跨模态时空注意力与上下文门控的情感分析

李丽红[1,2], 李志勋[1,2], 刘威伟[1,2], 秦肖阳[1,2]

(1.华北理工大学 理学院,河北 唐山 063210;2.华北理工大学 河北省数据科学与应用重点实验室,河北 唐山 063210)

摘　要:多模态情感分析因模态异质导致的交互不一致性、语言场景复杂性及静态跨模态注意力难以捕捉多模态数据时序动态性,限制了深层模态关联挖掘与情感分类性能。针对以上难题,提出一种多模态情感分析框架,引入跨模态时空注意力(CM-STA)捕获文本、图像与音频的时空依赖,增强跨模态交互;上下文门控(CG)动态筛选情感表达强相关的特征,突出关键情感信息;Transformer 跨模态融合交互(TCMFI)通过多头自注意力与双线性池化实现深层跨模态融合,提升融合效率。所提模型在公开数据集 TESS(音频)和 MVSA-Multiple(文本、图像)上的实验准确率为 81.45%、*F*1 分数为 80.84%、*AUROC* 为 96.40%,较最佳基线模型 MISA 分别提升 0.95、0.24 和 7.91 百分点,计算复杂度的实验结果显示,所提模型占用 GPU 内存 7.8 GB,GPU 利用率 98%。所提模型以低空间复杂度和高 GPU 利用率实现高效融合,性能优于对比基线模型。实验结果验证了所提模型在复杂多模态情感分析场景中展现出优异的性能与鲁棒性。

关键词:多模态情感分析; 时空注意力; 上下文门控; Transformer 跨模态融合; 跨模态交互

中图分类号: TP391.1;TN912.3;TP18　　**文献标志码**:A　　**doi**:10.13705/j.issn.1671-6833.2026.04.002

随着科技进步和信息化程度的深化,社交媒体及短视频平台迎来爆发式发展,用户通过发布文本、图像和音频等多模态数据表达情感和观点[1]。情感分析作为人工智能的重要分支,旨在挖掘与理解此类数据中的隐式情感信息。传统情感分析多聚焦于单一模态,采用特征拼接或单一模态注意力机制[2],难以捕捉模态间的高阶依赖和动态时序信息,导致识别性能受限。如何提升多模态情感分析的准确性成为研究热点。近年来,深度学习和 Transformer 架构的快速发展推动了多模态情感分析的进步。Morency 等[3]率先提出多模态情感分析框架,Wang 等[4]提出 MTAMW 模型,通过将多模态序列特征直接拼接,实现多模态情感分析;Liu 等[5]提出 MTMS 模型,通过模态翻译将多模态数据特征对齐后进行情感分析;Kim 等[6]提出的 ViLT 通过视觉-语言 Transformer 架构,实现模态间的无缝交互;Liu 等[7]增强文本信息并融合视频和音频特征,引入注意力机制设计情感分析框架;Dou 等[8]引入 CLIP-ViT 利用对比学习预训练视觉-语言模型实现多模态情感分类任务。此外,Zeng 等[9]通过跨模态数据分布对齐挖掘模态间关系;陈燕等[10]结合 CLIP 和交叉注意力机制提出 CLIP-CA-MSA 模型;Khan 等[11]的 MemoCMT 利用跨模态 Transformer 整合全局和局部特征;Baevski 等[12]的语音模型 data2vec 通过自监督学习提取高质量特征提升单模态特征质量。然而,现有研究在处理模态冲突与语义一致性建模方面易引发语义偏差问题,在模态间动态交互和情感一致性建模方面仍显不足,限制了模型对复杂情感场景的适应能力。

针对上述挑战,本文提出一种多模态情感分析框架,通过跨模态时空注意力、上下文门控和 Transformer 跨模态融合交互,深度捕获模态间的动态交互,优化时空序列特征的自适应提取,实现局部与全局特征的高效整合。相较于现有方法,本文框架更注重模态间动态交互和情感一致性的协同建模,能够有效应对多模态复杂情感表达需求,主要贡献如下。

(1)提出跨模态时空注意力(cross-modal spatio-temporal attention,CM-STA)模块,深度挖掘文本、图像和音频模态之间的空间与时间依赖关系,增强

收稿日期:2025-09-01;**修订日期**:2025-10-03

基金项目:河北省数据科学与应用重点实验室项目(10120201)

作者简介:李丽红(1979—),女,辽宁锦州人,华北理工大学教授,主要从事数据挖掘和三支决策研究,E-mail:22687426@qq.com。

不同模态间交互,提升模型对数据异质性的适应能力。

(2)引入上下文门控(contextual gating,CG)模块,通过深度注意力机制自适应调整时间步特征权重,突出关键情感信息,提升动态情感建模能力和模型对复杂情感场景的建模精度。

(3)设计 Transformer 跨模态融合交互(Transformer cross-modal fusion interaction,TCMFI)模块,基于多层 Transformer 和双线性交互策略,整合 CM-STA 和 CG 的优化特征,实现高效跨模态融合并协同构建全局与局部信息,增强情感表征能力。

1 模型描述

本文提出的多模态情感分析框架由 3 个关键模块组成:跨模态时空注意力模块、上下文门控模块与 Transformer 跨模态融合交互模块在信息流上层层递进、协同优化,构建了一个从特征提取、交互建模到情感分类的完整流程,模型框架如图 1 所示。

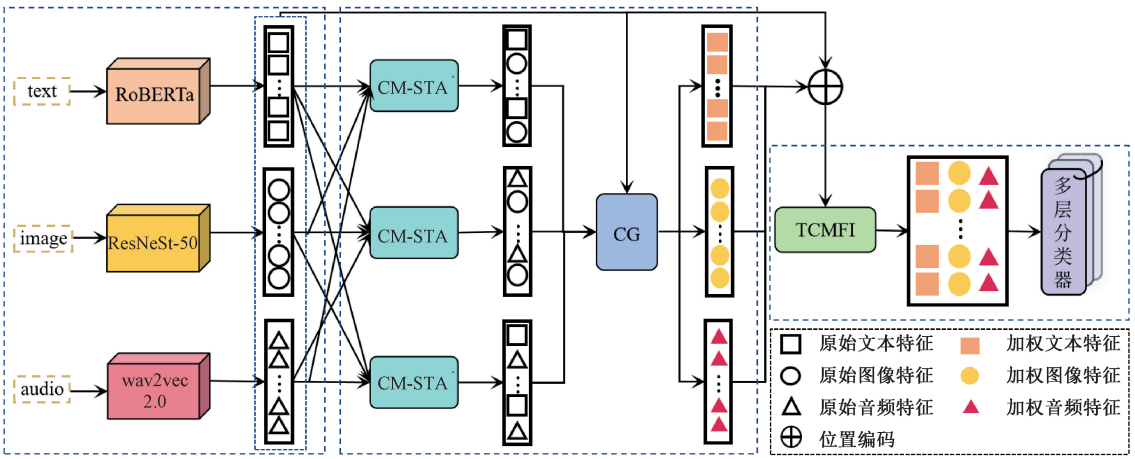

图 1 多模态情感分析模型框架

Figure 1 Multimodal sentiment analysis model framework

1.1 特征提取与投影

为实现多模态数据的有效融合,需首先提取各模态特征并投影至统一特征空间。对于文本模态,引入 RoBERTa-wwm-ext 预训练语言模型[13]提取文本特征,保留最后一层隐藏状态表示上下文信息;对于图像模态,采用基于 ResNet 改进的 ResNeSt-50[14]网络,提升对空间和多尺度特征的建模能力,保留最后一层特征向量作为图像特征;对于音频模态,使用预训练语音模型 wav2vec 2.0[15]捕捉时间动态特征,选取最后一层隐藏状态作为音频特征,保留完整的序列动态信息。

1.2 跨模态时空注意力

基于 STCA 注意力机制[16],本文提出了一种跨模态注意力机制,用于更有效地建模时空依赖关系。对于图像模态,利用空间注意力捕捉像素间的空间依赖;对于文本和音频模态,结合时间注意力建模序列的动态变化,增强模态间的关联。支持序列输入并通过多层堆叠结构和残差连接实现模态间的深度交互。跨模态时空注意力模块架构如图 2 所示,包含以下 3 个部分。

(1)多头注意力计算。CM-STA 模块以统一维度后的模态特征 $\boldsymbol{F}_{\mathrm{t}}$、$\boldsymbol{F}_{\mathrm{i}}$、$\boldsymbol{F}_{\mathrm{a}}$(文本、图像、音频)作为

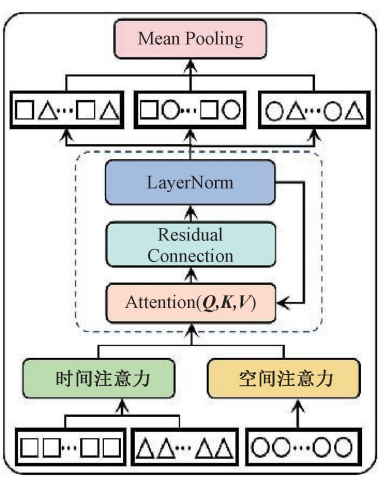

图 2 跨模态时空注意力

Figure 2 Cross-modal spatio-temporal attention

输入。以文本-图像交互为例,通过多头注意力机制计算不同模态之间的交互特征,将文本特征作为查询 $\boldsymbol{Q}=\boldsymbol{F}_{\mathrm{t}}\boldsymbol{W}_{Q}$,图像特征作为键 $\boldsymbol{K}=\boldsymbol{F}_{\mathrm{i}}\boldsymbol{W}_{K}$ 和值 $\boldsymbol{V}=\boldsymbol{F}_{\mathrm{i}}\boldsymbol{W}_{V}$。其中,$\boldsymbol{W}_{Q}$、$\boldsymbol{W}_{K}$、$\boldsymbol{W}_{V}$ 均为可学习的参数矩阵,计算注意力分数以生成交互特征:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}。\quad(1)$$

式中:$\sqrt{d_{\mathrm{k}}}$ 为缩放因子;d_{k} 为每个注意力头中键向

量的维度。

(2)深层堆叠与残差连接。使用多层堆叠逐步增强交互特征,结合残差连接保留原始特征信息:

$$\boldsymbol{F}_{\mathrm{t}}^{(i+1)} = \boldsymbol{F}_{\mathrm{t}}^{(i)} + \alpha_i \cdot \mathrm{Attention}(\boldsymbol{F}_{\mathrm{t}}^{(i)}, \boldsymbol{F}_{\mathrm{i}}, \boldsymbol{F}_{\mathrm{a}})。 \tag{2}$$

式中:i 为层($i=0,1$);α_i 为可学习的残差权重。初始时目标模态特征 $\boldsymbol{F}_{\mathrm{t}}^{(0)} = \boldsymbol{F}_{\mathrm{t}}$。为保证特征分布稳定,加入层归一化对特征进行处理:

$$\boldsymbol{F}_{\mathrm{t}}^{(i+1)} = \mathrm{LayerNorm}(\boldsymbol{F}_{\mathrm{t}}^{(i+1)})。 \tag{3}$$

经过 *num_layers* 次迭代后,得到文本-图像交互特征 $\boldsymbol{F}_{\mathrm{ti}}$:

$$\boldsymbol{F}_{\mathrm{ti}} = \boldsymbol{F}_{\mathrm{t}}^{(num\_layers)} \in \mathbf{R}^{B\times L_{\max\_\mathrm{text}}\times D}。 \tag{4}$$

式中:B 为批量大小;$L_{\max\_\mathrm{text}}$ 为文本序列的最大长度;D 为线性变换维度。

(3)跨模态交互特征计算与池化处理:得出 3 对模态的交互特征后,采用平均池化将序列特征压缩为单一向量:

$$\boldsymbol{F}'_{\mathrm{ti}}, \boldsymbol{F}'_{\mathrm{ta}}, \boldsymbol{F}'_{\mathrm{ai}} = \mathrm{Mean}(\boldsymbol{F}_{\mathrm{ti}}, \boldsymbol{F}_{\mathrm{ta}}, \boldsymbol{F}_{\mathrm{ai}}, dim = 1)。 \tag{5}$$

池化后的特征为后续 CG 和 TCMFI 模块提供优化的交互表示,提升多模态情感分类的精度。

1.3 上下文门控

CG 模块融合 Transformer 注意力机制[17]、动态融合策略[18]和模态权重调节[19]的思想并在其基础上进行改进。传统多模态方法易忽略模态间的动态交互和语义依赖关系,同时容易引入噪声且难以筛选情感相关特征,而 CG 通过上下文感知的门控机制动态调整特征权重强化多模态协同作用的建模能力。上下门控如图 3 所示。

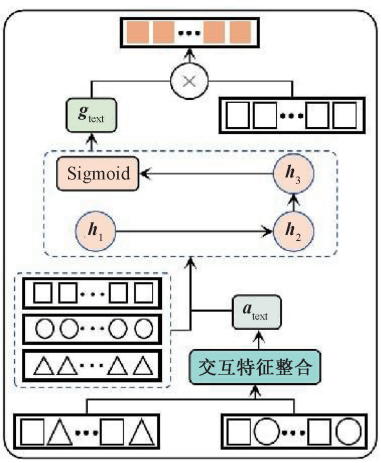

图 3 上下文门控

Figure 3 Contextual gating

CG 模块以模态原始特征和 CM-STA 模块生成的交互特征作为输入,为给每个模态生成综合的交互信息[17],融合不同模态的交互信息。以文本-图像和文本-音频为例:

$$\boldsymbol{a}_{\mathrm{text}} = \boldsymbol{F}'_{\mathrm{ti}} + \boldsymbol{F}'_{\mathrm{ta}} \in \mathbf{R}^{B\times D}。 \tag{6}$$

式中:$\boldsymbol{a}_{\mathrm{text}}$ 表示与文本模态相关的综合交互特征,结合了文本-图像与文本-音频的交互信息。随后通过多层感知机(MLP)生成非线性门控信号[18],动态调整原始特征权重。门控信号计算过程如下:

$$\boldsymbol{h}_1 = \mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{a}_{\mathrm{text}} + \boldsymbol{b}_1); \tag{7}$$

$$\boldsymbol{h}_2 = \mathrm{ReLU}(\boldsymbol{W}_2\boldsymbol{h}_1 + \boldsymbol{b}_2); \tag{8}$$

$$\boldsymbol{h}_3 = \mathrm{ReLU}(\boldsymbol{W}_3\boldsymbol{h}_2 + \boldsymbol{b}_3); \tag{9}$$

$$\boldsymbol{g}_{\mathrm{text}} = \mathrm{Sigmoid}(\boldsymbol{W}_4\boldsymbol{h}_3 + \boldsymbol{b}_4)。 \tag{10}$$

式中:$\boldsymbol{W}_1 \sim \boldsymbol{W}_4$ 均为可学习的权重矩阵;$\boldsymbol{b}_1 \sim \boldsymbol{b}_4$ 均为偏置项;ReLU 和 Sigmoid 分别为中间层和输出层的激活函数;$\boldsymbol{g}_{\mathrm{text}}$ 为输出门控信号,利用门控信号对原始特征加权调整[19],突出情感相关特征,抑制无关噪声。加权文本特征计算公式为

$$\boldsymbol{F}_{\mathrm{t}}^{\mathrm{gated}} = \boldsymbol{g}_{\mathrm{text}} \cdot \boldsymbol{F}'_{\mathrm{t}} \in \mathbf{R}^{B\times D}。 \tag{11}$$

式中:$\cdot$ 表示逐元素乘积;$\boldsymbol{F}_{\mathrm{t}}^{\mathrm{gated}}$ 为加权文本特征。

1.4 Transformer 跨模态融合交互

基于 Transformer 架构[20]、跨模态学习[21]和动态融合策略[22]的思想设计,TCMFI 接收 CG 模块生成的加权特征与原始多模态特征作为输入,添加可训练的位置编码以区分模态。包含 6 层 Transformer 编码器,每层包括多头自注意力、前馈网络、两次残差连接和层归一化。Transformer 跨模态融合交互如图 4 所示。

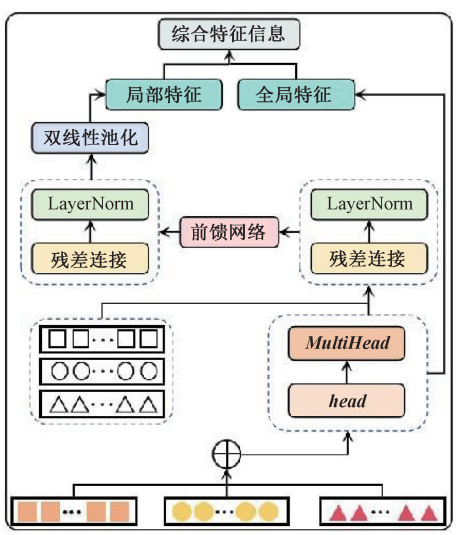

图 4 Transformer 跨模态融合交互

Figure 4 Transformer cross-modal fusion interaction

(1)输入特征与位置编码[20]:

$$\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{P}。 \tag{12}$$

式中:$\boldsymbol{X}$ 为 CG 模块生成的加权特征和原始模态特征;$\boldsymbol{P}$ 为可训练的位置编码;$\boldsymbol{X}'$ 为添加位置编码的输入加权和原始模态特征。

(2)多头自注意力[20]:

$$\boldsymbol{head}_i = \mathrm{Softmax}\left(\frac{(\boldsymbol{QW}_i^Q)(\boldsymbol{KW}_i^K)^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{VW}_i^V; \quad (13)$$

$$\boldsymbol{MultiHead} = \mathrm{Concat}(\boldsymbol{head}_1,\cdots,\boldsymbol{head}_i)\boldsymbol{W}^{\mathrm{o}}。 \quad (14)$$

式中:$\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$、$\boldsymbol{W}_i^V$ 均为第 i 个头的权重矩阵;$\boldsymbol{head}_i$ 为第 i 个头的输出;$\boldsymbol{W}^{\mathrm{o}}$ 为输出权重矩阵;$\boldsymbol{MultiHead}$ 为最终输出,捕捉全局依赖关系。

(3)层归一化[20]:

$$\mathrm{LayerNorm}(\boldsymbol{x}) = \boldsymbol{\gamma}\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2+\boldsymbol{\varepsilon}}} + \boldsymbol{\beta}。 \quad (15)$$

式中:$\boldsymbol{x}$ 为层归一化的输入;μ、σ^2 分别为输入的均值和方差;$\boldsymbol{\varepsilon}$ 为数值稳定性常数;$\boldsymbol{\gamma}$、$\boldsymbol{\beta}$ 均为可学习的缩放和平移参数。

(4)第一次残差连接和层归一化[20-21]:

$$\boldsymbol{O}_1 = \mathrm{LayerNorm}(\boldsymbol{X}' + \boldsymbol{MultiHead})。 \quad (16)$$

式中:$\boldsymbol{O}_1$ 为残差连接与层归一化生成的初始特征。

(5)前馈网络[20-21]:

$$\mathrm{FFN}(\boldsymbol{O}_1) = \mathrm{ReLU}(\boldsymbol{O}_1\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2。 \quad (17)$$

式中:$\boldsymbol{W}_1$、$\boldsymbol{b}_1$ 为第一层全连接参数;ReLU 为激活函数;$\boldsymbol{W}_2$、$\boldsymbol{b}_2$ 为第二层全连接参数。

(6)第二次残差连接和层归一化[20-21]:

$$\boldsymbol{O}_2 = \mathrm{LayerNorm}(\boldsymbol{O}_1 + \mathrm{FFN}(\boldsymbol{O}_1))。 \quad (18)$$

第二次残差连接和层归一化进一步稳定训练,生成中间特征 $\boldsymbol{O}_2$。

(7)双线性池化[22]:

$$\begin{cases}\boldsymbol{B}_{TI} = (\boldsymbol{TW}_b\boldsymbol{I}^{\mathrm{T}});\\ \boldsymbol{B}_{TA} = (\boldsymbol{TW}_b\boldsymbol{A}^{\mathrm{T}});\\ \boldsymbol{B}_{IA} = (\boldsymbol{IW}_b\boldsymbol{A}^{\mathrm{T}})。\end{cases} \quad (19)$$

式中:$\boldsymbol{T}$、$\boldsymbol{I}$、$\boldsymbol{A}$ 分别表示第二次残差连接和层归一化后得到的文本、图像和音频特征;$\boldsymbol{W}_b$ 为可学习权重矩阵;$\boldsymbol{B}_{TI}$,$\boldsymbol{B}_{TA}$,$\boldsymbol{B}_{IA}$ 分别为计算得到的双线性交互特征,表示模态间的局部信息。双线性特征降维至特定维度后,与 Transformer 得到的全局特征通过可学习权重加权融合,生成综合特征表示,经多层分类器生成 logits 完成情感分类。TCMFI 融合全局与局部特征,提升情感分类准确性。

2 实验结果与分析

2.1 数据集

实验选择公开数据集 TESS[23] 与 MVSA-Multiple[24] 评估所提多模态情感分析模型的性能。TESS 数据集为音频数据集,包含 4 240 个样本。MVSA-Multiple 是一个多模态数据集,包含 19 600 条文本-图像对数据,涵盖正向、中性、负向 3 种情感类型,由 3 位独立标注者标注,增加标注结果的可靠性。为提升数据质量,本文对获取得到的公开数据集样本进行预处理操作,最终得到 4 151 条音频数据和 17 507 条文本-图像对数据。本文对 TESS 数据集的 7 种情感类别进行了系统性重新划分,将其映射至正向、中性、负向 3 类情感类型。随后,将两种数据集均按照 8∶1∶1 的比例划分为训练集、验证集与测试集,所得数据集样本划分信息如表 1 所示。

表 1 数据集样本划分信息

Table 1 Dataset sample division information

数据集	训练数量	验证数量	测试数量	总计
TESS[23]	3 321	415	415	4 151
MVSA-Multiple[24]	14 005	1 751	1 751	17 507

2.2 实验配置

模型训练环境为 CUDA 12.4,GPU 使用 RTX 4070S(显存 12 GB),CPU 为 i5-14600KF。优化器采用 AdamW,所有模块的初始学习率为 0.002,引入 0.003 权重衰减,使用学习率调度器动态调整学习率,并加入混合精度训练应对多模态情感分析任务中复杂的时间复杂度问题。Batch 设为 16,Dropout 设置为 0.2。为确保实验的可重复性,本文在数据划分、样本采样、参数初始化及数据增强等关键环节统一设置 42 个随机种子。加入早停机制,当验证集损失连续 5 个 epoch 未下降时停止训练。为提高实验稳定性,实验在 5 个不同随机种子下分别训练,并对性能指标取平均值,以降低随机初始化的影响。

性能评估采用准确率、精确率、召回率、$F1$ 分数和 $AUROC$。准确率(Acc)衡量模型总体预测正确的比例;召回率($Recall$)衡量模型正确识别正类样本的比例;精确率(Pre)衡量模型预测为正类的样本中真正为正类的比例;$F1$ 分数为精确率和召回率的调和平均值,综合评估模型在正类和负类上的表现。计算公式如下所示:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}; \quad (20)$$

$$Recall = \frac{TP}{TP+FN}; \quad (21)$$

$$Pre = \frac{TP}{TP+FP}; \quad (22)$$

$$F1 = \frac{2Precision \times Recall}{Precision + Recall}。 \quad (23)$$

$AUROC$ 评估模型对模态不一致性的鲁棒性,可通过对 ROC 曲线下各部分的面积求和而得。ROC 曲线以假阳性率 FPR 为横轴,真阳性率 TPR 为

纵轴:

$$\begin{cases} FPR = \dfrac{FP}{FP + TN}; \\ TPR = \dfrac{TP}{TP + FN}。 \end{cases} \tag{24}$$

2.3 对比实验

为充分验证提出模型的有效性,选取了多种多模态情感分析领域的基准模型进行对比分析。ALMT[25]通过不同尺度提取语言特征,利用 Transformer 融合多模态特征。MFN[26]利用记忆网络捕获模态间的时序交互,通过记忆单元存储和更新数据特征。MulT[27]基于 Transformer 架构,通过跨模态注意力机制建模文本、图像和音频模态的交互。TETFN[28]基于文本的成对融合模型,通过张量融合网络建模文本与其他模态的交互。TFN[29]通过张量外积融合文本、图像和音频模态特征,生成高维张量捕捉模态间交互关系。LMF[30]通过低秩分解优化 TFN 的张量融合,降低计算复杂度,同时保留模态间交互信息 MMGCN[31]基于图神经网络,将模态特征构建为图结构,捕捉模态间依赖关系。DialogueGCN[32]结合多模态和多话语信息,利用跨模态注意力处理模态间交互,增强情感预测。MISA[33]通过模态不变和模态特定表示学习,结合注意力机制融合特征来提升准确性。EMO-GCN[34]基于图神经网络,通过构建模态间图结构,捕捉模态内和模态间的潜在关系。

表 2 展示了各模型在多模态数据集上的实验结果。结果表明,本文提出的多模态融合模型在 MV-SA-Multiple 和 TESS 数据集上的分类性能优于基线模型。相较于最佳基线 MISA,本文模型的 *Acc*、*F*1 和 *AUROC* 分别提升 0.95 百分点、0.24 百分点和 7.91 百分点;但 *Pre* 略低于 MISA,是因为图卷积网络和模态不变表示更擅长概率分布校准,而本文模型则聚焦于分类性能。本文模型以低空间复杂度和高 GPU 利用率实现高效融合,优于主流模型,在资源受限和高精度需求场景中展现出应用价值。

2.4 消融实验

为探究 CM-STA、CG 和 TCMFI 模块对模型性能的具体影响,开展了消融实验。实验结果如表 3 所示。

实验结果表明,移除 CM-STA 模块后,模型的 *Acc* 从 81.45%降至 77.67%,*F*1 分数从 80.84%降至 77.84%,*AUROC* 从 96.40%降至 89.21%。此时,模态间的时空交互信息未能被有效提取和利用,削弱了模型对跨模态特征的表示能力,影响了后续融合和分类性能。实验结果验证了 CM-STA 模块在跨模态交互建模能力方面的关键作用。

表 2 本文模型较基准模型对比实验结果

Table 2 Experimental results comparing the proposed model to baseline models

模型	*Acc*/%	*Recall*/%	*F*1/%	*AUROC*/%	*Pre*/%	占用 GPU 内存/GB	GPU 利用率/%	单次训练花费时间/s
EMO-GCN[34]	63.50	63.50	63.26	75.31	63.02	11.50	91.00	48.00
MFN[26]	65.50	65.50	63.31	65.54	61.26	11.50	85.00	45.00
MMGCN[31]	67.50	67.50	67.05	83.44	66.61	11.50	86.00	45.00
LMF[30]	69.50	69.50	68.67	89.57	78.01	11.60	86.00	46.00
ALMT[25]	70.00	70.00	65.57	87.17	77.57	11.50	90.00	47.00
TETFN[28]	73.00	73.00	72.79	87.67	78.13	11.60	83.00	51.00
MulT[27]	77.00	77.00	75.74	82.65	74.52	11.60	83.00	45.00
TFN[29]	77.50	77.50	76.07	82.79	74.69	9.90	78.00	55.00
DialogueGCNv[32]	79.50	79.50	79.61	87.99	79.72	11.60	89.00	46.00
MISA[33]	80.50	80.50	80.60	88.49	**80.70**	11.60	85.00	**44.00**
本文模型	**81.45**	**81.45**	**80.84**	**96.40**	80.23	**7.800**	**98.00**	73.00

表 3 模型消融实验结果

Table 3 Model ablation experiment results

CM-STA	CG	TCMFI	*Acc*/%	*Recall*/%	*F*1/%	*AUROC*/%	*Pre*/%
	√	√	77.67	77.67	77.84	89.21	79.89
√		√	73.00	73.00	73.01	80.37	73.00
√	√		72.00	72.00	72.23	83.43	73.71
		√	59.00	59.00	43.79	48.53	34.81
√	√	√	**81.45**	**81.45**	**80.84**	**96.40**	**80.23**

移除 CG 模块后,模型的 *Acc* 从 81.45%降至 73.00%,*F*1 分数从 80.84%降至 73.01%,*AUROC* 从 96.40%降至 80.37%。此时,模型无法通过门控机制动态调整各模态特征的权重,导致不同模态的信息未能根据其重要性进行有效加权和优化,从而降低了特征表示的质量。实验结果验证了 CG 模块通过动态调整模态权重显著提升多模态融合能力的有效作用。

移除 TCMFI 模块后,模型的 *Acc* 从 81.45%降至 72.00%,*F*1 分数从 80.84%降至 72.23%,*AUROC* 从 96.40%降至 83.43%。此时,模型失去了基于 Transformer 编码器和双线性交互的深层跨模态融合能力,无法充分挖掘和整合模态间的交互信息,削弱了多模态特征的融合效果和综合建模能力,进而影响分类性能。实验结果验证了 TCMFI 模块在提升模型性能中的重要性。

同时移除 CM-STA 和 CG 模块后,模型的 *Acc* 从 81.45%降至 59.00%,*F*1 分数从 80.84%降至 43.79%,*AUROC* 从 96.40%降至 48.53%,分类性能显著下降。此时,模型无法有效捕捉模态间的时空交互信息并根据模态重要性动态调整特征权重。CM-STA 模块通过模态间的时空交互计算生成交互特征;CG 模块通过门控机制对模态特征进行动态特征加权调整,二者通过模态间交互和动态特征加权协同增强模型的多模态特征表示。结果验证了 CM-STA 和 CG 模块在模态间交互和特征优化中的作用。

3 模型分析与评估

3.1 模型效率分析

实验统计了 5 个指标评估模型效率。其中模型参数总量为 2 497 666,参数规模较大,反映了多模态融合与时空建模所带来的较高计算复杂性。训练与推理的平均内存使用大小分别为 3 535.55 MB 与 3 530.91 MB,内存需求较为稳定。

实验结果显示,训练时间误差约为 2~3 s,反映了训练过程的时间性能稳定性;约 95%的测试集样本推理时间小于 0.004 s,表明模型具备较高的实时处理能力;训练与推理吞吐量范围分别为每秒 0.27%~0.32%与每秒 0.28%~0.33%,二者效率接近,验证了模型在训练与推理阶段的优化一致性。

3.2 超参数敏感性分析

本文对比了不同学习率 *LR*(0.003 0,0.001 0,0.000 5,0.000 3)对模型性能的影响,结果如图 5 所示。实验结果表明,不同学习率下,通过合理调整训练轮次,各性能指标均呈现收敛趋势,表明所提模型对学习率参数的超参数敏感性较低,具有较好的稳定性。

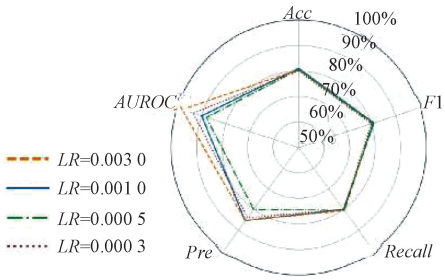

图 5 学习率对模型性能影响

Figure 5 Impact of learning rate on model performance

本文设计了不同模块个数配置的对比实验,结果如图 6 所示。实验结果表明,随着模块个数的增加,模型性能未呈现显著提升,较少模块个数与注意力头的配置更具效率优势,能够在维持性能稳定性的同时优化计算资源利用率。

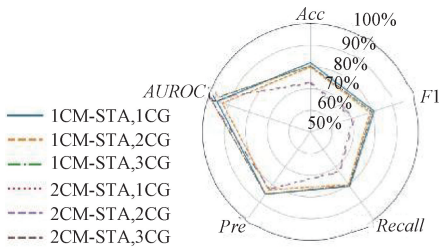

图 6 模块个数对模型性能的影响

Figure 6 Impact of module number on model performance

4 结论

本文提出了多模态情感分析模型。其中 CM-STA 模块通过跨模态注意力机制增强模态间的时空交互计算,提升模态相关性表示;CG 模块通过门控机制动态加权调整各模态特征,优化特征表示并突出重要特征信息;TCMFI 模块利用 Transformer 编码器和双线性交互机制进行深层跨模态融合,生成综合特征表示。通过在公开数据集上的实验结果显示,模型的 *Acc* 为 81.45%,*F*1 分数为 80.84%,*AUROC* 为 96.40%,占用 GPU 内存 7.8 GB,GPU 利用率达 98%,优于基线对比模型,验证了本文所提模型在复杂多模态情感分析场景中的优异性能和鲁棒性。该模型通过多层次模态交互和特征优化有效提升分类性能,在多模态情感分析任务中表现出色,为多模态情感分析提供了新的解决方案。

未来研究可以探索更高效的模态交互机制,以增强模态一致性表示能力;引入先进模态融合和优化策略,以进一步提升模型性能。

参考文献:

[1] CHANDRASEKARAN G, NGUYEN T N, HEMANTH D J. Multimodal sentimental analysis for social media applications: a comprehensive review[J]. WIREs Data Mining and Knowledge Discovery, 2021, 11(5): e1415.

[2] 吕学强, 田驰, 张乐, 等. 融合多特征和注意力机制的多模态情感分析模型[J]. 数据分析与知识发现, 2024, 8(5): 91-101.
LYU X Q, TIAN C, ZHANG L, et al. Multimodal sentiment analysis model integrating multi-features and attention mechanism[J]. Data Analysis and Knowledge Discovery, 2024, 8(5): 91-101.

[3] MORENCY L P, MIHALCEA R, DOSHI P. Towards multimodal sentiment analysis: harvesting opinions from the web[C]// The 13th International Conference on Multimodal Interfaces. New York: ACM, 2011: 169-176.

[4] WANG Y F, HE J H, WANG D, et al. Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis[J]. Neurocomputing, 2024, 572: 127181.

[5] LIU Z Z, ZHOU B, CHU D H, et al. Modality translation-based multimodal sentiment analysis under uncertain missing modalities [J]. Information Fusion, 2024, 101: 101973.

[6] KIM W, SON B, KIM I. ViLT: vision-and-language transformer without convolution or region supervision[EB/OL]. (2021-02-05)[2025-08-09]. https://arxiv.org/abs/2102.03334v2.

[7] LIU Z J, CAI L, YANG W J, et al. Sentiment analysis based on text information enhancement and multimodal feature fusion[J]. Pattern Recognition, 2024, 156: 110847.

[8] DOU Z Y, XU Y C, GAN Z, et al. An empirical study of training end-to-end vision-and-language transformers [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 18145-18155.

[9] ZENG Y, YAN W J, MAI S J, et al. Disentanglement Translation Network for multimodal sentiment analysis [J]. Information Fusion, 2024, 102: 102031.

[10] 陈燕, 赖宇斌, 肖澳, 等. 基于 CLIP 和交叉注意力的多模态情感分析模型[J]. 郑州大学学报(工学版), 2024, 45(2): 42-50.
CHEN Y, LAI Y B, XIAO A, et al. Multimodal sentiment analysis model based on CLIP and cross-attention [J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(2): 42-50.

[11] KHAN M, TRAN P N, PHAM N T, et al. MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion [J]. Scientific Reports, 2025, 15: 5473.

[12] BAEVSKI A, HSU W N, XU Q T, et al. data2vec: a general framework for self-supervised learning in speech, vision and language[EB/OL]. (2022-02-07)[2025-08-09]. https://arxiv.org/abs/2202.03555.

[13] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[14] ZHANG H, WU C R, ZHANG Z Y, et al. ResNeSt: split-attention networks[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2022: 2735-2745.

[15] BAEVSKI A, ZHOU H, MOHAMED A, et al. wav2vec 2.0: a framework for self-supervised learning of speech representations[EB/OL]. (2020-06-20)[2025-08-09]. https://arxiv.org/abs/2006.11477.

[16] BHUIYAN A, HUANG J X. STCA: Utilizing a spatio-temporal cross-attention network for enhancing video person re-identification[J]. Image and Vision Computing, 2022, 123: 104474.

[17] BAGHER ZADEH A, LIANG P P, PORIA S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph[C]// The 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 2236-2246.

[18] SUN H, LIU J Q, CHEN Y W, et al. Modality-invariant temporal representation learning for multimodal sentiment classification[J]. Information Fusion, 2023, 91: 504-514.

[19] GOLAGANA V, ROW S V, RAO P S. Adaptive multimodal sentiment analysis: improving fusion accuracy with dynamic attention for missing modality[J]. Journal of Electrical Systems, 2024, 20(S1): 134-147.

[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-06-12)[2025-08-09]. https://arxiv.org/abs/1706.03762.

[21] XIAO L W, WU X J, YANG S W, et al. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis[J]. Information Processing & Management, 2023, 60(6): 103508.

[22] ZHAO F, ZHANG C C, GENG B C. Deep multimodal data fusion [J]. ACM Computing Surveys, 2024, 56 (9): 1-36.

[23] LOK E J. Toronto emotional speech set (TESS) [DB/OL]. [2025-08-09]. https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess.

[24] DIEM L, ZAHARIEVA M. Video content representation using recurring regions detection[J]. Lecture Notes in

Computer Science, 2016, 9516: 16-28.
[25] ZHANG H Y, WANG Y, YIN G H, et al. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis[EB/OL]. (2023-10-09)[2025-08-09]. https://arxiv.org/abs/2310.05804.
[26] ZADEH A, LIANG P P, MAZUMDER N, et al. Memory fusion network for multi-view sequential learning[EB/OL]. (2018-02-03)[2025-08-09]. https://arxiv.org/abs/1802.00927.
[27] TSAI Y H, BAI S J, PU LIANG P, et al. Multimodal transformer for unaligned multimodal language sequences[C]// Proc Conf Assoc Comput Linguist Meet. Stroudsburg: ACL, 2019: 6558-6569.
[28] WANG D, GUO X T, TIAN Y M, et al. TETFN: a text enhanced transformer fusion network for multimodal sentiment analysis [J]. Pattern Recognition, 2023, 136: 109259.
[29] ZADEH A, CHEN M H, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis [EB/OL]. (2017-07-23)[2025-08-09]. https://arxiv.org/abs/1707.07250.
[30] LIU Z, SHEN Y, LAKSHMINARASIMHAN V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[EB/OL]. (2018-05-31)[2025-07-09]. https://arxiv.org/abs/1806.00064.
[31] HU J W, LIU Y C, ZHAO J M, et al. MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation[EB/OL]. (2021-07-14)[2025-08-09]. https://arxiv.org/abs/2107.06779.
[32] GHOSAL D, MAJUMDER N, PORIA S, et al. DialogueGCN: a graph convolutional neural network for emotion recognition in conversation[EB/OL]. (2019-08-30)[2025-08-09]. https://arxiv.org/abs/1908.11540.
[33] 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. New York: ACM, 2020: 1122-1131.
[34] XING T, DOU Y T, CHEN X L, et al. An adaptive multi-graph neural network with multimodal feature fusion learning for MDD detection [J]. Scientific Reports, 2024, 14: 28400.

Sentiment Analysis with Cross-modal Spatio-Temporal Attention and Contextual Gating

LI Lihong[1,2], LI Zhixun[1,2], LIU Weiwei[1,2], QIN Xiaoyang[1,2]

(1. Department of Science, North China University of Science and Technology, Tangshan 063210, China; 2. Hebei Province Key Laboratory of Data Science and Application, North China University of Science and Technology, Tangshan 063210, China)

Abstract: Multimodal sentiment analysis is hindered to capture the temporal dynamics of multimodal data by interaction inconsistencies due to modality heterogeneity, the complexity of linguistic scenarios, and the inability of static cross-modal attention, limiting deep modality correlation mining and sentiment classification performance. To address these challenges, a multimodal sentiment analysis framework was proposed incorporating cross-modal spatio-temporal attention (CM-STA) to capture spatio-temporal dependencies among text, image, and audio, enhancing cross-modal interactions. Contextual gating (CG) dynamically filtered features strongly correlated with emotional expressions, emphasizing key sentiment information. A Transformer cross-modal fusion interaction (TCMFI) leveraged multi-head self-attention and bilinear pooling for efficient deep cross-modal fusion. Experiments on the TESS (audio) and MVSA-Multiple (text, image) datasets yielded an accuracy of 81.45%, an $F1$ score of 80.84%, and an *AUROC* of 96.40%, outperforming the best baseline model MISA by 0.95 percentage points, 0.24 percentage points, and 7.91 percentage points, respectively. Computational complexity analysis revealed that the proposed model occupied 7.8 GB of GPU memory with a 98% GPU utilization rate, achieving efficient fusion with low spatial complexity and high GPU utilization, surpassing baseline models in performance. These results demonstrated the superior performance and robust effectiveness of the proposed model in complex multimodal sentiment analysis scenarios.

Keywords: multimodal sentiment analysis; spatio-temporal attention; contextual gating; Transformer cross-modal fusion; cross-modal interaction