计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2026,35(3):59−68 [doi: 10.15888/j.cnki.csa.010107] [CSTR: 32024.14.csa.010107]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 H-GEM 模型的多模态情感分析①

杨新航, 王晶晶, 陈思宇, 田　宏

(大连交通大学 轨道智能工程学院, 大连 116052)
通信作者: 田　宏, E-mail: th@djtu.edu.cn

摘　要: 传统多模态情感分析方法在特征拼接和融合中易产生信息冗余, 难以捕捉细粒度复杂情感特征, 在模态缺失和跨域迁移场景下鲁棒性不足. 同时, 现有混合专家 (MoE) 方法大多为单层结构, 专家分工不明确, 存在功能重叠和泛化性欠佳的问题. 本文提出一种分层自适应混合专家模型 H-GEM (hierarchical gated expert mixture). 通过构建 3 层分级专家体系: 模态专家层提炼模态特征; 融合与抽象专家层自适应选择融合策略; 情感极性专家层进行细粒度建模. 同时引入信息论与判别性约束提升专家选择的语义区分性和稀疏性. 通过分层门控实现逐级决策, 保证专家差异化分工与跨任务建模. 在 CMU-MOSI 和 CMU-MOSEI 数据集上的实验结果表明, H-GEM 在一系列指标上均优于基线模型. 与单层 MoE 架构相比, 显著降低的路由熵表明其能够有效缓解专家冗余问题. 该模型在低资源和模态缺失复杂任务中表现出更高的鲁棒性, 展现出良好的应用潜力.

关键词: 多模态情感分析; 分层门控机制; 混合专家模型; 互信息约束; 鲁棒性

引用格式: 杨新航,王晶晶,陈思宇,田宏.基于 H-GEM 模型的多模态情感分析.计算机系统应用,2026,35(3):59–68. http://www.c-s-a.org.cn/1003-3254/10107.html

Multimodal Sentiment Analysis with H-GEM Model

YANG Xin-Hang, WANG Jing-Jing, CHEN Si-Yu, TIAN Hong

(School of Intelligent Railway Engineering, Dalian Jiaotong University, Dalian 116052, China)

Abstract: Traditional multimodal sentiment analysis methods often suffer from information redundancy during feature concatenation and fusion, making it difficult to capture fine-grained and complex emotional features, while also exhibiting limited robustness in modality-missing and cross-domain transfer scenarios. Meanwhile, most existing mixture of experts (MoE) methods adopt a single-layered structure with ambiguous expert specialization, leading to functional overlap and suboptimal generalization. To address these issues, this study proposes a hierarchical gated expert mixture (H-GEM) model. A three-layer hierarchical expert architecture is constructed: a modality expert layer extracts modal features, a fusion and abstraction expert layer adaptively selects fusion strategies, and a sentiment polarity expert layer performs fine-grained modeling. In addition, information-theoretic and discriminative constraints are incorporated to enhance the semantic discriminability and sparsity of expert selection. By leveraging hierarchical gating for progressive decision-making, H-GEM ensures differentiated expert specialization and cross-task modeling. Experimental results on CMU-MOSI and CMU-MOSEI datasets demonstrate that H-GEM outperforms baseline models across a series of metrics. Compared with single-layer MoE architectures, the significantly reduced routing entropy indicates effective mitigation of expert redundancy. Moreover, the proposed model demonstrates higher robustness in low-resource and modality-missing scenarios, highlighting its strong practical applicability.

Key words: multimodal sentiment analysis; hierarchical gating mechanism; mixture of experts (MoE); mutual information constraint; robustness

① 基金项目: 国家自然科学基金 (622711491)
收稿时间: 2025-08-11; 修改时间: 2025-09-10, 2025-10-17; 采用时间: 2025-10-29; csa 在线出版时间: 2026-01-19
CNKI 网络首发时间: 2026-01-20

随着社交媒体、论坛等平台上多模态数据的激增, 情感分析任务逐渐由单一文本向图像、音频等模态扩展, 多模态情感分析[1]因更全面的感知能力成为研究热点. 然而, 现有模型多采用静态交互策略, 难以适配模态异质性, 易出现跨模态语义干扰和冗余特征淹没判别性线索的问题[2−4]. 主流方法对信息流的建模不足, 常出现模态不平衡与信息丢失[5,6]. 同时, 传统多模态融合方法通常无法有效解决模态不平衡的问题, 使得一些模态信息被低估导致整体识别性能受限[7,8]. 近年来, 混合专家 (MoE) 技术被引入多模态情感分析, 通过门控机制动态选择专家以提升特征建模与融合能力[9,10]. 但现有 MoE 多为单层结构, 存在专家分工不清、功能重叠的问题, 难以兼顾泛化性与鲁棒性.

在各种问题及前人研究的基础上, 本文设计了一种分层自适应混合专家模型 H-GEM (hierarchical gated expert mixture) 作为核心融合组件, 实现多模态特征的高效交互与自适应融合, 主要工作如下.

(1) 提出一种 3 层分级的自适应门控专家混合模块 H-GEM. 模态专家层负责提炼各模态特征并保留模态独有信息; 融合与抽象专家层通过分层门控网络自适应选择最优融合策略; 情感极性专家层在情感空间实现细粒度建模, 提升跨任务情感判别能力.

(2) 构建基于 H-GEM 的多模态融合框架, 引入分层稀疏优化与互信息一致性约束机制, 通过层级专家的动态调度实现任务导向的特征选择与语义一致性建模, 从而在理论上缓解传统单层 MoE 专家分工不清与激活冗余的问题.

通过上述工作, H-GEM 在多模态特征选择、信息流动、非线性特征表达以及跨域和模态缺失场景下均取得显著改进. 实验结果表明, 在 CMU-MOSI 和 CMU-MOSEI 数据集上, 模型在平均绝对误差 (MAE)、七分类准确率 (Acc-7) 等关键指标均优于现有方法. 与传统单层 MoE 相比, H-GEM 在融合效率与情感判别性能上均实现明显提升, 路由熵显著降低, 有效缓解专家冗余的问题. 在复杂任务中表现出更高的鲁棒性与泛化性, 其设计的有效性得到验证.

1 相关工作

多模态情感分析通过整合文本、语音和视觉信息来识别用户情感, 已成为自然语言处理和计算机视觉领域的重要研究方向. 早期工作如 TFN[11]和 LMF[12]采用张量融合等方法进行简单的模态交互, 但这些方法往往忽略模态间的动态关联. 近年来, 研究者们开始关注更复杂的跨模态交互机制, 例如 MulT[13]通过跨模态注意力实现模态间的信息流动, MMGCN[14]则利用图卷积网络建模模态间的关系. 然而, 现有的全局依赖建模不充分、关键信息被稀释等问题, 限制模型性能的进一步提升.

1.1 多模态情感识别

多模态情感识别是指融合语音、文本和视觉等多源数据推断人类情感状态的技术. 其核心挑战在于实现模态间的有效协同, 以精确捕捉情感的连续性和复杂性. 七分类准确率 (Acc-7) 因其对情感强度的精确量化能力成为当前细粒度情感识别的关键评估指标之一. Li 等人[15]指出现有方法在融合话语级全局表示时忽略细粒度特征与模态关联, 为此提出 MMTA, 通过细粒度对齐模块实现单词级特征对齐以弥合模态差距, 并采用多级融合模块捕获局部与全局的跨模态交互. 注意力融合网络 (AFN) 用于建模模态间与模态内相关性, 生成一致多模态表示. 而 Verma 等人[16]发现现有模型在处理分布不平衡的数据集时易存在偏差. 为此, 他们通过 BERT 结合 BiLSTM 和 ResNeXt SE 提取文本和图像特征并使用文本-图像交互 (TII) 的附加模块促进动态跨模态交互. 与此同时, Aruna Gladys 等人[17]提出一种多模态表示学习框架, 使用多模态自动编码器来学习底层异构模态的综合表示. 王楠等人[18]设计一种基于知识蒸馏与动态调整机制的多模态情感分析模型, 使用动态权重调整模块和多模态掩码 Transformer 来平衡特征贡献并捕获模态间的细微交互. 用知识蒸馏和动态调整机制弥补模态缺失场景下的研究缺陷. 这些方法使建立模态间的动态互补机制与细粒度对齐在提高多模态情感分析性能方面的有效性得到证明.

1.2 模态融合方法

模态融合是多模态情感分析的核心挑战. 早期方法多在特征层拼接或加权, 易造成信息损失; 后来的方法则在模态独立建模后进行决策融合, 忽略了模态间交互. 近年来, 混合专家 (mixture of experts, MoE) 机制被广泛用于多模态任务, 以应对异质性强、冗余高和任务复杂度大的问题. 与固定融合策略不同, MoE 通过门控网络自适应选择专家, 实现稀疏激活与动态融合, 从而在降低计算开销的同时提升跨模态适应性. 研究表明, 该机制在多模态翻译与视频理解等任务中表现

出更高的灵活性与泛化性. 针对资源受限场景下面临的高计算复杂度等问题, Zhao 等人[19]在 SAM (segment anything model) 的基础上引入稀疏注意力动态激活与病斑稀疏分布相匹配的关键通道, 使用 MoE 解码器进行粗粒度边界定位. Deng 等人[20]开发的 Stereo-talker 系统引入先验引导的双 MoE 结构, 并辅以掩码预测模块提升局部一致性. Shi 等人[21]提出一种基于 LERT-MoE 的受害过程分析模型. 使用 MoE 抽取触发词, 并结合点积注意力完成角色分类. 该方法在精度、准确率和 $F1$ 值上均优于基线模型, 并能提取更细粒度的诈骗模式. 但该方法训练样本有限, 泛化能力不足. 在脑影像分析中, 现有深度学习方法多聚焦于特定任务或疾病, 难以构建统一模型. Ding 等人[22]提出 DenseFormer-MoE, 结合 DenseNet 与 Vision Transformer, 通过掩码自编码与自监督预训练获取通用表征, 并在多任务学习中引入混合专家机制以缓解任务间优化冲突. 该模型在 UKB、ADNI 和 PPMI 数据集上表现优异, 但其训练复杂度较高、计算资源需求较大. Mengara 等人[23]提出一种基于 MoE 的融合框架, 利用多组 Transformer 子专家提取模态特征, 并通过稀疏 Top-k 门控实现动态选择. 跨模态注意力模块促进信息交互, 增强噪声与缺失条件下的鲁棒性, 引入多样性损失以鼓励专家分工, 专家冗余问题得到缓解.

现有多模态融合方法虽通过特征拼接、加权或注意力机制在一定程度上缓解了模态异构与信息冗余, 但在处理不同长度和复杂度的输入时仍缺乏自适应性. 传统 MoE 方法虽能动态选择专家, 却受限于单层或固定路由策略, 难以充分利用层次化模态信息, 在模态缺失、噪声干扰和跨域场景下鲁棒性仍显不足.

2 基于 H-GEM 的多模态情感分析方法

为解决现有方法中融合策略固定、专家分工模糊及模态适应性差等问题, 本文提出一种分层自适应混合专家模型 (H-GEM), 显著缓解了专家冗余的问题.

与传统单层 MoE 仅在输入空间进行平面式专家选择不同, H-GEM 在理论上基于“分层稀疏优化与互信息一致性约束”思想, 通过多层门控机制在模态、融合与情感层级上实现专家分工协作与动态激活. 该机制使模型在不同任务与场景下自适应分配特征表示路径, 显著提升可解释性、鲁棒性与泛化能力. H-GEM 框架由模态特征编码层、多模态表征与融合层、情感分类输出层这 3 部分构成, 其整体结构如图 1 所示.

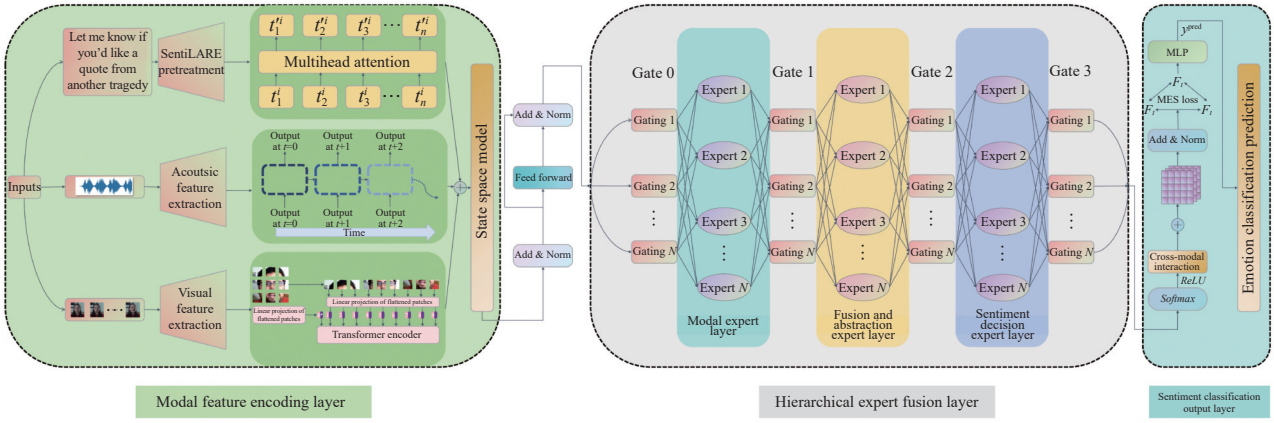

图 1 基于 H-GEM 模块的多模态情感分析模型整体结构图

首先, 文本、视觉和音频这 3 种模态原始特征被分别嵌入到统一的向量空间中, 具体而言, H-GEM 在多模态特征与融合阶段引入 4 层级联的门控结构 (Gate 0–Gate 3), 每一 Gate 层均由 N 个并行 Gating 单元组成, 用于对候选专家子空间进行条件选择与权重分配. 不同层级的 Gate 分别侧重于不同粒度的语义调控. Gate 0 门控网络动态计算各模态权重, 对模态特征进行缩放. 随后, 各专家提取的模态特征输入多模态表征与融合层, Gate 1 在稀疏、密集与鲁棒性专家间动态选择, 实现多粒度跨模态融合. 融合特征经 Gate 2 根据情感倾向自适应加权情感极性专家, 生成针对性情感语义. 最后, Gate 3 对情感专家输出进行加权融合, 得到最终情感预测结果.

该分层门控机制在结构上突破了传统 MoE 的平

面瓶颈, 形成稀疏选择与信息一致性的统一优化框架.

2.1 任务描述

在多模态情感分析 (MSA) 任务中, 模型输入为同一视频片段的 3 个模态原始序列:

$$X_t \in R^{T_t \times D_t},\ X_v \in R^{T_v \times D_v},\ X_a \in R^{T_a \times D_a} \tag{1}$$

分别表示文本、视觉与声学模态的输入特征, 其中T_t、T_v、T_a表示各模态时间步长, D_t、D_v、D_a表示向量维度.

2.2 特征提取

首先, 我们对多模态输入序列进行编码, 以获得对应模态单位长度的模态特征E_t、E_v和E_a. 具体来说, 对于文本模态, 经 SentiLARE 预处理后, 采用 RoBERTa 模型对输入句子进行编码, 将输入的文本 ID 转换为统一的语义表示. 其嵌入由多种子嵌入矩阵相加构成, 包括词汇嵌入、位置嵌入、句子类型嵌入, 以及可选的情感、词性和极性嵌入, 复合形式如下所示:

$$E_t = E_{\text{word}} + E_{\text{pos}} + E_{\text{type}} + E_{\text{senti}} + E_{\text{pos-tag}} + E_{\text{polarity}} \tag{2}$$

对于视觉和声学模态, 我们将 ID 嵌入结合 Gate 0 进行加权选择以增强鲁棒性. 通过嵌入层将原始特征映射到统一的语义空间. 其中, E_v, $E_a \in R^{(L+1)\times D}$分别表示视觉与声学模态的嵌入矩阵, D为对应模态特征维度, 额外的索引$L+1$用于缺失模态的填充.

2.3 模态表征学习

模态表征学习旨在为每种模态构建具备判别力与鲁棒性的深层特征表示, 为后续的跨模态交互与融合奠定基础. H-GEM 在模态专家层通过门控网络自适应选择最优专家组合, 该层专家配置如下.

文本专家网络: 基于 Transformer 架构, 利用多头自注意力机制和前馈层实现上下文建模.

视觉专家网络: 结合卷积特征提取与 Vision Transformer, 同时捕捉局部空间模式和全局依赖.

声学专家网络: 融合双向 LSTM 与卷积-时序注意力机制, 兼顾语音的时序动态与频域信息.

2.4 多模态融合

多模态融合是指将 3 种模态特征融合, 以获取更具有全面性和表征力的信息. Gate 1 和多模态表征与融合层通过选择不同融合专家在统一的语义空间中完成模态间交互建模. 稀疏专家捕捉关键模态间的关联, 密集专家提供全局层面信息整合, 鲁棒性专家抵御模态不一致与噪声干扰. 通过分工合作, 模型能获得兼顾精细交互与整体一致性的跨模态抽象表征.

稀疏专家使用多头注意力机制 (multihead attention, MHA) 与稀疏特征提取来建模跨模态关键交互. 标准多头注意力可能分散在所有 token 上, 易引入冗余和噪声. 通过引入 Top-k 技术, 确保稀疏专家专注于最重要的跨模态交互点, 提高计算效率和可解释性. 注意力计算公式及 Top-k 定义如下:

$$Attention_{\text{sparse}}(Q,K,V) = G_k\left(Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)\right)\cdot V \tag{3}$$

$$G_k(X_i) = \begin{cases} X_i, & X_i \in \text{Top-}k(X) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

为进一步控制稀疏性, 模型只保留前$k = r\cdot d$个最重要的 token, 其中, r为稀疏比例, d为特征维度, 专家激活示例见图 2.

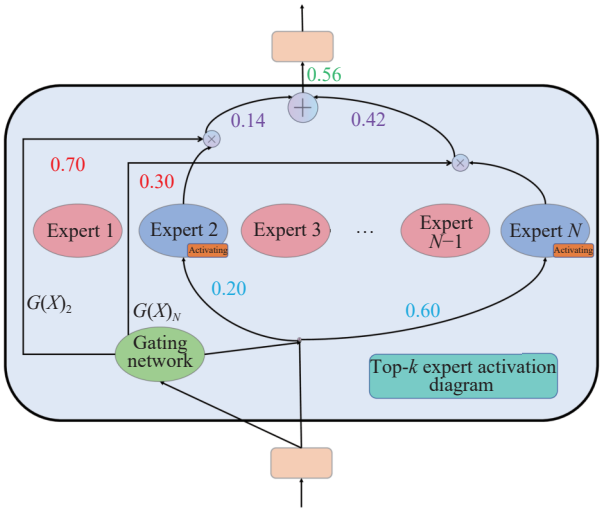

图 2 稀疏融合专家中的 Top-k 专家选择与权重计算图

密集专家首先通过全连接网络对输入拼接特征进行非线性映射:

$$Z' = \sigma(W_2\cdot\phi(W_1 H_{\text{concat}} + b_1) + b_2) \tag{5}$$

随后利用全局平均池化 (global average pooling) 进一步压缩时间或空间维度, 获得全局特征表示:

$$Z_{\text{dense}} = \frac{1}{N}\sum_{i=1}^{N} Z_i' \tag{6}$$

Z_{dense}保证输出提供整体一致性与全局信息整合, 确保融合后表示具备全面性.

鲁棒性专家在训练过程中通过向输入加入高斯噪声来增强鲁棒性:

$$\bar{H} = H_{\text{concat}} + \varepsilon,\ \varepsilon \sim N\left(0,\sigma^2 I\right) \tag{7}$$

为控制各模态在不同特征维度上的可靠性, 我们设计模态可信度矩阵 $C \in [0,1]^{\text{modal}\times\text{feature}}$ 进行加权融合:

$$Z_{\text{robust}} = f_{\text{robust}}(C \odot T_{\text{concat}} + (1-C)\odot \hat{T}) \tag{8}$$

该矩阵将鲁棒性转化为数学公式, 使每个特征维度的可信度可显式表达和优化. 对于受噪声影响的输入, 鲁棒性专家通过去噪器 (denoiser) 进行恢复.

2.5 情感预测

我们设计引入信息论和判别性约束的 Gate 2 来负责对情感专家进行选择与加权, 其网络结构定义为:

$$h = ReLU\left(W_{1\text{fused}}^{\text{T}} + b_1\right) \tag{9}$$

其中, $W_{1\text{fused}}^{\text{T}} \in R^d$ 表示 Gate 1 输出的融合特征, b_1 为第 1 层全连接层的偏置项, h 为隐藏层维度.

为避免 Gate 2 输出的路由分布在不同类别间高度重叠, 我们引入类内紧凑、类间分离的约束. 该约束在理论上可视为对任务可分性的优化约束, 使专家选择空间在分布层面趋于正则化. 类内判别损失定义如下:

$$L_{\text{disc}} = \frac{1}{|P|}\sum_{(i,j)\in P}\|w(i)-w(j)\|_2^2 - \frac{1}{|N|}\sum_{(i,j)\in N}\|w(i)-w(j)\|_2^2 \tag{10}$$

其中, $P = \{(i,j): y_i = y_j\}$, $N = \{(i,j): y_i \neq y_j\}$, 该约束确保同类别样本更倾向于路由到相似专家集合, 不同类别样本分散至差异化专家, 提升判别性和任务相关性. 因判别性不足以保证路由选择对标签信息敏感, 因此我们基于信息论的优化目标进一步最大化路由分布与标签的互信息约束. 通过最小化条件熵 $H(Y|Z)$ (即在给定路由分布 Z 的情况下减少标签 Y 的不确定性), 提高专家激活与标签语义的互依程度, 形成理论层面的信息一致性机制. 直接鼓励路由分布携带更多类别标签信息, Gate 2 的选择进一步增强路由与语义的耦合, 该约束采用 InfoNCE 下界近似:

$$\hat{I}_{\text{NCE}}(w;Y) \approx E_n\left[\log\frac{\exp\left(\frac{s\left(w^{(n)},y^{(n)}\right)}{\tau_{\text{NCE}}}\right)}{\sum_{m\in B}\exp\left(\frac{s\left(w^{(n)},y^{(m)}\right)}{\tau_{\text{NCE}}}\right)}\right] \tag{11}$$

其中, $s(\cdot,\cdot)$ 为相似度函数, 负样本来自 batch 内采样, 损失项定义为 $L_{\text{MI}} = -\hat{I}_{\text{NCE}}(w;Y)$, 温度 $\tau_{\text{NCE}} \in [0.08, 0.18]$. Gate 2 门控自适应选择最优专家组合.

情感表达的识别不仅依赖模态内部特征, 还取决于跨模态间的动态语义关联. 我们在 Gate 3 引入信息驱动的稀疏聚合机制 (information-guided sparse fusion, IGSF). IGSF 以 Gate 2 优化得到的路由分布 Z 作为稀疏门控调节因子, 对跨模态稀疏注意力输出进行动态调控:

$$H_{\text{fusion}} = G_{\text{info}}(Z)\odot H_{\text{sparse}} \tag{12}$$

其中, H_{sparse} 为式 (12) 所定义的稀疏注意力, $G_{\text{info}}(Z)$ 表示由互信息强度驱动的动态权重生成函数, 用于自适应调整不同模态片段的保留比例与融合强度. 该设计使用信息论一致性强化模态间的语义对齐与判别, 避免情感识别因单一模态偏差而失真.

H-GEM 在情感预测阶段建立分层稀疏优化与互信息一致性约束的统一框架. 与传统单层 MoE 仅依赖输入特征进行平面式专家选择不同, H-GEM 通过层级化门控在模态、融合与情感空间中递进实现专家分工与梯度分解优化. 该分层结构在理论上缓解单层 MoE 中存在的梯度竞争与专家重叠问题, 实现从“结构改进”到“优化机理创新”的转变.

3 实验分析

3.1 数据集

本文选用两个广泛应用于多模态情感分析研究的公开数据集 CMU-MOSI 和 CMU-MOSEI 对所提出的模型进行验证.

CMU-MOSI[24]数据集由 CMU MultiComp Lab 发布, 数据来源于 YouTube 网站, 共包含 93 个视频, 划分为 2 199 个语音视频样本. 每个样本均包含文本、语音与视觉这 3 种模态信息.

CMU-MOSEI[25]是 MOSI 的扩展版本, 拥有更大的数据规模与更高的多样性, 涵盖超过 1 000 名演讲者及 250 个话题, 包含 16 265 条训练样本、1 869 条验证样本和 4 643 条测试样本, 同样包含文本、语音和视觉这 3 种模态, 具有多标签特性.

在后续章节中, 我们分别以 MOSI 和 MOSEI 代指上述两个数据集.

3.2 参数设置及评价指标

本实验基于 Python 3.8 和 PyTorch 1.11.0 框架构建模型, 在 RTX 4070 显卡上进行训练与测试. 文本模态隐藏特征维度为 768, 音频和视频模态在 MOSI 数据集下的维度分别为 74 和 27; 对齐后序列长度为 50, 融合后隐藏特征长度为 256. 输出层头部隐藏维度设为

64, dropout 比例为 0.1. 训练过程中, 学习率为 2E–5, 注意力模块 dropout 比例为 0.25, 训练轮数为 40, batch size 设置为 32, 隐藏层维度 256, 优化器采用 Adam.

为评估模型在多模态情感分析任务的性能, 回归任务以平均绝对误差 (MAE) 和皮尔逊相关系数 (Corr) 为主要指标. MAE 衡量预测值与真实值的平均绝对差值, 数值越小表示预测越精确; Corr 衡量二者的线性相关性, 越接近 1 表明拟合效果越好. 分类任务中额外采用二分类准确率 (Acc-2)、七分类准确率 (Acc-7) 及 $F1$ 分数. Acc-2 将情感标签划分为正负两类; $F1$ 分数综合考虑精确率与召回率, 适用于类别不平衡场景; Acc-7 将情感细分为 7 个等级, 衡量模型对情绪强度的区分能力.

3.3 效率分析与基线对比方法

为全面评估 H-GEM 模型的实际应用可行性, 本文对模型的参数量、计算复杂度、训练耗时及推理延迟进行系统分析, 并与多种基线模型进行对比. 结果表明, H-GEM 总参数量约为 61.88M, 其中 Gating 网络占 16.6% (约 10.25M), 专家网络占 76.7% (约 47.45M). 在 3 层 MoE 架构中, 模态专家层约 19.32M, 融合与抽象专家层约 27.56M, 情感极性专家层约 4.73M.

相比于 MUG (总参数量约 110M)、MISA/MAG-BERT (总参数量约 110M) 及 Self-MM (总参数量约 109–110M) 等基线模型, H-GEM 参数量明显较少. 这主要归因于 3 层 MoE 的分层稀疏激活机制, 在互信息一致性约束与批内判别性约束的共同驱动下, 门控网络能够学习更具判别力且更稀疏的专家激活模式. 每次前向传播仅激活部分关键专家, 保证模型表达能力的同时降低实际计算负担. 因 H-GEM 采用模块化设计, 未引入多任务预测或额外投影网络等冗余模块, 参数规模进一步减小.

在训练和推理效率上, H-GEM 单步迭代耗时约 0.79 s, 相比于 Self-MM (约 1.16 s)、MUG (约 0.86 s), 其分层稀疏 MoE 设计显著提高了计算效率.

综上, H-GEM 通过将优化约束与分层 MoE 结构相结合, 实现模型性能与推理效率的协同提升. 通过稀疏且判别性强的专家激活机制, 结合模块化设计, 在多模态特征处理与情感建模任务中表现出高效性与实用性, 验证了模型在实际应用中的可行性和优势.

3.4 对比实验

为验证本模型的有效性, 本文将实验结果与多种主流基线方法进行了对比分析.

TFN: 通过张量外积实现多模态交互建模, 提升情感识别性能.

LMF: 在 TFN 基础上引入低秩张量分解, 以降低计算复杂度并提升融合效率.

MulT: 基于跨模态注意力机制, 实现无需显式对齐的自适应特征融合, 增强模态间信息交互能力.

MISA: 通过多重损失函数联合优化模态不变与特定表示, 提升情感分类的融合效果和性能.

FMT: 一种基于分解式多模态自注意力机制的新型序列建模框架 (FMS), 有效捕捉模态间交互, 提升信息融合能力.

Self-MM: 基于自监督伪标签生成机制, 为各模态独立构建标签并联合优化多模态任务, 增强模态间一致性与差异性.

MAG-BERT: 引入多模态自适应门机制 (MAG), 以方向与强度向量形式编码非语言行为, 并通过注意力调节实现自适应融合.

CENet: 通过跨模态增强模块注入视觉与声学线索语言信息, 结合模态特征转换策略, 提升多模态情感特征的融合与文本表征能力.

表 1 是本文模型与其他模型的性能对比结果. 在 MOSI 数据集上, 本文方法所有指标均显著高于基线模型, 其中 Corr 达到 0.862, MAE 降至 0.574, Acc-2、Acc-7、$F1$ 分数分别为 89.2%、51.9% 和 88.3%. 相比于 CENet 模型, 情感极性专家层成功捕获了不同情感极性的细微复杂特征, Acc-7 提升 3%, 验证了改进机制在细粒度情感辨别中的有效性.

在 MOSEI 数据集上, 本文方法仍取得领先性能, Corr 达到 0.863、MAE 降至 0.524, 同时 Acc-2、Acc-7、$F1$ 分数分别为 88.2%、55.1% 和 88.1%. 其中, Acc-2 和 Acc-7 与最接近的 Self-MM (85.2% 和 53.4%) 相比, 本文模型情感二分类与七分类识别更准确. Corr 较 Self-MM 增长 12.81%, 进一步体现其综合性能优势.

为验证所提出的分层 MoE 架构在不同任务下的有效性, 我们设计了在 MOSEI 数据集上的专家激活实验, 结果如图 3 所示.

实验结果表明, 该架构在二分类 (Acc-2) 与细粒度七分类 (Acc-7) 任务中呈现差异化优势: 在 Acc-2 中, 模型主要依赖文本模态区分情感极性, 密集专家占主导, 正负情感专家激活均衡而中性专家几乎不被使用.

在 Acc-7 中, 视觉与声学模态对情感强度辨别的作用增强, 密集专家与鲁棒性专家激活趋于平衡, 情感专家层呈现多专家协同激活特征. 整体来看, 分层 MoE 能根据任务粒度自适应调整专家使用策略.

表 1 对照模型在 CMU-MOSI 和 CMU-MOSEI 上的实验结果

模型	CMU-MOSI					CMU-MOSEI				
	Corr	MAE	Acc-2 (%)	Acc-7 (%)	*F*1 (%)	Corr	MAE	Acc-2 (%)	Acc-7 (%)	*F*1 (%)
TFN	0.698	0.901	80.8	34.9	80.7	0.700	0.593	82.5	50.2	82.1
LMF	0.698	0.917	82.5	33.2	82.4	0.677	0.623	82.0	48.0	82.1
MulT	0.698	0.871	83.0	40.0	82.8	0.703	0.580	82.5	51.8	82.3
MISA	0.764	0.804	82.1	42.3	82.0	0.724	0.568	84.2	52.2	84.0
FMT	0.744	0.837	83.5	—	83.5	—	—	—	—	—
Self-MM	0.798	0.713	86.0	44.9	86.0	0.765	0.530	85.2	53.4	85.3
MAG-BERT	0.782	0.784	84.3	43.6	84.3	0.755	0.543	84.8	52.6	84.7
CENet	0.851	0.597	86.1	48.9	85.9	0.806	0.578	83.3	51.0	86.6
本文模型	**0.862**	**0.574**	**89.2**	**51.9**	**88.3**	**0.863**	**0.524**	**88.2**	**55.1**	**88.1**

注: 加粗表示该指标下的最优结果, 下划线表示次优结果, “—”表示原论文未报告该项数据

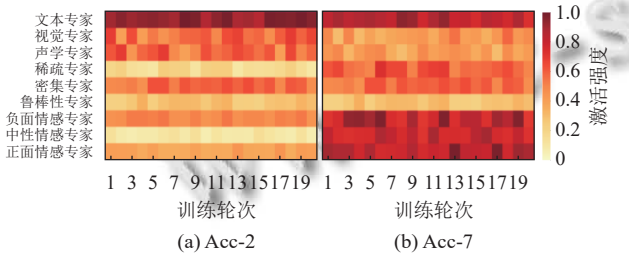

图 3 MoE 专家激活热力图

为检验分层门控机制的选择性与专家分工效果, 我们计算不同 MoE 架构的路由熵 (routing entropy) 进行对比实验. 表 2 显示, H-GEM 的路由熵仅为 0.95, 显著低于 CAG-MoE (1.48)、DenseFormer-MoE (1.45)、CuMo (1.23) 和 Metis-HOME (1.54), 表明其分层稀疏优化机制能够有效抑制冗余激活并实现稳定专家分工. 具体而言, CAG-MoE 通过专用子专家网络与稀疏门控, 有效融合异质情感信号, 但门控分布平滑, 易导致多专家同时被激活, 从而增加路由熵; DenseFormer-MoE 融合 DenseNet 与 Transformer 提取局部与全局特征, 采用带噪声 Top-*k* 门控动态分配专家, 多任务特征差异促使专家激活适度分散, 路由熵上升; CuMo 通过视觉 MoE 与预训练专家初始化结合辅助平衡损失提升多模态 LLM 能力, 推理激活参数低, 在一定程度上缓解了专家冗余的问题, 因此其路由熵仅次于本文模型. 但其复杂视觉特征与均衡负载导致路由熵仍有进一步优化空间; Metis-HOME 利用双分支 MoE 与动态路由器, 其高路由熵的结果源于多模态特征和专家负载平衡. 相比之下, H-GEM 通过分层门控在决策链上逐级筛选专家, 使语义抽取、模态融合与情感建模阶段形成稳定分工, 有效提升专家利用率并实现最低路由熵水平. 而 H-GEM 在 MAE (0.524) 和 *F*1 分数 (88.1%) 等性能指标上也表现优异. 综合结果表明, 分层门控、互信息一致性约束与批内判别性约束的设计互为补充, 不仅保持了模型表达能力, 还显著缓解了梯度竞争与专家重叠问题.

表 2 对照模型在 CMU-MOSEI 上的实验结果

模型	MAE	Acc-2 (%)	Acc-7 (%)	*F*1 (%)	路由熵
CAG-MoE	0.553	87.9	53.8	87.8	1.48
DenseFormer-MoE	—	—	—	—	1.45
CuMo	**0.496**	85.2	52.8	84.6	1.23
Metis-HOME	0.561	87.2	52.7	87.0	1.54
H-GEM	0.524	**88.2**	**55.1**	**88.1**	**0.95**

3.5 消融实验

为验证 H-GEM 的分层模型在多模态信息融合中的有效性, 本文在 MOSEI 数据集上进行了消融实验, 结果如表 3 所示.

表 3 H-GEM 在 MOSEI 上的消融实验结果

方法	Corr	MAE	Acc-2 (%)	Acc-7 (%)	*F*1 (%)
平均连接	0.864	0.597	88.5	51.9	88.5
固定权重	0.864	0.594	88.5	51.7	88.5
MoE (单层)	0.861	0.549	87.7	53.4	87.6
MoE (多层)	0.863	0.524	88.2	55.1	88.1

实验结果显示, 平均连接与固定权重基线表现相近 (Corr 均为 0.864, MAE 分别为 0.594 和 0.597), 说明简单融合效果有限. 单层 MoE 的 MAE 为 0.549, Acc-7 为 53.4%, 明显更优. 多层 MoE 进一步提升性能, Corr 达 0.863, MAE 降至 0.524, Acc-7 提升至 55.1%, *F*1 分数达 88.1%, 验证了分层专家架构在捕捉跨模态多极性情感特征上的有效性. 为验证鲁棒性专家的有效性, 我们在 MOSEI 数据集上进一步实验 (结果见

图 4). 文本部分缺失时激活强度下降, 视觉与声学专家相应上升, 以弥补语义信息的不足. 融合层鲁棒性专家权重同步增强, 提升模型适配性. 音频或视频模态缺失时同理, 保证判别能力稳定. 情感专家的正负向专家分布基本保持稳定, 仅中性专家激活略有提升, 以吸收语义不确定性. 实验结果表明, 分层门控 MoE 能够通过动态调节不同层次的专家激活分布, 实现对多模态缺失的鲁棒适应, 保持不完整输入下的情感识别性能.

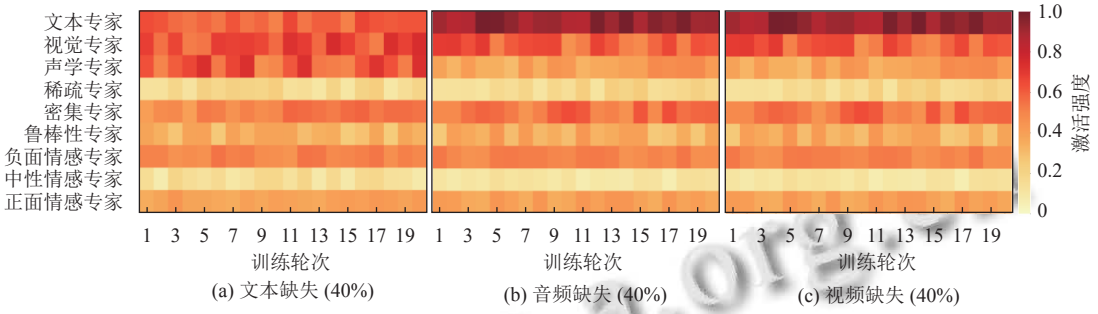

图 4　模态缺失情况下 MoE 专家激活热力图

为验证鲁棒性专家在模态缺失场景下的作用, 我们将本文模型在模态完整、文本缺失 40% 和文本缺失 40% 且移除鲁棒性专家的 3 种情况下进行消融实验.

图 5 显示, 当仅文本缺失 40% 时, 模型性能略有下降 (Corr 由 0.863 降至 0.825, MAE 由 0.524 上升至 0.554, Acc-2 由 88.2% 降至 84.3%, Acc-7 由 55.1% 降至 50.9%, *F*1 由 88.1% 降至 83.8%). 当同时移除鲁棒性专家时, 各项指标进一步下降 (Corr 为 0.803, MAE 为 0.569, Acc-2 为 82.1%, Acc-7 为 47.2%, *F*1 为 81.5%), 表明鲁棒性专家能够有效缓解模态缺失带来的负面影响, 提升多模态情感分析的稳定性和多类别预测能力.

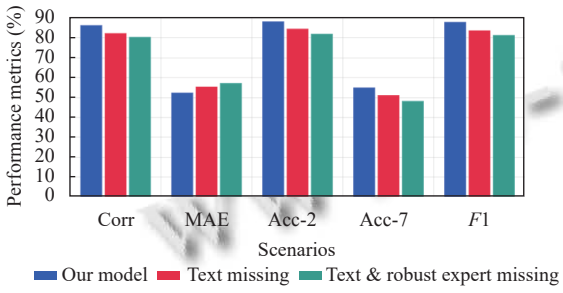

图 5　不同场景下性能比较柱状图

为评估模型在低资源场景的泛化能力, 我们在不同训练数据比例下进行了实验. 结果如表 4 所示.

从实验结果可以看出, 随着训练数据从 100% 降至 50%、20% 和 10%, 模型性能呈平滑下降: Corr 从 0.863 降至 0.782, MAE 从 0.524 上升至 0.664, Acc-2 从 88.2% 降至 78.9%, Acc-7 从 55.1% 降至 42.2%, *F*1 从 88.1% 降至 77.5%. 模型在低资源下仍能保持较强的情感预测能力, 数据量下降只导致整体精度和多类别分类能力轻微降低, 体现出稳健性和可泛化性.

表 4　H-GEM 在不同资源场景下的泛化实验结果

训练数据比例 (%)	Corr	MAE	Acc-2 (%)	Acc-7 (%)	*F*1 (%)
100	0.863	0.524	88.2	55.1	88.1
50	0.843	0.592	87.4	50.1	87.3
20	0.815	0.626	83.7	46.9	83.7
10	0.782	0.664	78.9	42.2	77.5

3.6　泛化实验

为验证其在跨领域数据集上的泛化性, 我们使用南加州大学的多模态情感识别数据集 IEMOCAP[26]进行额外的情绪识别实验. 该数据集涵盖 10 名演员的 302 个视频会话, 时长达 11 h. 数据集为每个语句提供离散情感类别及连续维度标签. 实验结果如表 5 所示.

表 5　H-GEM 在 IEMOCAP 上的泛化实验结果 (%)

模型	Happy		Sad		Angry		Neutral	
	Acc-2	*F*1	Acc-2	*F*1	Acc-2	*F*1	Acc-2	*F*1
BC-LSTM	83.1	81.7	82.1	81.7	85	84.2	66.1	64.1
MFN	90.2	85.8	88.4	86.1	87.5	86.7	72.1	68.1
RAVEN	87.3	85.8	83.4	83.1	87.3	86.7	69.7	69.3
MulT	90.7	88.6	86.7	86.0	87.4	87.0	72.4	70.7
本文模型	90.3	88.7	85.1	85.5	87.1	86.7	73.2	71.3

实验结果表明, 本文模型在 IEMOCAP 分类实验中的整体表现虽略低于部分针对分类任务优化的现有 SOTA 方法. 但在 Neutral 类别上取得了 73.2% 的准确率和 71.3% 的 *F*1 分数. 体现出模态模糊场景下的优势. 这与模型的设计初衷相关——分层 MoE 架构主要面向 MOSI、MOSEI 等连续型情感分析任务, 旨在建模情感强度的细粒度变化. 结果表明, 模型在跨域环境

中仍保持稳定表现, 并在 Neutral 等难分类别上实现突破, 验证了其在复杂分布差异下的适应能力与稳健性.

4 结论与展望

在本研究中, 我们提出了一种基于分层门控专家混合 (H-GEM) 模块的多模态情感分析模型, 通过构建 3 层专家体系与多层门控机制, 实现从模态特征提取到情感极性建模的递进式优化. 同时通过引入信息论与判别性约束, 显著增强了专家选择的语义区分性与稀疏性. 实验结果表明, H-GEM 多项评价指标均优于现有基线模型, 在细粒度情感分类与回归任务中表现突出的同时, 与单层 MoE 对比显示路由熵显著降低. 表明模型在保持高准确率之际能有效减少专家冗余激活. 此外, 模型在低资源训练和模态缺失场景下均展现出良好的鲁棒性与泛化能力, 验证了其在实际应用中的潜力. 未来的研究将致力于探索更精细的多模态特征融合策略. 例如通过改进信息引导的稀疏聚合机制或引入多种约束方法, 以进一步提升细粒度情感建模性能和专家选择效率.

参考文献

1 Shou YT, Meng T, Zhang FC, *et al*. Revisiting multi-modal emotion learning with broad state space models and probability-guidance fusion. arXiv:2404.17858, 2024.

2 Zhang YC, Zhong H, Alhusaini N, *et al*. Multilevel information compression and textual information enhancement for multimodal sentiment analysis. Knowledge-based Systems, 2025, 312: 113121. [doi: 10.1016/J.KNOSYS.2025.113121]

3 Liu MQ, Li ZX. A dissimilarity feature-driven decomposition network for multimodal sentiment analysis. Multimedia Systems, 2025, 31(1): 68. [doi: 10.1007/S00530-024-01660-X]

4 Subbaiah B, Murugesan K, Saravanan P, *et al*. An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network. Artificial Intelligence Review, 2024, 57(2): 34. [doi: 10.1007/S10462-023-10645-7]

5 Geethanjali R, Valarmathi A. A novel hybrid deep learning IChOA-CNN-LSTM model for modality-enriched and multilingual emotion recognition in social media. Scientific Reports, 2024, 14(1): 22270. [doi: 10.1038/s41598-024-73452-2]

6 Yang J, Xiao YL, Du X. Enhancing aspect-based sentiment analysis with multiple-knowledge promotion and multi-perspective noise filtering. Complex & Intelligent Systems, 2025, 11(9): 404. [doi: 10.1007/S40747-025-02034-0]

7 Yang L, Zhong JH, Wen T, *et al*. CCIN-SA: Composite cross modal interaction network with attention enhancement for multimodal sentiment analysis. Information Fusion, 2025, 123: 103230. [doi: 10.1016/j.inffus.2025.103230]

8 Tuerhong G, Fu FF, Wushouer M. Adaptive multimodal Transformer based on exchanging for multimodal sentiment analysis. Scientific Reports, 2025, 15(1): 27265. [doi: 10.1038/s41598-025-11848-4]

9 Li LJ, Wu ZY, Ji Y. MoTE: Mixture of task-specific experts for pre-trained model-based class-incremental learning. Knowledge-based Systems, 2025, 324: 113795. [doi: 10.1016/J.KNOSYS.2025.113795]

10 Li YX, Jiang SY, Hu BT, *et al*. Uni-MoE: Scaling unified multimodal LLMs with mixture of experts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(5): 3424–3439. [doi: 10.1109/TPAMI.2025.3532688]

11 Yuan XM, Zhang ZK, Liang PF, *et al*. A fusion TFDAN-based framework for rotating machinery fault diagnosis under noisy labels. Applied Acoustics, 2024, 219: 109940. [doi: 10.1016/J.APACOUST.2024.109940]

12 Liu Z, Shen Y, Lakshminarasimhan VB, *et al*. Efficient low-rank multimodal fusion with modality-specific factors. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 2247–2256. [doi: 10.18653/v1/P18-1209]

13 Tsai YHH, Bai SJ, Liang PP, *et al*. Multimodal Transformer for unaligned multimodal language sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 6558–6569. [doi: 10.18653/v1/P19-1656]

14 Hu JW, Liu YC, Zhao JM, *et al*. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, 2021. 5666–5675. [doi: 10.18653/v1/2021.acl-long.440]

15 Li XG, Ma YN, An XC, *et al*. Multi-level fusion with fine-

grained alignment for multimodal sentiment analysis. Journal of King Saud University Computer and Information Sciences, 2025, 37(5): 82. [doi: 10.1007/S44443-025-00094-3]

16 Verma B, Meel P, Vishwakarma KD. MHAM: A novel framework for multimodal sentiment analysis in memes. Knowledge and Information Systems, 2025, 67(11): 10355–10394. [doi: 10.1007/s10115-025-02535-x]

17 Aruna Gladys A, Vetriselvi V. Sentiment analysis on a low-resource language dataset using multimodal representation learning and cross-lingual transfer learning. Applied Soft Computing, 2024, 157: 111553. [doi: 10.1016/J.ASOC.2024.111553]

18 王楠, 王淇, 欧阳丹彤. 基于知识蒸馏与动态调整机制的多模态情感分析模型. 计算机学报, 2025, 48(8): 1923–1942. [doi: 10.11897/SP.J.1016.2025.01923]

19 Zhao BH, Kang XL, Zhou H, *et al*. Sparse-MoE-SAM: A lightweight framework integrating MoE and SAM with a sparse attention mechanism for plant disease segmentation in resource-constrained environments. Plants, 2025, 14(17): 2634. [doi: 10.3390/plants14172634]

20 Deng X, Pang YX, Zhao XC, *et al*. Stereo-talker: Audio-driven 3D human synthesis with prior-guided mixture-of-experts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025. [doi: 10.1109/TPAMI.2025.3596160]

21 Shi T, Hu J, Li DY, *et al*. JMoE-FAP: A novel model for telecom network fraud victimization pattern analysis. Journal of Safety Science and Resilience, 2025, 6(3): 100199. [doi: 10.1016/J.JNLSSR.2025.01.006]

22 Ding RZ, Lu H, Liu MH. DenseFormer-MoE: A dense Transformer foundation model with mixture of experts for multi-task brain image analysis. IEEE Transactions on Medical Imaging, 2025, 44(10): 4037–4048. [doi: 10.1109/TMI.2025.3551514]

23 Mengara AGM, Moon YK. CAG-MoE: Multimodal emotion recognition with cross-attention gated mixture of experts. Mathematics, 2025, 13(12): 1907. [doi: 10.3390/MATH13121907]

24 Zadeh A, Zellers R, Pincus E, *et al*. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems, 2016, 31(6): 82–88. [doi: 10.1109/MIS.2016.94]

25 Zadeh AAB, Liang PP, Poria S, *et al*. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 2236–2246. [doi: 10.18653/v1/P18-1208]

26 Busso C , Bulut M , Lee C , *et al*. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2009, 42(4): 335–359. [doi: 10.1007/s10579-008-9076-6]

(校对责编: 张重毅)