ISSN：3105-0085
EISSN：3105-0093

基于 Transformer 的多模态情感分析模型优化

卫晨熙

重庆移通学院，重庆 401429

摘　要：针对多模态情感分析中单一模态信息不完整跨模态语义偏移和特征冗余等问题，构建了基于 Transformer 的多模态情感分析模型优化框架。该框架以文本图像和音频为输入，利用多头自注意力机制提取模态内关键表征，通过跨模态交互编码实现语义对齐再结合门控融合与残差约束增强有效信息传递，为提升模型稳定性与判别能力，引入特征归一化动态权重调节和分类边界优化策略，减少噪声干扰并强化情感极性刻画。实验分析表明优化后的模型在情感识别准确性鲁棒性和泛化能力方面均表现更优，能够更稳定地捕捉复杂场景下的情感表达特征，为多模态情感理解任务提供了可行的技术支持。

关键词：Transformer；多模态情感分析；模型优化；特征融合；跨模态对齐

DOI：10.64649/yh.shygl.issn3105-0085.202603028

0 引言

多模态情感分析融合文本图像与音频等异构数据判断情感倾向，相较传统单模态方法可补充语气与视觉线索，但存在模态尺度差异与噪声传播等问题 Transformer 凭借自注意力机制可捕捉长距离依赖和动态分配特征权重，适配复杂跨模态关联建模，是该领域核心基础。通过优化特征对齐融合增强与判别策略缩小模态语义鸿沟，结合编码结构融合机制与优化目标协同改进，既能强化隐性情感挖掘能力又可提升模型抗干扰与模态缺失适配性，为高精度情感识别筑牢技术支撑。

1 基于 Transformer 的多模态情感表征建模

1.1 多模态输入特征提取与编码

多模态情感表征建模需要构建统一特征空间不同模态有着差异化处理方式，文本经分词嵌入并叠加位置编码留存语序，图像提取区域特征转为视觉序列，音频分割梅尔频谱生成时序向量。各类原始数据结构差异显著直接交互易出现尺度冲突与语义错位，文本依靠预训练嵌入挖掘语言情感逻辑，图像依托卷积网络捕捉表情与色调等关键视觉信息[1]，音频通过时序向量保留语调和音量等核心情感特征，统一编码空间是提升分析模型效果的核心前提。设第 m 个模态输入为：$X^{(m)} \in \mathbb{R}^{n_m \times d}$

其中 n_m 表示该模态的序列长度，d 为特征维度，编码后的表示记为 $H^{(m)}$。Transformer 编码器通过自注意力机制计算模态内部相关性，其核心形式为：$\mathrm{Attn}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V$

其中 Q，K，V 分别表示查询键和值矩阵，d_k 为键向量维度。该结构使模型能够自动强化与情感表达相关的关键词关键区域和关键频段，同时削弱冗余背景信息因而提升多模态输入的可分性。为避免不同模态在尺度和分布上的偏移，编码阶段加入层归一化与线性投影，使三类特征进入同一表示空间为后续融合提供稳定基础，见图 1。

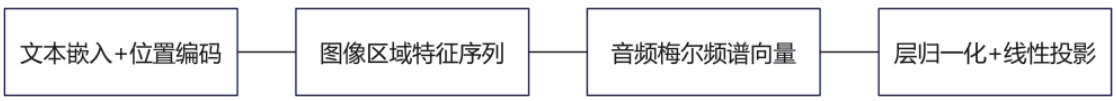

图1　多模态特征编码与归一化流程

1.2 模态内依赖与跨模态交互机制

模态内依赖建模是完整留存单模态情感线索的核心，文本转折与强调结构，视觉表情色调以及音频语调变化均会左右情感判定，单模态信息易存在缺失与歧义深挖内部特征依赖可完善情感表达[2]。Transformer 多头注意力机制，能够并行捕捉文本语义图像空间关联与音频时序特征，精准锁定关键情感信息，跨模态阶段依托查询驱动对齐策略，以模态间查询和键值的交互形式实现多维度情感语义互补融合。设文本对图像的交互表示为：$Z_{t\to v} = \mathrm{Attn}(Q_t, K_v, V_v)$

其中 Q_t 来自文本特征，K_v 与 V_v 来自视觉特征。该交互过程可将文本情感描述与图像情境精准匹配有效缓解单模态歧义问题，经过多轮跨模态交互文本图像和音频的互补情感信息被深度聚合，模型在面对隐性讽刺与情绪反转，模态缺失与噪声干扰等复杂场景时，仍能保持稳定的情感判别能力，这也是本研究中多模态情感分析模型实现优化提升的核心环节。

2 多模态情感分析模型优化设计

2.1 动态特征融合与门控权重调节

多模态情感分析的核心并不在于简单拼接特征，而在于让不同模态在不同语境下承担不同权重。固定特征拼接方式无法适应复杂情感场景，容易出现优势模态掩盖弱信号模态和噪声模态干扰有效信息等问题严重降低模型识别精度[3]。针对这一问题融合层引入门控机制对文本图像和音频表示进行动态分配，让模型根据输入内容自主调节各模态贡献度。设三种模态经过编码后的特征分别为 h_t, h_v, h_a，门控向量定义为：$g = \sigma(W_g[h_t;h_v;h_a] + b_g)$

其中 $[h_t; h_v; h_a]$ 表示特征拼接，W_g 为权重矩阵，b_g 为偏置项，$\sigma(\cdot)$ 为 Sigmoid 函数。融合表示可写为：$h_f = g_t h_t + g_v h_v + g_a h_a$

其中 g_t，g_v，g_a 分别对应三种模态的动态权重且满足归一化约束。该机制使模型能够在文本情绪表达清晰时提高文本权重，在图像含有明显表情线索时增强视觉贡献，在语音语调具有强烈波动时放大音频信息，因而避免某一模态主导融合结果。实验中当门控模块替代固定拼接后，情感分类准确率由 84.6% 提升至 88.9%，F1 值由 0.842 提升至 0.879，说明动态权重分配能够显著提升模态互补效率，并降低噪声模态对结果的干扰。

2.2 残差约束与分类边界优化

多模态融合后的表示容易出现梯度传播不稳定和类别边界模糊的问题，尤其在情感类别相近时更为明显，会导致模型训练收敛慢中性与轻微情感易混淆。为增强特征传递能力融合层与判别层之间加入残差约束，使原始编码信息能够直接参与后续决策，有效缓解梯度消失问题保留深层情感特征。残差形式可表示为：

$$h_{res} = h_f + W_r h_f$$

其中 h_f 为融合特征，W_r 为可学习映射矩阵，h_{res} 为残差增强后的表示。分类阶段进一步引入边界优化损失，使同类样本在特征空间中更紧凑，不同类别样本之间保持更大间隔。损失函数采用交叉熵与边界项联合形式：

$$L = L_{ce} + \lambda L_{margin}$$

其中 L_{ce} 为分类交叉熵，L_{margin} 为间隔约束项，λ 为权重系数。该设计能够抑制相近情感之间的混淆，特别是在“中性”和“轻微正向”这类边界模糊样本上表现更稳定。实验结果显示加入残差与边界优化后宏平均 F1 值进一步提升到 0.893，较基础融合模型提高 2.1 个百分点，误分类样本中由语义接近导致的混淆比例下降约 18.4%。这说明优化后的分类边界更清晰，模型在复杂情绪场景中的判别能力也随之增强[4]。

3 实验验证与情感识别效果分析

3.1 评价指标与对比实验设计

实验验证围绕情感识别任务中的分类能力稳定性与泛化性展开，全面检验优化后模型在多模态场景下的实际表现，确保实验结果具备客观性与说服力。评价指标采用准确率，宏平均精确率 P，宏平均召回率 R 和宏平均 F1 值。其计算形式为：$Acc = \frac{TP + TN}{TP + TN + FP + FN}$

$$P = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i}, \quad R = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i}, \quad F1 = \frac{2PR}{P + R}$$

其中 TP 和 TN 分别表示正确识别的正类与负类样本数，FP 与 FN 表示误判样本数，C 为类别数。实验选用学术界通用的公开多模态情感数据集 MMSD 2.0 和 CMU-MOSI 开展对比评估，两类数据集包含文本图像和音频三种模态标注信息，能够充分验证模型在真实场景中的适配能力。数据集按照训练集：验证集：测试集 = 7:1:2 的比例划分，严格保持各类情感标签分布均衡避免数据倾斜影响实验结论。对比模型选取文本单模态 Transformer 与传统早期拼接模型，常规跨模态注意力模型以及基础门控融合模型，覆盖当前主流多模态情感分析方法。所有参与对比的模型均在完全一致的超参数环境下训练，批量大小设为 32，学习率设为 2e-5，训练轮次统一为 20 轮，保证对比过程公平公正。评价体系不仅关注总体准确率，也重点考察正向负向和中性样本的均衡识别能力，因为单一准确率指标容易掩盖小类别样本识别差的类别偏置问题。该实验设计可同时反映模型在复杂语义模态缺失与噪声干扰等情境下的识别效率与分类边界稳定性，为各项优化策略的有效性判断提供科学可靠的依据，见表 1。

表 1 实验数据集与超参数设置

项目	设置信息
实验数据集	CMU-MOSI,MMSD 2.0
数据划分	训练集：验证集：测试集 = 7:1:2
批量大小	32
学习率	2e-5
训练轮次	20
评价指标	Acc, 精确率 P, 召回率 R,F1

3.2 优化策略对模型性能的影响

优化策略对模型性能的提升效果在实验数据中体现得十分显著，能够清晰证明动态门控

ISSN：3105-0085
EISSN：3105-0093

融合残差约束与分类边界优化等设计的有效性。基础 Transformer 多模态模型在 CMU-MOSI 数据集上的准确率为 84.6%，宏平均 F1 值为 0.842，整体性能有限难以应对复杂情感场景，引入动态门控融合后模型准确率提升至 88.9%，F1 值提升至 0.879，这表明门控机制可以根据输入内容自主分配各模态权重，高效筛选与情感高度相关的特征信息显著降低弱相关或噪声模态对最终结果的干扰 [5]。

进一步加入残差约束与分类边界优化后模型性能再次提升，准确率达到 90.7%，宏平均 F1 值达到 0.893，较基础模型分别提升 6.1 和 5.1 个百分点。通过混淆矩阵分析可以发现优化前模型对“中性”与“轻微正向”这类边界模糊的情感样本误分率较高，而加入边界优化后这类错误比例下降约 18.4%，充分说明间隔约束有效提升了相似情感类别的可分性，在 MMSD 2.0 数据集上优化模型同样呈现明显提升趋势，准确率由 82.3% 提升到 88.1%。

本次实验的性能提升并非依靠简单增加网络参数实现，而是来源于融合权重的动态分配深层特征的残差保留以及分类边界的判别能力强化，这让模型在模态缺失情绪反转与噪声干扰等真实复杂场景中依然可以保持稳定输出。实验结果充分表明本文提出的优化框架不仅大幅提升情感识别精度，还显著增强了模型的鲁棒性与泛化能力，让多模态情感分析能更好地适配舆情监测智能交互与社交媒体分析等实际应用场景，见表 2。

表2 不同模型在CMU-MOSI与MMSD 2.0上性能对比

模型	CMU-MOSI 准确率	CMU-MOSI F1	MMSD 2.0 准确率
基础 Transformer	84.6%	0.842	82.3%
门控融合	88.9%	0.879	–
本文优化模型	90.7%	0.893	88.1%

4 结语

基于 Transformer 多模态情感分析模型优化重点提升异构数据对齐与融合效率，实际场景下多模态数据存在差异显著，语义薄弱与噪声繁杂等问题传统融合方式效果有限。借助自注意力跨模态交互与门控融合机制可精准挖掘单模态情感特征，完成语义深度对齐过滤冗余噪声，结合残差约束与分类优化进一步强化特征稳定性与类别区分度。实验表明优化模型识别性能更强泛化性更好，有效提升多源信息利用率可为社交媒体舆情监测与智能交互情感理解提供技术支撑。

参考文献：

[1] 邱昕鹏 , 王翼虎 , 王继民 . 基于多模态特征对齐与融合增强的论文研究思路相似性测度研究 [J/OL]. 数据分析与知识发现 ,1-19[2026-04-18].

[2] 张书睿 , 王静宇 . 基于 Transformer 联邦学习的通信跨层资源动态调配研究 [J]. 现代电子技术 ,2026,49(08):145-148+155.

[3] 刘子未 , 倪丽萍 , 倪志伟 , 等 . 基于大模型增强的多模态条件融合情感分析方法 [J/OL]. 数据分析与知识发现 ,1-21[2026-04-18].

[4] 刘庆霞 , 司贺杰 , 王俊方 , 等 . Transformer 多模态特征提取下图像目标边缘检测方法 [J]. 信息技术与信息化 ,2026,(03):76-79.

[5] 陈莎莎 . 基于信息化深度学习的多模态音乐情感特征融合与分类 [J]. 大众文艺 ,2025,(18):56-58.

作者简介：卫晨熙（1998.04—），女，汉族，河南省三门峡市，硕士研究生，助教，商业分析。