中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

基于听说知识融合网络的多模态对话情绪识别

刘琴, 谢珺, 胡勇, 郝戍峰, 郝雅卉

引用本文:
刘琴,谢,胡勇,郝戍峰,郝雅卉. 基于听说知识融合网络的多模态对话情绪识别[J]. 控制与决策, 2024, 39(6): 2031−2040.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2023.0033

基于听说知识融合网络的多模态对话情绪识别

刘 琴[1], 谢 珺[1†], 胡 勇[2], 郝戍峰[3], 郝雅卉[1]

(1. 太原理工大学 信息与计算机学院，太原 030024；2. 北京航空航天大学 新媒体艺术与设计学院，北京 100191；3. 太原理工大学 大数据学院，太原 030024)

摘 要: 多模态对话情绪识别旨在根据多模态对话语境判别出目标话语所表达的情绪类别,是构建共情对话系统的基础任务. 现有工作中大多数方法仅考虑多模态对话本身信息,忽略了对话中与倾听者和说话者相关的知识信息,从而限制了目标话语情绪特征的捕捉. 为解决该问题,提出一种基于听说知识融合网络的多模态对话情绪识别模型 (LSKFN), 引入与倾听者和说话者相关的外部常识知识, 实现多模态上下文信息和知识信息的有机融合. LSKFN包含多模态上下文感知、听说知识融合、情绪信息汇总和情绪决策4个阶段,分别用于提取多模态上下文特征、融入听说知识特征、消除冗余特征和预测情绪分布. 在两个公开数据集上的实验结果表明,与其他基准模型相比,LSKFN能够为目标话语提取到更加丰富的情绪特征,并且获得较好的对话情绪识别效果.

关键词: 情感计算；对话情绪识别；多模态特征；外部常识知识；上下文语义；知识特征融合

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2023.0033

引用格式: 刘琴, 谢珺, 胡勇, 等. 基于听说知识融合网络的多模态对话情绪识别[J]. 控制与决策, 2024, 39(6): 2031-2040.

Listening and speaking knowledge fusion network for multi-modal emotion recognition in conversation

LIU Qin[1], *XIE Jun*[1†], *HU Yong*[2], *HAO Shu-feng*[3], *HAO Ya-hui*[1]

(1. College of Information and Computer，Taiyuan University of Technology，Taiyuan 030024，China；2. College of New Media Art and Design，Beihang University，Beijing 100191，China；3. College of Data Science，Taiyuan University of Technology，Taiyuan 030024，China)

Abstract: Multi-modal emotion recognition in conversation aims to identify the emotion of the target utterance according to the multi-modal conversation context, which is the primary task of building empathetic dialogue systems (EDS). Existing works only consider multi-modal conversation itself while ignoring the knowledge information about the listener and the speaker, leading to the limit in capturing the emotional features of the target utterance. To solve this problem, a listening and speaking knowledge fusion network (LSKFN) is proposed, which introduces the external common sense knowledge and fuses it with multi-modal context efficiently. The proposed LSKFN consists of four stages, which are used to extract multi-modal context features, integrate listening and speaking knowledge features, eliminate redundant features, and predict emotional probability distribution. Experimental results on two public multi-modal conversation datasets demonstrate that the LSKFN can extract richer emotional features for the target utterance, and obtain better emotional recognition performance compared with other benchmark models.

Keywords: effective computing；emotion recognition in conversation；multi-modal features；external common sense knowledge；context semantics；knowledge feature fusion

0 引 言

情绪作为一种信息线索直接影响个体的决策与判断,在日常人际交往中具有重要作用[1-2], 在对话中准确识别对方的情绪是利用情绪进行交流的前提. 心理学中用于情绪识别的信息来源于两方面[3], 一是视听等多模态觉知信息,二是储存在记忆中与

收稿日期: 2023-01-09；**录用日期:** 2023-05-08.

基金项目: 虚拟现实技术与系统国家重点实验室(北京航空航天大学)开放课题基金项目(VRLAB2022C11)；山西省基础研究计划青年科学研究项目(20210302124168)；山西省留学人员科技活动择优资助项目(20220009)；山西省重点研发计划项目(202102020101004).

†通讯作者. E-mail: xiejun@tyut.edu.cn.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

情绪推断相关的知识信息. 随着深度学习技术的发展和社交网络上对话数据的激增,利用深度学习算法进行多模态对话情绪识别 (multi-modal emotion recognition in conversation, MMERC) 逐渐成为研究热点. 该研究不仅对于共情对话系统[4] 的发展具有重大意义,而且在社交网络分析[5]、在线评论分析[6]、智能推荐[7] 等领域也有潜在的应用价值. 综上,如何利用深度学习网络的表征能力实现多模态感知信息与知识信息的融合和多模态对话情绪特征的抽取, 进而提升多模态对话情绪识别的性能,是本文的主要研究问题.

相比于评论等陈述型内容的多模态情绪分析, MMERC任务除了要考虑多模态上下文依赖以外,还要考虑对话中说话人之间复杂的情绪交互[8],因此更具挑战性. 现阶段主要采用基于图神经网络的方法[9-10] 和基于预训练模型的方法[11] 来捕捉说话者内部的自我依赖和说话人之间的交互依赖. 这些方法忽略了对话中与倾听者和说话者相关的知识信息对于情绪交互建模的作用,因此识别效果有限. 图1是一个多模态对话实例,用于表明听说知识在理解和检测话语情绪方面的重要性. 首先定义“听说知识”这一概念: 在对话过程中,每个对话参与者存在两种身份,分别为倾听者和说话者,本文将与倾听者相关的知识称为倾听者知识,与说话者相关的知识称为说话者知识,二者合称为听说知识. 接着分析听说知识对于情绪识别的作用: $\#U_4$ 是待预测情绪的目标话语,从图1中可以看出, $\#U_4$ 的情绪受到倾听者知识 $\#L_1$、$\#L_3$ 和说话者知识 $\#S_2$、$\#S_4$ 不同程度的影响,且距离目标话语越近影响越明显. 这种反映对话者感受、意图等信息的听说知识能够为对话情绪识别提供更多的情绪线索.

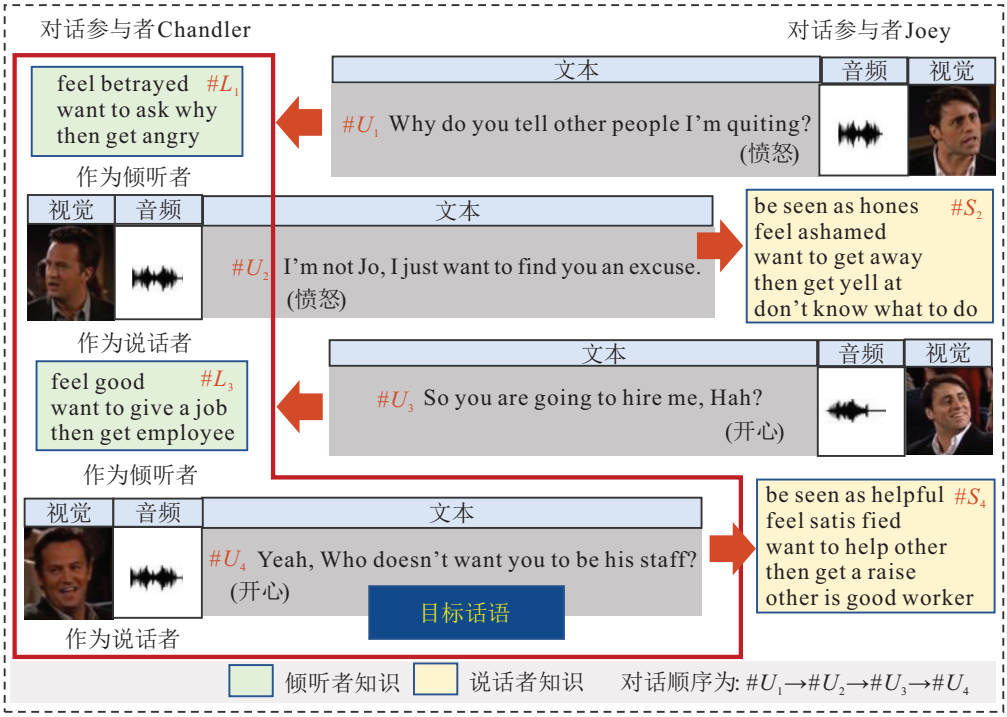


图 1 多模态对话实例

针对现阶段MMERC任务在对话情绪交互建模方面的不足,提出一种基于听说知识融合网络的多模态对话情绪识别模型(LSKFN),用于有效结合多模态上下文信息和听说知识信息,丰富目标话语的情绪特征,进一步提升模型识别效果. 模型包含4个阶段: 在多模态上下文感知阶段,利用多模态融合机制[12] 捕获模态增强后的话语特征,在此基础上通过BiGRU捕获话语级别的上下文依赖信息,使得当前话语的特征包含其前后话语的信息; 在听说知识融合阶段,利用知识融合模块和跨信息平均池化将听说知识特征有效融入上一步提取出的多模态上下文特征中,为目标话语提供更丰富的情绪特征; 在情绪信息汇总阶段,利用门控输出模块给不同特征分配不同权重,消除冗余信息,提取出能够最大程度表征情绪类型的数据特征; 在情绪决策阶段,通过多层感知机和归一化指数函数得到目标话语的情绪分布.

1 相关研究

近年来,许多学者设计了多种深度神经网络模型提取目标话语的情绪特征并进行情绪类型预测,这些模型整体上可以分为3类: 基于文本的模型、基于外

部知识的模型和基于多模态的模型.

1.1 基于文本的对话情绪识别模型

早期研究仅考虑文本模态的信息,重点关注目标话语的上下文依赖. Poria等[13]采用BiLSTM捕获话语间的序列关系. Hu等[14]提出语境感知网络,同时捕捉情景级和说话者级语境,经过多轮推理模块对情绪线索进行迭代提取和整合. Li等[15]采用双线性网络和双向机制将上下文特征融入当前话语. 以上方法大多基于序列模型,难以建模多个说话人间的交互,将对话结构表示为图中不同节点之间的连接,通过图神经网络不断聚合邻居节点的信息,能够得到更全面的情绪表征. Shen等[16]基于说话者的身份和位置关系设计了一种有向无环图,模拟对话内在结构信息. Lee等[17]通过异构对话图实现不同说话人之间的信息交互. 这些深度网络模型在提取话语情绪特征上展现了一定的能力,但效果仍然不佳,可能是由于模型输入仅考虑文本对话本身的信息,缺少了能够提供丰富情绪线索的外部知识信息.

1.2 基于外部知识的对话情绪识别模型

人类常依赖外部常识知识表达情绪,考虑到对话中话语语句短小且信息量有限,部分学者融入外部常识知识来丰富目标话语的语义信息. Ghosal等[18]认为常识知识对于对话结构的建模以及对话参与者的情感动态建模是必要的,因此提出一种结合心理状态、事件、因果关系等常识元素的话语级情绪识别框架,结合外部常识知识设计更好的上下文表示. Zhao等[19]考虑到情绪原因对于对话情绪识别的影响,提出因果感知交互网络,利用常识知识作为因果线索,捕捉会话语境中的深层情绪表征. Xie等[20]认为常识知识和情绪词汇都有助于情绪检测,提出多任务学习知识交互网络,强调话语与知识之间显性互动,并引入短语级情绪极性强度预测任务来辅助对话情绪识别. 以上工作表明知识在文本对话情绪识别中起着关键作用. 但已有研究[21]表明,相较于单个模态的对话情绪识别,在文本信息的基础上融入语音、视觉等其他模态的特征能够有效提升情绪识别的效果.

1.3 基于多模态的对话情绪识别模型

随着多模态人机交互(multi-modal human-computer interaction, MMHCI)设备的普及和多模态技术的发展,工业界和学术界逐渐认识到多模态对话场景的分析需求. 为满足研究需求,Busso等[22]提出包含多模态信息的大规模二元对话数据集IEMOCAP. Poria等[23]提出MELD数据集,填补了多模态多方对话数据集的空白. 自此,MMERC研究得到迅速发展. Zhang等[24]提出双向动态双重影响网络,同时考虑模态内和模态间的影响. Mao等[11]通过跨模态门控注意力机制融合多种模态信息. Hu等[10]采用图网络探索对话中的单模态交互和跨模态交互. Hu等[25]提出多模态动态融合网络,解决了传统的基于图的融合方法在每一层积累冗余信息的问题,促进了跨模态的上下文理解. 虽然引入其他模态能够提升对话情绪识别的性能,但目前几乎没有相关工作考虑在多模态语境下融入知识,这是限制目标话语情绪特征提取的一个原因.

综上,提出一种听说知识融合网络,实现多模态感知信息和听说知识信息的有效融合. 在两个公开基准数据集上进行实验,分析模型的实际效果.

2 模型描述

本节将详细介绍基于听说知识融合网络的多模态对话情绪识别模型(简称LSKFN模型). LSKFN的整体结构如图2所示,主要包含4个阶段.

2.1 LSKFN的任务定义

给定一个对话序列$C=\{U_1^M,U_2^M,\dots,U_N^M\}$. 其中: N为对话所包含的话语(Utterance)个数, $M=\{t,a,v\}$为文本、音频和视觉模态信息,元素$U_i^M=\{U_i^t,U_i^a,U_i^v\}$为对话序列中的第$i$句话语. LSKFN模型的目标是根据3种模态的对话序列信息预测出对话中每句话语U_i^M的情绪类别标签.

2.2 多模态上下文感知阶段

目标话语的情绪与其多模态上下文信息具有相关性. 设一个对话包含N句话语,文本模态的原始特征矩阵表示为$U^T=[U_1^t,U_2^t,\dots,U_N^t],U_i^t\in\mathbb{R}^{d_t},d_t$为文本特征维度; 音频模态的原始特征矩阵表示为$U^A=[U_1^a,U_2^a,\dots,U_N^a],U_i^a\in\mathbb{R}^{d_a},d_a$为音频特征维度; 视觉模态的原始特征矩阵表示为$U^V=[U_1^v,U_2^v,\dots,U_N^v],U_i^v\in\mathbb{R}^{d_v},d_v$为视觉特征维度. 首先,引入多模态早期融合策略获得模态增强后的话语特征矩阵G_M,具体公式为

$$f_{\text{con}}=f_T(U^T)\oplus f_A(U^A)\oplus f_V(U^V),\tag{1}$$

$$G_M=W_{\text{mul}}^T\cdot f_{\text{con}}+b_{\text{mul}}=[g_1,g_2,\dots,g_N].\tag{2}$$

其中: $\oplus$表示拼接运算; $f_A(\cdot)$、$f_T(\cdot)$和$f_V(\cdot)$分别为各模态的线性变换,用于将不同模态原始特征映射至同一向量空间中; $W_{\text{mul}}\in\mathbb{R}^{3d_h\times d_h}$为线性层的权重矩阵, $b_{\text{mul}}\in\mathbb{R}^{d_h\times N}$为偏置项; $g_i\in\mathbb{R}^{d_h}$为第i句话语的多模态特征表示; d_h为隐藏层维度.

将多模态话语特征序列$[g_1,g_2,\dots,g_N]$输入到BiGRU中,得到捕获了上下文依赖后的多模态话语特征矩阵H_{BiTAV},其公式为

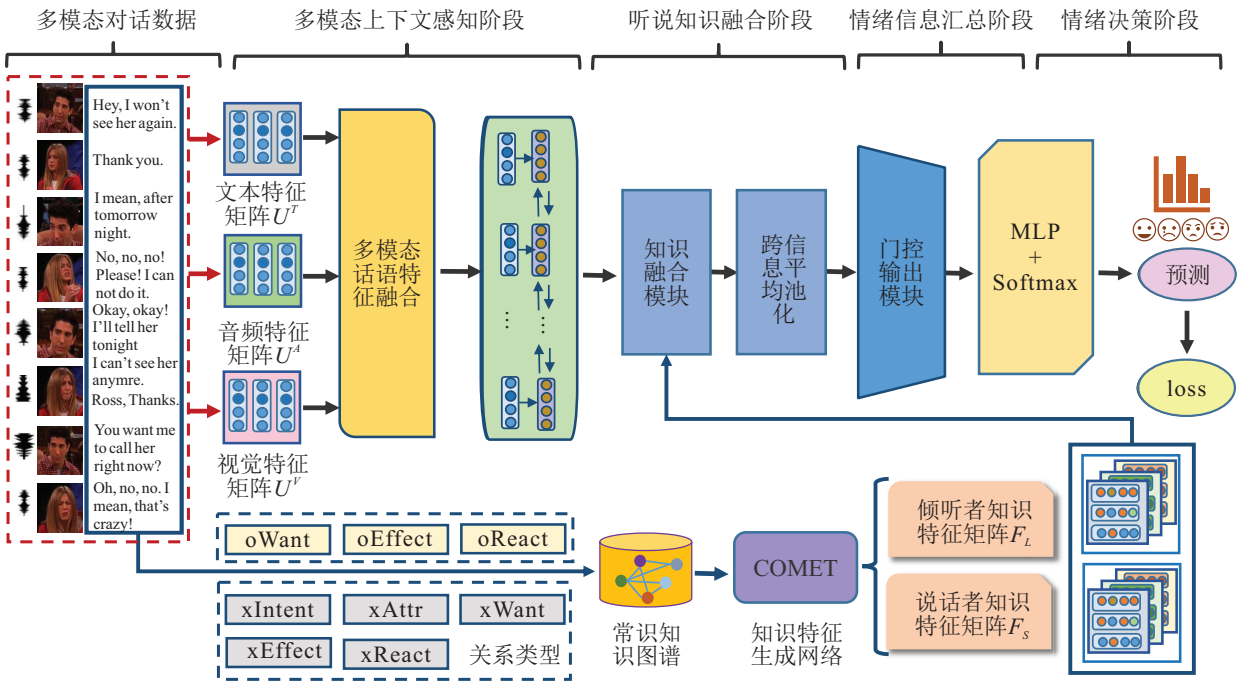


图 2 LSKFN 模型整体结构

$$\overrightarrow{H_m} = \overrightarrow{\mathrm{GRU}}([g_1, \ldots, g_N]) = [\overrightarrow{h_1}, \ldots, \overrightarrow{h_N}], \quad (3)$$

$$\overleftarrow{H_m} = \overleftarrow{\mathrm{GRU}}([g_1, \ldots, g_N]) = [\overleftarrow{h_1}, \ldots, \overleftarrow{h_N}], \quad (4)$$

$$H_{\mathrm{BiTAV}} = \overrightarrow{H_m} \oplus \overleftarrow{H_m} = [h_1, h_2, \ldots, h_N]. \quad (5)$$

其中: $\overrightarrow{H_m} \in \mathbb{R}^{d_h \times N}$ 为前向隐藏状态特征序列, $\overleftarrow{H_m} \in \mathbb{R}^{d_h \times N}$ 为反向隐藏状态特征序列, $H_{\mathrm{BiTAV}} \in \mathbb{R}^{2d_h \times N}$ 为引入上下文信息后的话语特征序列, $h_i \in \mathbb{R}^{2d_h}$ 为包含多模态上下文信息的第 i 句话语的特征向量.

2.3 听说知识融合阶段

融入对话者的听说知识能够捕获更丰富的话语情绪特征. 为此, 本文在第 2.2 节获取的多模态上下文感知特征的基础上设计了听说知识融合阶段, 该阶段主要包括知识融合模块和跨信息平均池化模块.

2.3.1 听说知识特征的抽取

在目标话语情绪特征的表示学习过程中, 引入常识知识特征能够增强语义理解, 一般通过微调在 ATOMIC[26] 知识库中预训练的知识生成模型 COMET[27] 获取外部知识特征. 为抽取听说知识特征, 本文探讨了 ATOMIC 知识库中的 8 种关系: 对于说话者, 考虑说话者意图 (xIntent)、说话者属性 (xAttr)、说话者期望 (xWant)、说话者反应 (xEffect) 和说话者感受 (xReact) 这 5 种关系; 对于倾听者, 考虑倾听者期望 (oWant)、倾听者反应 (oEffect) 和倾听者感受 (oReact) 这 3 种关系. 具体地, 在模型计算中, 将话语 U_i^t 和关系类型输入预训练的 COMET 模型, 提取 COMET 最后一层编码器的隐藏状态向量作为常识知识特征的表示, 每个知识特征的嵌入维度为 $d_k = 768$. 采用 xIntent、xAttr、xWant、xEffect、xReact 这 5 类关系得到的知识表征向量代表 5 种不同类型的说话者知识特征, 记为 $F_S = \{F_{S_1}, F_{S_2}, F_{S_3}, F_{S_4}, F_{S_5}\}$, 其中 $F_{S_i} \in \mathbb{R}^{d_k \times N}$; 采用 oWant、oEffect、oReact 这 3 类关系得到的知识表征向量代表 3 种不同类型的倾听者知识特征, 记为 $F_L = \{F_{L_1}, F_{L_2}, F_{L_3}\}$, 其中 $F_{L_i} \in \mathbb{R}^{d_k \times N}$.

2.3.2 知识融合模块

为了在多模态话语特征表示 H_{BiTAV} 的基础上融入听说知识信息, 设计一个听说知识融合模块. 以说话者知识融合模块为例, 该模块输入信息为 $(H_{\mathrm{BiTAV}}, F_{S_i}, M_S)$. 首先利用线性映射将说话者知识特征 F_{s_i} 映射到 H_{BiTAV} 的向量表示空间; 接着计算说话者的注意力分布矩阵 F_{α_i}, 公式为

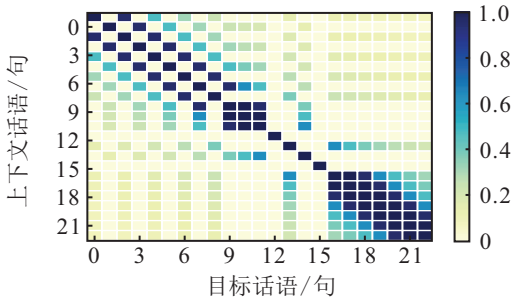
$$F_{k_i} = H_{\mathrm{BiTAV}} + (W_s^{\mathrm{T}} F_{s_i} + b_s), \quad (6)$$

$$F_{\alpha_i} = \mathrm{softmax}\left[\frac{H_{\mathrm{BiTAV}}^{\mathrm{T}} \times F_{k_i}}{\sqrt{2d_h}} \odot M_S\right]. \quad (7)$$

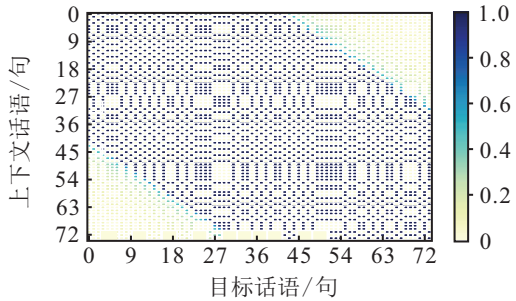
其中: $W_s \in \mathbb{R}^{d_k \times 2d_h}$ 和 $b_s \in \mathbb{R}^{2d_h \times N}$ 为可学习的权重矩阵和偏置项; $\odot$ 表示哈达玛积; $M_S \in \mathbb{R}^{N \times N}$ 为一种特殊的加权掩码矩阵, 用于搜索对话中与目标话语属于相同的说话人的其他话语, 并根据其他话语距离目标话语的远近程度赋予其相应的权重. $M_{S_{i,j}}$ 表示 M_S 掩码矩阵中每个位置的具体值, 即对话中话语 j 与待预测情绪的目标话语 i 之间的关系, 公式为

$$M_{S_{i,j}} = \begin{cases} \dfrac{1.9}{A}, & d > \mathrm{DTU} \text{ and } u(i) = u(j); \\ 1, & d \leqslant \mathrm{DTU} \text{ and } u(i) = u(j). \end{cases} \quad (8)$$

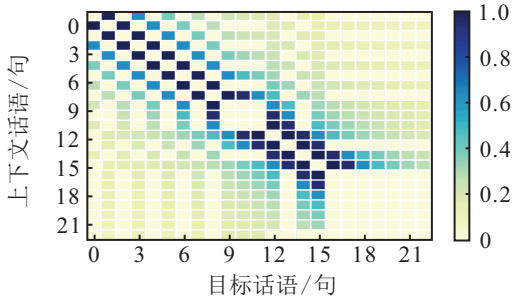
其中: $A = d - \text{DTU} + 1, u(\cdot)$ 为话语到说话人的映射, $d = |j - i|$ 为其他话语与目标话语的距离绝对值. 参数DTU (distance from target utterance) 是预先设定的一个参数值, 表示其他话语与目标话语的距离阈值. 图3(a)和(b)是DTU参数分别为1和40时 M_S 的可视化图, 各位置的值从主对角线向左右两侧辐射递减, DTU参数越大主对角线越宽.



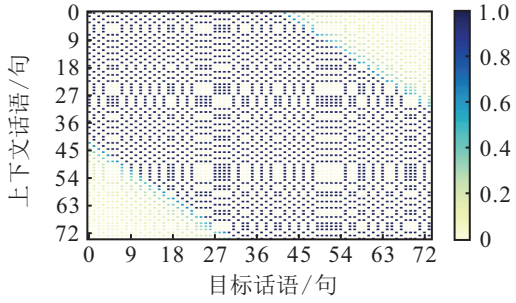
(a)　说话者掩码图(DTU=1)



(b)　说话者掩码图(DTU=40)



(c)　倾听者掩码图(DTU=1)



(d)　倾听者掩码图(DTU=40)

图 3　掩码矩阵可视化图

通过以上操作, 模型计算了同一说话人不同话语与知识之间的语义相关性, 相关性越大则由式(7)计算出的注意力权重越高. 下面通过加权融合的方式得到注意力结果矩阵 F_{o_i}, 具体过程如下:

$$F_{o_i}^1 = (F_{\alpha_i} \odot M_S) \times H_{\text{BiTAV}}^{\text{T}}, \tag{9}$$

$$F_{o_i}^2 = (F_{\alpha_i} \odot M_S) \times F_{k_i}^{\text{T}}, \tag{10}$$

$$F_{o_i} = W_s(F_{o_i}^1 \oplus F_{o_i}^2)^{\text{T}} + b_s. \tag{11}$$

其中: $F_{o_i}^1 \in \mathbb{R}^{N\times 2d_h}$ 为对多模态话语特征加权后的结果, $F_{o_i}^2 \in \mathbb{R}^{N\times 2d_h}$ 为对知识特征加权后的结果, 将这两个结果拼接后送入全连接层; $W_S \in \mathbb{R}^{2d_h\times 4d_h}$ 和 $b_s \in \mathbb{R}^{2d_h\times N}$ 分别为全连接层的权重矩阵和偏置项, 最终得到多模态话语特征与知识特征融合之后的结果 $F_{o_i} \in \mathbb{R}^{2d_h\times N}$. 同样地, 对于倾听者知识融合模块, 输入信息为 $(H_{\text{BiTAV}}, F_{L_i}, M_L)$, 其中 M_L 为倾听者加权掩码. 图3(c)和(d)是DTU参数分别为1和40时 M_L 的形式, 其中每一个元素 $M_{L_{i,j}}$ 的计算公式为

$$M_{L_{i,j}} = \begin{cases} \dfrac{1.9}{A}, & d > \text{DTU and } u(i) \neq u(j); \\ 1, & d \leqslant \text{DTU and } u(i) \neq u(j). \end{cases} \tag{12}$$

2.3.3　跨信息平均池化

受到池化中将小邻域内特征点整合得到新特征这一思想的启发, 设计一种跨信息平均池化方法, 分别将5个说话者特征矩阵和3个倾听者特征矩阵融合为一个特征矩阵, 计算过程如图4所示. 输入为 $D \times 2d_h \times N$ 的三维数据, 表示有 D 个 $2d_h \times N$ 的对话特征矩阵. 以说话者为例, 在信息数量 D 这一维度上进行平均池化, 最终得到一个融合了说话者知识信息的话语特征矩阵 $K_S \in \mathbb{R}^{2d_h\times N}$. 对倾听者进行同样的操作, 通过跨信息平均池化得到融合了倾听者知识的话语特征矩阵 $K_L \in \mathbb{R}^{2d_h\times N}$.

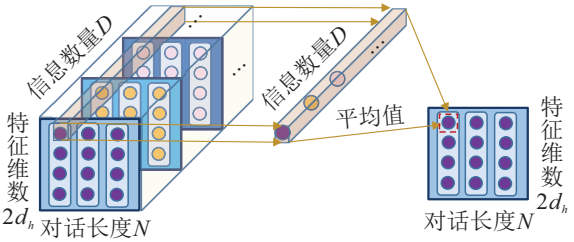


图 4　跨信息平均池化示意图

该方法将原三维空间中的特征点转化到二维空间中, 对知识融合模块输出的特征图进行特征保留和降维. 同时, 该操作没有引入外部参数, 减少了模型下一阶段的参数量和计算量, 能够防止过拟合.

2.4　情绪信息汇总阶段

为消除冗余信息, 最大程度提取出与情绪分类最相关的特征, 受多模态门控机制[28]的启发, 设计一种门控输出模块, 给不同特征矩阵分配不同的权重, 并通过加权融合的方式从融合知识后的信息和引入多模态上下文后的信息中筛选出有用特征.

首先将K_S、K_L和H_{BiTAV}两两组合,得到(K_S, K_L)、(K_S, H_{BiTAV})、(K_L, H_{BiTAV})三个特征对,将它们拼接后分别依次经过两层前馈神经网络,学习得到各个特征对应的权重向量z_1、z_2和z_3. 以(K_S, K_L)为例,具体计算过程如下:

$$K_{SL} = K_S \oplus K_L \oplus (K_S \odot K_L), \tag{13}$$

$$\text{Hidden}_{SL} = \tanh(W_{H_1}^{\text{T}} \cdot K_{SL} + b_z), \tag{14}$$

$$z_1 = \text{sigmoid}(W_z^{\text{T}} \cdot \text{Hidden}_{SL}). \tag{15}$$

其中: W_{H_1}和b_z分别为全连接层的权重矩阵和偏置项, W_z为sigmoid层的权重参数, $z_1 \in \mathbb{R}^{2d_h \times N}$. 接着通过加权融合的方式完成$K_S$、$K_L$和$H_{\text{BiTAV}}$三类特征的有效融合. 设$I \in \mathbb{R}^{2d_h \times N}$表示全1矩阵,权重分配的具体公式为

$$F_{fu} = (z_1 + z_2) \odot K_S + (I - z_1 + z_3) \odot K_L + (2I - z_3 - z_2) \odot H_{\text{BiTAV}}. \tag{16}$$

2.5 情绪决策阶段

为预测目标话语的情绪类别,将第2.4节得到的情绪汇总信息F_{fu}经过MLP层进一步激活,学习用于最终分类的情绪特征$E \in \mathbb{R}^{d_h \times N}$. 接着经过全连接层和Softmax归一化层,得到目标话语的情绪概率分布,计算过程如下:

$$\text{MLP}(X) = \text{ReLU}(W^{\text{T}} X + b), \tag{17}$$

$$E = \text{MLP}[\text{MLP}(F_{fu})], \tag{18}$$

$$P_{\text{pred}} = \text{Softmax}(W_{\text{class}}^{\text{T}} E + b_{\text{class}}). \tag{19}$$

其中: ReLU为MLP层的激活函数, W和b为MLP层的权重矩阵和偏置项, W_{class}和b_{class}分别为全连接层中可学习的权重矩阵和偏置项, $P_{\text{pred}} \in \mathbb{R}^{d_{\text{class}} \times N}$为对话中$N$句话语对应的情绪标签概率分布, d_{class}为情绪类别数量.

3 实验分析

所有实验均在同一台服务器上完成,具体环境为: CPU, 64×Xeon 6226R; GPU, RTX 3090 Ti; Python 3.7.0; CUDA 11.6; cuDNN 8.6.0; Pytorch 1.11.0.

3.1 实验数据集

选取两个公开基准数据集IEMOCAP和MELD进行实验. IEMOCAP中每个视频都包含一个二元对话,共7 380个片段,每个片段被标注为6类情绪中的一种: 快乐(happy)、悲伤(sad)、中性(neutral)、愤怒(angry)、兴奋(excited)和沮丧(frustrated),即$d_{\text{class}} = 6$. MELD来自美国情景喜剧Friends,每个会话包含多个说话人,每句话语被标注为7类情绪中的一种: 愤怒(anger)、厌恶(disgust)、悲伤(sadness)、喜悦(joy)、中立(neutral)、惊讶(surprise)和恐惧(fear),即$d_{\text{class}} = 7$.两个数据集的统计情况如表1所示.

表1 数据集统计说明

数据集	对话样本个数(话语个数)		
	训练集	验证集	测试集
IEMOCAP	100 (4 778)	20 (980)	31 (1 622)
MELD	1 038 (9 989)	114 (1 109)	280 (2 610)

3.2 实验设置

3.2.1 各模态原始特征提取

对于文本模态的原始特征,直接采用文献[18]提供的原始特征文件获取,即采用RoBERTa预训练语言模型抽取特征,特征嵌入维度为$d_t = 1\,024$. 对于音频模态原始特征和视觉模态原始特征,直接采用文献[13]提供的原始特征文件获取,即分别采用OpenSMILE语音分析框架和3D-CNN模型抽取特征. 对于IEMOCAP数据集,特征嵌入维度分别为$d_a = 100, d_v = 512$;对于MELD数据集,特征嵌入维度分别为$d_a = 300, d_v = 342$.

3.2.2 实验设置和评价指标

与其他多分类任务[29]一样,使用交叉熵损失函数作为训练的目标函数. 使用基于随机梯度下降的Adam优化器[30]优化模型参数. 超参数设置的具体情况如表2所示. 为衡量所提模型的性能,实验采用分类准确率Accuracy和Weighted-Average-F_1值评估模型的效果,简记为ACC和F_1,值越大分类效果越好. 在两个数据集上分别选取10个固定的随机种子实验,以减少实验过程的随机性,取10次实验结果的平均值及其标准差作为最终实验结果.

表2 模型超参数设置

超参数描述	参数值	
	IEMOCAP	MELD
批大小	32	32
学习率	1e-4	1e-4
DTU参数	40	1
权重衰减比率	1e-5	3e-4
隐藏层维度	300	200
Dense层的Dropout率	0.3	0.4
BiGRU层的Dropout率	0.15	0.15

3.3 对比实验及其结果分析

针对多模态对话情绪识别任务提出一种听说知识融合网络,为验证所提出模型的优越性,与其他先进模型在两个公开基准数据集上进行对比. 这些模型整体上可以分为三组: 1) 基于多模态的对话情绪

识别模型,包括双向动态双影响网络BiDDIN[24]、多模态深度图卷积网络MMGCN[10]、多模态动态融合网络MM-DFN[25]; 2)基于外部知识的对话情绪识别模型,包括因果感知交互网络CauAIN[19]、常识知识引导网络COSMIC[18]和知识驱动的双向情感循环网络BiERU-CSK[15]; 3)基于文本的对话情绪识别模型,包括基于BiLSTM和注意力的模型BcLSTM-Att[13]、基于3类GRU的序列模型DialogRNN[8]和对话情境推理网络DialogCRN[14].

针对以上3组基准模型分别进行3组对比实验,实验结果如表3所示. 相较于其他基准模型,LSKFN模型在3个分组、2个数据集上均取得了最优的结果,表明所提模型能够捕捉更有判别性的情绪特征,提升情绪识别效果,相较于其他模型具有优越性. 此外,所提模型的标准差普遍小于其他模型,表明模型具有较好的稳定性,多次运行仍然能够取得集中且优异的指标. 从表3的分组情况看, 引入外部知识的模型效果基本上优于仅使用文本模态的模型. 具体而言,由于对话双方存在大量隐含的背景知识,可能会出现情绪的跳跃,仅考虑上下文信息可能造成情绪的误判,因此引入知识可以丰富目标话语的语义信息,为识别提供更多的情绪线索. 基于多模态的模型效果整体优于基于文本的模型,表明音频、视觉等其他模态能够增强文本模态的情绪表达. LSKFN模型的识别效果在各分组下均超越其他基线方法,表明结合多模态感知信息和听说知识信息能够提升对话情绪识别效果.

表3　不同模型对比实验结果

模型分组	模型	IEMOCAP		MELD	
		F_1/%	ACC/%	F_1/%	ACC/%
基于多模态的模型	BiDDIN[24]	62.79(±0.62)	62.85(±0.71)	58.16(±0.17)	60.58(±0.40)
	MMGCN[10]	65.66(±0.42)	65.86(±0.53)	57.87(±0.27)	60.01(±0.51)
	MM-DFN [25]	66.99(±0.79)	67.40(±0.81)	57.73(±0.38)	61.01(±0.41)
	LSKFN(ours)	72.30(±0.45)	72.21(±0.52)	65.25(±0.22)	66.07(±0.26)
	Improve	5.31 ↑	4.81 ↑	7.09 ↑	5.06 ↑
基于外部知识的模型	COSMIC [18]	61.74(±1.11)	62.91(±0.94)	63.52(±0.46)	65.77(±0.28)
	BiERU-CSK [15]	63.21(±0.75)	63.13(±0.98)	56.64(±0.94)	59.94(±0.59)
	CauAIN [19]	66.09(±0.71)	65.93(±0.69)	64.76(±0.26)	65.45(±0.37)
	LSKFN(ours)	72.30(±0.45)	72.21(±0.52)	65.25(±0.22)	66.07(±0.26)
	Improve	6.21 ↑	6.28 ↑	0.49 ↑	0.30 ↑
基于文本的模型	DialogRNN [8]	61.87(±0.81)	62.12(±0.92)	57.30(±0.25)	59.50(±0.43)
	BcLSTM-Att [13]	62.25(±1.14)	62.91(±1.07)	57.11(±0.13)	59.40(±0.40)
	DialogCRN [14]	65.63(±0.44)	65.65(±0.54)	56.85(±0.35)	58.87(±0.57)
	LSKFN(ours)	69.67(±0.32)	69.54(±0.38)	65.03(±0.34)	65.79(±0.26)
	Improve	4.04 ↑	3.89 ↑	7.73 ↑	6.29 ↑

以上分析了模型的整体分类效果,为了进一步验证模型的优越性,在IEMOCAP数据集上进行单类情绪对比实验,结果如图5所示. LSKFN模型在所有单类情绪上的识别效果均超越了其他基线模型,尤其是在训练数据相对较少的happy情绪类别上,F_1值达到了61.22%,比第2高的模型提高了约21个百分点. 由此可知,LSKFN模型能够提升对话中单类情绪的识别效果.

3.4　消融实验及其结果分析

为验证所提模型中各模块对模型效果的影响以及不同模态的重要性,在其他条件保持不变的情况下分别设计了模块消融实验和模态消融实验.

3.4.1　模块消融

以原始LSKFN模型为基准模型,分别在该基准模型的基础上移除相应的模块后进行消融实验,结果如表4所示. 表中,I表示移除倾听者融合模块,直接将倾听者知识F_{L_1}、F_{L_2}和F_{L_3}简单拼接; II表示移除说话者融合模块,直接将说话者知识F_{S_1}、F_{S_2}、F_{S_3}、F_{S_4}和F_{s_5}简单拼接; III表示移除门控输出模块,直接将3种特征矩阵K_S、K_L和H_{BiTAV}拼接后预测情绪分布; IV表示移除上下文感知模块中的BiGRU,将多模态融合后的特征直接输入知识融合模块; V表示仅保留上下文感知模块和门控输出模块.

表4的实验结果显示,缺少任一模块时模型性能

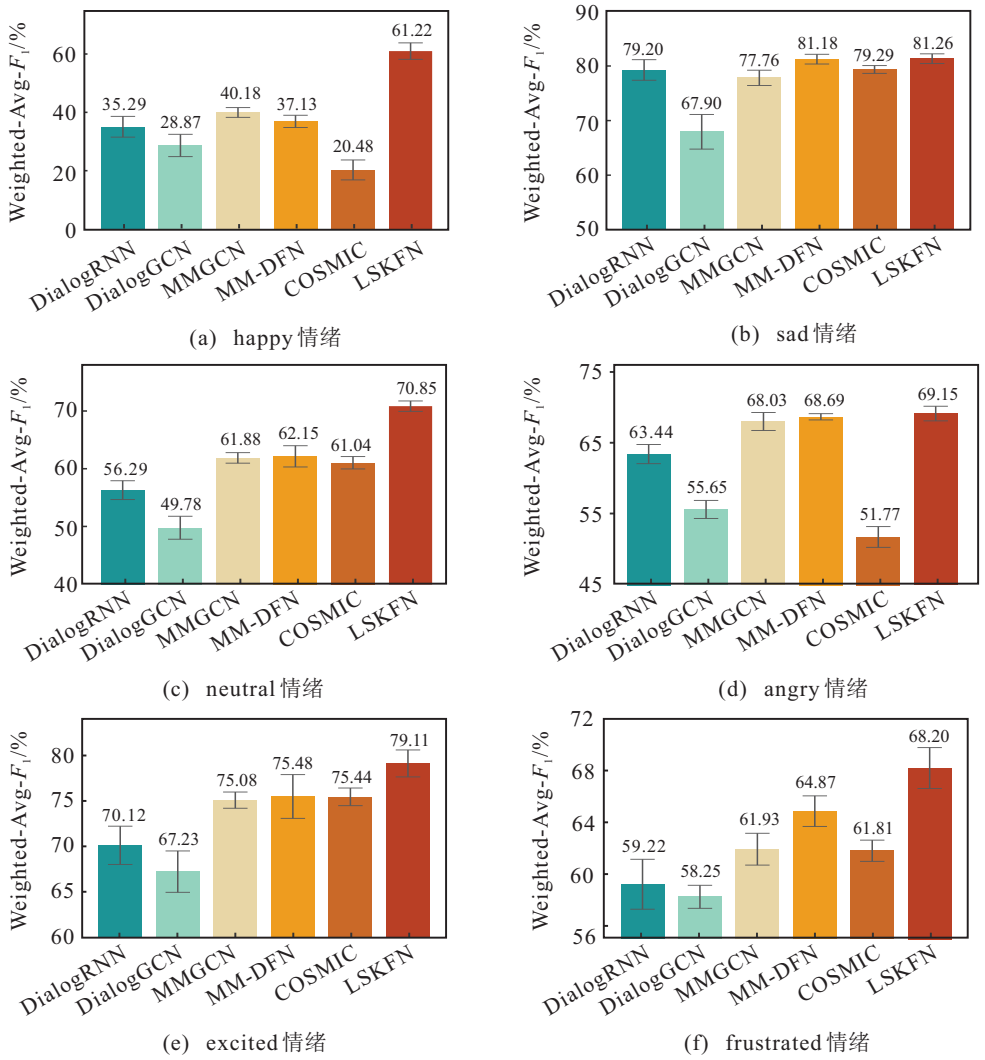


图 5 IEMOCAP 数据集上单类情绪对比实验结果

表 4 模块消融实验结果

模型	IEMOCAP		MELD	
	F_1/%	ACC/%	F_1/%	ACC/%
LSKFN	72.30	72.21	65.25	66.07
I	70.79	70.66	64.55	65.68
II	70.06	69.96	64.61	65.50
III	69.57	69.52	64.99	65.66
IV	69.54	69.29	64.44	65.18
V	66.66	66.54	64.25	65.41

均出现不同程度的下降,表明这些模块均对提升情绪识别效果起到一定的作用. 另外,如果只保留上下文感知模块和门控输出模块,F_1 值相比基线模型分别下降了 5.64% 和 1%;分类准确率 ACC 相比基线模型分别下降了 5.67% 和 0.66%. 由此可知,有效融合听说知识特征有助于增强模型的情绪特征捕捉能力,提高模型的识别性能.

3.4.2 模态消融

已有研究证明,有效融合多种模态信息可以提升情绪识别的效果[21]. 为探索不同模态信息对于多模态对话情绪识别任务的影响以及多模态融合策略对于情绪识别的作用,本文设计了 3 组消融实验进行分析,结果如表 5 所示. 其中:I 为单模态实验,仅以文本、音频或视觉特征作为模型的输入;II 利用早期融合策略分别进行双模态实验和三模态实验;III 利用成对注意力机制[31]融合不同模态信息,同样包括双模态实验和三模态实验.

由表 5 结果可知: 1)有文本模态参与的模型会取得更优异的结果,表明在对话情绪识别任务中,不同模态特征对于任务的贡献程度是不同的,文本特征能够给模型提供更丰富的语义信息且包含的情绪特性最显著,这与文献[28]所得的实验结论一致. 2)对比 II 和 III 的双模态实验可知,对于 IEMOCAP 数据集,以早期融合策略融合双模态信息可能引入噪声干扰,基于成对注意力的双模态融合机制更有效. 对于 MELD 数据集两种融合策略均会引入噪声特征,影响识别效果. 3)对比 II 和 III 的三模态实验结果,基于早期融合策略的三模态情绪识别效果均优于基于成对

注意力机制的三模态情绪识别效果,表明基于早期融合的三模态融合机制更有效. 此外,与单模态相比,融入音频和视觉模态可以带来模型性能的改进,表明相较于单模态信息,同时利用3种模态的信息能提供更多的情绪线索.

表 5　模块消融实验结果

分组	模态	IEMOCAP		MELD	
		F_1/%	ACC/%	F_1/%	ACC/%
I	T	69.67	69.54	65.03	65.79
	A	62.57	62.60	52.55	55.26
	V	58.75	58.99	51.34	55.32
II	TA	69.16	69.09	65.06	65.85
	TV	66.88	67.02	64.87	65.52
	AV	58.31	59.37	53.13	55.76
	TAV	72.30	72.21	65.25	66.07
III	TA	71.85	71.67	65.02	65.93
	TV	70.03	69.98	64.99	65.80
	AV	62.86	62.77	53.07	55.58
	TAV	70.99	70.90	65.04	65.71

综上所述,为有效融合多模态的信息,针对不同数据集、不同模态组合情况,应当选择不同的多模态融合策略. LSKFN模型输入3种模态的信息,采用早期融合策略获得了最优的分类准确率和F_1值.

4　结　论

针对多模态对话情绪识别任务,本文提出了一种听说知识融合网络(简称LSKFN). LSKFN通过知识融合模块和跨信息平均池化有效融合了多模态感知信息和听说知识信息,为目标话语捕获到了更有效的情绪特征. 此外,通过门控输出模块,消除了融合特征中的冗余信息,进一步提升了对话情绪识别模型的性能. 在两个经典数据集上的对比实验和消融实验结果表明,所提出LSFKN模型具有优越性.

然而,对于有多个对话者参与的MELD数据集,识别效果提升不明显,在后续的工作中将改进现有模型以提升多人对话情境下的多模态情绪识别效果. 此外,通过消融实验验证了外部知识对于多模态对话情绪识别任务具有积极作用,然而现有模型并未考虑外部知识选择的重要性,选择不当可能会给模型带来噪声,影响情绪识别效果. 因此,下一步工作将考虑针对该任务的外部知识选择机制.

参考文献(References)

[1] Simon H A. Motivational and emotional controls of cognition[J]. Psychological Review, 1967, 74(1): 29-39.

[2] Dolan R J. Emotion, cognition, and behavior[J]. Science, 2002, 298(5596): 1191-1194.

[3] 傅小兰. 情绪心理学[M]. 上海: 华东师范大学出版社, 2016: 136-174.
(Fu X L. Psychology of emotion[M]. Shanghai: East China Normal University Press, 2016: 136-174.)

[4] Ma Y K, Nguyen K L, Xing F Z, et al. A survey on empathetic dialogue systems[J]. Information Fusion, 2020, 64: 50-70.

[5] Andalibi N, Buss J. The human in emotion recognition on social media: Attitudes, outcomes, risks[C]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu, 2020: 1-16.

[6] 李杨, 徐泽水, 王新鑫. 基于在线评论的情感分析方法及应用[J]. 控制与决策, 2023, 38(2): 304-317.
(Li Y, Xu Z S, Wang X X. Methods and applications of sentiment analysis with online reviews[J]. Control and Decision, 2023, 38(2): 304-317.)

[7] Jannach D, Manzoor A, Cai W L, et al. A survey on conversational recommender systems[J]. ACM Computing Surveys, 2021, 54(5): 1-36.

[8] Majumder N, Poria S, Hazarika D, et al. Dialogue RNN: An attentive RNN for emotion detection in conversations[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 6818-6825.

[9] Ghosal D, Majumder N, Poria S, et al. Dialogue GCN: A graph convolutional neural network for emotion recognition in conversation[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, 2019: 154-164.

[10] Hu J W, Liu Y C, Zhao J M, et al. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation[C]. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, 2021: 5666-5675.

[11] Mao Y Z, Liu G, Wang X J, et al. Dialogue TRM: Exploring multi-modal emotional dynamics in a conversation[C]. Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, 2021: 2694-2704.

[12] 熊宇, 张一飞, 冯时, 等. 基于多模态特征深度融合的微博流事件检测与跟踪[J]. 控制与决策, 2019, 34(7): 1409-1416.
(Xiong Y, Zhang Y F, Feng S, et al. Event detection and tracking in microblog stream based on multimodal feature deep fusion[J]. Control and Decision, 2019, 34(7): 1409-1416.)

[13] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos[C]. Proceedings of the 55th Annual Meeting of the

Association for Computational Linguistics. Vancouver, 2017: 873-883.

[14] Hu D, Wei L W, Huai X Y. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations[C]. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, 2021: 7042-7052.

[15] Li W, Shao W, Ji S X, et al. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis[J]. Neurocomputing, 2022, 467: 73-82.

[16] Shen W Z, Wu S Y, Yang Y Y, et al. Directed acyclic graph network for conversational emotion recognition[C]. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processin. Stroudsburg, 2021: 1551-1560.

[17] Lee B, Choi Y S. Graph based network with contextualized representations of turns in dialogue[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, 2021: 443-455.

[18] Ghosal D, Majumder N, Gelbukh A, et al. COSMIC: Commonsense knowledge for emotion identification in conversations[C]. Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, 2020: 2470-2481.

[19] Zhao W X, Zhao Y Y, Lu X. CauAIN: causal aware interaction network for emotion recognition in conversations[C]. Proceedings of the 21st International Joint Conference on Artificial Intelligence. Vienna, 2022: 4524-4530.

[20] Xie Y H, Yang K L, Sun C J, et al. Knowledge-interactive network with sentiment polarity intensity-aware multi-task learning for emotion recognition in conversations[C]. Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg, 2021: 2879-2889.

[21] 王雨竹, 谢珺, 陈波, 等. 基于跨模态上下文感知注意力的多模态情感分析[J]. 数据分析与知识发现, 2021, 5(4): 49-59.
(Wang Y Z, Xie J, Chen B, et al. Multi-modal sentiment analysis based on cross-modal context-aware attention[J]. Data Analysis and Knowledge Discovery, 2021, 5(4): 49-59.)

[22] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42(4): 335-359.

[23] Poria S, Hazarika D, Majumder N, et al. MELD: A multimodal multi-party dataset for emotion recognition in conversations[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019: 527-536.

[24] Zhang D, Zhang W S, Li S S, et al. Modeling both intra- and inter-modal influence for real-time emotion detection in conversations[C]. Proceedings of the 28th ACM International Conference on Multimedia. Seattle, 2020: 503-511.

[25] Hu D, Hou X L, Wei L W, et al. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, 2022: 7037-7041.

[26] Sap M, Le Bras R, Allaway E, et al. ATOMIC: An atlas of machine commonsense for if-then reasoning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 3027-3035.

[27] Bosselut A, Rashkin H, Sap M, et al. COMET: commonsense transformers for automatic knowledge graph construction[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019: 4762-4779.

[28] 张峰, 李希城, 董春茹, 等. 基于深度情感唤醒网络的多模态情感分析与情绪识别[J]. 控制与决策, 2022, 37(11): 2984-2992.
(Zhang F, Li X C, Dong C R, et al. Deep emotional arousal network for multimodal sentiment analysis and emotion recognition[J]. Control and Decision, 2022, 37(11): 2984-2992.)

[29] 张哲益, 曹卫华, 朱蕊, 等. 基于脉冲卷积神经网络稀疏表征的高分辨率遥感图像场景分类方法[J]. 控制与决策, 2022, 37(9): 2305-2313.
(Zhang Z Y, Cao W H, Zhu R, et al. Sparse representation with spike convolutional neural networks for scene classification of remote sensing images of high resolution[J]. Control and Decision, 2022, 37(9): 2305-2313.)

[30] Diederik P K, Jimmy B. Adam: A method for stochastic optimization[C]. Proceedings of the 3rd International Conference on Learning Representations. San Diego: Open Review, 2015: 1-15.

[31] Ghosal D, Akhtar M S, Chauhan D, et al. Contextual inter-modal attention for multi-modal sentiment analysis[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, 2018: 3454-3466.

作者简介

刘琴 (1997−), 女, 硕士生, 从事多模态情感计算的研究, E-mail: lq_cugb@163.com;

谢珺 (1979−), 女, 副教授, 博士, 从事数据挖掘、推荐系统等研究, E-mail: xiejun@tyut.edu.cn;

胡勇 (1979−), 男, 教授, 博士, 从事虚拟现实/混合现实内容智能生成与交互设计等研究, E-mail: huyong@buaa.edu.cn;

郝戍峰 (1991−), 男, 讲师, 博士, 从事信息检索、情感分析等研究, E-mail: haoshufeng@tyut.edu.cn;

郝雅卉 (1999−), 女, 硕士生, 从事情感分析的研究, E-mail: 2101675508@qq.com.