中图法分类号:TP18;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2026)02-0525-16

论文引用格式:Deng T S, Cai G Y, Dong K and Wang S J. 2026. Decoupled emotion dependencies and cross-modal awareness for emotion recognition in conversations. Journal of Image and Graphics, 31(2):0525-0540(邓天生, 蔡国永, 董凯, 王顺杰. 2026. 解耦情绪依赖关系的跨模态感知对话情绪识别. 中国图象图形学报, 31(2):0525-0540)[DOI:10.11834/jig.250309]

解耦情绪依赖关系的跨模态感知对话情绪识别

邓天生,蔡国永*,董凯,王顺杰

1. 桂林电子科技大学计算机与信息安全学院,桂林　541004; 2. 广西可信软件重点实验室,桂林　541004

摘　要:**目的**　对话情绪识别旨在准确捕捉对话中各话语的情绪状态,但面临两大挑战。其一,情绪状态随对话语境动态演变,且个体内部与说话者间存在结构异质的情绪依赖关系,增加了建模难度。其二,多模态信息融合时常出现语义错位和模态冗余,影响跨模态语义对齐与情绪线索的准确捕捉。**方法**　提出一种解耦情绪依赖关系的跨模态感知对话情绪识别模型(decoupled emotion dependencies and cross-modal awareness network,DECANet)。该模型通过结构解耦策略,将个体内部和说话者间的情绪依赖分别建模为两个独立子图,并设计启发式动态交互机制实现差异化建模与协同融合,精准捕捉多层次情绪演化特征。在多模态建模方面,设计跨模态上下文感知自注意力机制强化模态间深层语义关联,辅以语义一致性驱动的特征选择模块,有效筛选语义偏离和冗余信息,提升情绪表征的判别力。**结果**　在IEMOCAP(interactive emotional dyadic motion capture database)和MELD(multimodal emotion lines dataset)数据集上的实验结果表明,DECANet在准确率与加权F1值两个关键指标上均优于多种主流方法。在IEMOCAP上分别提升1.74%和1.77%;在MELD上亦取得了具有竞争力的性能增幅。消融实验进一步验证了情绪依赖解耦建模和跨模态交互机制的有效性。**结论**　DECANet能够更清晰地区分并建模对话中异构的情绪依赖结构,增强多模态语义对齐,减少信息冗余,在复杂交互场景中展现出较强的情绪识别能力和良好的泛化性能。

关键词:对话情绪识别(ERC);图注意力网络(GAT);情绪依赖图解耦与融合;多模态融合;跨模态交互机制

Decoupled emotion dependencies and cross-modal awareness for emotion recognition in conversations

Deng Tiansheng, Cai Guoyong*, Dong Kai, Wang Shunjie

1. College of Computer and Information Security, Guilin University of Electronic Technology, Guilin　541004, China;
2. Key Laboratory of Guangxi Trusted Software, Guilin　541004, China

Abstract: **Objective** Emotion recognition in conversations (ERC) aims to identify the emotional state of each utterance within dialogues. Unlike traditional emotion recognition, which typically classify emotions in isolated utterances, ERC must account for the dynamic evolution of emotions shaped by contextual cues throughout the dialogue. Two key emotional dependencies are crucial: intra-speaker dependencies, representing emotional continuity or variation within the same speaker, and inter-speaker dependencies, reflecting emotional influence between speakers. Properly modeling both dependencies is essential for tracking emotional flow and transitions over time. However, existing methods often conflate these

收稿日期:2025-07-16;**修回日期**:2025-09-02;**预印本日期**:2025-09-09

*通信作者:蔡国永　ccgycai@guet.edu.cn

基金项目:国家自然科学基金项目(62366010);广西自然科学基金项目(2024GXNSFAA010374);广西研究生教育创新计划项目(YCBZ2025158)

Supported by: National Natural Science Foundation of China (62366010); Natural Science Foundation of Guangxi, China (2024GXNSFAA010374); Innovation Project of Guangxi Education (YCBZ2025158)

structures, failing to distinguish their unique characteristics, thereby limiting their performance. Concurrently, ERC relies heavily on multimodal fusion of textual, acoustic, and visual cues to capture the full emotional context. Yet, current fusion strategies frequently suffer from semantic misalignment, noisy or conflicting modality signals, and inadequate discrimination of relevant information. These shortcomings result in suboptimal multimodal representations and impair recognition accuracy. Together, these challenges point to the need for a unified framework that can simultaneously disentangle emotional dependencies and enable robust cross-modal semantic integration. **Method** A decoupled emotion dependencies and cross-modal awareness network, named DECANet, is proposed for ERC to address the aforementioned challenges. The core idea of DECANet lies in the structural disentanglement and dynamic integration of emotional dependencies. Specifically, two distinct subgraphs are constructed to separately model intra-speaker and inter-speaker emotional relationships. The intra-speaker subgraph tracks emotional continuity or fluctuation within the same speaker's dialogue turn, while the inter-speaker subgraph captures emotion shifts triggered by interaction among different speakers. Each subgraph is processed using graph attention networks augmented with learnable relation-type embeddings. These embeddings encode various temporal and emotional relations, enabling more precise and context-sensitive propagation of affective cues across the graph structure. A heuristic dynamic interaction strategy is introduced to bridge these two emotional structures while preserving their independence. Drawing inspiration from Siamese architecture, the model concatenates original node features with the outputs of both subgraphs to preserve foundational semantic information and ensure representation alignment. Element-wise difference operations are used to quantify semantic divergence between representations, whereas element-wise multiplication captures their semantic consistency. These composite features are then passed through a gating mechanism, which adaptively learns fusion weights on the basis of the current conversational context. This selective integration strategy enhances emotional discrimination by emphasizing the more informative dependency pathway under different conditions. A context-aware self-attention mechanism is developed to further improve cross-modal semantic alignment. This component performs iterative refinement of audio and visual modality representations through sequential integration of contextual cues from other modalities. For example, audio features are initially refined via context-aware interaction with textual representations, followed by a second-stage fusion with visual features, conditioned on the updated audio-text representation. This sequential alignment strengthens inter-modal cohesion and reduces modality gaps. Additionally, speaker embeddings and positional embeddings are incorporated to capture structural cues inherent in multiturn, multispeaker dialogues. These embeddings help encode speaker-specific emotional tendencies and temporal structure within the conversation, facilitating context-aware emotion recognition. A semantic consistency-driven feature selection mechanism is employed to address the issue of semantic noise and inconsistencies often introduced during multimodal fusion. This mechanism selectively preserves only those multimodal representations that maintain high semantic similarity with their corresponding unimodal features. By filtering out semantically deviating or redundant signals, this process helps maintain the integrity of emotion-related information and enhances the robustness of the final emotion representations. Finally, the entire model is implemented under the supervision of multimodal and unimodal objectives, ensuring that the learned features retain discriminative capacity while maintaining robustness across modality configurations. **Result** DECANet was evaluated on two widely-used ERC benchmarks, IEMOCAP and MELD, consistently demonstrating competitive performance across multiple evaluation metrics. On IEMOCAP, it achieved improvements of 1.74% in accuracy and 1.77% in weighted F1 score. On MELD, it attained gains of 0.63% in accuracy and 0.52% in weighted F1. These results demonstrate the effectiveness and generalizability of DECANet across diverse conversational scenarios. Comprehensive ablation studies further validate the functional contribution of each proposed component. The removal of either the intra-speaker or inter-speaker subgraph leads to notable performance degradation, underscoring the complementary roles of the two emotional dependency structures. Comparative experiments with single-graph modeling, parallel fusion, and hierarchical fusion strategies reveal that the proposed heuristic dynamic interaction strategy consistently achieves superior results by enabling adaptive and context-sensitive subgraph integration. Furthermore, the cross-modal context-aware self-attention mechanism is essential for effective multimodal fusion. Replacing it with conventional cross-attention or removing speaker and positional embeddings significantly reduces model performance, highlighting the importance of fine-grained inter-modal alignment and contextual sensitivity. In addition, the semantic consistency-based feature selection module effectively filters out semantically incon-

sistent signals, thereby enhancing the robustness and discriminative quality of the resulting multimodal representations. **Conclusion** DECANet effectively disentangles and models intra- and inter-speaker emotional dependencies in multispeaker conversations. It further achieves fine-grained semantic alignment across modalities through an iterative and selective fusion mechanism, enhancing the robustness and accuracy of multimodal emotion recognition. Extensive experiments verify that DECANet offers a generalizable and interpretable solution for understanding emotions in complex dialogue scenarios.

Key words: emotion recognition in conversations(ERC); graph attention network(GAT); emotion dependency graph disentanglement and fusion; multimodal fusion; cross-modal interaction mechanism

0 引　言

情感是人类与生俱来的核心特质，不仅反映个体的潜在认知过程，还深刻影响其认知判断与决策行为。情感识别作为理解人类意图与行为的重要支撑技术，近年来在自然语言处理、认知计算及人机交互等多个领域得到了广泛关注。尤其随着对话式人工智能的迅猛发展，对话情绪识别(emotion recognition in conversations，ERC)逐渐成为研究的前沿热点(Bhat 和 Modi，2023；Tao 等，2024)。准确识别对话中细微的情绪变化对于实现更深层次的语义理解至关重要，已在智能客服、舆情监测、心理健康与医疗诊断等实际应用场景中展现出广泛价值。

相较于传统针对独立语句的情绪识别任务，ERC需要充分建模对话上下文，以捕捉情绪随时间演化和话语交互所形成的动态特征。具体而言，ERC任务不仅要求建模语句之间的时序依赖，还需同时关注个体内部与说话人间的情绪关联，从而理解情绪的传播路径和演变机制。其中，说话人内部的情绪依赖体现了个体情绪在连续对话中呈现的惯性演化和波动趋势，而说话人间的情绪依赖则揭示了不同个体情绪状态间的交互影响与传播规律(Hazarika 等，2018；Ishiwatari 等，2020；Zhang 和 Chai，2021；Zhao 等，2022)。这两类情绪依赖在对话中相互交织，共同驱动情绪变化，因此，精细区分并建模其结构特性，对于提升情绪表征的层次性与建模精度至关重要。当前情绪依赖建模方法主要包括序列结构和图结构两类。在序列结构建模方面，Majumder 等人(2019)通过构建全局 GRU(gated recurrent unit)、角色 GRU 和情绪 GRU 3 种机制，分别捕捉跨说话人的上下文、个体情绪演化及话语级情绪表示，从而增强情绪预测能力。Ghosal 等人(2020)则引入常识知识库 ATOMIC(an atlas of everyday commonsense reasoning)(Sap 等，2019)，通过建模个体属性、事件、心理状态、意图与情绪之间的因果关系，提升对情绪演变的因果推理能力。在图结构建模方面，Joshi 等人(2022)提出全局—局部联合建模策略，利用 Transformer 编码全局对话上下文保持语义完整性，同时借助 RGCN(relational graph convolutional network)与 GraphTransformer 建模局部情绪依赖，以挖掘更细粒度的情绪驱动关系。Zhang 等人(2023)通过语篇感知图注意力网络(graph attention network，GAT)建模话语间结构依赖，通过说话者感知 GAT 刻画角色间的语义互动，最终通过双图注意力机制实现协同融合。Ai 等人(2025)针对说话者情绪易受上下文与外部事件双重驱动的特点，使用 Doc2EDAG(end-to-end document-level framework for Chinese financial event extraction)模型(Zheng 等，2021)进行事件抽取并建模情绪驱动因素，从而缓解因情绪表达隐晦或缺乏显性语义线索带来的信息缺失问题。尽管上述方法在提升 ERC 性能方面取得积极进展，情绪依赖建模仍面临两大挑战：1)情绪状态的动态变化尚未被充分刻画，尤其是在复杂对话情境下的非线性演化趋势；2)长距离语义交互的建模能力仍显不足。序列结构模型在长距离建模中容易出现信息衰减，而图结构模型虽然缓解了该问题，却通常采用统一的图结构对整段对话进行建模，未能区分说话人内部与说话人间的情绪依赖，忽略了它们在结构与语义上的根本差异，如个体情绪惯性演化与群体情绪传播路径。这种混合建模策略容易引发信息干扰与结构模糊，削弱模型对情绪传播机制的辨别能力与可解释性。因此，针对这两类情绪依赖结构解耦的建模策略将为模型提供更具辨别能力的语义基础。

另外，多模态融合是 ERC 任务中的关键研究方

向，其核心目标是有效整合文本、音频和视觉等多种模态信息，从而提升情绪识别的准确性和鲁棒性。在真实对话场景中，情绪往往通过文本、语音语调和面部表情等多种模态信号共同传达。单一模态难以全面捕捉情绪的复杂表征（Li等，2022）。然而，由于不同模态数据在表示空间、信息密度、表达形式及语义结构上的显著异质性，模态间往往存在信息冗余、表达冲突与结构不一致等问题，严重制约了多模态协同建模的有效性与稳定性。因此，如何缓解模态间的异质性，充分挖掘其潜在互补性，并实现协调高效的深度融合机制，已成为当前亟待突破的核心挑战。为应对上述问题，近年来研究者提出多种融合策略，致力于提升多模态情绪识别的建模能力与表达性能。例如，Mao等人（2021）采用层次化Transformer结构捕捉模态内的情绪动态特征，同时引入多粒度融合策略，建模模态间的原型依赖与表示依赖关系，从而提升跨模态交互的精准性。此外，Hu等人（2022）则结合图卷积与门控机制，在多模态聚合过程中对模态内与模态间的上下文信息进行选择性过滤，以减少冗余并增强模态间的有效交互。Shi和Huang（2023）引入双向多头交叉注意力机制强化文本、音频和视觉特征的跨模态关联，并使用Soft-HGR（soft Hirschfeld-Gebelein-Renyi）损失以最大化多模态特征间的相关性。Li等人（2024）构建多模态异质有向图，通过跨模态配对补全策略保持特征的一致性与多样性，从而提升多模态融合效果。Meng等人（2024）进一步提出基于图的掩蔽学习与递归对齐机制，利用GAT并结合递归记忆增强的跨模态注意力机制实现单模态特征对齐，去除模态噪声，最终通过图卷积网络（graph convolutional network，GCN）完成多模态融合，生成鲁棒性更强的情绪表示。尽管上述方法在多模态ERC任务中取得了显著进展，但仍存在一定局限性：多数现有方法在融合阶段主要聚焦于模态特征的整体对齐与整合，缺乏对不同模态间语义依赖的细粒度建模，导致融合过程中易引入语义歧义与结构偏差，模态间的协同效率与信息利用率仍有待提升。

为应对上述挑战，本文提出一种解耦情绪依赖关系的跨模态感知对话情绪识别模型（decoupled emotion dependencies and cross-modal awareness network，DECANet）。该模型构建多视角对话图，将个体内部与说话人间的情绪依赖关系解耦为两个独立子图，分别建模不同层次的情绪传播机制。同时，引入可学习的关系类型嵌入，以编码多样的时间和情绪关系，增强边的语义表达能力；并设计启发式动态交互策略，能够同时捕捉两类依赖的动态演化过程，从而更精准地反映情绪传播机制。在多模态建模方面，DECANet融合文本、音频和视觉3种模态特征，提出跨模态上下文感知自注意力机制，用于学习模态间的深层语义关联；并结合语义一致性驱动的特征选择模块，对融合过程中的冗余信息与语义冲突进行过滤，优化最终的情绪表示。

本文的主要贡献包括：1）提出基于子图的动态交互策略，分别建模个体内部情绪依赖与说话人间情绪依赖，精确刻画对话中的情绪传播与演化模式，避免混合建模带来的信息混淆，提升情绪识别的准确性；2）设计跨模态上下文感知自注意力机制，增强不同模态间的深层语义对齐能力，同时引入语义一致性驱动的特征选择模块，有效缓解多模态融合过程中产生的语义歧义与模态冲突，从而增强多模态特征的融合效果与表示鲁棒性；3）在IEMOCAP和MELD数据集上进行广泛的实验，结果验证了所提方法的有效性与鲁棒性。

1 方法描述

1.1 问题定义

在ERC任务中，一段对话可以形式化地表示为一个话语序列$\boldsymbol{U}=\{u_1,u_2,\cdots,u_N\}$与对应的说话人序列$\boldsymbol{S}=\{s_1,s_2,\cdots,s_M\}$。其中，$N$表示话语数量，$M$表示说话人数，两者之间存在映射关系$[u_i;\ s_j]$，即第$i$条话语由第$j$位说话人发出。每句话语由3种模态组成，表示为$\boldsymbol{H}=[\boldsymbol{H}_t,\boldsymbol{H}_a,\boldsymbol{H}_v]$。其中，$t$、$a$和$v$分别表示文本、音频和视觉信息。对于任意模态$m\in\{t,a,v\}$，其原始特征为

$$\boldsymbol{H}_m=[\boldsymbol{h}_m^1,\boldsymbol{h}_m^2,\cdots,\boldsymbol{h}_m^N],\quad \boldsymbol{h}_m^i\in\mathbf{R}^{d_m} \tag{1}$$

式中，$\boldsymbol{h}_m^i$表示第i条话语在模态m下的原始特征，d_m是模态m的原始特征维度。ERC任务的目标是准确预测每条话语u_i的情绪标签。

由于对话中的情绪状态不仅由当前话语所表达的内容决定，还受到上下文中情绪演化轨迹的影响。为更全面地建模对话中的情绪动态变化，本文提出

DECANet。DECANet 主要由 4 个核心模块组成：多模态特征提取模块、情绪依赖子图动态交互模块、跨模态上下文感知融合模块和情绪分类模块，模型结构如图 1 所示。

1.2 多模态特征提取模块

1.2.1 文本模态

借鉴 COSMIC（commonsense knowledge for emotion identification in conversations）（Ghosal 等，2020）的做法，采用预训练语言模型 RoBERTa-Large 对文本模态进行特征编码。具体而言，对 RoBERTa 进行任务相关微调后，提取其最后一层中[CLS]标记对应的隐藏状态作为每条话语的文本表示，特征维度为 1 024。

1.2.2 音频模态

参考 MMGCN（multimodal fusion via deep graph convolution network for emotion recognition in conversation）（Hu 等，2021）的处理方式，本文使用加载 IS10 配置的 OpenSMILE 工具包对原始语音信号进行低级特征提取。考虑不同数据集音频分布差异，分别将 IEMOCAP（interactive emotional dyadic motion capture database）和 MELD（multimodal emotion lines dataset）数据集的特征维度设为 1 582 和 300，以更好地适应各自的模态特性。

1.2.3 视觉模态

采用在 FER+（facial expression recognition plus）数据集上预训练的 DenseNet 网络作为特征编码器，对每帧视频图像提取静态面部表征。经处理后，每条话语对应的视觉特征维度设定为 342。

由于 3 种模态的原始特征维度不一致，为每种模态设计一个独立的线性变换层，用于统一映射至相同维度的公共空间表示。具体为

$$\boldsymbol{H}_m^{\text{in}} = Linear(\boldsymbol{H}_m) \in \mathbf{R}^{N\times d} \tag{2}$$

1.3 情绪依赖子图动态交互模块

1.3.1 情绪依赖子图构建

DECANet 将一段对话表示为图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$，其中每个话语对应一个节点 $v_i \in \boldsymbol{V}$，边集 $\boldsymbol{E}$ 表示话语之间的情绪依赖关系。为细化情绪传播建模，整体图分为两类子图：说话人内部情绪依赖子图 $\boldsymbol{G}_{\text{intra}}$ 和说话人间情绪依赖子图 $\boldsymbol{G}_{\text{inter}}$。每个子图中的节点特征为其对应话语的文本模态表示，边则包含关系类型与

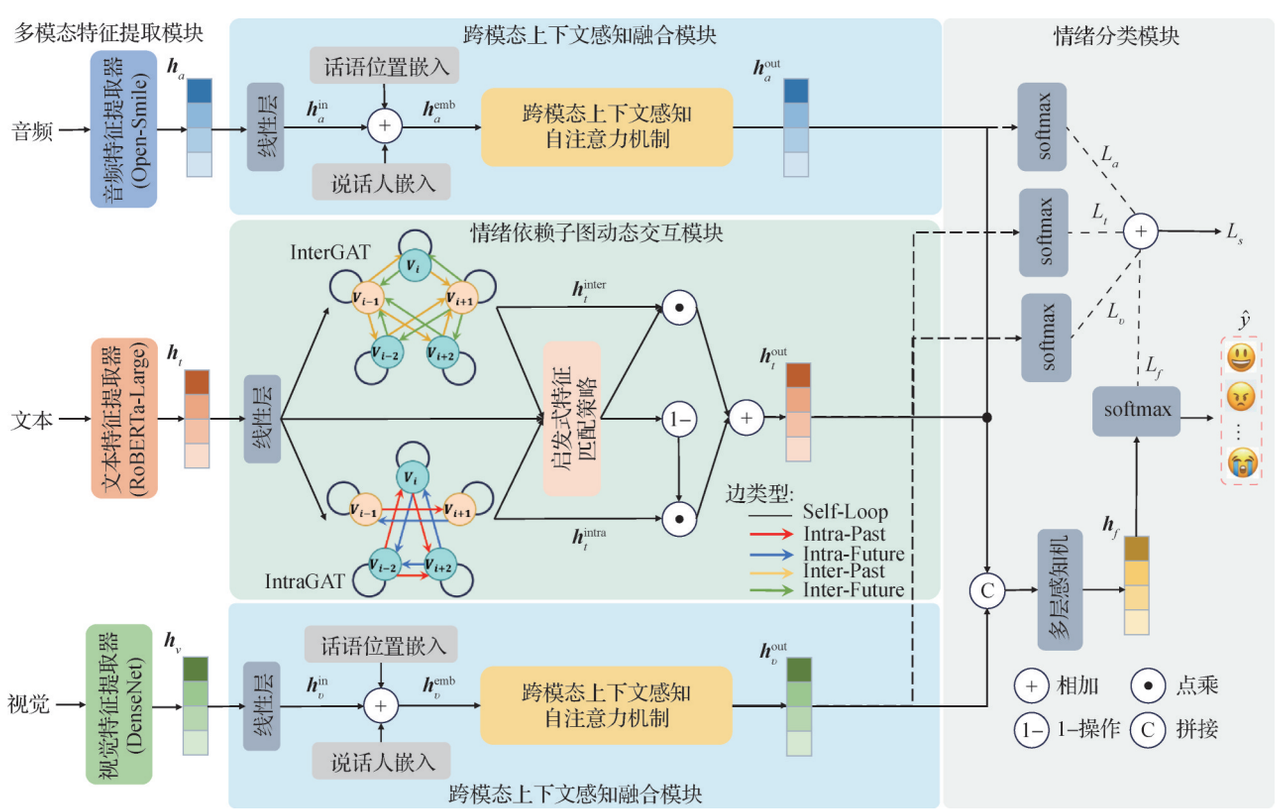

图 1 DECANet 总体架构

Fig. 1 Overall architecture of DECANet

权重两类属性，用于捕捉不同语义结构及关系强度。

考虑到情绪演化既受时间顺序影响，又因说话人角色差异呈现复杂依赖，DECANet在构图阶段引入5类边，以精细建模多维依赖关系。如图1所示，不同说话人用不同颜色的圆圈区分：Intra-Past（红色箭头）与Intra-Future（蓝色箭头）分别表示当前话语与同一说话人的历史话语和未来话语的依赖，用于刻画个体内部情绪的延续与转变；Inter-Past（黄色箭头）与Inter-Future（绿色箭头）分别表示当前话语与其他说话人的历史话语和未来话语的依赖，用于刻画跨角色情绪传播与交互；Self-Loop（黑色曲线）表示话语对自身状态的保持与反馈，以保证局部信息不丢失，并增强局部节点特征在复杂图结构中的稳定性。

为控制图结构的信息传播范围，引入滑动窗口k，仅保留时间范围内的依赖连接，避免远距离连接带来的冗余干扰。

1.3.2 情绪依赖子图动态交互策略

为实现边权重的动态学习，在GAT基础上，本文引入关系类型嵌入，对不同话语间的情绪依赖关系进行建模，从而增强情绪信息在图结构中的传播与聚合能力。具体而言，节点v_i与其相邻节点v_j之间存在边关系r_{ij}，则首先通过可学习的关系嵌入矩阵$\boldsymbol{R}_{\text{embedding}} \in \mathbf{R}^{N_r \times d_r}$将one-hot编码后的关系类型$\boldsymbol{\mathcal{O}}(r_{ij}) \in \mathbf{R}^{N_r}$映射为边类型表示。其中，$N_r$为关系类型的数目，$d_r$为关系类型嵌入的维度。

$$\boldsymbol{e}_{ij} = \boldsymbol{\mathcal{O}}(r_{ij}) \times \boldsymbol{R}_{\text{embedding}} \in \mathbf{R}^{d_r} \tag{3}$$

在此基础上，对任意一对中心节点v_i与邻接节点v_j，其注意力权重的计算式为

$$\boldsymbol{\alpha}_{ij} = f_{\text{softmax}}\left(f_{\text{LeakyReLU}}\left(\boldsymbol{w}^{\text{T}}[\boldsymbol{W}\boldsymbol{h}_t^i \| \boldsymbol{W}\boldsymbol{h}_t^j \| \boldsymbol{e}_{ij}]\right)\right) \tag{4}$$

式中，$\boldsymbol{h}_t^i$，$\boldsymbol{h}_t^j$分别表示节点v_i与邻接节点v_j的文本特征表示，$\boldsymbol{W}$为共享的线性变换矩阵，$\boldsymbol{w}$为可训练的注意力参数，$\|$表示向量的拼接操作，$f_{\text{LeakyReLU}}$表示非线性激活函数。关系类型嵌入$\boldsymbol{e}_{ij}$的引入有助于显式建模不同语义结构下的话语依赖差异，提升模型对边关系语义的表达能力。节点v_i表示通过聚合其邻居节点的加权信息获得，具体计算过程为

$$\boldsymbol{h}_t^{\text{out},i} = \sigma\left(\sum_{j \in \boldsymbol{N}_i} \boldsymbol{\alpha}_{ij} \boldsymbol{W}\boldsymbol{h}_t^j\right) \in \mathbf{R}^d \tag{5}$$

式中，σ表示非线性激活函数，$\boldsymbol{N}_i$为节点v_i的邻居集合。基于上述机制，情绪依赖子图分别通过两组独立的GAT模块进行更新。具体为

$$\boldsymbol{H}_t^{\text{intra}} = f_{\text{intraGAT}}(\boldsymbol{H}_t^{\text{in}}, \boldsymbol{E}^{\text{intra}}) \tag{6}$$

$$\boldsymbol{H}_t^{\text{inter}} = f_{\text{interGAT}}(\boldsymbol{H}_t^{\text{in}}, \boldsymbol{E}^{\text{inter}}) \tag{7}$$

式中，$\boldsymbol{H}_t^{\text{in}}$表示输入文本特征表示，$\boldsymbol{E}^{\text{intra}}$和$\boldsymbol{E}^{\text{inter}}$分别表示说话人内部和说话人间的边集合。为充分挖掘说话人内部与说话人间情绪依赖子图的结构互补性，本文引入启发式门控机制以实现两类情绪依赖表示的动态融合。该机制借鉴“孪生网络”架构的思想，通过启发式特征匹配策略估计信息的保留比例。其核心动机在于：不同类型的情绪依赖具有语义差异，统一融合时若处理不当可能引入噪声，因此需通过对依赖表示与当前语义输入之间的差异性与交互性进行建模，引导模型自适应调整融合权重，从而提升整体情绪建模能力与鲁棒性。

具体而言，首先对当前节点的原始输入表示与说话人内部依赖表示进行拼接，并计算它们的元素级差值与乘积，分别衡量二者之间的语义差异性与交互强度。具体为

$$\bar{\boldsymbol{H}}_t = ReLU\left(FC\left(\left[\boldsymbol{H}_t^{\text{in}} \| \boldsymbol{H}_t^{\text{intra}} \| \boldsymbol{H}_t^{\text{in}} - \boldsymbol{H}_t^{\text{intra}} \| \boldsymbol{H}_t^{\text{in}} \odot \boldsymbol{H}_t^{\text{intra}}\right]\right)\right) \tag{8}$$

式中，FC（fully connected layer）表示全连接层，$ReLU$（rectified linear unit）为非线性激活函数，$\odot$为元素级乘法。类似地，原始输入表示与说话人间的依赖表示$\boldsymbol{H}_t^{\text{inter}}$也进行同样操作，得到另一路语义融合表示。即

$$\hat{\boldsymbol{H}}_t = ReLU\left(FC\left(\left[\boldsymbol{H}_t^{\text{in}} \| \boldsymbol{H}_t^{\text{inter}} \| \boldsymbol{H}_t^{\text{in}} - \boldsymbol{H}_t^{\text{inter}} \| \boldsymbol{H}_t^{\text{in}} \odot \boldsymbol{H}_t^{\text{inter}}\right]\right)\right) \tag{9}$$

在上述设计中，拼接操作引入当前节点的原始输入表示$\boldsymbol{H}_t^{\text{in}}$，其核心目的是保留节点自身的基础语义信息，为情绪判别提供稳定的语义参考，避免依赖表示在图结构传播过程中产生的语义偏移对融合判断产生干扰。同时，将依赖表示（如$\boldsymbol{H}_t^{\text{intra}}$和$\boldsymbol{H}_t^{\text{inter}}$）与原始输入拼接，不仅为后续的差值项与乘积项提供了特征对齐基础，还增强了模型对语义差异与结构共性的感知能力。具体而言，差值项刻画了原始输入表示与情绪依赖表示之间的语义距离，反映它们在特定语义维度上的偏离程度，从而度量信息结构上的“差异性”；而乘积项则表达了两者在各维度上的协同激活程度，用于捕捉语义上的“相似性”与潜在关联强度。通过这种组合方式构建的融合表达具

有更高的语义容量与判别性，能够使后续的全连接层在保留局部语义特征的同时，学习出更为精准的融合权重策略。随后，将两路融合表示拼接，通过全连接层与 sigmoid 激活函数生成融合门控权重 P。具体为

$$P = f_{\text{sigmoid}}\left(FC\left(\left[\bar{\boldsymbol{H}}_t \| \hat{\boldsymbol{H}}_t\right]\right)\right) \tag{10}$$

最终，依据该权重动态调整两类依赖特征在融合表示中的占比，实现情绪信息的自适应整合。具体为

$$\boldsymbol{H}_t^{\text{out}} = P \odot \boldsymbol{H}_t^{\text{intra}} + (1 - P) \odot \boldsymbol{H}_t^{\text{inter}} \tag{11}$$

上述启发式门控机制不仅保持了依赖子图在建模不同情绪传播机制时的结构优势，还通过精细的语义对齐与关联建模，实现了更具选择性的特征融合策略，从而有效提升模型对情绪演化过程的表达能力与对话情绪识别任务中的鲁棒性。

1.4 跨模态上下文感知融合模块

1.4.1 说话人和话语位置嵌入

为增强音频与视觉模态中对说话人身份的建模能力，本文将说话人嵌入显示引入音频与视觉模态的特征表示中。具体地，给定话语 u_i 的说话人为 s_j，通过如下方式生成说话人嵌入。具体为

$$\boldsymbol{spk}_{u_i} = \mathcal{O}(s_j) \times \boldsymbol{S}_{\text{embedding}} \in \mathbf{R}^{d_{\text{spk}}} \tag{12}$$

$$\boldsymbol{SE} = [\boldsymbol{spk}_{u_1}, \boldsymbol{spk}_{u_2}, \cdots, \boldsymbol{spk}_{u_N}] \tag{13}$$

式中，$\boldsymbol{S}_{\text{embedding}} \in \mathbf{R}^{M \times d}$ 是可学习的说话人嵌入矩阵，$\mathcal{O}(s_j) \in \mathbf{R}^M$ 表示说话人标签 s_j 经过独热编码后的向量，M 为说话人数，d_{spk} 为嵌入维度。最终得到的 $\boldsymbol{spk}_{u_i}$ 表示第 i 条话语的说话人嵌入，$\boldsymbol{SE}$ 为整段对话的说话人嵌入集合。

此外，为了进一步捕捉对话中的时序结构信息，本文引入基于正余弦函数构建的位置嵌入，具体为

$$PE(pos, 2i) = \sin\left(\frac{pos}{10\,000^{\frac{2i}{d}}}\right) \tag{14}$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10\,000^{\frac{2i}{d}}}\right) \tag{15}$$

式中，pos 表示话语在对话序列中的索引位置，i 是话语特征的维度索引。该嵌入方式可将绝对位置信息编码进模型中，有助于模型理解话语之间的顺序与相对位置。最后，通过将统一维度的音频或视觉特征、说话人嵌入与位置嵌入相加，用于后续建模。具体为

$$\boldsymbol{H}_m^{\text{emb}} = \boldsymbol{H}_m^{\text{in}} + \boldsymbol{SE} + \boldsymbol{PE} \in \mathbf{R}^{N \times d}, m \in \{a, v\} \tag{16}$$

式中，$\boldsymbol{H}_m^{\text{in}}$ 表示模态 m（音频或视觉）的输入特征，$\boldsymbol{H}_m^{\text{emb}}$ 为融合中用于后续建模的多模态输入表示。

1.4.2 跨模态上下文感知自注意力机制

传统的跨模态交互通常依赖交叉注意力机制，以促进模态间的信息融合。以文本与音频的交互为例，常以文本模态作为查询（Query），音频模态作为键（Key）和值（Value），实现跨模态特征提取。然而，由于模态间嵌入空间的异质性差异较大，这种直接交互方式往往无法充分捕捉语义关联，容易引入冗余或噪声信息，进而削弱特征的判别能力和模型的泛化性能。为缓解上述问题，提出跨模态上下文感知自注意力机制，其结构如图 2 所示。

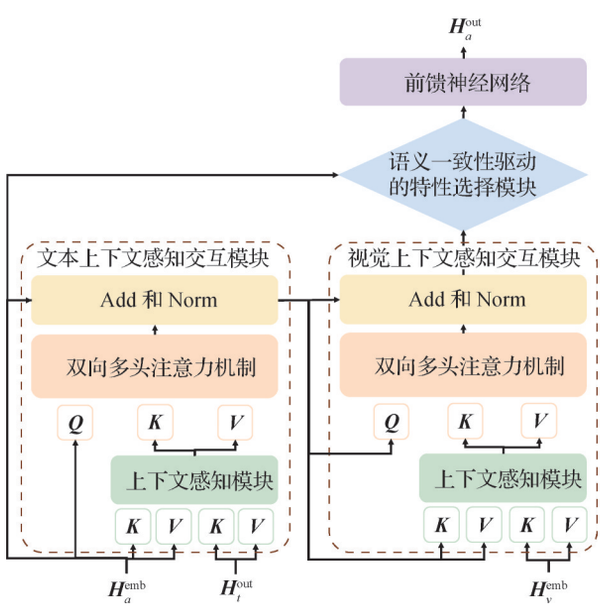

图 2 跨模态上下文感知自注意力机制

Fig. 2 The details of the cross-modal context-aware self-attention mechanism

由于音频和视觉的模态融合模块具有相同的结构，仅在 Query、Key 和 Value 的来源上有所区别。下文以音频模态为例说明该机制的整体流程。首先，通过线性映射将音频模态嵌入 $\boldsymbol{H}_a^{\text{emb}}$ 投影为初始的 Query、Key 与 Value。即

$$\begin{bmatrix} \boldsymbol{Q} \\ \boldsymbol{K} \\ \boldsymbol{V} \end{bmatrix} = \boldsymbol{H}_a^{\text{emb}} \begin{bmatrix} \boldsymbol{W}_Q \\ \boldsymbol{W}_K \\ \boldsymbol{W}_V \end{bmatrix} \tag{17}$$

为增强模态间上下文感知能力，进一步引入来自文本模态的补充信息，生成模态上下文感知的 Key 和 Value。具体为

$$\begin{bmatrix}\hat{\boldsymbol{K}}_a\\ \hat{\boldsymbol{V}}_a\end{bmatrix}=\left(1-\begin{bmatrix}\lambda_K\\ \lambda_V\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{K}\\ \boldsymbol{V}\end{bmatrix}+\begin{bmatrix}\lambda_K\\ \lambda_V\end{bmatrix}\left(\boldsymbol{H}_t^{\text{out}}\begin{bmatrix}U_K^t\\ U_V^t\end{bmatrix}\right) \tag{18}$$

式中，U_K^t和U_V^t是可训练参数，λ_K和λ_V为模态间融合的权重系数，用以控制跨模态信息的引入程度。不同于设定固定的超参数，本文设计门控机制自适应地学习该权重。具体为

$$\begin{bmatrix}\lambda_K\\ \lambda_V\end{bmatrix}=\sigma\left(\begin{bmatrix}\boldsymbol{K}\\ \boldsymbol{V}\end{bmatrix}\begin{bmatrix}W_k^a\\ W_v^a\end{bmatrix}+\boldsymbol{H}_t^{\text{out}}\begin{bmatrix}U_K^t\\ U_V^t\end{bmatrix}\begin{bmatrix}W_k^t\\ W_v^t\end{bmatrix}\right) \tag{19}$$

式中，W_k^a，W_v^a，W_k^t和W_v^t均为可训练参数。在获得模态感知的Key和Value后，使用标准的多头注意力机制进行跨模态建模，并结合残差连接与层归一化。具体为

$$\boldsymbol{H}_{at}=f_{\text{LayerNorm}}\left(f_{\text{softmax}}\left(\frac{\boldsymbol{Q}(\hat{\boldsymbol{K}}_a)^{\text{T}}}{\sqrt{d}}\right)\hat{\boldsymbol{V}}_a+\boldsymbol{H}_a^{\text{emb}}\right) \tag{20}$$

随后，将上述融合后的$\boldsymbol{H}_{at}$作为新一层的Query，进一步引入视觉模态信息，重复上述步骤，可得

$$\begin{bmatrix}\boldsymbol{Q}'\\ \boldsymbol{K}'\\ \boldsymbol{V}'\end{bmatrix}=\boldsymbol{H}_{at}\begin{bmatrix}\boldsymbol{W}_Q'\\ \boldsymbol{W}_K'\\ \boldsymbol{W}_V'\end{bmatrix} \tag{21}$$

$$\begin{bmatrix}\hat{\boldsymbol{K}}_{at}\\ \hat{\boldsymbol{V}}_{at}\end{bmatrix}=\left(1-\begin{bmatrix}\lambda_K'\\ \lambda_V'\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{K}'\\ \boldsymbol{V}'\end{bmatrix}+\begin{bmatrix}\lambda_K'\\ \lambda_V'\end{bmatrix}\left(\boldsymbol{H}_v^{\text{emb}}\begin{bmatrix}U_K^v\\ U_V^v\end{bmatrix}\right) \tag{22}$$

$$\begin{bmatrix}\lambda_K'\\ \lambda_V'\end{bmatrix}=\sigma\left(\begin{bmatrix}\boldsymbol{K}'\\ \boldsymbol{V}'\end{bmatrix}\begin{bmatrix}W_k^{at}\\ W_v^{at}\end{bmatrix}+H_v^{\text{emb}}\begin{bmatrix}U_K^v\\ U_V^v\end{bmatrix}\begin{bmatrix}W_k^v\\ W_v^v\end{bmatrix}\right) \tag{23}$$

$$\boldsymbol{H}_{\text{atv}}=f_{\text{LayerNorm}}\left(f_{\text{softmax}}\left(\frac{\boldsymbol{Q}'(\hat{\boldsymbol{K}}_{at})^{\text{T}}}{\sqrt{d}}\right)\hat{\boldsymbol{V}}_{at}+\boldsymbol{H}_{at}\right) \tag{24}$$

通过上述分阶段的感知融合，模型能够逐步整合文本、音频与视觉模态中的上下文信息，提升跨模态表示的上下文感知能力、判别能力与鲁棒性，为下游的情绪识别任务提供更具表现力的多模态表示。

为缓解多模态融合过程中产生语义偏移问题，本文设计了一种基于语义一致性的特征选择机制。该机制基于如下假设：有效的增强特征应在提升上下文表达能力的同时，最大程度保留原始模态的语义信息。具体而言，通过计算增强特征与原始特征之间的余弦相似度作为语义一致性指标。即

$$\boldsymbol{H}'_{\text{atv}}=\begin{cases}\boldsymbol{H}_{\text{atv}} & \dfrac{\boldsymbol{H}_{\text{atv}}\cdot\boldsymbol{H}_a^{\text{emb}}}{\|\boldsymbol{H}_{\text{atv}}\|\ \|\boldsymbol{H}_a^{\text{emb}}\|}\geqslant limit\\ \boldsymbol{H}_a^{\text{emb}} & \text{其他}\end{cases} \tag{25}$$

式中，*limit*为预设的阈值。当二者之间的余弦相似度超过该阈值时，认为增强过程在丰富上下文信息的同时保留了原始语义，因此保留增强特征；反之，则认为增强可能引入了模态冲突或语义漂移，回退至原始特征。

最后，为了提高模型的表达能力和稳定性，将$\boldsymbol{H}'_{\text{atv}}$输入到两层前馈网络中，得到最后的输出$\boldsymbol{H}_a^{\text{out}}$。具体为

$$\boldsymbol{H}''_{\text{atv}}=ReLU(\boldsymbol{H}'_{\text{atv}}\boldsymbol{W}_1+b_1) \tag{26}$$

$$\boldsymbol{H}_a^{\text{out}}=f_{\text{LayerNorm}}(\boldsymbol{H}''_{\text{atv}}\boldsymbol{W}_2+b_2+\boldsymbol{H}_{\text{atv}}) \tag{27}$$

式中，$\boldsymbol{W}_1$，$\boldsymbol{W}_2$和b_1，b_2分别为前馈网络的权重矩阵和偏置项。

1.5 情绪分类器

在完成情绪依赖建模与多模态融合后，DECANet将学习到的文本$\boldsymbol{H}_t^{\text{out}}$、音频$\boldsymbol{H}_a^{\text{out}}$和视觉$\boldsymbol{H}_v^{\text{out}}$表示进行拼接，然后送入两层感知机中。最后，使用softmax计算情绪类别上的概率分布，其中具有最高概率的情绪标签被选择作为话语u_i的预测标签，具体计算过程为

$$\boldsymbol{H}_f=\boldsymbol{W}_3[\boldsymbol{H}_t^{\text{out}}\|\boldsymbol{H}_a^{\text{out}}\|\boldsymbol{H}_v^{\text{out}}]+b_3 \tag{28}$$

$$\boldsymbol{H}_f'=ReLU(\boldsymbol{H}_f\boldsymbol{W}_4+b_4) \tag{29}$$

$$\hat{\boldsymbol{Y}}_f=f_{\text{softmax}}(\boldsymbol{H}_f')=[\hat{\boldsymbol{y}}_f^1;\hat{\boldsymbol{y}}_f^2;\cdots;\hat{\boldsymbol{y}}_f^N]\in\mathbf{R}^{N\times c} \tag{30}$$

式中，$\hat{\boldsymbol{Y}}_f\in\mathbf{R}^{N\times c}$表示整段对话的所有$N$条话语的情绪预测概率分布，$\hat{\boldsymbol{y}}_f^i\in\mathbf{R}^c$是话语$u_i$在$c$个情绪类别上的预测概率向量。

为了进一步提升多模态学习的一致性与鲁棒性，设计辅助单模态监督机制，对每种模态单独施加分类约束，增强模态自身的判别能力。具体为

$$L_u=-\frac{1}{N}\sum_{m\in\{t,a,v\}}\sum_{i=1}^{N}\sum_{j=1}^{C}\boldsymbol{y}^{i,j}\log(\hat{\boldsymbol{y}}_m^{i,j}) \tag{31}$$

式中，$\hat{\boldsymbol{y}}_m^{i,j}$是模态$m$对话语$i$在类别$j$上的预测概率，$\boldsymbol{y}^{i,j}$为真实标签。最终，总损失函数由融合表示主分类损失$L_f$与3种单模态辅助损失$L_u$相加构成。具体为

$$L_f=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\boldsymbol{y}^{i,j}\log(\hat{\boldsymbol{y}}_f^{i,j}) \tag{32}$$

$$L_s=L_f+L_u \tag{33}$$

2 实验分析

2.1 数据集

本文在两个广泛使用的多模态对话数据集上进行了实验评估，数据集分别为IEMOCAP和MELD，数据集的统计信息如表1所示。

IEMOCAP数据集(Busso等,2008)由两名演员通过剧本演绎或即兴表演生成,共包含151段对话和7 433个话语。每个话语均被标注为6种情绪类别之一,分别为高兴、悲伤、中立、愤怒、兴奋和沮丧。MELD数据集(Poria等,2019)来源于《老友记》中的多方对话场景片段,包含1 433段对话和13 708个话语。每个话语均被标注为7种情绪类之一,分别为中立、惊讶、恐惧、悲伤、开心、厌恶和愤怒。

表1　数据集的统计信息

Table 1　Statistical of datasets

数据集	对话[话语]数量		
	训练集	验证集	测试集
IEMOCAP	100 [5 163]	20 [647]	31 [1 623]
MELD	1 039 [9 989]	114 [1 109]	280 [2 610]

2.2　参数设置

所有实验均基于PyTorch框架实现,并在单张NVIDIA RTX 3060 GPU上完成训练与测试。优化器使用AdamW,训练的批次大小设为16,最大迭代轮数为50。在不同数据集上,根据模型收敛情况分别设置学习率:IEMOCAP数据集为0.000 1,MELD数据集为0.000 01,以获得更稳定的训练效果。

2.3　基线方法

1)DialogueRNN(Majumder等,2019)。通过3个不同的GRU分别捕捉对话中的上下文信息和说话人信息。

2)DialogueGCN(Ghosal等,2019)。将整段对话建模为有向完全图,并通过GCN进行长距离对话上下文信息的建模。

3)MMGCN(Hu等,2021)。构建多模态图卷积神经网络,捕捉包含上下文信息和说话人信息的多模态特征。

4)MM-DFN(Hu等,2022)。结合图卷积操作与门控机制,对模态内以及跨模态上下文信息进行动态筛选,有效缓解信息冗余并促进模态互补。

5)COGMEN(Joshi等,2022)。基于图神经网络(graph neural network,GNN)结构建模对话中的复杂依赖关系,同时捕捉局部和全局信息。

6)GraphCFC(Li等,2024)。采用跨模态配对补全策略进行特征融合,以缓解多模态融合中的异质性差异。

7)RL-EMO(Zhang等,2024)。将MMGCN与强化学习相结合,以在语义和情绪层面建模对话上下文。

8)DER-GCN(Ai等,2025)。引入Doc2EDAG进行事件抽取,以识别关键事件并建模情感驱动因素,从而弥补因隐含表达或缺乏明确语义特征而导致的信息缺失。

9)DQ-Former(Jing和Zhao,2024)。引入基于Q-Former的注意力机制,在不同模态间提取并整合情绪线索。

10)SEDC(Yang等,2025)。采用情绪与语义双通道处理策略,分别建模情绪与语义信息,有效缓解情绪干扰与语义误导问题。

2.4　对比实验

为确保实验结果的公平性与有效性,所有对比基线模型均采用与DECANet相同的多模态特征提取方式进行统一实现,从源头上消除由输入特征差异导致的性能偏差,保障对比结论的客观性与可信度。鉴于IEMOCAP与MELD数据集中存在情绪类别分布不均的问题,选用准确率(accuracy, Acc)与加权F1值(weighted F1 score, W-F1)作为模型整体性能的主要评估指标。

表2和表3分别展示了DECANet与多个主流基线模型在IEMOCAP和MELD两个数据集上的性能对比结果,涵盖不同情绪类别的详细评估指标。

实验结果表明,DECANet在两个数据集上均取得了最佳整体性能。在IEMOCAP数据集中,相较于表现最优的基线模型DQ-Former,DECANet在Acc和W-F1上分别提升1.74%和1.77%。此外,在大多数情绪类别上,DECANet的F1得分均实现显著提升,验证了模型在处理多类别情绪识别任务中的综合能力。在MELD数据集上,DECANet同样表现出稳定的性能优势。尽管整体提升幅度略小,但在大多数情绪类别上仍优于基线模型,进一步体现出DECANet良好的适应性与鲁棒性。从具体情绪类别的识别效果来看,DECANet在正向情绪(如“开心”)与负向情绪(如“沮丧”)的识别上均表现出更优性能。这表明,引入结构解耦的情绪依赖建模机制有助于提升模型对复杂情绪状态的感知与区分能力。总体而言,DECANet通过显式区分并动态建模说话人内部与说话人间的情绪依赖关系,显著增强了模型对多模态对话语境中情绪演化路径的理解能力,从而有效提升了对话情绪识别的整体性能。

表2　所有模型在IEMOCAP数据集上的整体表现

Table 2　The overall performance of all models on the IEMOCAP dataset

/%

模型	F1						Acc	W-F1
	高兴	悲伤	中立	愤怒	兴奋	沮丧		
DialogueRNN	52.61	<u>81.24</u>	69.07	59.15	64.01	60.45	66.72	65.45
DialogueGCN	46.28	77.84	61.90	54.38	68.09	59.47	64.51	62.70
MMGCN	41.58	81.20	66.14	64.12	73.72	60.84	66.50	66.17
MM-DFN	45.31	**81.88**	70.48	64.22	73.76	58.27	68.31	67.05
COGMEN	47.20	77.67	68.79	63.47	69.85	62.89	67.57	66.47
GraphCFC	52.80	78.39	70.97	64.40	75.30	61.04	69.35	68.26
RL-EMO	47.38	79.41	67.94	61.67	69.50	63.10	67.35	66.55
DER-GCN	58.80	79.80	61.50	<u>72.10</u>	73.30	67.80	69.70	69.40
DQ-Former	**64.09**	69.92	<u>71.87</u>	**79.74**	**83.39**	<u>69.48</u>	<u>72.63</u>	<u>72.54</u>
SEDC	–	–	–	–	–	–	71.60	71.68
DECANet(本文)	<u>60.42</u>	80.89	**75.23**	68.38	<u>82.54</u>	**70.57**	**74.37**	**74.31**

注：加粗、下划线字体表示各列最优、次优结果。“–”表示该结果在原始论文不可获得。

表3　所有模型在MELD数据集上的整体表现

Table 3　The overall performance of all models on the MELD dataset

/%

模型	F1							Acc	W-F1
	中立	惊讶	恐惧	悲伤	开心	厌恶	愤怒		
DialogueRNN	79.03	56.36	**27.65**	40.57	60.94	25.90	53.16	64.67	64.95
DialogueGCN	78.94	55.85	8.06	41.55	61.35	13.78	53.51	64.39	64.35
MMGCN	78.65	57.00	17.85	**43.08**	61.11	19.19	55.14	65.29	64.96
MM-DFN	78.91	57.57	21.53	<u>42.16</u>	62.37	25.50	54.75	65.28	65.45
GraphCFC	78.10	56.16	21.73	37.76	60.73	24.76	53.86	64.74	64.18
RL-EMO	78.81	56.49	24.44	40.78	62.33	23.48	54.48	65.04	65.14
DER-GCN	**80.60**	51.00	10.40	41.50	<u>64.30</u>	10.30	<u>57.40</u>	<u>66.80</u>	66.10
DQ-Former	78.31	**59.13**	21.90	39.93	60.84	<u>26.54</u>	**59.13**	64.88	64.70
SEDC	–	–	–	–	–	–	–	**67.43**	<u>66.16</u>
DECANet(本文)	<u>79.83</u>	<u>58.86</u>	<u>25.64</u>	41.95	**64.99**	**30.36**	54.67	**67.43**	**66.62**

注：加粗、下划线字体表示各列最优、次优结果。“–”表示该结果在原始论文不可获得。

2.5　消融实验

2.5.1　情绪依赖子图动态交互模块消融实验

表4展示了关系子图交互模块及其内部组件对模型性能的影响。实验结果表明，移除任意一个子图组件均会导致模型性能下降，进一步验证了该模块在ERC任务中的重要作用。

此外，为了深入讨论关系子图在动态情绪依赖建模对正负向情绪识别中的作用，进行了多组对比实验。如图3所示，当移除InterGAT组件后，模型在识别负向情绪（如“悲伤”/“厌恶”）方面出现显著退化，说明缺乏说话人间的情绪依赖建模会削弱模型对负面情绪传播机制的捕捉能力。相反，当移除

表 4　关系子图交互模块及其组件对模型的影响

Table 4　Influence of relational subgraph interaction module and its component on the overall effect of model

/%

InterGAT	IntraGAT	IEMOCAP		MELD	
		Acc	W-F1	Acc	W-F1
×	×	69.75	69.73	66.48	65.32
√	×	72.58	72.50	66.70	65.98
×	√	72.77	72.93	66.86	65.78
√	√	**74.37**	**74.31**	**67.43**	**66.62**

注：加粗字体表示各列最优结果。“×”表示不包含该子图；“√”表示包含该子图。

IntraGAT 组件时，模型对正向情绪（如“高兴”/“开心”）的识别能力减弱，表明说话人内部的情绪演化信息对于正向情绪的建模至关重要。上述分析表明，完整保留并区分这两类依赖结构，能有效提升模型对情绪传播路径和动态演化机制的建模能力，从而增强整体性能与情绪识别的细粒度辨别能力。

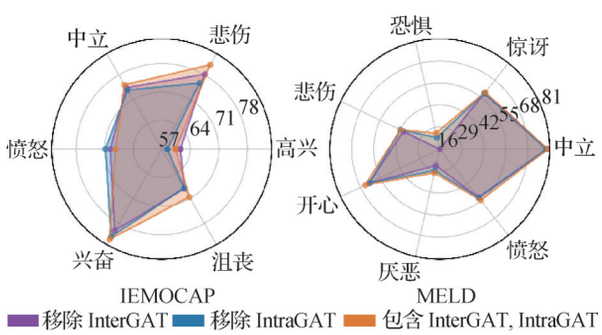

图 3　移除子图的影响

Fig. 3　Impact of subgraph removal

为了验证启发式动态并行交互策略的有效性，在 DECANet 框架上对多种子图交互方式进行实验对比，相关结果如表 5 所示。首先，尝试将所有子图合并为单一图结构进行建模，发现该策略无法区分不同情绪依赖关系的结构差异，导致情绪信息混淆，整体性能显著下降。随后，对比了 3 种并行融合策略，包括相加、拼接和双图注意力机制。尽管这些策略可以在一定程度上整合多种情绪依赖信息，但由于缺乏层次感知机制，仍无法充分建模对话中的结构化情绪动态，情绪信息的表示能力受到限制。相比之下，采用增量交互策略，即先建模说话者之间依赖再建模说话者内部情绪的依赖，表现较为优异。该方式符合对话中情绪演化的自然层次结构，能有效缓解情绪建模中的误差传播与信息干扰。此外，进一步分析不同子图交互顺序对模型性能的影响，实验结果表明，交互顺序对增量交互策略至关重要。若前置模块未能正确建模其对应依赖关系，将导致后续情绪传播路径建模产生偏差，进而造成错误累积并削弱整体识别性能。综合来看，启发式动态交互策略不仅提升了模型的结构解耦能力，还实现了多层次情绪依赖关系的精准建模，在处理复杂对话情绪动态方面展现出显著优势。

表 5　不同交互策略下的性能

Table 5　Performance under different interaction strategies

/%

交互策略		IEMOCAP		MELD	
		Acc	W-F1	Acc	W-F1
单一图结构		70.67	70.84	66.51	65.56
多子图结构	相加	71.41	71.49	67.24	66.38
	拼接	72.15	72.13	67.01	66.09
	双图注意力	70.55	70.58	65.63	64.14
	增量交互	72.34	72.54	66.55	65.89
	不同子图交互顺序	72.40	72.10	67.20	65.60
DECANet（本文）		**74.37**	**74.31**	**67.43**	**66.62**

注：加粗字体表示各列最优结果。

2.5.2　跨模态上下文感知融合模块消融实验

说话人嵌入、话语位置嵌入、跨模态上下文感知自注意力机制以及辅助模态损失构成了本文提出的多模态融合方法的 4 个关键组成部分。为评估每个组件的有效性，依次移除各个组件，并观察模型性能的变化。

表 6 展示了详细的实验结果，由此可得以下结论：1）当移除位置嵌入和说话人嵌入时，DECANet 在 IEMOCAP 和 MELD 数据集上的性能均有所下降，说明两者在提升模型对话结构理解能力方面发挥了重要作用。说话人嵌入显著提升了模型对角色交互模式与角色身份变化的感知能力；而话语位置嵌入则有助于模型识别话语的时序关系，增强对情绪演化过程的建模能力。2）当移除跨模态上下文感知自注意力机制时，在 IEMOCAP 数据集上加权 F1 分数

表6　跨模态上下文感知融合模块消融实验结果

Table 6　Results of ablation studies on the cross-modal context-aware fusion module

/%

方法		IEMOCAP数据集		MELD数据集	
		Acc	W-F1	Acc	W-F1
多模态融合	w/o 位置编码	73.94	73.80	66.70	65.87
	w/o 说话人嵌入	72.95	72.78	66.82	65.93
	w/o 跨模态上下文感知自注意力机制	72.09	72.22	67.43	65.96
	w 交叉注意力机制	72.77	72.90	66.82	66.13
	w/o 辅助模态融合损失	70.79	70.86	67.20	66.22
模态设置	文本	69.07	69.22	66.50	66.01
	音频	53.23	52.70	47.09	41.32
	视觉	39.56	38.48	48.12	32.23
	文本 + 音频	72.95	73.00	67.21	66.44
	文本 + 视觉	71.29	71.22	67.12	66.27
	音频 + 视觉	53.97	53.38	47.05	41.41
DECANet(本文)		**74.37**	**74.31**	**67.43**	**66.62**

注：加粗字体表示各列最优结果，“w/o”表示移除该组件，“w”表示使用该组件。

下降2.09%，在MELD数据集上下降0.66%。若使用交叉注意力机制替代，IEMOCAP数据集上的加权F1分数下降1.41%，MELD数据集上下降0.49%。这一结果充分验证了本文提出的上下文感知机制在跨模态信息对齐与判别特征提取方面的优势，其通过动态构建模态间的语义依赖与重要性权重，有效抑制冗余干扰信息，提升融合表达的精度与鲁棒性。3）当移除辅助模态融合损失时，DECANet在两个数据集上的准确率和加权F1值均显著下降，说明在多模态学习过程中，依赖最终融合特征可能会导致某些模态的特征学习不足甚至退化。通过引入辅助损失，模型在各模态间实现更为平衡的表征学习，从而强化模态间协同建模效果。

为进一步分析各模态对情绪识别的贡献，分别移除一种或两种模态进行实验。表6展示了实验结果，可以得出以下结论：1）在所有单模态实验中，文本模态取得了最高的性能，远超音频和视觉模态，这表明文本模态包含比其他两种模态更多的情绪信息，这一发现与现有研究结论一致（Hu等，2021；Hu等，2022；Mao等，2021；王顺杰 等，2023）。相较于视觉模态，音频模态在IEMOCAP和MELD数据集上均取得了更好的单模态性能。这可能是因为视觉信息通常包含复杂的背景，容易受到环境噪声的干扰，导致特征提取效果下降。此外，本文所采用的视觉特征提取方法较为简单，难以捕捉细粒度的视觉特征，如面部表情和肢体动作，也在一定程度上限制了视觉模态的表现。为提升非文本模态利用率，可采用更精细的视觉特征建模方法、增强语音特征表示能力，并通过跨模态语义对齐或模态可信度加权，使其在多模态融合中发挥更大作用。非文本模态虽然单独性能有限，但能够捕捉文本难以通过上下文建模获得的瞬时情绪特征，从而在融合中有效补充文本信息，提升整体识别能力。2）在双模态融合实验中，文本+音频和文本+视觉均优于视觉+音频。在IEMOCAP数据集上，文本+音频模态的准确率达到72.95%，比文本+视觉模态高1.66%，MELD数据集上也呈现类似趋势。3）图4展示了不同模态组合在ERC任务中的表现，引入3种模态的信息进一步提升了模型性能，说明多模态融合有助于提升对情绪特征的建模能力与对话中多元情绪线索的理解深度。综上，本文所提出的多模态融合方法在结构上有效集成了说话人信息、时序结构与模态间上下文语义关联，并通过辅助损失项调节融合过程中的偏差问题，显著提升了多模态情绪识别的表现能力与

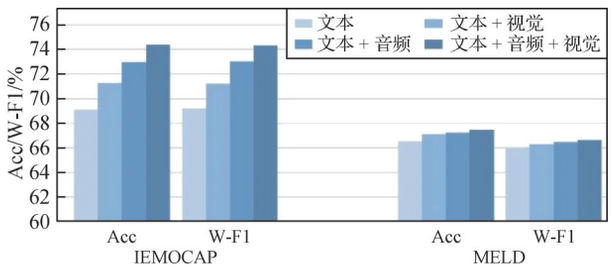

图 4　基于文本的模态组合对比

Fig. 4　Comparison of modality combinations with text

鲁棒性，各模块在模型中的协同作用共同构成了 DECANet 在 ERC 任务中优越性能的基础。

2.6　超参数分析

2.6.1　不同窗口大小的影响

为了分析对话窗口大小对模型性能的影响，在实验中调整窗口大小，图 5 展示了相应结果。当只关注当前对话节点且不进行任何上下文交互（即对话窗口大小 = 0）时，模型表现最差。这表明，上下文信息在对话情绪识别任务中至关重要，仅依赖单个话语的情绪信息不足以精准建模对话情境。通过调整对话窗口的大小，可以逐步找到最适合模型的窗口大小。此外，从图 5 中还可以看出，不同数据集的最优窗口大小不同，IEMOCAP 数据集中最佳窗口大小为 20，MELD 数据集中最佳窗口大小为 5。这一现象可能源于两者对话长度的显著差异。IEMOCAP 数据集包含较长的对话，较大的窗口可捕捉更丰富的上下文信息，帮助模型更全面地理解情绪变化趋势。而 MELD 数据集由较短的对话组成，较小的窗口已足以提供关键的语境信息，过大的窗口可能引入冗余信息甚至噪声，影响模型性能。

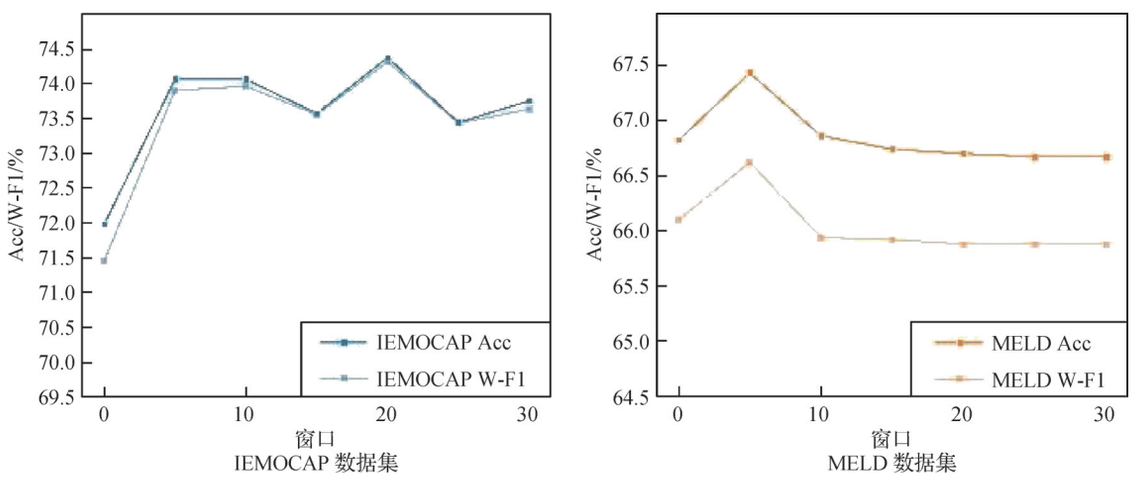

图 5　不同窗口大小对模型性能的影响

Fig. 5　The effect of different window sizes on the model

2.6.2　不同阈值大小的影响

为探究不同阈值大小对多模态融合效果的影响，本文在 IEMOCAP 和 MELD 数据集上分别进行了参数敏感性实验，结果如图 6 所示。在 IEMOCAP 数据集中，当阈值设定为 0.2 时，模型性能达到最优。此时模型倾向于保留通过跨模态上下文感知机制增强后的多模态特征作为最终表征输入。该结果表明，IEMOCAP 中文本、音频与视觉模态之间具有较强的语义协同性与一致性，较低的判定阈值能够充分利用深层融合特征，从而提升情绪信息的表达完整性与判别能力。随着阈值的升高，融合特征的保留比例减少，模型逐渐退回至模态独立表示，导致融合优势减弱，性能随之下降。

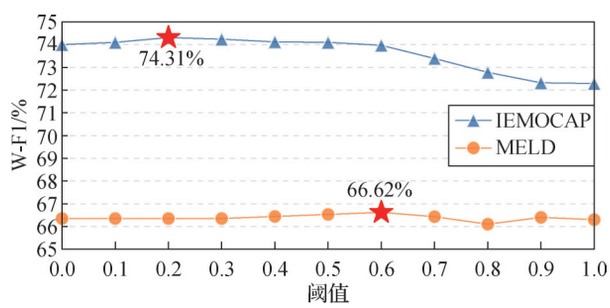

图 6　不同阈值大小对模型性能的影响

Fig. 6　The effect of different threshold sizes on the model

相比之下，在 MELD 数据集中，模型在阈值为 0.6 时取得最优效果。MELD 数据集中模态间异质性更为显著，存在一定程度的语义不一致与信息冗

余。若设定较低阈值强制使用所有融合后的模态特征，可能引入冲突信息，降低模型的判别稳定性与泛化能力。适当提高阈值，则使模型在语义不一致场景下能主动“退回”至模态独立表示，有助于增强对异常模态的容忍性，从而在复杂场景中保持更优性能。

综合来看，最优阈值的差异主要源于数据集的模态异构性与语义一致性：IEMOCAP数据集采集于封闭实验环境，说话者数量有限且情绪表达规范，音频和视觉信号清晰，模态一致性较高，因此低阈值即可充分保留融合特征，最大化情绪信息的表达与判别能力。相比之下，MELD数据集采集于现实环境，说话者多样、情绪表达自然随意，且常伴随背景噪声、光照变化和模态缺失等问题，模态间存在语义冲突和信息冗余。此时低阈值融合容易引入干扰，而较高阈值则使模型能在冲突场景下退回至模态独立表示，从而提升容错性与鲁棒性。因此，语义一致性驱动的特征选择模块作为多模态交互过程中的关键调节参数，需根据具体任务和数据集特性灵活设置，以在“信息整合”与“保留模态差异性”之间实现最优平衡，从而获得最佳融合效果。

2.7 抗干扰能力分析

为验证跨模态上下文感知自注意力机制在面对噪声干扰时的鲁棒性，设计了噪声扰动实验，即在多模态输入中分别引入相同比例的高斯噪声，观察模型性能在IEMOCAP数据集上随噪声强度变化的趋势，实验结果如图7所示。与对比模型MM-DFN和DQ-Former相比，DECANet在各个噪声水平下的性能下降幅度更缓，表现出更强的抗干扰能力，且在噪声干预下仍保持优于对比模型的整体性能。综上所述，DECANet在多模态输入存在噪声扰动时，依然能够有效提取关键信息并抑制冗余干扰，通过引导不同模态间的深层语义交互，动态调整模态间的注意力权重，突出语义一致的模态特征，同时能在面对模态噪声时，自适应地削弱语义不一致的多模态信息，从而提升模型的鲁棒性与泛化能力。

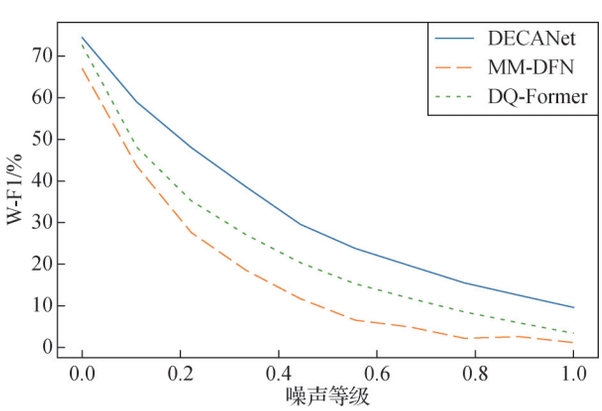

图7 不同噪声干扰对模型性能的影响

Fig. 7 Impact of different noise levels on the model

3 结 论

本文提出一种解耦情绪依赖关系的跨模态感知对话情绪识别模型DECANet，通过构建区分说话者内外情绪依赖的子图结构，并引入启发式动态交互策略，精确建模对话中的情绪传播路径与演化模式，克服混合建模带来的信息混淆问题。同时，设计跨模态上下文感知自注意力机制，有效捕捉多模态间的深层语义关联，缓解模态异质性带来的融合偏差；并结合语义一致性驱动的特征选择模块，过滤冗余或冲突信息，提升融合质量。实验结果表明，DECANet在IEMOCAP与MELD数据集上均优于对比的主流方法，展现出更强的情绪建模能力与跨模态泛化性能。进一步分析表明，建模说话者之间的情绪依赖有助于识别负向情绪，而建模说话者内部依赖则更有利于正向情绪的识别。未来的研究将继续深入挖掘对话中的动态情绪结构特征，并探索更高效、鲁棒的多模态融合策略，以提升模型在复杂现实场景下的应用能力。

参考文献（References）

Ai W, Shou Y T, Meng T and Li K Q. 2025. DER-GCN: dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. IEEE Transactions on Neural Networks and Learning Systems, 36(3): 4908-4921 [DOI: 10.1109/TNNLS.2024.3367940]

Bhat A and Modi A. 2023. Multi-task learning framework for extracting emotion cause span and entailment in conversations//Proceedings of the 1st Transfer Learning for Natural Language Processing Workshop. New Orleans, USA: PMLR: 33-51

Busso C, Bulut M, Lee C C, Kazemzadeh A, Mower E, Kim S, et al. 2008. IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4): 335-359 [DOI: 10.1007/s10579-008-9076-6]

Ghosal D, Majumder N, Gelbukh A, Mihalcea R and Poria S. 2020. COSMIC: commonsense knowledge for emotion identification in conversations//Proceedings of the Findings of the Association for

Computational Linguistics: EMNLP 2020. Association for Computational Linguistics: 2470-2481 [DOI: 10.18653/v1/2020.findings-emnlp.224]

Ghosal D, Majumder N, Poria S, Chhaya N and Gelbukh A. 2019. DialogueGCN: a graph convolutional neural network for emotion recognition in conversation//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics: 154-164 [DOI: 10.18653/v1/D19-1015]

Hazarika D, Poria S, Mihalcea R, Cambria E and Zimmermann R. 2018. ICON: interactive conversational memory network for multimodal emotion detection//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics: 2594-2604 [DOI: 10.18653/v1/D18-1280]

Hu D, Hou X L, Wei L W, Jiang L X and Mo Y. 2022. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations//Proceedings of 2022 ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 7037-7041 [DOI: 10.1109/ICASSP43922.2022.9747397]

Hu J W, Liu Y C, Zhao J M and Jin Q. 2021. 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 (Volume 1: Long Papers). Association for Computational Linguistics: 5666-5675 [DOI: 10.18653/v1/2021.acl-long.440]

Ishiwatari T, Yasuda Y, Miyazaki T and Goto J. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations//Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [s.l.]: Association for Computational Linguistics: 7360-7370 [DOI: 10.18653/v1/2020.emnlp-main.597]

Jing Y and Zhao X P. 2024. DQ-Former: querying transformer with dynamic modality priority for cognitive-aligned multimodal emotion recognition in conversation//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: Association for Computing Machinery: 4795-4804 [DOI: 10.1145/3664647.3681599]

Joshi A, Bhat A, Jain A, Singh A and Modi A. 2022. COGMEN: contextualized GNN based multimodal emotion recognition//Proceedings of 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA: Association for Computational Linguistics: 4148-4164 [DOI: 10.18653/v1/2022.naacl-main.306]

Li J, Wang X P, Lv G Q and Zeng Z G. 2024. GraphCFC: a directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. IEEE Transactions on Multimedia, 26: 77-89 [DOI: 10.1109/TMM.2023.3260635]

Li Z J, Tang F X, Zhao M and Zhu Y S. 2022. EmoCaps: emotion capsule based model for conversational emotion recognition//Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Irelcrnd: Association for Computational Linguistics: 1610-1618 [DOI: 10.18653/v1/2022.findings-acl.126]

Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A and Cambria E. 2019. DialogueRNN: an attentive RNN for emotion detection in conversations//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI Press: 6818-6825 [DOI: 10.1609/aaai.v33i01.33016818]

Mao Y Z, Liu G, Wang X J, Gao W G and Li X. 2021. DialogueTRM: exploring multi-modal emotional dynamics in a conversation//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics: 2694-2704 [DOI: 10.18653/v1/2021.findings-emnlp.229]

Meng T, Zhang F C, Shou Y T, Shao H, Ai W and Li K Q. 2024. Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32: 4298-4312 [DOI: 10.1109/TASLP.2024.3434495]

Poria S, Hazarika D, Majumder N, Naik G, Cambria E and Mihalcea R. 2019. MELD: a multimodal multi-party dataset for emotion recognition in conversations//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics: 527-536 [DOI: 10.18653/v1/P19-1050]

Sap M, Le Bras R, Allaway E, Bhagavatula C, Lourie N, Rashkin H, et al. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI Press: 3027-3035 [DOI: 10.1609/aaai.v33i01.33013027]

Shi T and Huang S L. 2023. MultiEMO: an attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics: 14752-14766 [DOI: 10.18653/v1/2023.acl-long.824]

Tao J H, Fan C H, Lian Z, Lyu Z, Shen Y and Liang S. 2024. Development of multimodal sentiment recognition and understanding. Journal of Image and Graphics, 29(6): 1607-1627 (陶建华, 范存航, 连政, 吕钊, 沈莹, 梁山. 2024. 多模态情感识别与理解发展现状及趋势. 中国图象图形学报, 29(6): 1607-1627) [DOI: 10.11834/jig.240017]

Wang S J, Cai G Y, Lyu G R and Tang W B. 2023. Aspect-level multimodal co-attention graph convolutional sentiment analysis model. Journal of Image and Graphics, 28(12): 3838-3854 (王顺杰,

蔡国永，吕光瑞，唐炜博. 2023. 方面级多模态协同注意图卷积情感分析模型. 中国图象图形学报，28(12)：3838-3854)［DOI：10.11834/jig.221015］

Yang Z Y，Zhang Z B，Cheng Y H，Zhang T and Wang X S. 2025. Semantic and emotional dual channel for emotion recognition in conversation. IEEE Transactions on Affective Computing，16(3)：1885-1902［DOI：10.1109/TAFFC.2025.3544608］

Zhang C W，Zhang Y H and Cheng B. 2024. RL-EMO：a reinforcement learning framework for multimodal emotion recognition//Proceedings of 2024 ICASSP IEEE International Conference on Acoustics，Speech and Signal Processing（ICASSP）. Seoul Korea（South）：IEEE：10246-10250［DOI：10.1109/ICASSP48485.2024.10446459］

Zhang D Z，Chen F L and Chen X Y. 2023. DualGATs：dual graph attention networks for emotion recognition in conversations//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics（Volume 1：Long Papers）. Toronto，Canada：Association for Computational Linguistics：7395-7408［DOI：10.18653/v1/2023.acl-long.408］

Zhang H D and Chai Y K. 2021. COIN：conversational interactive networks for emotion recognition in conversation//Proceedings of the 3rd Workshop on Multimodal Artificial Intelligence. Mexico：Association for Computational Linguistics：12-18［DOI：10.18653/v1/2021.maiworkshop-1.3］

Zhao W X，Zhao Y Y and Qin B. 2022. MuCDN：mutual conversational detachment network for emotion recognition in multi-party conversations//Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju，Korea（South）：International Committee on Computational Linguistics：7020-7030

Zheng S，Cao W，Xu W and Bian J. 2021. Revisiting the evaluation of end-to-end event extraction//Proceedings of the Findings of the Association for Computational Linguistics：ACL-IJCNLP 2021［s.l.］. Association for Computational Linguistics：4609-4617［DOI：10.18653/v1/2021.findings-acl.405］

作者简介

邓天生，男，硕士研究生，主要研究方向为情感分析。
E-mail：23032304004@mails.guet.edu.cn
蔡国永，通信作者，男，教授，主要研究方向为社交媒体挖掘、情感分析和可信AI技术。E-mail：ccgycai@guet.edu.cn
董凯，男，硕士研究生，主要研究方向为情感分析。
E-mail：dongkai042@163.com
王顺杰，女，博士研究生，主要研究方向为社交媒体挖掘和情感分析。E-mail：wsj12130@foxmail.com