文章编号:1671-9352(2026)03-0086-10 DOI:10.6040/j.issn.1671-9352.0.2024.230

基于自注意力机制的中心距差异多模态情感分析

陈忠源,路翀*
(新疆财经大学信息管理学院, 新疆 乌鲁木齐 830012)

摘要:为了解决现有模型在模态间相关性挖掘、特征融合方式和标签更新机制上存在的问题,提出一种基于自注意力机制的中心距差异多模态情感分析方法(center moment discrepancy multimodal sentiment analysis based on self-attention mechanism, SA-CMD)。首先,使用编码器对提取的特征序列进行编码,并通过自注意力机制动态调整各模态特征的权重,捕捉模态间复杂的依赖关系。然后,引入中心距差异方法动态优化特征表示和标签分布,增强模型的鲁棒性。在特征融合过程中,通过计算模态特征与其正负中心的距离差异,生成更准确的特征标签,进一步提高融合特征的质量。最终,使用线性层将融合特征投影到低维空间进行预测。实验结果表明,SA-CMD 在公开数据集 CMU-MOSI 和 CMU-MOSEI 上的各项评价指标均优于现有基准模型,特别是在相关系数、二分类精度和七分类精度指标上表现优越。进一步验证自注意力机制和中心距差异方法在提升模型性能中的关键作用,充分说明 SA-CMD 模型在多模态情感分析任务中的有效性和鲁棒性。

关键词:多模态情感分析;自注意力机制;中心距差异;多模态特征融合
中图分类号:TP391 **文献标志码**:A
引用格式:陈忠源,路翀. 基于自注意力机制的中心距差异多模态情感分析[J]. 山东大学学报(理学版),2026,61(3):86-95,110.

Center moment discrepancy multimodal sentiment analysis based on self-attention mechanism

CHEN Zhongyuan, LU Chong*
(College of Information Management, Xinjiang University of Finance and Economics, Urumqi 830012, Xinjiang, China)

Abstract: A center moment discrepancy multimodal sentiment analysis based on self-attention mechanism (SA-CMD) is proposed, aiming to address issues related to modality correlation mining, feature fusion strategies, and label updating mechanisms in existing models. First, an encoder is used to encode the extracted feature sequences, and the weights of each modality's features are dynamically adjusted through a self-attention mechanism to capture the complex dependencies between modalities. Next, the center moment discrepancy method is introduced to dynamically optimize feature representations and label distributions, enhancing the model's robustness. During the feature fusion process, the model calculates the distance discrepancy between modality features and their respective positive and negative centers to generate more accurate feature labels, further improving the quality of the fused features. Finally, a linear layer is used to project the fused features onto a lower-dimensional space for prediction. Experimental results show that SA-CMD outperforms existing baseline models in the public CMU-MOSI and CMU-MOSEI datasets across various evaluation metrics, especially in terms of the Pearson correlation coefficient, binary classification accuracy, and seven-class classification accuracy. Ablation experiments further verify the key role of the self-attention mechanism and the center moment discrepancy method in enhancing model performance, fully demonstrating the effectiveness and robustness of the SA-CMD model in multimodal sentiment analysis tasks.

Key words: multimodal sentiment analysis; self-attention mechanism; center moment discrepancy; multimodal feature fusion

0 引言

多模态情感分析(multimodal sentiment analysis, MSA)在计算机科学和人工智能领域受到广泛关注,应

收稿日期:2024-07-04; **网络出版时间**:2025-07-16
基金项目:国家自然科学基金资助项目(62166039)
第一作者:陈忠源(1999—),男,硕士研究生,研究方向为多模态情感分析、数据处理. E-mail:15864206265@163.com
*通信作者:路翀(1966—),男,教授,硕士生导师,博士,研究方向为人工智能、多模态情感分析、图像处理. E-mail:498841300@qq.com

用场景包括人机交互、社交媒体分析以及心理健康监测等[1-2]。例如,智能语音助手 Siri 通过语音识别和自然语言处理(natural language processing, NLP)技术与用户进行互动,执行设置提醒、播放音乐或查询天气等任务;企业使用社交媒体监测工具分析用户在平台上的反馈和情绪,以优化产品和服务;心理健康应用通过聊天机器人与用户交流,提供情绪支持和认知行为疗法建议,帮助用户管理焦虑和抑郁情绪。

如何提高多模态情感分析模型的性能始终是自然语言处理领域的热点问题。Tsai 等[3]提出的多模态变换器(multimodal Transformer, MulT)通过跨模态注意力机制实现模态间的信息交互,提升了多模态情感分析的性能。Liu 等[4]提出了跨模态注意力网络,通过在时间不一致的多模态情感分析中利用注意力机制实现模态间的特征交互。罗渊贻等[5]提出一种学习情感语义表达一致的多模态情感分析方法,通过交叉注意力机制获取模态间辅助信息,并利用软注意力加权连接情感一致的特征,增强强模态表达、抑制弱模态影响。然而,当前的研究在处理模态间相关性、优化特征融合方式和更新标签机制方面仍存在诸多挑战[6-7]。1) 现有模型在处理模态间相关性时,多采用简单的特征拼接或对每个模态独立处理的方法,导致模型无法充分捕捉模态之间的复杂依赖关系和互补信息[7]。2) 现有方法在特征融合时缺乏动态调整能力,导致融合特征的表达能力有限[8]。3) 现有模型在训练过程中忽略了标签分布的动态变化,无法有效地对特征进行重新标注和调整,导致模型的鲁棒性和泛化能力不足[9]。

为解决上述问题,本文提出一种新的多模态情感分析模型——基于自注意力机制的中心距差异多模态情感分析模型(center moment discrepancy multimodal sentiment analysis based on self-attention mechanism, SA-CMD)。该模型利用自注意力机制(self-attention mechanism, SA),在模态融合之前根据每个模态的重要性为其分配权重,从而更有效地捕捉模态间的复杂依赖关系,并提高模型的鲁棒性。此外,在特征融合的基础上,SA-CMD 模型引入了中心距差异(center moment discrepancy, CMD)方法,动态更新每个模态的特征中心,从而优化特征表示和标签分布。具体而言,中心距差异方法通过计算每个模态特征与其正负中心的距离差异,并基于此差异来更新特征标签,使得特征表示逐步逼近最优标签分布。通过这些方法,SA-CMD 模型能够有效地捕捉文本、音频和视觉模态之间的复杂依赖关系,并通过动态调整特征权重和优化标签分布,提升多模态情感识别的性能。

1　相关工作

1.1　多模态情感分析

MSA 是一个跨学科的研究领域,旨在通过综合分析文本、音频和视觉等多模态数据,准确识别和预测人类的情感状态。现有研究主要关注如何更高效地融合和利用不同模态的信息,以提升情感识别的精度和鲁棒性。Zadeh 等[10]提出了张量融合网络(tensor fusion network, TFN),通过计算输入模态的外积来捕捉模态间的高阶关系。Liu 等[11]提出了低秩多模态融合(low-rank multimodal fusion, LMF)方法,通过将高阶张量分解为低秩因子,实现高效地多模态融合。Sun 等[12]构建了 2 个外积矩阵来表示文本与视频及文本与音频之间的交互,并利用典型相关分析(canonical correlation analysis, CCA)层进行特征优化。Han 等[13]通过分层结构,最大化单模态输入与多模态融合结果之间的互信息。田昌宁等[14]将不同模态映射到私有和共享子空间,以获得各模态的私有和共享表示,从而学习每种模态的差异化信息和统一信息。尽管这些方法在多模态情感分析中取得了显著进展,但仍无法充分捕捉模态间的复杂依赖关系。本文以自注意力为基础,在模态融合之前按照重要性为每个模态分配权重,解决了模态间复杂依赖关系的问题,并提高了模型的精度和鲁棒性。

1.2　自注意力机制

自注意力机制最早由 Bahdanau 等[15]提出,用于神经机器翻译(neural machine translation, NMT)中。这种机制通过计算输入序列中每个词与其他词的关系权重,有效地捕捉了长距离依赖关系。Vaswani 等[16]提出完全依赖自注意力机制的 Transformer 模型,避免了传统循环神经网络(recurrent neural network, RNN)的顺序计算瓶颈,提高了计算效率和性能。Devlin 等[17]提出了 BERT(bidirectional encoder representations from Transformers)预训练模型,利用双向 Transformer 对文本进行编码,显著提升了多种 NLP 任务的性能。随着自注意力机制在单模态 NLP 任务中的成功,研究者们开始将其应用于多模态情感分析任务中,如文本增强 Transformer 融合网络(text enhanced Transformer fusion network, TETFN),利用自注意力机制实现不同模态

间的有效信息融合[18]。

2 SA-CMD 模型方法

多模态情感分析任务旨在针对给定的视频话语预测情感分数。本文提出的 SA-CMD 模型的输入由文本($\boldsymbol{Z}_{\mathrm{t}} \in \mathbf{R}^{l_{\mathrm{t}} * d_{\mathrm{t}}}$)、视觉($\boldsymbol{Z}_{\mathrm{v}} \in \mathbf{R}^{l_{\mathrm{v}} * d_{\mathrm{v}}}$)和音频($\boldsymbol{Z}_{\mathrm{a}} \in \mathbf{R}^{l_{\mathrm{a}} * d_{\mathrm{a}}}$)3 种模态组成。具体而言,文本模态 $\boldsymbol{Z}_{\mathrm{t}}$ 是数据集的初始特征序列,视觉 $\boldsymbol{Z}_{\mathrm{v}}$ 和音频 $\boldsymbol{Z}_{\mathrm{a}}$ 模态特征序列则通过特征提取器从视频中提取,l_m, $m \in \{\mathrm{t},\mathrm{v},\mathrm{a}\}$ 表示各输入模态的序列长度;d_m, $m \in \{\mathrm{t},\mathrm{v},\mathrm{a}\}$ 表示输入模态的特征维度。图 1 为 SA-CMD 的整体框架图,包括模态特征提取、自注意力机制、特征融合及预测、中心距差异方法和损失函数 5 个部分。首先,使用编码器对初始特征序列进行编码,通过自注意力机制对编码后的时间序列特征进行加权。然后,将 3 种模态的特征表示通过融合层进行融合,并通过线性层投影到低维空间进行预测。在此过程中,利用中心距差异方法生成标签,并通过融合模态与单一模态的损失来优化模型。

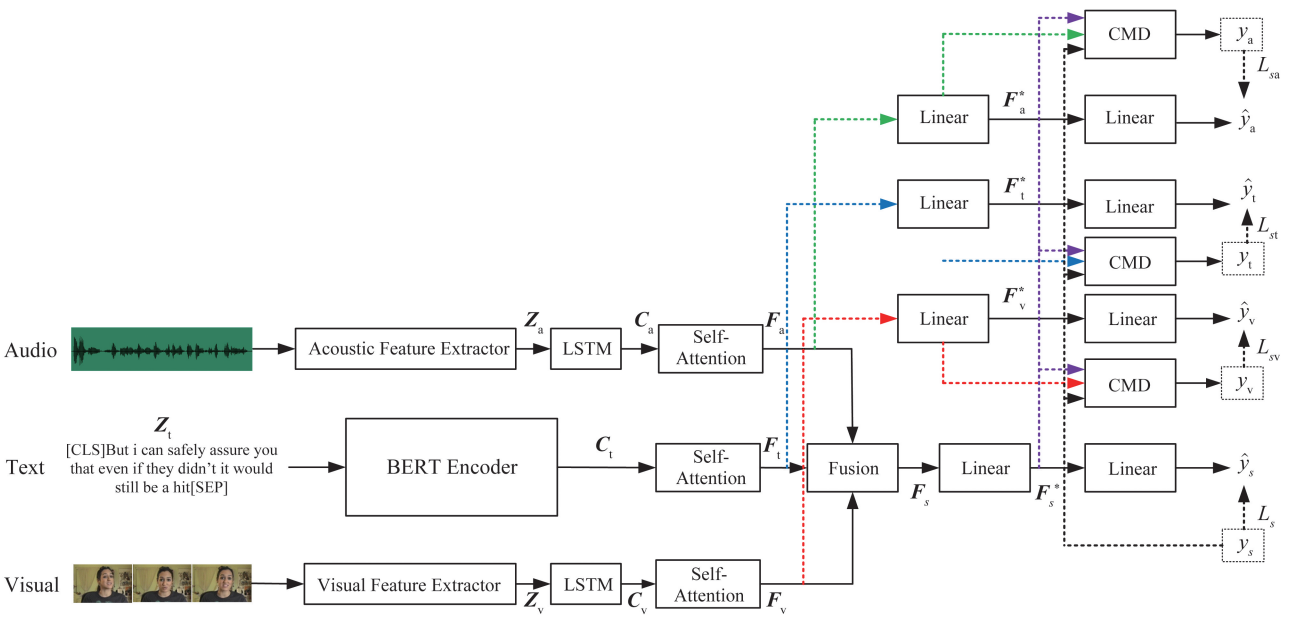

图 1 SA-CMD 模型的整体框架

Fig.1 Overall framework of the SA-CMD model

2.1 模态特征提取

首先,使用模态编码器对初始特征序列 $\boldsymbol{Z}_m$, $m \in \{\mathrm{t},\mathrm{v},\mathrm{a}\}$ 进行编码,并将编码后的特征序列定义为相应的模态表示 $\boldsymbol{C}_m$, $m \in \{\mathrm{t},\mathrm{v},\mathrm{a}\}$。具体来说,SA-CMD 模型使用预先训练的工具包提取初始的视频和音频特征序列 $\boldsymbol{Z}_{\mathrm{v}}$ 和 $\boldsymbol{Z}_{\mathrm{a}}$。对于文本模态,使用 BERT[17] 对输入的句子进行编码,并从最后一层的输出中提取[CLS]嵌入作为文本特征表示 $\boldsymbol{C}_{\mathrm{t}}$。在视觉和音频模态方面,使用单层单向 LSTM [19] 对模态的特征序列进行编码,以获取视觉特征序列和音频特征序列 $\boldsymbol{C}_{\mathrm{v}}$ 和 $\boldsymbol{C}_{\mathrm{a}}$。编码过程为

$$\boldsymbol{C}_{\mathrm{t}} = \mathrm{BERT}(\boldsymbol{Z}_{\mathrm{t}};\theta^{\mathrm{BERT}}),$$

$$\boldsymbol{C}_{\mathrm{v}} = \mathrm{sLSTM}(\boldsymbol{Z}_{\mathrm{v}};\theta^{\mathrm{LSTM}}),$$

$$\boldsymbol{C}_{\mathrm{a}} = \mathrm{sLSTM}(\boldsymbol{Z}_{\mathrm{a}};\theta^{\mathrm{LSTM}}),$$

其中,sLSTM 为单层单向 LSTM,θ^{BERT}为 BERT 的参数集合,θ^{LSTM}为 LSTM 的参数集合。

尽管通过编码器生成的特征序列包含了各模态丰富的情感信息,但不同模态中的噪声和无关特征仍可能对情感分析的准确性产生负面影响。

2.2 自注意力机制

为了有效应对噪声和无关信息对情感分析的干扰,本文采用了自注意力机制。该机制能够动态调整各模态特征的权重,从而强化重要的特征表示并抑制无关或噪声特征。Zadeh 等[10] 证明了自注意力机制在捕

捉序列数据的长距离依赖和处理噪声方面的有效性。Vaswani 等[16]则提出了 Transformer 模型,该模型完全依赖自注意力机制,通过多头自注意力取代了传统编码器(解码器)架构中的循环层,从而在多个任务中实现了显著的性能提升。此外,Tsai 等[3]在多模态情感分析中引入了跨模态注意力机制,验证了注意力机制在捕捉模态间依赖关系方面的有效性。

基于上述研究成果,本文设计了一种自注意力机制用于优化文本、视觉和音频模态特征的表示。具体而言,自注意力机制通过对每个模态的输入序列 $\boldsymbol{C}_m$ 进行加权求和,生成加权后的特征表示 $\boldsymbol{F}_m$, $m \in \{\mathrm{t},\mathrm{v},\mathrm{a}\}$。这种加权机制旨在强化重要特征,同时抑制无关或噪声特征,从而提升情感分析的准确性。自注意力机制模块的具体结构如图 2 所示。

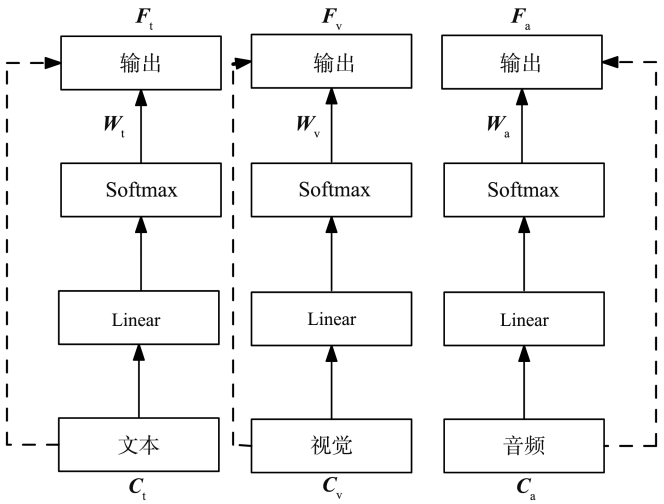

图 2　自注意力机制的结构

Fig.2　Structure of the self-attention mechanism

在 SA-CMD 模型中,自注意力机制用于模态融合之前,具体实现步骤如下。

第 1 步,计算注意力权重。对于输入序列 $\boldsymbol{C}_m=[\boldsymbol{C}_m^1,\boldsymbol{C}_m^2,\cdots,\boldsymbol{C}_m^l]$,使用线性变换层计算每个时间步的注意力权重为

$$\boldsymbol{W}_m^i=\mathrm{Softmax}(\boldsymbol{w}_m\boldsymbol{C}_m^i),\quad m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\},$$

其中 $\boldsymbol{w}_m$ 是一个线性层的权重矩阵,i 小于等于序列长度 l。

第 2 步,使用计算得到的注意力权重对输入序列进行加权求和,得到加权后的特征表示,计算过程为

$$\boldsymbol{F}_m=\sum_{i=1}^{l_m}\boldsymbol{W}_m^i*\boldsymbol{C}_m^i,$$

其中:l_m 表示模态 m 的序列长度,$\boldsymbol{C}_m^i$ 表示 m 模态输入序列的第 i 个元素。

2.3　特征融合及预测

在得到每个模态加权求和的特征表示后,将 3 种模态的特征表示 $\boldsymbol{F}_\mathrm{t}$、$\boldsymbol{F}_\mathrm{v}$ 和 $\boldsymbol{F}_\mathrm{a}$ 进行简单的拼接,形成统一的特征表示 $\boldsymbol{F}_s$。接着,通过线性层将统一的特征表示投影到低维空间得到 $\boldsymbol{F}_s^*$,最后通过另一个线性层对低维度的特征表示 $\boldsymbol{F}_s^*$ 进行处理,得到最终的预测结果 $\hat{y}_s$,计算过程为

$$\boldsymbol{F}_s=\mathrm{Concat}[\boldsymbol{F}_\mathrm{t},\boldsymbol{F}_\mathrm{v},\boldsymbol{F}_\mathrm{a}],$$

$$\boldsymbol{F}_s^*=\mathrm{ReLU}((\boldsymbol{W}_{l_1}^s)^\mathrm{T}*\boldsymbol{F}_s+\boldsymbol{b}_{l_1}^s),$$

$$\hat{y}_s=\mathrm{ReLU}((\boldsymbol{W}_{l_2}^s)^\mathrm{T}*\boldsymbol{F}_s^*+\boldsymbol{b}_{l_2}^s),$$

其中:$\boldsymbol{W}_{l_1}^s$、$\boldsymbol{W}_{l_2}^s\in\mathbf{R}^{(d_t+d_v+d_a)*d_s}$, ReLU 为激活函数,$\boldsymbol{b}_{l_1}^s$、$\boldsymbol{b}_{l_2}^s$为线性层的偏置项。

2.4　中心距差异

已有研究证明,中心距差异方法在多模态特征优化中具有显著的效果[5,20-21]。中心距差异方法通过在特征空间中引入中心距离的概念,有效地优化了模态间的共性表示,提高了模型的泛化能力和鲁棒性。

在本文提出的 SA-CMD 模型中,中心距差异方法不仅用于生成单模态标签,还在动态优化各模态特征表示中发挥关键作用。具体来说,在特征融合过程中,中心距差异方法首先对每个模态的加权求和特征 $\boldsymbol{F}_m$ 进行处理,通过线性层将其投影到低维空间,得到低维度的特征表示 $\boldsymbol{F}_m^*$, $m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$,确保了每个模态的

特征在降维后仍然保持其重要信息。

接着,低维度的特征表示通过另一个线性层得到单模态的预测结果 $\hat{y}_m$, $m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$。这些单模态的预测结果虽然仅在训练阶段使用,但它们对于优化特征表示和标签分布起到了关键作用。计算过程如下:

$$\boldsymbol{F}_m^* = \mathrm{ReLU}(\boldsymbol{W}_{l_1}^m)^{\mathrm{T}} * \boldsymbol{F}_m + \boldsymbol{b}_{l_1}^m),$$

$$\hat{y}_m = \mathrm{ReLU}(\boldsymbol{W}_{l_2}^m)^{\mathrm{T}} * \boldsymbol{F}_m^* + \boldsymbol{b}_{l_2}^m)。$$

在训练过程中,中心距差异方法通过计算每个模态特征与其正负中心的距离差异,生成单模态标签。具体实现步骤如下。

第 1 步,确定不同模态表示的正负中心。对于单一模态 $\boldsymbol{F}_m^*$ 和融合模态 $\boldsymbol{F}_s^*$,通过计算各模态特征在正向和负向样本上的均值,来获得正负中心,计算过程为

$$c_m^{\mathrm{pos}} = \frac{\sum_{j=1}^{l} I(\boldsymbol{F}_m^*(j)>0)\cdot \boldsymbol{F}_{mj}^{\mathrm{g}}}{\sum_{j=1}^{l} I(\boldsymbol{F}_m^*(j)>0)}, \quad m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\},$$

$$c_m^{\mathrm{neg}} = \frac{\sum_{j=1}^{l} I(\boldsymbol{F}_m^*(j)<0)\cdot \boldsymbol{F}_{mj}^{\mathrm{g}}}{\sum_{j=1}^{l} I(\boldsymbol{F}_m^*(j)<0)}, \quad m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\},$$

$$c_s^{\mathrm{pos}} = \frac{\sum_{j=1}^{l} I(\boldsymbol{F}_S^*(j)>0)\cdot \boldsymbol{F}_{sj}^{\mathrm{g}}}{\sum_{j=1}^{l} I(\boldsymbol{F}_S^*(j)>0)},$$

$$c_s^{\mathrm{neg}} = \frac{\sum_{j=1}^{l} I(\boldsymbol{F}_S^*(j)<0)\cdot \boldsymbol{F}_{sj}^{\mathrm{g}}}{\sum_{j=1}^{l} I(\boldsymbol{F}_S^*(j)<0)},$$

其中:$\sum_{j=1}^{l}$ 表示对所有样本进行求和操作;$I(\cdot)$为示性函数,当括号内的条件为真时取值为 1,否则取值为 0;$\boldsymbol{F}_{mj}^{\mathrm{g}}$为 m 模态第 j 个样本的全局表示;$\boldsymbol{F}_{sj}^{\mathrm{g}}$为 s 模态第 j 个样本的全局表示。

第 2 步,计算模态特征与正负中心的距离。首先将单一模态 $\boldsymbol{F}_m^*$ 和融合模态 $\boldsymbol{F}_s^*$ 执行归一化操作,然后计算模态特征与正负中心的欧氏距离 D_m^{pos}, $m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$, D_m^{neg}, $m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$, D_s^{pos} 和 D_s^{neg} 分别为

$$D_m^{\mathrm{pos}} = \| \hat{\boldsymbol{F}}_m^* - \hat{\boldsymbol{c}}_m^{\mathrm{pos}} \|_2,$$

$$D_m^{\mathrm{neg}} = \| \hat{\boldsymbol{F}}_m^* - \hat{\boldsymbol{c}}_m^{\mathrm{neg}} \|_2,$$

$$D_s^{\mathrm{pos}} = \| \hat{\boldsymbol{F}}_s^* - \hat{\boldsymbol{c}}_s^{\mathrm{pos}} \|_2,$$

$$D_s^{\mathrm{neg}} = \| \hat{\boldsymbol{F}}_s^* - \hat{\boldsymbol{c}}_s^{\mathrm{neg}} \|_2,$$

其中:$\hat{\boldsymbol{F}}_m^*$ 和 $\hat{\boldsymbol{F}}_s^*$ 表示对单一模态 $\boldsymbol{F}_m^*$ 和融合模态 $\boldsymbol{F}_s^*$ 执行归一化后得到的结果;$\hat{\boldsymbol{c}}_m^{\mathrm{pos}}$,$\hat{\boldsymbol{c}}_m^{\mathrm{neg}}$ 和 $\hat{\boldsymbol{c}}_s^{\mathrm{pos}}$,$\hat{\boldsymbol{c}}_s^{\mathrm{neg}}$ 分别表示对单一模态的正负中心以及融合模态的正负中心执行归一化后得到的结果;$\|\cdot\|_2$ 表示欧几里得范数。

第 3 步,计算相对距离。计算特征与正负中心的相对距离,以及当前模态的权重 δ_m, $m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$ 和 δ_s 为

$$\delta_m = \frac{D_m^{\mathrm{neg}} - D_m^{\mathrm{pos}}}{D_m^{\mathrm{pos}} + \varepsilon},$$

$$\delta_s = \frac{D_s^{\mathrm{neg}} - D_s^{\mathrm{pos}}}{D_s^{\mathrm{pos}} + \varepsilon},$$

其中,ε 是一个很小的常数,用于防止除零错误。

第 4 步,生成单一模态标签 y_m, $m\in\{\mathrm{t},\mathrm{v},\mathrm{a}\}$

$$y_m = y_s + \frac{\delta_m - \delta_s}{2} * \frac{y_s + \delta_s}{\delta_s},$$

其中 y_s 为实际的融合模态标签。

第 5 步,更新单一模态标签。对于每个模态和对应的索引,标签更新的过程如下:

$$\begin{cases} y_m^n = \dfrac{n-1}{n+1} y_m^{n-1} + \dfrac{2}{n+1} y_m^{n-1}, & n>1, \\ y_m^n = y_s, & n=1, \end{cases}$$

其中 n 为训练的轮次。

最后,为了有效地将中心距差异方法整合到模型的训练过程中,本文将中心距差异作为一种关键的度量方法引入到损失函数的设计中。具体而言,中心距差异方法不仅用于衡量模态间特征的距离差异,还在损失函数中起到优化模态间特征的作用。通过在损失函数中引入中心距差异方法,模型能够在训练过程中动态地最小化不同模态特征之间的中心距差异,从而提升多模态特征的表达能力和情感预测的准确性。损失函数的详细计算见 2.5 节。

2.5　损失函数

为了将中心距差异方法与模型的整体训练目标有机结合,本文设计了一种综合性的损失函数 L_{total}。该损失函数综合了传统分类损失和基于中心距差异方法的损失,从而为多模态特征的优化提供依据。通过结合中心距差异方法,损失函数能够同时优化特征表达的质量和模型的整体性能。具体而言,该损失函数的计算结合了当前预测值 $\hat{y}_s$、$\hat{y}_m$ 和实际标签 y_s、y_m,确保模态间特征差异的最小化。对于融合模态,赋予其固定的权重 1;而对于单一模态,则通过与融合模态的实际标签值进行计算,得到 m 模态第 i 个样本的权重 $\boldsymbol{W}_{sm}^i$。最终的损失值 L_{total}是预测误差与权重乘积的均值,计算过程如下:

$$L_s = |\hat{y}_s^i - y_s^i|,$$

$$L_{sm} = \sum_{m \in \{\text{t,v,a}\}} \boldsymbol{W}_{sm}^i \, |\hat{y}_m^i - y_m^i|,$$

$$L_{\text{total}} = \frac{1}{N} \sum_{i=1}^{N} (L_{sm} + L_s),$$

其中 N 为训练样本的数量。

3　实验

本章将介绍实验所用的数据集、评价标准、基准模型、实验参数、实验结果分析、消融研究以及案例分析。

3.1　数据集

本次试验在自然语言处理研究中的 CMU-MOSI 和 CMU-MOSEI 2 个公开数据集上进行。

CMU-MOSI[22]:CMU-MOSI 数据集是多模态情感分析研究中的经典基准数据集。该数据集由YouTube 独白视频片段组成,包含 93 个视频,总计 2 199 个主观话语片段,这些片段由 89 位不同的演讲者分享他们对电影或其他感兴趣主题的观点。每个话语片段被手动注释为介于[-3,3]之间的连续意见得分,其中-3 表示强烈的负面情绪,3 表示强烈的正面情绪。

CMU-MOSEI[1]:CMU-MOSEI 数据集与 CMU-MOSI 相似,但规模更大。它包含来自 3 228 个视频和 1 000位不同演讲者的 23 453 个带注释话语,涉及在线视频中的 250 个常用主题。该数据集的标注与 CMU-MOSI 相同,每个话语都可以被视为一个独立的多模态示例。表 1 展示了 CMU-MOSI 和 CMU-MOSEI 数据集的具体划分。

表 1　CMU-MOSI 和 CMU-MOSEI 数据集的数据划分

Table 1　Data split of the CMU-MOSI and CMU-MOSEI datasets　　单位:个

数据集	训练集	验证集	测试集	总数
CMU-MOSI	1 284	229	686	2 199
CMU-MOSEI	16 326	1 871	4 659	22 856

3.2　评价标准

根据文献[9,13,21],本文在 MOSI 和 MOSEI 数据集上进行实验,采用以下评价指标来评估模型的分类和回归性能:对于分类任务,使用 Acc-7 衡量模型从强烈负面(-3)到强烈正面(3)的 7 个区间中预测正确

的比例,其中,7 个区间包括:[-3,-2.5],(-2.5,-1.5],(-1.5,-0.5],(-0.5,0.5),[0.5,1.5),[1.5,2.5),[2.5,3];使用 Acc-2 和加权分数(*F*1 Score, *F*1)评估模型在正/负情感分类中的准确性和整体性能;对于回归任务,采用平均绝对误差(mean absolute error, MAE)计算预测值与实际值之间的平均绝对差值,并使用相关系数(correlation coefficient, Corr)评估预测值与实际值之间的相关性。除 MAE 外,其他指标值越高代表模型性能越好。

本文通过上述评价指标对 SA-CMD 模型的性能进行验证,以全面衡量其在多模态情感分析任务中的有效性。

3.3 基准模型

为了验证本文模型的性能,本文将其与以下 8 种基准模型进行比较。

1) TETFN[18]:通过文本导向的跨模态映射和多头注意力机制,获取有效的统一多模态表示。

2) ICCN[12]:最大限度地减少模态表示间的规范损失,以改善融合结果。

3) MISA[9]:模态表示和特定表示将特征投影到具有特殊限制的单独的 2 个空间中,然后在这些特征上完成融合。

4) SELF-MM[21]:为每个模态分配一个带有自动生成标签的单模态训练任务,旨在优化梯度反向传播。

5) MMIM[13]:通过分层结构最大化单模态输入对及其多模态融合结果与单模态输入之间的互信息。

6) PS-Mixer[23]:通过混合极性向量和强度向量,实现不同模态数据之间的有效通信。

7) MSTFN[11]:将不同模态映射到私有和共享子空间,获取不同模态的私有表示和共享表示,从而学习每种模态的差异化信息和统一信息。

8) 语义一致的模态特征表示[5]:通过学习每个模态的共性特征表示,利用交叉注意力机制使各模态从其他模态的共性特征中获取辅助信息,提高多模态协同决策的效果,同时保留模态的原始信息。

3.4 参数设置

本次实验使用未对齐的原始数据[21],在 RTX 3090 GPU 上利用深度学习框架 PyTorch 进行实现。为了优化模型性能,本文对超参数进行了网格搜索。具体的超参数设置如下:批量大小(batch size)设置为 16、32、64;学习率(learning rate)设置为 1e-3、1e-4、1e-5、2e-5、3e-4、3e-5、4e-5、5e-3、5e-5;视觉和音频模态的隐藏层大小(V_hidden size 和 A_hidden size)设置为 16、32、64、128;节点丢弃率(dropout)设置为 0.1、0.2、0.3、0.4、0.5。经过网格搜索,选择了最优的超参数组合,其详细信息如表 2 所示。

表 2 最优超参数

Table 2 Optimal hyperparameters

数据集	batch size	learning rate bert	learning rate audio	learning rate video	learning rate other	V-LSTM hidden size	A-LSTM hidden size	dropout
CMU-MOSI	16	1e-5	5e-5	5e-5	3e-4	16	32	0.1
CMU-MOSEI	32	2e-5	1e-3	1e-3	1e-4	32	64	0.1

3.5 实验结果分析

本文模型与基准模型在 CMU-MOSI 和 CMU-MOSEI 数据集上的对比结果如表 3 所示。

从表 3 的实验结果可以看出,本文模型在 CMU-MOSI 数据集上的各项指标均优于现有基准模型。其中,本文模型的 Corr 指标较基准模型提升了 0.288%~12.37%; Acc-2 指标提升了 0.127%~5.652%; Acc-7 指标提升了 0.570%~22.199%。上述分析结果表明,本文模型在多个评价指标上均表现出稳健的性能。

在 CMU-MOSEI 数据集上,本文模型在 Corr、Acc-2、*F*1 和 Acc-7 方面同样超越了现有基准模型。例如,本文模型在 Acc-2 指标相较于同样使用标签生成方法的 SELF_MM 提升了 0.901%,在使用类似模态特征提取方法的条件下,相较于 MSTFN 提升了 0.326%,而在结合标签生成和类似模态特征提取方法的情况下,较 TETFN 提升了 1.280%。值得注意的是,随着数据集规模的扩大,本文模型的性能提升更为显著。

这些结果表明,本文提出的自注意力机制与中心距差异方法,通过更有效地捕捉模态间的复杂依赖关系并优化特征表示,提高了模型在处理噪声和不一致特征时的鲁棒性,并且在处理更大规模数据集时也表现出良好的扩展性和稳定性。总体而言,本文模型在 CMU-MOSI 和 CMU-MOSEI 数据集上的表现均优于现有基准模型,证明了其在多模态情感分析任务中的有效性。

表 3　多个模型在 CMU-MOSI 和 CMU-MOSEI 数据集上的结果对比
Table 3　Comparison of multiple model results on the CMU-MOSI and CMU-MOSEI dataset

模型	CMU-MOSI					CMU-MOSEI				
	Corr	MAE	Acc-2	*F*1	Acc-7	Corr	MAE	Acc-2	*F*1	Acc-7
TETFN[a]	0.800 0	0.717 0	86.10	86.07	—	0.748 0	0.551 0	85.18	85.27	—
ICCN[a]	0.714 0	0.862 0	83.07	83.02	39.01	0.714 0	0.565 0	84.18	84.36	51.58
MISA[a]	0.761 0	0.783 0	83.40	83.60	42.30	0.756 0	0.555 0	85.50	85.30	52.20
SELF-MM[a]	0.798 0	0.713 0	85.98	85.95	—	0.765 0	0.530 0	85.17	85.30	—
MMIM[a]	0.800 0	0.700 0	86.06	85.98	46.65	0.772 0	**0.526 0**	85.97	85.94	54.24
PS-Mixer[a]	0.748 0	0.794 0	82.10	82.10	44.31	0.765 0	0.537 0	86.10	85.77	53.00
MSTFN[a]	0.800 0	0.705 0	86.63	86.63	—	0.762 0	0.537 0	85.99	85.92	—
语义一致的模态特征表示[a]	0.796 0	0.709 0	86.60	**86.70**	47.40	0.767 0	0.537 0	85.40	85.40	53.10
TETFN[b]	0.792 0	0.714 0	84.60	84.53	45.63	0.769 6	0.537 3	86.21	86.11	53.90
SELF_MM[b]	0.795 0	0.718 3	85.37	85.35	46.65	0.768 0	0.530 8	85.50	85.34	53.66
本文模型	**0.802 3**	**0.695 1**	**86.74**	86.69	**47.67**	**0.772 3**	0.533 6	**86.27**	**86.28**	**54.67**

注:“a”表示数据来源于已发表的文献,“b”表示根据开源代码获得的结果。

3.6　消融研究

为了深入探究本文模型在情感分析任务中的有效性,本文在 CMU-MOSI 数据集上进行了系列消融实验,设计如下。

1）w/o 自注意力机制:移除自注意力机制模块,使模型无法通过自注意力为每个模态动态分配权重。

2）w/o 文本模态:移除文本模态生成标签和预测标签的损失计算,使模型无法利用文本模态的生成标签来优化目标。

3）w/o 视觉模态:移除视觉模态生成标签和预测标签的损失计算,使模型无法利用视觉模态的生成标签来优化目标。

4）w/o 音频模态:移除音频模态生成标签和预测标签的损失计算,使模型无法利用音频模态的生成标签来优化目标。

5）w/o 文本、视觉和音频模态:移除文本、视觉和音频模态生成标签和预测标签的损失计算,使模型无法利用文本、视觉和音频模态的生成标签来优化目标。

为了全面观察各模块对 SA-CMD 模型的影响,表 4 展示了 SA-CMD 模型与消融后各个模型的对比结果。这些对比结果将有助于详细分析每个模块在整体模型性能中的贡献,从而验证本文提出的方法在多模态情感分析任务中的有效性和关键组件的重要性。

表 4　基于 CMU-MOSI 数据集的消融实验
Table 4　Ablation experiments based on CMU-MOSI dataset

模型	Corr	MAE	Acc-2	*F*1	Acc-7
w/o 自注意力机制	0.799 3	0.704 4	84.91	84.87	48.25
w/o 文本模态	0.798 8	0.689 9	85.98	85.95	**49.42**
w/o 视觉模态	0.800 6	**0.689 3**	85.52	85.42	48.83
w/o 音频模态	**0.802 8**	0.702 9	86.13	86.14	47.81
w/o 文本、视觉和音频模态	0.802 1	0.690 4	85.82	85.77	48.54
本文模型	0.802 0	0.695 1	**86.74**	**86.69**	47.67

从表 4 中可以看出,移除自注意力机制模块后,模型性能显著下降,尤其是在 Acc-2、*F*1 和 MAE 这 3 个评价指标上,表现最为明显。这一结果表明,自注意力机制模块是确保模型达到最优性能的关键,对模型的贡献程度最高。

此外,在移除单一模态生成标签和预测标签损失函数的消融实验中,移除任一模态的损失对模型的影响相对较小。例如,移除文本模态生成标签和预测标签的损失计算后,虽然其他指标均有所下降,但模型的 Acc-7 提升了 1.75。进一步地,移除文本、视觉和音频模态的损失计算后,模型性能显著降低,这也说明了引入中心距差异方法的必要性。

总体而言,无论移除哪一个模块,都使得本文模型的性能变得不稳定。尤其是在移除自注意力机制模块后,这一现象最为明显。这说明自注意力机制模块对模型性能的影响大于其他模块。这些结果验证了本文提出的 SA-CMD 方法的有效性和鲁棒性以及自注意力机制与中心距差异方法结合的必要性。通过这些实验可以清楚地看到各个模块对整体模型性能的不同贡献,从而在未来的工作中更好地理解和优化多模态情感分析模型。

3.7 案例分析

本节从 CMU-MOSI 数据集中随机选取若干原始数据作为输入,进一步验证本文模型的有效性。案例分析结果如表 5 所示。其中,“文本”列表示本文模型对文本模态生成的单标签预测结果,“视觉”列表示对视觉模态生成的单标签预测结果,“音频”列表示对音频模态生成的单标签预测结果,而“多模态预测”则展示了本文模型结合自注意力机制和中心距差异方法的最终预测结果。

表 5 CMU-MOSI 数据集案例分析结果

Table 5 The results of case analysis on CMU-MOSI dataset

例号	多模态信息	实际值	文本	视觉	音频	多模态预测
1	文本:And he was still **boring.** 视觉:轻微叹气 音频:语气平缓	−1.8	−1.737 4	−0.491 0	−0.338 9	−1.804 5
2	文本:“**I liked all** the, um, **I like** the cast. 视觉:表情平缓,摇头 音频:音调略高,有停顿	1.6	1.654 9	−0.459 3	−0.457 4	1.589 60
3	文本:In fact, the first time she goes in, it looks like a Japanese temple. 视觉:皱眉 音频:语速较快	0	0.159 5	−0.455 4	−0.476 8	0.030 70

表 5 中的分析结果表明,本文模型不仅能够在多模态语义一致的情况下准确预测情感,还能在模态间语义不一致时,通过中心距差异方法有效调整预测结果。在例 1 中,文本模态包含明显的消极词汇“boring”,视觉和音频模态也表现出轻微的消极倾向(如轻微叹气和平缓的语气)。在情感语义一致的情况下,自注意力机制能够在模态融合之前为文本模态分配较高权重,从而更好地捕捉消极情感,展现了自注意力机制在模态融合中的关键作用。同时,中心距差异方法通过进一步优化各模态特征表示,确保了在模态间语义一致时对情感信息的准确捕捉。在例 2 和例 3 中,尽管文本模态的情感倾向与实际值接近,但视觉和音频模态中包含了一定的噪声特征。例如,例 2 中的视觉模态表现出表情平缓和摇头动作,音频模态则检测到音调略高且存在停顿;例 3 中的视觉模态表现出皱眉,音频模态中语速较快。在这些语义不一致的情况下,自注意力机制为强特征(文本模态)分配了较高的权重,中心距差异方法则有效抑制了噪声特征,确保了最终情感预测的准确性。这表明本文提出的方法能够有效处理多模态数据中的噪声问题和情感信息不一致问题。

4 结论

本文提出了多模态情感分析模型 SA-CMD,通过引入自注意力机制和中心距差异方法,动态调整各模态特征的权重,优化特征表示和标签分布,从而显著提高多模态情感识别的性能。对比实验结果表明,SA-CMD模型在 CMU-MOSI 和 CMU-MOSEI 数据集上的情感分析任务中表现出优越的性能,尤其在回归任务的评价指标(Corr 和 MAE)以及分类任务的评价指标(Acc-2 和 Acc-7)上显著优于现有的多模态情感分析模型。通过消融实验,进一步验证了自注意力机制和中心距差异方法在模型中的关键作用。去除自注意力机制模块后,模型性能显著下降,特别是在 Acc-2、$F1$ 和 MAE 指标上,表明自注意力机制在捕捉模态间复杂依赖关系方面至关重要。移除单一模态生成标签和预测标签损失函数的消融实验显示,虽然对某些指标的影响较小,但去除所有模态的损失计算后,模型性能整体下降,这进一步验证了引入中心距差异方法的必要性。

总的来说,SA-CMD 模型在多模态情感分析任务中展示了显著的优势和潜力。未来的研究可以继续优化模型结构,并在更多的多模态数据集上进行验证,以进一步提升模型的泛化能力和应用价值。

参考文献:

[1] ZADEH A A B, LIANG P P, PORIA S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: ACL, 2018:2236-2246.

[2] TSAI Y H H, LIANG P P, ZADEH A, et al. Learning factorized multimodal representations[EB/OL]. (2018-06-16)[2024-07-04].https://arxiv.org/abs/1806.06176.

[3] TSAI Y H H, BAI S, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019:6558-6569.

[4] LIU Weide, ZHAN Huijing, CHEN Hao, et al. Multimodal sentiment analysis with missing modality: a knowledge-transfer approach[EB/OL]. (2023-11-28)[2024-07-04]. https://arxiv.org/abs/2401.10747.

[5] 罗渊贻,吴锐,刘家锋,等. 面向情感语义不一致的多模态情感分析方法[J/OL]. 计算机研究与发展.(2024-03-09)[2024-07-04]. http://kns.cnki.net/kcms/detail/11.1777.tp.20240305.1731.006.html.
LUO Yuanyi, WU Rui, LIU Jiafeng, et al. Multimodal sentiment analysis method for sentimental semantic inconsistency[J/OL]. Journal of Computer Research and Development. (2024-03-09)[2024-07-04]. http://kns.cnki.net/kcms/detail/11.1777.tp.20240305.1731.006.html.

[6] BALTRUŠAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: a survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2):423-443.

[7] SUN Z, SARMA P K, SETHARES W, et al. Multi-modal sentiment analysis using deep canonical correlation analysis[EB/OL]. (2019-07-15)[2024-07-04]. https://arxiv.org/abs/1907.08696.

[8] MAI Sijie, HU Haifeng, XING Songlong. Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2020:164-172.

[9] HAZARIKA D, ZIMMERMANN R, PORIA S. Misa: modality-invariant and-specific representations for multimodal sentiment analysis[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020:1122-1131.

[10] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[EB/OL]. (2017-07-23)[2024-07-04]. https://arxiv.org/abs/1707.07250.

[11] LIU Z, SHEN Y, LAKSHMINARASIMHAN V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[EB/OL]. (2018-05-31)[2024-07-04]. https://arxiv.org/abs/1806.00064.

[12] SUN Z, SARMA P, SETHARES W, et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2020:8992-8999.

[13] HAN W, CHEN H, PORIA S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis[EB/OL]. (2019-09-01)[2024-07-04]. https://arxiv.org/abs/2109.00412.

[14] 田昌宁,贺昱政,王笛,等. 基于 Transformer 的多子空间多模态情感分析[J/OL]. 西北大学学报(自然科学版).(2024-04-03)[2024-07-04]. https://xdxbzk.nwu.edu.cn/thesisDetails#10.16152/j.cnki.xdxbzr.2024-02-002&lang=zh.
TIAN Changning, HE Yuzheng, WANG Di, et al. Multi-subspace multimodal sentiment analysis method based on Transformer[J/OL]. Journal of Northwest University (Natural Science Edition). (2024-04-03)[2024-07-04]. https://xdxbzk.nwu.edu.cn/thesisDetails#10.16152/j.cnki.xdxbzr.2024-02-002&lang=zh.

[15] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. (2014-09-03)[2024-07-04]. https://arxiv.org/abs/1409.0473.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30:5998-6008.

[17] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11)[2024-07-04]. https://arxiv.org/abs/1810.04805.

[18] WANG Di, GUO Xutong, TIAN Yumin, et al. TETFN: a text enhanced transformer fusion network for multimodal sentiment analysis[J]. Pattern Recognition, 2023, 136:109259.

[19] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

(下转第 110 页)