之间的情感关系，从而提高对多模态数据的处理能力。卷积交叉注意力模块 (convolutional cross-attention module, CCAM) 结构如图 2 所示。

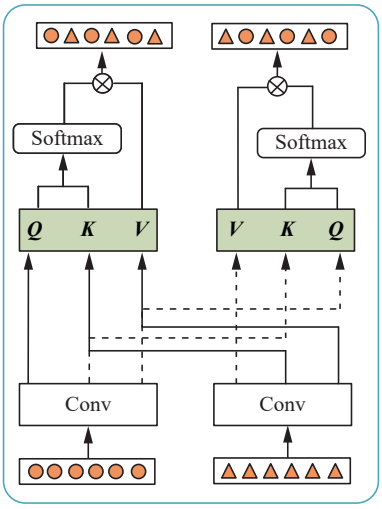


图 2　CCAM 结构

Fig. 2　CCAM structure

对于文本词级特征 $\boldsymbol{T}$，经过卷积层卷积提取后可以捕捉到更细致的文本语义特征 $\boldsymbol{T}_{\mathrm{Conv}}$。图像局部特征 $\boldsymbol{W}$，经卷积网络得到卷积特征 $\boldsymbol{W}_{\mathrm{Conv}}$ 可以学习到图像中空间层次结构，从而凸显图像中显著特征，图文特征提取公式为

$$\boldsymbol{T}_{\mathrm{Conv}} = \mathrm{Conv}(\boldsymbol{T})$$

$$\boldsymbol{W}_{\mathrm{Conv}} = \mathrm{Conv}(\boldsymbol{W})$$

在注意力机制中，缩放点积注意力机制处理的信息由 3 个关键部分构成：查询 $\boldsymbol{Q}$（query）、键 $\boldsymbol{K}$（key）和值 $\boldsymbol{V}$（value）。对于文本向图像进行交互融合特征，$\boldsymbol{Q}$ 值为图像特征 $\boldsymbol{W}_{\mathrm{conv}}$，$\boldsymbol{K}$ 值和 $\boldsymbol{V}$ 值为文本特征 $\boldsymbol{T}_{\mathrm{conv}}$。

首先，计算文本与图像的相似度矩阵 $\boldsymbol{A}$：

$$\boldsymbol{A} = (\boldsymbol{W}_{\mathrm{Conv}}\boldsymbol{W}_w)(\boldsymbol{T}_{\mathrm{Conv}}\boldsymbol{W}_t)^{\mathrm{T}}$$

式中：$\boldsymbol{W}_t$ 和 $\boldsymbol{W}_w$ 是一个可训练的权重矩阵，$\boldsymbol{A}_{ij}$ 表示第 j 条文本与第 i 张图片相似度矩阵。

得到相似度矩阵后，通过交叉注意力进一步计算图像与文本关联的注意力权重分布 $\bar{\boldsymbol{A}}$：

$$\mathrm{Attention}(\bar{\boldsymbol{A}}) = \mathrm{Softmax}\left(\frac{\boldsymbol{A}^{\mathrm{T}}}{\sqrt{d}}\right)$$

最后通过将图片区域所有单词特征加权聚合，得到文本相关联的图像特征 $\boldsymbol{W}_1$：

$$\boldsymbol{W}_1 = \bar{\boldsymbol{A}} \odot \boldsymbol{T}$$

式中：$\boldsymbol{W}_{1i}$ 表示第 i 张图像的文本关联特征，“⊙”表示 Hadamard 乘积。同理，可计算出图像关联的文

本特征$\boldsymbol{T}_1$, $\boldsymbol{T}_{1i}$表示第i条文本的图像关联特征。

2.3　跨模态动态门控模块

在图像特征与文本特征进行多模态融合时, 考虑到图像与文本是不完全对齐的, 直接融合可能获取无意义的融合特征, 从而影响下游情感分析任务。本文通过跨模态动态门控来控制不同模态数据之间的融合, 动态调整不同模态数据之间的融合权重, 增强模型多模态融合交互作用, 跨模态动态门控模块 (cross-modal dynamic gating module, CDGM) 结构如图 3 所示。

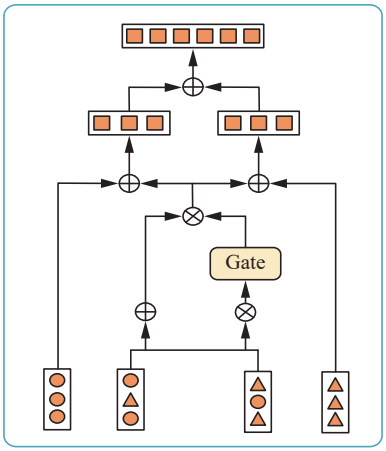


图 3　CDGM 结构
Fig. 3　CDGM structure

对于图像第i区域的特征$\boldsymbol{w}_i$与其相关的文本特征$\bar{\boldsymbol{t}}_i$, 计算门控权重来评估相关程度:

$$\boldsymbol{g}_{wi} = \text{Sigmoid}(\boldsymbol{w}_i \cdot \bar{\boldsymbol{t}}_i)$$

式中$\boldsymbol{g}_{wi}$表示第i个关于图像的门控权重向量。将门控权重向量进一步归一化处理, 使所有权重之和为 1:

$$\boldsymbol{G}_W = \frac{\mathrm{e}^{\boldsymbol{g}_j}}{\sum_{j=1}^{n}\mathrm{e}^{\boldsymbol{g}_j}},\ i = 1,2\cdots,n$$

式中$\boldsymbol{G}_W$是关于图像的门控权重值。同理, 计算得到关于文本的门控权重向量$\boldsymbol{G}_T$。如果图像与文本数据对齐, $\boldsymbol{G}$将赋予一个高门控值权重, 促进模态特征融合。如果没有很好地对齐, $\boldsymbol{G}$将会得到一个低门控权重, 特征融合将会被抑制, 原始特征重要性将会增加。最终, 得到文本与图像融合后的融合特征$\boldsymbol{F}_1$:

$$\boldsymbol{W}_T = (\boldsymbol{G}_W(\boldsymbol{W} \oplus \boldsymbol{T}_1) + (1 - \boldsymbol{G}_W)\boldsymbol{W})$$

$$\boldsymbol{T}_W = (\boldsymbol{G}_T(\boldsymbol{T} \oplus \boldsymbol{W}_1) + (1 - \boldsymbol{G}_T)\boldsymbol{T})$$

$$\boldsymbol{F}_1 = \boldsymbol{W}_T \oplus \boldsymbol{T}_W$$

2.4　全局特征联合模块

为了联合文本与图像上下文信息实现更可靠预测, 分别对文本和图像进行了全局特征提取。对于文本上下文特征, 将每条文本的词级特征$\boldsymbol{T}$平均池化, 得到整个句子上下文文本特征$\boldsymbol{S}$:

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{t}_i$$

式中$\boldsymbol{S}_i$表示第i条文本的上下文特征。

对于图像$\boldsymbol{R}$全局特征, 采用 Resnet18 去除顶部全连接层[22], 提取 512 维的图像全局特征$\boldsymbol{V}$:

$$\boldsymbol{V} = \text{Resnet18}(\boldsymbol{R})$$

将全局图像特征$\boldsymbol{V}$与上下文文本特征$\boldsymbol{S}$拼接, 得到全局融合特征$\boldsymbol{F}_2$。通过多层感知机将图文关联融合特征$\boldsymbol{F}_1$与全局融合特征$\boldsymbol{F}_2$拼接得到最终特征$\boldsymbol{F}$, 特征$\boldsymbol{F}$包含图文局部对齐特征和上下互补特征:

$$\boldsymbol{F}_2 = \boldsymbol{S} \oplus \boldsymbol{V}$$

$$\boldsymbol{F} = \lambda \times \text{MLP}(\boldsymbol{F}_1) \oplus (1-\lambda) \times \text{MLP}(\boldsymbol{F}_2) \tag{1}$$

式中λ是分配的相关权重。

将特征向量$\boldsymbol{F}$通过Softmax层, 预测出最终情感$\hat{y}$, 公式为

$$\hat{y} = \text{Softmax}(\boldsymbol{\omega}\boldsymbol{F} + \boldsymbol{b})$$

式中: $\boldsymbol{\omega}$和$\boldsymbol{b}$是可训练权重矩阵, $y \in \{1,0,-1\}$分别表示积极、中性和消极情感。

针对三分类问题, 本文采用多类交叉熵损失函数进行模型训练, 公式为

$$L_i = -\sum_{j=1}^{3} y_{ij}\log(\hat{y}_{ij})$$

式中: y_{ij}表示样本i属于第j类, $\hat{y}_{ij}$表示样本i预测为第j类。

3　实验和结果分析

3.1　数据集

为了验证所提出的 CCA-CDG 的有效性, CCA-CDG 在公开的 MVSA-single 和 MVSA-multi 数据集上进行实验。MVSA(multi-view sentiment analysis)[23] 数据集来源于社交媒体 Twitter, 其中的子集 MVSA-single 包含了 4 869 对图文数据, 每一对图文都经过 1 位标注员标注情感。扩展的 MVSA-multi 数据集, 则收纳了 19 600 对图文信息, 针对每一条图文对, 3 位独立的标注员进行逐一标注, 每位标注员都会基于自己的判断对图文内容进行标记情感。

为了便于比较, 本文对两个数据集进行了预处理。首先, 去除了数据集中文本数据为空以及不完整的图文对。针对 MVSA-multi 数据集, 删除了 3 组图文标签完全对立的图文对数据, 采取投票策略以多数标注一致的情感标签作为 MVSA-multi 图文标签。对于 MVSA 数据集中的图文对,

如果图像与文本的标签不同，则人工重新判断情感加以标注。经过上述的预处理工作，MASV-single 数据集得到了 4 511 条图文对。MVSA-multi 数据集得到 16 779 条图文对。新的数据集如表 1 所示。

表 1　处理后的 MVSA 数据集

Table 1　MVSA dataset after processing

数据集	积极	消极	中性	总数
MVSA-single	2 688	467	1 356	4 511
MVSA-multi	11 285	4 315	1 179	16 779

3.2　实验设置

实验的硬件配置和软件环境如表 2 所示。

表 2　硬件配置与软件环境

Table 2　Hardware configuration and software environment

硬件	型号	软件	版本
CPU	16 vCPU Intel(R) Xeon(R) Gold 6430	Torch Python	2.1 3.8
GPU	RTX 4090 24 GB	CUDA	11.8

模型优化器选择 Adam，BERT_GRU 隐藏层维度设置 512，BERT_GRU 的 Dropout 设置为 0.25。其他超参数根据验证集表现设置如表 3 所示。

表 3　模型参数

Table 3　Model parameter

参数	数值
学习率	2×10^{-5}
句子最大长度	100
批处理大小	64
注意力头数量	8
卷积核大小	5×5
卷积核数量	512
Dropout	0.05
权重λ	0.8

实验中，将 MVSA 数据集按照 8∶1∶1 的比例划分训练集、验证集和测试集，数据集详细信息如表 4 所示。

表 4　数据集划分

Table 4　Data set partitioning

数据集	训练集	验证集	测试集
MVSA-single	3 609	451	451
MVSA-multi	13 423	1 678	1 678

实验评估采用准确率 (accuracy, Acc) 和 F1 分数 (F1-score, F1) 为评价指标，计算过程为

$$A_{cc}=\frac{N_{TP}+N_{TN}}{N_{TP}+N_{TN}+N_{FP}+N_{FN}}$$

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}}$$

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}}$$

$$F_1=\frac{2\times P\times R}{P+R}$$

式中：A_{cc} 表示准确率，P 表示精确率，R 表示召回率，F_1 表示 F1 分数，N_{TP}为真正例的数量，N_{TN}为真反例的数量，N_{FP}为假正例的数量，N_{FN}为假反例的数量。

3.3　对比模型

采用 MVSA 数据集，针对不同的情感分析模型预测结果比较，来验证 CCA-CDG 的有效性。

1）BERT：基于 BERT-base 的文本分析方法，通过模型微调，提取文本特征实现情感分析。

2）BERT_GRU：使用预训练的 BERT 获得文本词向量，作为双向 GRU 的输入。使用双向 GRU 进一步提取文本特征用于情感分析。

3）Resnet18：使用预训练的 Resnet18 模型，去除最后一层全连接层，保留前面的卷积层部分，获取图像的全局特征向量用于图像情感分析。

4）OSDA[10]：基于多视图注意网络的图像情感分析模型，着重于物体特征和场景特征的综合考虑。模型不仅关注单个对象在画面中的细节表达，同时也重视整体场景氛围的影响。

5）MSNM[24]：通过图像特征引导的 LSTM 方法，重点提取文本中情感重要的词，然后对文本特征与图像特征进行聚合。

6）COMM[25]：基于共记忆网络的视觉内容和文本交互迭代。

7）MVAN[10]：基于多视图注意网络的多模态情感分析模型，利用记忆网络模块迭代获取语义图像文本特征。

8）MGNNS[26]：引入多通道图神经网络获取数据集的全局特征学习多模态表示，利用多头注意力机制实现多模态深度融合来预测图像文本对的情感。

9）MII[27]：模态信息交互模型，利用注意力机制获得模态局部语义特征和关联特征用于情感分类。

10）CLCAF[28]：基于监督对比学习的多模态情绪分析模型，利用卷积神经网络和 Transformer 解决模态特征的冗余问题，采用监督对比学习来增

强其从数据中学习标准情绪特征的能力。

11)MAMF[29]: 基于多层注意力机制的图文双模态情感分析模型, 利用注意力机制引导视觉特征与文本特征重构, 加权融合区分其影响。

3.4　实验结果与分析

为了评估 CCA-CDG 有效性, 分别从文本、图像和图文 3 个方面在 MVSA-single 和 MVSA-multi 数据集进行对比实验, 结果如表 5 所示。

表 5　对比实验结果

Table 5　Comparative experimental results

模态	模型	MVSA-single		MVSA-multi	
		Acc	F1	Acc	F1
文本	BERT	0.651	0.643	0.647	0.623
	BERT_GRU	0.673	0.651	0.668	0.659
图像	Resnet18	0.623	0.612	0.626	0.625
	OSDA	0.667	0.665	0.666	0.662
多模态	MSNM	0.681	0.677	0.681	0.663
	COMM	0.705	0.700	0.699	0.698
	MVAN	0.729	0.701	0.723	0.723
	MGNNS	0.738	0.727	0.725	0.693
	MII	0.740	0.733	0.708	0.698
	CLCAF	0.764	0.756	0.705	0.679
	MAMF	0.765	0.753	0.743	0.741
	CCA-CDG	**0.771**	**0.767**	**0.758**	**0.749**

注: 加粗代表本列最优结果。

由表 5 实验结果可知, 在同一数据集中, 多模态的情感分析方法相对于单模态情感分析方法分类效果更好。说明同时考虑文本特征与图像特征, 能够充分利用模态之间的互补优势, 从而提供更为精细和可靠的情感预测。

本文提出的 CCA-CDG 在 MVSA-single 数据集上 Acc 达到 0.771, F1 值为 0.767。在 MVSA-multi 数据集上 Acc 值为 0.758, F1 值为 0.749。Acc 和 F1 的值都优于当前主流多模态情感分析模型。MSNM 模型重点关注文本特征, 忽略了图像存在重要情感。COMM 模型在图文交互时容易产生冗余信息, 影响模态融合。MVAN 模型和 MGNNS 模型考虑图像情感信息, 但未能关注到图文之间情感权重。MII 模型和 MAMF 模型通过注意力机制获得图文情感关联, 解决了图文情感对齐问题, 但未能考虑图文未对齐时原始特征的重要性。CLCAF 模型依据卷积神经网络和 Transformer 优势, 解决了模态特征的冗余问题, 但未能考虑到全局特征对情感分析补充作用。针对上述模型存在问题, CCA-CDG 考虑到图像与文本之间的关联性, 通过卷积交叉注意力模块对齐图文情感特征; 通过跨模态动态门控模块, 动态权重分配促进图像与文本的情感对齐融合, 抑制负对融合, 保留原始特征信息; 同时兼顾全局特征作为补充, 进行情感预测。相对于主流的多模态分析模型, 在情感分类效果更加准确, 评估指标均有所上升。

3.5　消融实验

为了验证 CCA-CDG 的有效性, 在 MVSA-single 和 MVSA-multi 数据集上进行消融实验, 实验方法如下。

1)-Conv_Cross 表示去除卷积交叉注意力模块;

2)-Conv 表示去除卷积层;

3)-Gate 表示去除跨模态动态门控模块;

4)-OriLink 表示去除保留原始特征;

5)-Content 表示去除全局特征联合。

消融实验结果如表 6 所示。从表 6 结果中发现, CCA-CDG 去掉任意模块都会导致分类结果的下降。

表 6　消融实验结果

Table 6　Ablation results

方法	MVSA-single		MVSA-multi	
	Acc	F1	Acc	F1
-Conv_Cross	0.701	0.694	0.682	0.676
-Conv	0.761	0.754	0.742	0.736
-Gate	0.714	0.695	0.708	0.683
-OriLink	0.723	0.703	0.720	0.696
-Content	0.756	0.741	0.738	0.724
CCA-CDG	**0.771**	**0.767**	**0.758**	**0.749**

注: 加粗代表本列最优结果。

1)-Conv_Cross 和 CCA-CDG 相比分类效果大幅度下降, 说明 CCA-CDG 关注图像和文本之间

的关联性，获得图像与文本的对齐特征，能够提升多模态情感分类效果。

2）-Conv 和 CCA-CDG 对比精度略微下降，说明图文特征对齐前，特征经过卷积层输出后能够提升显著特征，利于情感特征对齐。

3）-Gate 和 CCA-CDG 相比效果下降，证明了跨模态动态门控模块能够关注图文的对齐融合，不同图文实例对分配不同融合权重，抑制图文不匹配融合。

4）-OriLink 和 CCA-CDG 相比分类效果高于 -Gate。说明在抑制未对齐的多模态融合时，保留的原始特征更能代表情感特征。

5）-Content 和 CCA-CDG 相比，去除全局特征联合分类效果有所降低。说明全局特征有一定的特征互补作用，能够提升预测精度。

3.6　重要超参数实验

为了评估并对比不同超参数配置下模型的性能差异，比较了 CCA-CDG 在 MVSA-single 和 MVSA-multi 数据集上不同的卷积核大小、注意力头数量和全局特征融合权重值 λ 下的 Acc 和 F1 值。

考虑到卷积核大小对卷积特征提取的差异性，设置了 3、4、5、6、7 共 5 种卷积核大小的对比实验，验证对 CCA-CDG 情感分类性能影响，实验结果如图 4 所示。当卷积核大小为 5 时，分类效果最佳。卷积核为 3 时，分类效果低于卷积核为 4 的分类效果，这是因为过小的卷积核限制在局部特征关注，不能有效地捕获特征的依赖关系，限制了模型表达能力。然而，当卷积核大小分别为 6、7 时，分类效果逐渐降低，这是因为过大的卷积核太过关注全局而忽略局部细节，导致数据中存在无关特征，从而在未见过的数据上泛化能力下降，影响分类精度。

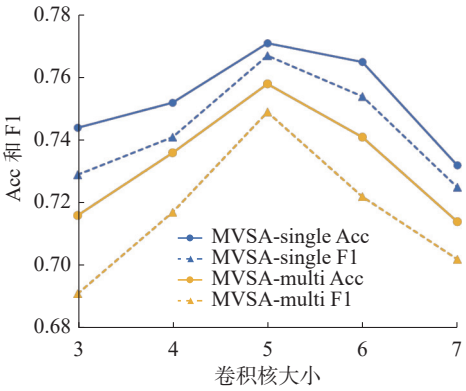


图 4　不同卷积核大小对应 Acc 和 F1

Fig. 4　Different convolution kernel sizes correspond to the Acc and F1

考虑到 CCAM 中注意力头数量对 CCA-CDG 性能影响，本文设置了 4、6、8、10 共 4 组注意力头数量的对比实验，实验结果如图 5 所示。随着头数量从 4 到 8 逐渐增加，模型的分类效果达到最优。然而，当头数量增加到 10 时，模型性能开始下降。上述现象的原因是，当注意力头数量较少时，模型处理不同的关注点受限，随着数量增加，模型更能充分捕获序列中复杂特征；头数量过大，可能导致模型学习到数据中的冗余特征，增加的头不会为模型性能带来实质性提升，增加了过拟合风险。所以，注意力头数量为 8 时，既能够捕获更复杂的特征，也不会造成信息冗余，分类效果达到最佳。

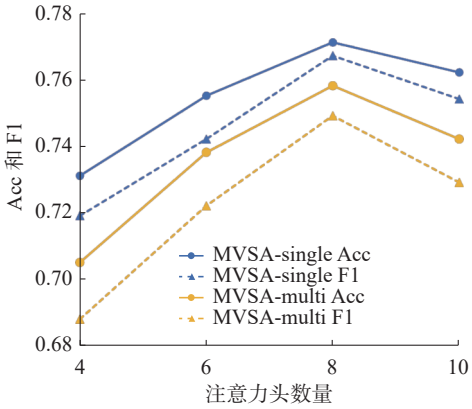


图 5　不同注意力头数量对应 Acc 和 F1

Fig. 5　The number of different attention heads corresponds to the Acc and F1

对于式（1）全局特征的融合权重 λ，分别设置了 λ=0.6,0.7,0.8,0.9 共 4 组对比实验，实验结果如图 6 所示。随着 λ 从 0.6 增加到 0.8，模型分类效果逐渐上升达到最优。而 λ=0.9 时，分类效果不如 λ=0.8。这是因为全局特征作为特征补充时，当 λ 过小，分配全局特征就会过大，导致了图文对齐特征关注度降低。当 λ 过大时，全局特征关注度就会降低，无法作为补充特征实现更全面的情感预测。

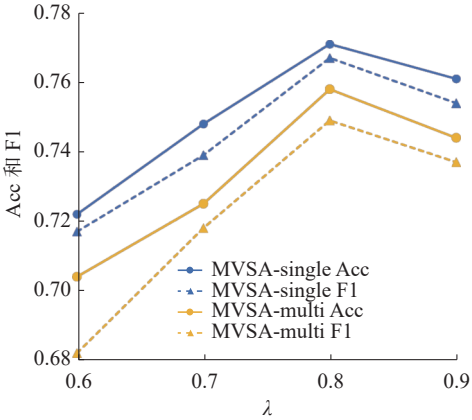


图 6　不同 λ 权重值对应 Acc 和 F1

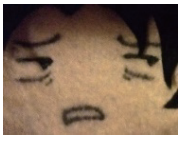
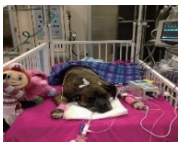
Fig. 6　Different weight values λ correspond to Acc and F1

3.7　案例研究

为了进一步证明 CCA-CDG 在图文情感分析的有效性, 随机抽取 MVSA-single 数据集多条数据进行案例示例分析。比较了 CCA-CDG、Resnet18 和 BERT 预测的情感标签。案例示例分析中从左往右依次是文本对 ID、图像信息、文本信息以及中文翻译、真实标签、Resnet18 情感预测结果、BERT 情感预测结果和 CCA-CDG 情感预测结果, 具体分类结果如表 7 所示。在抽取的 7 个图文对中, CCA-CDG 正确预测 6 个, 错误预测 1 个; Rsnet18 模型正确预测 5 个, 错误 2 个; BERT 模型正确预测 4 个, 错误预测 3 个。通过案例分析发现, 图像中存在丰富的视觉情感, 仅考虑文本数据或图像数据很难正确地预测用户情感倾向。例如 ID2 中, 文本描述偏向中性, 但是考虑到图像内容后发现, 用户真实情感是消极倾向。ID5 中, 仅考虑文本信息是中性情感, 但加入图像内容后, 用户情感倾向为积极。ID1 中图像表达为中性, 但文本积极情感非常突出, 所以 CCA-CDG 在特征融合时会重点关注文本情感特征, 正确预测为积极情感。对于案例 ID7, CCA-CDG、Resnet18 模型以及 BERT 模型都错误地预测了用户情感。ID7 真实情感为中性, BERT 模型将情感错误预测为消极, 可能因为 BERT 关注上下文特征, 捕捉深层语义特征, 对文本描述“获得第二名”理解成消极情感倾向。Resnet18 模型错误预测为积极情感, 是因为不考虑文本信息时, 图像内容的情感表达为积极。CCA-CDG 错误预测情感为积极, 这是因为图像中人物都在欢笑, 而文本描述的语义信息与图像信息情感未对齐。在图文特征融合时, 跨模态动态门控模块将过大的门控权重关注到原始特征中, 导致图像中人物欢笑区域被模型重点关注, 减少了对文本信息的关注, 导致模型情感预测错误。

表 7　案例示例

Table 7　Case example

ID	图像	文本	标签	Resnet18	BERT	CCA-CDG
1		Thank you to Eastwood 8th grader Sam L. for helping hang 7th grade #SafeDates Valentines Day cards. #Caring(感谢八年级学生Sam L.帮助挂起七年级情人节卡片。#爱心)	positive	**neutral(×)**	positive	positive
2		This year, for the first time I'm doting on kouhais... normally I'd be like(今年第一次这么宠后辈……平时的话)	negative	negative	**neutral (×)**	negative
3		The reason why this devoted dog is in critical condition will make you cry.(这只忠诚的狗处于危急状态的原因会让你落泪)	negative	negative	negative	negative
4		AWFUL...I'm the worst kind of person.(可怕的……我是最坏的那种人)	negative	negative	negative	negative
5		Taylor Lautner's Looking Huge!? His 2 Step Muscle Shredding System Is Flying Off The Shelves?(泰勒·洛特纳看起来壮硕了许多!？他的两步肌肉塑形正在走红)	positive	positive	**neutral(×)**	positive
6		1940 PTC Map of Philadelphia: Showing Street Car, Bus and Subway-Elevated Lines.(1940 PTC费城地图: 显示有轨电车、公共汽车和地铁高架线路)	neutral	neutral	neutral	neutral
7		Red Team finished 2nd today at the Lineman Challenge.(红队今天在Lineman Challenge中获得了第二名)	neutral	**positive (×)**	**negative(×)**	**positive (×)**

注: 加粗表示错误的预测结果。

4　结束语

本文提出了一种基于卷积交叉注意力与跨模态动态门控的多模态情感分析模型(CCA-CDG)。旨在解决传统多模态情感分析忽略图像和文本之间情感未对齐的问题。CCA-CDG 采用卷积交叉注意力机制，能够有效探索和捕获图像与文本之间的一致性情感表达。引入跨模态动态门控模块，通过权重值动态控制图像和文本情感特征的融合过程，能够有效地减少图文情感未对齐融合造成的特征冗余问题。考虑到图像和文本的上下文信息在情感分析中起互补作用，CCA-CDG 提取了图像全局特征与文本上下文特征作为特征补充，与图文情感交互特征融合，实现更可靠的情感预测。

同时本文研究仍然存在一定的局限性：1) CCA-CDG 解决了图文情感未对齐引起的特征冗余问题，但过度关注图文原始特征容易忽视图文交互特征。后续需进一步优化跨模态动态门控模块，解决案例 ID7 中图文情感表意不明显，模型陷入关注局部情感，导致预测错误问题。2) 模型结构复杂，训练时间较长，容易导致过拟合。后续研究重点放在网络层优化，降低模型复杂度。3) 当前社交媒体包含音频和视频等多种模态信息，未来研究将进一步融合更多模态数据。

参考文献：

[1] YUE Lin, CHEN Weitong, LI Xue, et al. A survey of sentiment analysis in social media[J]. Knowledge and information systems, 2019, 60: 617−663.

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

[3] 吴杰胜, 陆奎. 基于多部情感词典和规则集的中文微博情感分析研究[J]. 计算机应用与软件, 2019, 36(9): 93−99.
WU Jiesheng, LU Kui. Chinese weibo sentiment analysis based on multiple sentiment lexicons and rule sets[J]. Computer applications and software, 2019, 36(9): 93−99.

[4] 李洋, 董红斌. 基于 CNN 和 BiLSTM 网络特征融合的文本情感分析[J]. 计算机应用, 2018, 38(11): 3075−3080.
LI Yang, DONG Hongbin. Text sentiment analysis based on feature fusion of convolution neural network and bidirectional long short-term memory network[J]. Journal of computer applications, 2018, 38(11): 3075−3080.

[5] 刘继, 顾凤云. 基于 BERT 与 BiLSTM 混合方法的网络舆情非平衡文本情感分析[J]. 情报杂志, 2022, 41(4): 104−110.
LIU Ji, GU Fengyun. Unbalanced text sentiment analysis of network public opinion based on BERT and BiLSTM hybrid method[J]. Journal of intelligence, 2022, 41(4): 104−110.

[6] DATTA R, JOSHI D, LI Jia, et al. Studying aesthetics in photographic images using a computational approach[C]// Computer Vision–ECCV 2006. Berlin: Springer Berlin Heidelberg, 2006: 288−301.

[7] 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. Firenze: ACM, 2010: 83–92.

[8] ZHOU Yitao, REN Fuji, NISHIDE S, et al. Facial sentiment classification based on Resnet-18 model[C]//2019 International Conference on Electronic Engineering and Informatics. Nanjing: IEEE, 2019: 463−466.

[9] MEENA G, MOHBEY K K, INDIAN A, et al. Sentiment analysis from images using VGG19 based transfer learning approach[J]. Procedia computer science, 2022, 204: 411−418.

[10] YANG Xiaocui, FENG Shi, WANG Daling, et al. Image-text multimodal emotion classification via multi-view attentional network[J]. IEEE transactions on multimedia, 2020, 23: 4014−4026.

[11] 张继东, 张慧迪. 融合注意力机制的多模态突发事件用户情感分析[J]. 情报理论与实践, 2022, 45(11): 170−177.
ZHANG Jidong, ZHANG Huidi. Multimodal user emotion analysis of emergencies based on attention mechanism[J]. Information studies: theory & application, 2022, 45(11): 170−177.

[12] 杨力, 钟俊弘, 张赟, 等. 基于复合跨模态交互网络的时序多模态情感分析[J]. 计算机科学与探索, 2024, 18(5): 1318−1327.
YANG Li, ZHONG Junhong, ZHANG Yun, et al. Temporal multimodal sentiment analysis with composite cross modal interaction network[J]. Journal of frontiers of computer science and technology, 2024, 18(5): 1318−1327.

[13] ZADEH A, CHEN Minghai, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[EB/OL]. (2017−07−23)[2024−09−06]. https://arxiv.org/abs/1707.07250.

[14] HAN Wei, CHEN Hui, GELBUKH A, et al. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis[C]//Proceedings of the 2021 International Conference on Multimodal Interaction. Montréal: ACM, 2021: 6−15.

[15] 曾子明, 孙守强, 李青青. 基于融合策略的突发公共卫生事件网络舆情多模态负面情感识别[J]. 情报学报,

2023, 42(5): 611−622.
ZENG Ziming, SUN Shouqiang, LI Qingqing. Multimodal negative sentiment recognition in online public opinion during public health emergencies based on fusion strategy[J]. Journal of the China society for scientific and technical information, 2023, 42(5): 611−622.

[16] 杨颖, 钱馨雨, 王合宁. 结合多粒度视图动态融合的多模态方面级情感分析[J]. 计算机工程与应用, 2024, 60(22): 172−183.
YANG Ying, QIAN Xinyu, WANG Hening. Multimodal aspect-level sentiment analysis based on multi-granularity view dynamic fusion[J]. Computer engineering and applications, 2024, 60(22): 172−183.

[17] GAN Chenquan, FU Xiang, FENG Qingdong, et al. A multimodal fusion network with attention mechanisms for visual-textual sentiment analysis[J]. Expert systems with applications, 2024, 242: 122731.

[18] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Minneapolis: ACL, 2019: 4171−4186.

[19] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. (2014−09−01)[2024−09−06]. https://arxiv.org/abs/1409.0473.

[20] DENG Jia, DONG Wei, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248−255.

[21] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014−09−04)[2024−09−06]. https://arxiv.org/abs/1409.1556.

[22] ZHU Tong, LI Leida, YANG Jufeng, et al. Multimodal sentiment analysis with image-text interaction network[J]. IEEE transactions on multimedia, 2022, 25: 3375−3385.

[23] NIU Teng, ZHU Shiai, PANG Lei, et al. Sentiment analysis on multi-view social data[C]//MultiMedia Modeling. Cham: Springer International Publishing, 2016: 15−27.

[24] XU Nan, MAO Wenji. MultiSentiNet: a deep semant-ic network for multimodal sentiment analysis[C]//Proceedings of the 2017 ACM on Conference on Informa-tion and Knowledge Management. Singapore: ACM, 2017: 2399−2402.

[25] XU Nan, MAO Wenji, CHEN Guandan. A co-memory network for multimodal sentiment analysis[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Ann Arbor: ACM, 2018: 929–933.

[26] YANG Xiaocui, FENG Shi, ZHANG Yifei, 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). [S. l.]: ACL, 2021: 328−339.

[27] 蔡宇扬, 蒙祖强. 基于模态信息交互的多模态情感分析[J]. 计算机应用研究, 2023, 40(9): 2603−2608.
CAI Yuyang, MENG Zuqiang. Multimodal sentiment analysis based on modal information interaction[J]. Application research of computers, 2023, 40(9): 2603−2608.

[28] LU Wenjie, ZHANG Dong. Unified multi-modal multi-task joint learning for language-vision relation inference [C]//2022 IEEE International Conference on Multimedia and Expo. Taipei: IEEE, 2022: 1−6.

[29] 周婷, 杨长春. 基于多层注意力机制的图文双模态情感分析[J]. 计算机工程与设计, 2023, 44(6): 1853−1859.
ZHOU Ting, YANG Changchun. Image-text sentiment analysis based on multi-level attention mechanism[J]. Computer engineering and design, 2023, 44(6): 1853−1859.

作者简介:

仲兆满, 教授, 江苏海洋大学计算机工程学院院长, 中国矿业大学兼职博士生导师, 主要研究方向为互联网舆情大数据分析及管控。主持国家自然科学基金面上项目 1 项, 获中国自动化学会科技进步奖二等奖, 发表学术论文 50 余篇, 出版专著 1 部。E-mail: zhongzhaoman@163.com。

樊继冬, 硕士研究生, 主要研究方向为多模态情感分析、大数据采集与分析。E-mail: ffanjdong@163.com。

张渝, 硕士研究生, 主要研究方向为网络舆情分析、方面级情感分析。E-mail: zhou90616@gmail.com。

[责任编辑: 丁钰]