DOI:10.19651/j.cnki.emt.2519542

基于对话的多模态情感分析技术研究*

赵亚芳 梁志剑
(中北大学计算机科学与技术学院 太原 030051)

摘 要:针对多模态对话情感识别(MERC)中难以有效捕捉对话中跨模态语义关联以及对少数类和语义易混淆类情感的区分能力有限的问题,提出了一种新的多模态情感分析模型(FuseNet)。该模型采用双向注意力对话编码器(BiDRN)以捕捉对话上下文依赖,有效整合来自不同说话人的音频与视觉线索,并通过基于分层门控机制的融合模块实现动态多模态融合,同时引入类感知多模态对比(CAMC)损失以增强类间判别性,提升对少数类以及语义相近情感类别的区分能力。在 IEMOCAP 和 MELD 两个基准 ERC 数据集上的实验结果表明,与当前先进模型 CORECT 相比,FuseNet 的 F1 分数分别提升了 2.91%和 2.00%,在多数情感类别的分类性能上均优于现有基线模型,尤其在识别少数类和语义相近类情感上改进显著。

关键词:多模态对话情感识别;双向注意力;分层门控机制;动态多模态融合;对比损失
中图分类号:TP391;TN912.34 **文献标识码**:A **国家标准学科分类代码**:520.20

Research on multimodal sentiment analysis technology based on conversations

Zhao Yafang Liang Zhijian
(School of Computer Science and Technology, North University of China, Taiyuan 030051, China)

Abstract: Focused on the issue that multimodal emotion recognition in conversation (MERC) is difficult to effectively capture cross-modal semantic associations in conversation rounds and has limited discrimination ability for minority classes and semantically confusing classes of emotions, a new multimodal sentiment analysis model (FuseNet) is proposed. This model adopts the bidirectional attention dialogue encoder (BiDRN) to capture the context dependency of the dialogue, effectively integrates audio and visual cues from different speakers, and realizes dynamic multimodal fusion through the Hi-gated fusion module based on the hierarchical gated mechanism. Meanwhile, class-aware multimodal contrastive (CAMC) loss is introduced to enhance the inter-class discriminability and improve the discrimination ability of minority classes and semantically similar sentiment categories. Experimental results on the two benchmark ERC datasets of IEMOCAP and MELD show that compared with the current advanced model CORECT, the F1 score of the proposed framework has improved by 2.91% and 2.00%, respectively, which are better than the existing baseline model in terms of classification performance in most emotions, especially in identifying a few classes and semantic similar categories of emotions.

Keywords: multimodal emotion recognition in conversation; bidirectional attention; hierarchical gated mechanism; dynamic multimodal fusion; contrastive loss

0 引 言

对话中的多模态情感识别(multimodal emotion recognition in conversation, MERC)作为自然语言处理(natural language processing, NLP)领域内一个备受关注的研究方向,通过综合利用说话者的文本信息、音频特征以及视觉线索来准确识别对话中每一句话语所表达的情感状态。在情感脑机接口[1]、自闭症干预[2]、意见挖掘[3]、构建移情对话系统[4]等多个实际应用场景中展现出的巨大潜力,已然成为学术界与工业界共同关注的焦点。

根据是否利用说话人信息,可以将现有的研究方法划分为两大类别:不依赖说话人信息方法与说话人相关方法。

收稿日期:2025-08-05
* 基金项目:山西省 2024 年度研究生教育创新计划项目(2024SZ23)、2024 年山西省高等学校教学改革创新项目(J20240839)资助

不依赖说话人信息方法将对话视为普通文本序列,忽略说话者身份差异,直接对连续语句进行情感分类。早期主要基于单模态文本,采用层次化神经网络建模上下文。例如,BiLSTM[5]将对话视为时间序列,通过循环神经网络捕捉上下文依赖;HiGRU[6]包含两个门控循环单元,分别用于模拟单词和话语之间的上下文关系,而不建模说话者的状态。

说话人相关方法不仅会对上下文信息进行处理,还会对说话人敏感的依赖关系进行建模。例如,HiTrans[7]由两个分层转换器组成,用于捕获全局上下文信息,并利用一个辅助任务来建模说话人敏感的依赖关系。DialogueTRM[8]在Transformer层中加入说话人感知位置编码,区分同一说话人不同轮次的话语。EmoDM[9]为每个说话人维护独立的情感记忆模块,通过查询记忆库预测当前情感。

然而,大多数现有情感识别方法主要聚焦于简单的多模态拼接,难以有效捕捉跨模态间的复杂语义关联与对话轮次的双向依赖。目前,通过多模态特征融合实现高性能的目标识别分类仍是当前面临的一个难题[10],因此仍存在一些未解决的挑战:1)对话上下文表征单一[11]:单向循环神经网络架构难以捕捉对话中后续轮次对前文情感的修正(如反讽、隐含情绪),导致上下文连贯性建模不完整。与传统的基于上下文无关句子的情感识别不同,对上下文和说话人敏感的依赖关系进行建模是MERC的核心[12]。多模态情感分析模型通常只考虑到单个轮次的对话内容[13]。然而,在双人或多人对话的实际情境里,不同轮次对话所承载的情感与语义之间存在着紧密的内在联系。2)长尾情感类别识别不足[14]:对少数情感类别和语义易混淆情感类别的区分能力有限。现有的两个ERC基准数据集MELD[15]和IEMOCAP[16]都存在情感类别分布失衡的问题。目前最先进的方法在解决类别不平衡问题上,性能很差。此外,语义相近的情感类别常常具有相似的认知和情感表达,语义界限并不清晰,进一步增加了准确分类情感的难度。

针对以上问题,提出了一种新的多模态情感分析框架,称为FuseNet。FuseNet包含两个核心组件,即BiDRN和Hi-Gated Fusion模块。第1,对文本、音频、视觉模态进行单模态特征提取和上下文建模,其中引入对话上下文编码模块BiDRN,该提取器通过双向注意力机制能够同时考虑对话的历史和未来上下文音频以及视觉信息,增强对对话中复杂语义和情感转换的捕获能力;第2,提出了一种称为Hi-Gated Fusion的多模态融合模块,通过分层动态门控机制自适应地学习不同模态间的关联权重,实现从局部特征到全局语义的多尺度特征整合;第3,为了减轻识别少数和语义上易混淆的情感类别的困难,提出了一种基于类感知的多模态对比(CAMC)损失方法,其中,通过类内紧凑少数类样本彼此之间的接近程度,挖掘困难负样本以提高少数类的学习权重,并通过类间分离特征空间里易混淆类别的距离,提升语义相近情感类别的区分能力;第4,在两个基准ERC数据集上进行评估:IEMOCAP和MELD,实验表明FuseNet框架具有稳定性和优越性。

本文的贡献可概括为以下要点:

1)本文提出了一种基于双向注意力机制的对话上下文编码器BiDRN,它可以有效地捕获对话连贯性:通过分离前向历史状态和后向未来状态的注意力,对音频以及视觉模态的上下文进行显式建模,有效地整合来自不同说话人的音频以及视觉线索。

2)本文设计了一个基于分层门控的多模态动态融合模型Hi-Gated Fusion,该模型实现了多模态的细粒度特征对齐与自适应融合,从而有效捕捉跨模态的深层语义关联。

3)本文引入了CAMC损失来减轻识别少数和语义上易混淆的情感类别的困难。

4)在MELD和IEMOCAP两个基准数据集上进行了广泛的实验,结果表明,提出的FuseNet框架在两个数据集上都达到了先进的性能,在少数和语义易混淆情感类别方面的改进显著。

1 方 法

1.1 定 义

在对话中的多模态情感分析任务中,给定一段包含 n 个话语单元(记作 $\{\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_n\}$ 以及 m 个参与对话的说话者(记作 $\{s_1,s_2,\cdots,s_m\}$)的对话数据。针对每个话语单元 u_i,其信息呈现形式为多模态融合,具体涵盖文本模态(表示为 $\boldsymbol{u}_i^t$)、音频模态(表示为 $\boldsymbol{u}_i^a$)以及视觉模态(表示为 $\boldsymbol{u}_i^v$)3个维度。该任务的核心目标在于基于上述多模态信息精准预测每个话语单元 $\boldsymbol{u}_i$ 所对应的情感类别标签。

1.2 任务概述

图1给出了FuseNet框架的概述。FuseNet有两个重要模块:一个是在提取了话语级单模态特征后,用于捕获不同话语之间的模态内和模态间相互作用的模态编码器模块;另一个是用于自适应学习模态之间权重的分层门控动态融合模块。此外,引入了基于类感知的对比学习损失,以便在每个模态内从模型中转移知识,以学习更好的模态表示。

1.3 单模态特征提取

提取有效的特征是进行情绪分类的前提条件[17]。FuseNet提取话语级单模态特征过程包括:

1)文本特征提取:利用RoBERTa[18]大模型提取文本特征。对于特征表示,在RoBERTa的最后一层采用[CLS]标记的嵌入,输出的特征向量维度为256。

2)音频特征提取:利用OpenSMILE[19]工具提取音频特征。经过OpenSMlLE处理后,为每个语音音频提取6 373维的特征表示,然后采用一个全连接层来获取每个输入音频的512维特征。

3)视觉特征提取:利用DenseNet[20](densely connected convolutional networks)模型提取视觉特征,DenseNet是

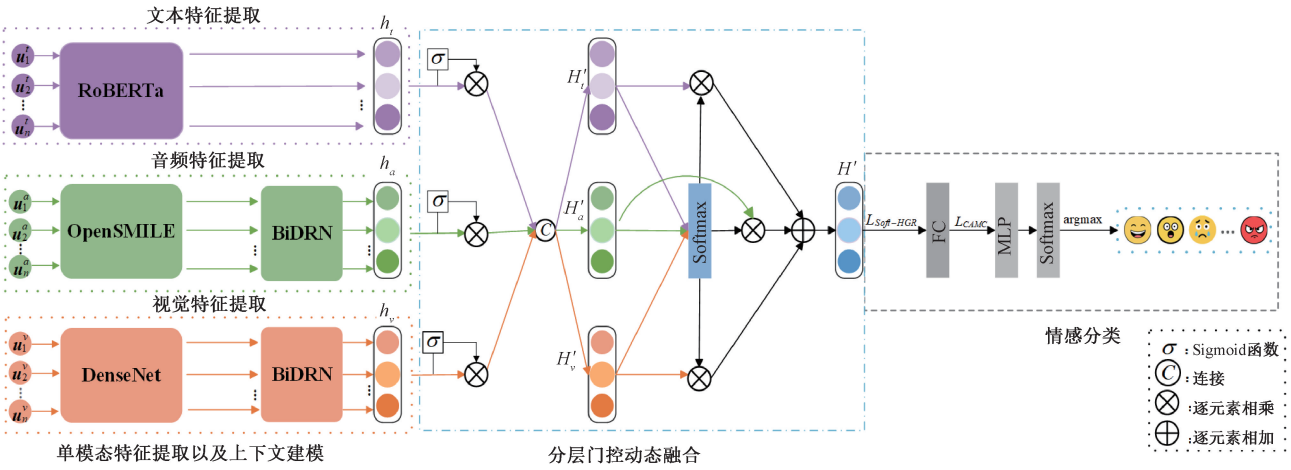

图 1　FuseNet 模型结构

Fig. 1　Network architecture of FuseNet

一种卷积神经网络(CNN)，通过密集连接(dense block)实现层间的特征复用。DenseNet 输出一个维度为 342 的特征表示。

1.4　对话上下文建模

单模态的特征识别精度较低，实时性差[21]。因此，在提取了话语级单模态特征后，引入基于双向响应注意力机制的对话上下文编码器 BiDRN，用于捕获不同话语之间在音频及视觉模态上的相互作用。与现有的 DialogueRNN 主要基于单向历史上下文进行建模不同，BiDRN 通过分离前向历史状态和后向未来状态的注意力，分别对音频和视觉模态的上下文进行协同建模，从而更完整地捕捉对话上下文依赖，最终获得音频及视觉模态上下文特征。

BiDRN 包含两个主要组件：前向 DialogueRNN 与后向 DialogueRNN，分别用于捕获对话的历史和未来上下文信息。这一双向架构突破了原始 DialogueRNN 仅沿时间顺序建模的局限，特别地，在后向 DialogueRNN 中，加入了响应注意力机制，该机制能够融合来自前向历史状态与后向未来状态的注意力权重，生成统一的上下文表征，从而更准确地理解当前话语的情感意图，最终将双向信息融合为模态特征，如图 2 所示。

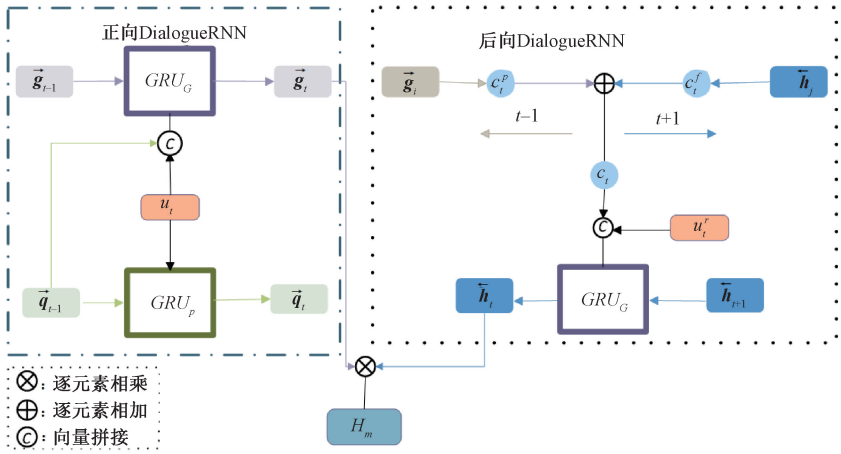

图 2　BiDRN 模型结构

Fig. 2　Network architecture of BiDRN

前向 DialogueRNN 处理原始对话序列 $\boldsymbol{U}=\{\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_T\}$，通过两个门控循环单元(GRU)对全局状态、说话者状态进行建模，输出时间步 t 下新的全局状态：

$$\vec{\boldsymbol{g}}_t=GRU_G([\boldsymbol{u}_t\oplus\vec{\boldsymbol{q}}_{t-1}],\vec{\boldsymbol{g}}_{t-1}) \tag{1}$$

$$\vec{\boldsymbol{q}}_t=GRU_P([\boldsymbol{u}_t\oplus\vec{\boldsymbol{g}}_t],\vec{\boldsymbol{q}}_{t-1}) \tag{2}$$

其中，$\vec{\boldsymbol{g}}_t$ 表示的前向全局状态，$\vec{\boldsymbol{q}}_t$ 表示说话人在对话中的状态，⊕ 为向量拼接。

后向 DialogueRNN 则处理反转的原始对话序列 $\boldsymbol{U}^r=\{\boldsymbol{u}_T,\boldsymbol{u}_{T-1},\cdots,\boldsymbol{u}_1\}$，引入响应注意力机制分离前向历史状态和后向未来状态的注意力分数，归一化权重后，生成上下文表征向量，对于时间步 t，有：

$$e_{t,i}^{p}=\boldsymbol{v}^{\top}\tanh(\boldsymbol{W}\vec{\boldsymbol{g}}_i+\boldsymbol{U}\overleftarrow{\boldsymbol{h}}_t) \tag{3}$$

$$e^{f}_{t,j} = \boldsymbol{v}^{\top} \tanh(\boldsymbol{W}\vec{\boldsymbol{h}}_j + \boldsymbol{U}\overleftarrow{\boldsymbol{h}}_t) \tag{4}$$

$$\alpha^{p}_{t,i} = \frac{\exp(e^{p}_{t,i})}{\sum_{k}^{t-1}\exp(e^{p}_{t,k})} \tag{5}$$

$$\alpha^{f}_{t,j} = \frac{\exp(e^{f}_{t,j})}{\sum_{k=t+1}^{T}\exp(e^{f}_{t,k})} \tag{6}$$

$$\boldsymbol{c}^{p}_{t} = \sum_{i=1}^{t-1}\alpha^{p}_{t,i}\vec{\boldsymbol{g}}_i \tag{7}$$

$$\boldsymbol{c}^{f}_{t} = \sum_{j=t+1}^{T}\alpha^{f}_{t,j}\overleftarrow{\boldsymbol{h}}_j \tag{8}$$

$$\boldsymbol{c}_t = \boldsymbol{c}^{p}_{t} + \boldsymbol{c}^{f}_{t} \tag{9}$$

其中,$i \in [1,t-1]$,$j \in [t+1,T]$,$e^{p}_{t,i}$ 表示当前后向状态 $\overleftarrow{\boldsymbol{h}}_t$ 对每个前向状态 $\vec{\boldsymbol{g}}_i$ 的注意力分数,$e^{f}_{t,j}$ 表示后向状态 $\overleftarrow{\boldsymbol{h}}_j$ 对当前状态 $\overleftarrow{\boldsymbol{h}}_t$ 的注意力分数,$\boldsymbol{c}_t$ 表示历史和未来上下文向量的融合。$\boldsymbol{W} \in \mathbb{R}^{d\times d}$ 和 $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ 表示权重矩阵。

上下文表征向量输入到门控循环单元中,输出时间步 t 下反转序列新的全局状态。最终,双向融合后得到模态特征,计算过程如式(10)、(11)所示。

$$\overleftarrow{\boldsymbol{h}}_t = GRU_G([\boldsymbol{u}^{r}_{t} \oplus \boldsymbol{c}_t], \overleftarrow{\boldsymbol{h}}_{t+1}) \tag{10}$$

$$\boldsymbol{H}_m = RELU(\boldsymbol{W}_h[\vec{\boldsymbol{g}}_t \otimes \overleftarrow{\boldsymbol{h}}_t] + \boldsymbol{b}_h) \tag{11}$$

其中,$\boldsymbol{H}_m$ 表示时间步 t 下的模态特征,$\otimes$ 为逐元素相乘。

1.5 多模态特征融合

在多模态特征融合部分,引入了一个基于分层门控机制的多模态动态融合模型 Hi-Gated Fusion,自适应地强化单模态序列的表征能力,生成增强后的单模态序列表示,并分别动态学习模态表示之间的权重,捕捉深层次的复杂语义关联。

1)单模态门控增强

每个模态的特征通过门控机制进行跨模态增强,过滤无关噪声并保留对目标模态有用的跨模态信息:

$$\boldsymbol{g}_m = sigmoid(\boldsymbol{W}_m \cdot \boldsymbol{H}_m) \tag{12}$$

$$\boldsymbol{H}'_m = \boldsymbol{H}_m \otimes \boldsymbol{g}_m \tag{13}$$

其中,$m \in \{t,a,v\}$ 表示文本、音频、视觉模态,$\boldsymbol{W}_m \in \mathbb{R}^{d\times d}$ 为权重矩阵,$\otimes$ 为逐元素相乘,$\boldsymbol{g}_m$ 表示门控权重。

然后,将过滤后的模态交互信息 $\boldsymbol{H}'_{m\to m}$,$\boldsymbol{H}'_{n_1\to m}$,$\boldsymbol{H}'_{n_2\to m}$ 拼接,并通过全连接层生成增强后的单模态表示:

$$\boldsymbol{H}'_m = \boldsymbol{W}_m \cdot [\boldsymbol{H}'_{m\to m};\boldsymbol{H}'_{n_1\to m};\boldsymbol{H}'_{n_2\to m}] + \boldsymbol{b}_m \tag{14}$$

其中,n_1 和 n_2 表示其他两种模态,$\boldsymbol{H}'_m \in \mathbb{R}^{N\times d}$,$\boldsymbol{W}_m \in \mathbb{R}^{3d\times d}$ 和 $\boldsymbol{b}_m \in \mathbb{R}^{d}$ 为可训练参数。

因此,增强后的单模态序列可表示为 $\boldsymbol{H}'_m = [\boldsymbol{h}'_{m_1};\boldsymbol{h}'_{m_2};\cdots;\boldsymbol{h}'_{m_N}]$,$\boldsymbol{h}'_{m_i}$ 是第 i 句话的增强模态表示。

2)多模态动态融合

通过 $Softmax$ 函数动态学习各模态的融合权重,可得到对话的多模态序列,计算过程如式(15)、(16)所示。

$$[\boldsymbol{g}_{ti};\boldsymbol{g}_{ai};\boldsymbol{g}_{vi}] = softmax([\boldsymbol{W}\cdot\boldsymbol{h}'_{ti};\boldsymbol{W}\cdot\boldsymbol{h}'_{ai};\boldsymbol{W}\cdot\boldsymbol{h}'_{vi};]) \tag{15}$$

$$\boldsymbol{h}'_i = \sum_{m\in\{t,a,v\}}\boldsymbol{h}'_{mi}\otimes\boldsymbol{g}_{mi} \tag{16}$$

其中,$\boldsymbol{W} \in \mathbb{R}^{d\times d}$ 是共享权重矩阵,$\boldsymbol{g}_{ti}$、$\boldsymbol{g}_{ai}$、$\boldsymbol{g}_{vi}$ 分别为文本、音频、视觉模态的动态权重。

因此,最终得到整个对话的多模态序列 $\boldsymbol{H}' = [\boldsymbol{h}'_1;\boldsymbol{h}'_2;\cdots;\boldsymbol{h}'_N]$。

1.6 情感分析

在多模态融合之后,将组合的多模态语义信息 H' 的特征输入到一个具有全连接层的多层感知器 MLP 中,然后使用 $RELU$ 激活函数进行非线性激活,并通过 $Softmax$ 函数对隐藏层的特征信息 P_i 进行归一化。计算过程如式(17)、(18)所示。

$$\boldsymbol{l}_i = RELU(\boldsymbol{W}_l\boldsymbol{H}' + \boldsymbol{b}_l) \tag{17}$$

$$\boldsymbol{P}_i = Softmax(\boldsymbol{W}_{smax}\boldsymbol{l}_i + \boldsymbol{b}_{smax}) \tag{18}$$

其中,$\boldsymbol{W}_l$ 和 $\boldsymbol{W}_{smax}$ 表示一个可学习的权重矩阵。最后,使用 $argmax$ 函数计算情感类别集上的概率分布,其中选择概率最高的情感标签作为第 i 个话语的预测 $\hat{y}_i$。

$$\hat{y}_i = argmax(\boldsymbol{P}_i[k]) \tag{19}$$

1.7 训练目标

给定由 M 个对话组成的一批 N 个样本,其中第 i 个对话包含 $C(i)$ 个话语,训练目标定义如下:

CAMC 损失:如图 3 和 4 所示,MELD 和 IEMOCAP 数据集存在情感类别分布失衡以及语义易混淆的问题。例如,MELD 中大多数类“中性”情感所占的比例远远大于少数类“厌恶”和“恐惧”情感。现有的模型会倾向于学习到多数类情感的特征,而对少数类情感的识别能力则相对较弱。IEMOCAP 中的“快乐”和“兴奋”这两种情绪,语义界限并不清晰,说话者常常会在相近的实际语境条件下表达出这两种情绪,进一步增加了准确分类的难度。

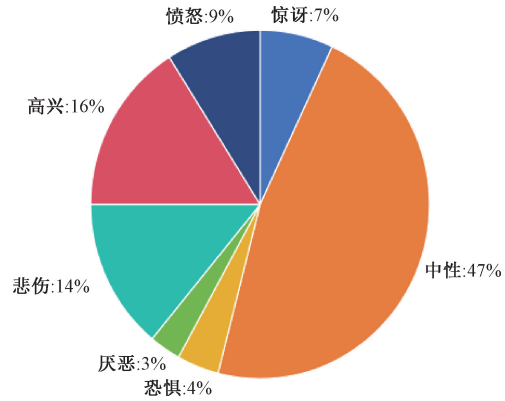

图 3 MELD 数据集上的情感类别分布

Fig. 3 The distribution ratio chart of emotion categories on the MELD dataset

为了减轻识别少数和语义上易混淆的情感类别的困难,引入基于类感知多模态对比损失(class-aware multimodal

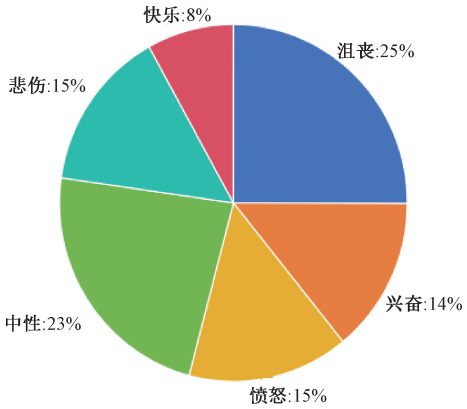

图 4　IEMOCAP 数据集上的情感类别分布

Fig. 4　The distribution ratio chart of emotion categories on the IEMOCAP dataset

contrastive，CAMC)方法。其中，通过类内紧凑少数类样本彼此之间的接近程度，并且增加了困难样本挖掘，主动加强少数类样本的对比学习权重，提高了模型对少数情绪类别识别能力。通过类间分离特征空间里易混淆类别的距离，加强对语义易混淆情绪的区分能力。CAMC 损失定义如下：

多模态融合的特征向量使用余弦相似度计算特征向量对之间的相似性矩阵 $\boldsymbol{S} \in \mathbb{R}^{N\times N}$，构造标签匹配矩阵 $\boldsymbol{M}$，若 $y_i = y_j$，$M_{i,j} = 1$，否则 $M_{i,j} = 0$，可以得到正样本(类内)以及负样本(类间)相似度：

$$S_{i,j} = \exp\left(\frac{\boldsymbol{f}_i \cdot \boldsymbol{f}_j}{\tau}\right) \tag{20}$$

$$P_i = \sum_{j=1}^{N} S_{i,j} M_{i,j} \tag{21}$$

$$N_i = \sum_{j=1}^{N} S_{i,j}(1 - M_{i,j}) \tag{22}$$

其中，$\boldsymbol{f}_i$ 为第 i 个样本的融合多模态特征向量，则 y_i 为相对应的情感标签，τ 表示温度参数。

此外，为增强对少数类的学习，引入困难负样本挖掘，从每个样本的负样本相似度中选择前 $k = max(1, N \cdot \rho)$ 个最大值，构成困难负样本相似度和：

$$H_i = \sum_{j\in\mathcal{H}_i} N_i \tag{23}$$

其中，$\mathcal{H}_i$ 为为第 i 个样本的 k 个最大负样本相似度的索引集合。

最后，通过对比正样本相似度与类内、类间及困难负样本的综合相似度得到 CAMC 损失：

$$L_{CAMC} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\sum_{j\in P_i} s_{i,j}}{\sum_{j\in P_i} s_{i,j} + \sum_{j\in N_i} s_{i,j} + \sum_{j\in H_i} s_{i,j}}\right) \tag{24}$$

Soft-HGR 损失[22]：利用 Soft-HGR 损失使多模态融合文本、音频和视觉特征之间的相关性最大化：

$$L_{Soft-HGR} = -\sum_{\boldsymbol{Q}\neq\boldsymbol{V},\boldsymbol{Q},\boldsymbol{V}\in F}\left(\mathbb{E}[\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{V}] - \frac{1}{2}\mathrm{tr}(\mathrm{cov}(\boldsymbol{Q})\,\mathrm{cov}(\boldsymbol{V}))\right)$$
$$s.t.\ \ E[\boldsymbol{Q}] = 0,\ \forall \boldsymbol{Q} \in F \tag{25}$$

其中，特征集 $F = \{\boldsymbol{F}^t, \boldsymbol{F}^a, \boldsymbol{F}^v\}$，文本特征 $\boldsymbol{F}^t = \{\boldsymbol{f}_1^t, \cdots, \boldsymbol{f}_N^t\}$，音频特征 $\boldsymbol{F}^a = \{\boldsymbol{f}_1^a, \cdots, \boldsymbol{f}_N^a\}$，视觉特征 $\boldsymbol{F}^v = \{\boldsymbol{f}_1^v, \cdots, \boldsymbol{f}_N^v\}$。通过样本均值和样本协方差来近似期望和协方差。

交叉熵损失：采用交叉熵损失来衡量预测概率与真实标签之间的差异：

$$L_{CE} = -\sum_{i=0}^{M}\sum_{j=1}^{C(i)}\log P_{i,j}[y_{i,j}] \tag{26}$$

其中，$P_{i,j}$ 是预测第 i 个样本属于类别 j 的概率，$y_{i,j}$ 是第 i 个样本的真实标签。

总损失函数：利用 CAMC 损失、Soft-HGR 损失和交叉熵损失的线性组合作为总损失函数：

$$L = \frac{1}{N}(\mu_1 L_{CAMC} + \mu_2 L_{Soft-HGR} + \mu_3 L_{CE}) + \lambda \| \boldsymbol{\theta} \|_2^2 \tag{27}$$

其中，μ_1 和 μ_2 为可调超参数，λ 为 L2 正则化权值，$\boldsymbol{\theta}$ 为所有可训练参数的集合。

2　实验设置

2.1　数据集和评价指标

本文使用 IEMOCAP 和 MELD 这两个多模态情感分析领域的基准数据集来评估所提出的 FuseNet 模型。两个数据集的统计数据详如表 1 所示。

表 1　在 IEMOCAP 和 MELD 数据集上的统计数据

Table 1　The statistics of IEMOCAP and MELD datasets

数据集	话语		对话		情感类别
	训练集和验证集	测试集	训练集和验证集	测试集	
IEMOCAP	5 810	1 623	120	31	6
MELD	11 098	2 610	1 153	280	7

IEMOCAP：数据集包括 10 位说话者参与的双向对话场景，涵盖 151 段对话，总计 7 433 条话语。它被分为 5 个会话，前 4 个会话用于训练，最后一个会话用于测试。每个话语都被标记为 6 种情绪中的一种：快乐、悲伤、中性、愤怒、兴奋和沮丧。

MELD：数据集来自电视剧《老友记》，从中收集了包含多个说话者的对话内容，共包含 1 433 段对话，涉及 13 708 条话语。在情感标注方面，每一条话语均被归为 7 种情感类别之一，分别为中性、惊讶、恐惧、悲伤、喜悦、厌恶和愤怒。

评价指标：采用加权平均 F1 分数(weighted average F1-Score)作为模型整体性能的评估指标。同时，为深入研

究模型在不同情感类别上的表现,还分别给出了每个情感类别对应的F1分数。

2.2 基线方法

BC-LSTM:通过双向LSTM对会话上下文进行建模,而不区分不同的说话者。

DialgueRNN[23]:通过3个不同的GRU对上下文信息和说话人状态进行建模。它通过3个门控循环单元对说话者、先前上下文和先前情绪进行建模,包括全局GRU、参与者GRU和情绪GRU。

DialogueGCN[24]:通过使用有向图对对话进行建模来捕获上下文。

SCMM[25]:旨在对每个话语的不同难度和每个情态的具体贡献进行建模。采用自适应路径选择策略,选择合适的路径来获得话语表征。

CORECT[26]:将图卷积网络(GCNs)与动态融合结合用于多模态上下文建模。

TS-GCL[27]:采用基于图对比学习的两阶段的分类方法,了解模态内部和模态之间的相似性和差异性,使得更好地关注不同层次的情感信息。

MGLRA[28]:采用具有记忆功能的循环迭代模块对齐多模态特征,然后使用掩码GCN进行多模态特征融合。

2.3 实验配置

在Pytorch3实现了所提出的模型,并使用Adam[29]作为优化器,其中,β_1为0.9,β_2为0.99,学习率初始化为0.000 1,每10 epoch以0.95的比例衰减,L2正则化系数λ为0.000 01。为了避免过拟合,应用Dropout[30]层以及丢弃率(dropout_rate)为0.1。由于MELD明显比IEMOCAP更不平衡,IEMOCAP的批大小为64,训练轮数为100轮,MELD的批大小为100,训练轮数为15轮。CAMC损失中的温度参数τ、困难样本挖掘比例ρ分别设计为0.8和0.3,训练目标的总损失函数L中的组合系数μ_1为0.4,μ_2和μ_3均为0.3。所有结果都是10次运行的平均值。

3 结果与分析

3.1 对比实验

FuseNet模型与基线方法和最先进的方法进行了比较。表2和3分别给出了所有模型在IEMOCAP和MELD数据集上的表现。

表2 在IEMOCAP测试集上的实验结果

Table 2 The experimental results on the IEMOCAP test set

模型	IEMOCAP						
	Happiness	Sadness	Neutral	Anger	Excitement	Frustration	Weighted-F1
BC-LSTM	34.43	60.87	51.81	56.73	57.95	58.92	54.95
DialgueRNN	33.18	78.80	59.21	65.28	71.86	58.91	62.75
DialogueGCN	51.87	76.76	76.76	62.26	72.71	58.04	63.16
SCMM	45.37	78.76	63.54	66.05	**76.70**	66.18	67.53
MGLRA	63.50	81.50	71.50	61.10	76.30	67.80	70.10
TS-GCL	70.00	81.70	64.20	61.40	76.50	64.60	70.20
CORECT	**58.74**	80.95	69.52	65.91	76.19	68.11	70.81
FuseNet(本文)	58.17	**84.02**	**72.63**	**67.86**	76.41	**70.42**	**72.87**

表3 在MELD测试集上的实验结果

Table 3 The experimental results on the MELD test set

模型	MELD							
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Weighted-F1
BC-LSTM	73.80	47.70	5.40	25.10	51.30	5.20	38.40	55.90
DialgueRNN	76.23	49.59	0.00	26.33	54.55	0.81	46.76	58.73
DialogueGCN	76.02	46.37	0.98	24.32	53.62	1.22	43.03	57.52
SCMM	—	—	—	—	—	—	—	59.44
MGLRA	80.60	56.40	5.20	43.70	66.30	2.60	48.50	64.10
TS-GCL	80.80	**59.50**	0.00	27.80	**66.50**	0.00	48.40	64.90
CORECT	**81.60**	49.60	26.47	**43.78**	63.32	31.58	51.64	65.92
FuseNet(本文)	79.98	59.35	**30.59**	43.45	65.48	**33.90**	**55.59**	**67.24**

实验结果表明,在IEMOCAP数据集上,FuseNet模型在所有基线模型中性能表现最好,在加权平均F1分数

上比最先进的基线模型 CORECT 提高了 2.91%，尤其是在两个语义易混淆的情感类别“悲伤”和“沮丧”上分别达到了 2.84%和 3.39%的提升，并且在“中性”和“愤怒”情感上分别提升了 1.58%和 2.74%。在 MELD 数据集上，与所有的基线模型相比，在加权 F1 分数方面比最先进的基线模型 CORECT 提高了 2.00%，并且在每个情感标签上的性能比其他模型更稳定。特别在两个少数情感类别“厌恶”和“愤怒”分别达到了 7.37%和 7.65%的提高，并且在“恐惧”情感上提升了 15.56%。

3.2　不同模态设置的效果

为了显示不同模式的效果，一次移除一个或两个模态。表 4 显示了 IEMOCAP 和 MELD 上不同模态设置下 F1 分数的比较。从表中可以看出：1）对于 3 种单模态结果，文本模态的表现显著好于其他两种模态；2）任何两种模态融合结果都比本身单模态结果好，并且融合文本模态和声学或视觉模态的表现优于声学和视觉模态的融合；3）同时使用 3 种模态表现最佳。结果说明了话语的文本情态在 ERC 中起着重要作用，而来自音频和视觉情态的互补线索可以带来比基于文本的 FuseNet 显著的改进，整合多模态信息对 ERC 至关重要。

表 4　不同模态设置下的实验结果

Table 4　The experiment results under different modal settings

模态	IEMOCAP	MELD
Text	65.48	62.23
Audio	57.42	49.83
Visual	48.34	43.16
Text＋Audio	69.20	64.34
Text＋Visual	67.73	63.36
Audio＋Visual	67.49	58.33
Text＋Audio＋Visual	72.87	67.24

3.3　消融实验

为了研究 FuseNet 中不同组分对模型性能的贡献，对 IEMOCAP 和 MELD 进行了消融研究，结果如表 5 和 6 所示。

表 5　在 IEMOCAP 测试集上的消融实验结果

Table 5　The ablation experiment results on the IEMOCAP test set

模型	IEMOCAP						
	Happiness	Sadness	Neutral	Anger	Excitement	Frustration	Weighted-F1
-BiDRN	51.83	83.00	71.32	67.66	75.04	68.62	71.02
-Hi-Gated Fusion	48.53	81.45	69.00	66.30	74.74	67.00	69.37
-CAMC	52.14	82.43	71.78	65.19	75.08	66.75	70.38

表 6　在 MELD 测试集上的消融实验结果

Table 6　The ablation experiment results on the MELD test set

模型	MELD							
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Weighted-F1
-BiDRN	79.24	58.24	21.69	41.93	63.85	22.68	54.62	65.81
-Hi-Gated Fusion	79.49	57.44	24.72	41.18	64.18	23.66	54.34	65.88
-CAMC	79.51	58.05	14.93	41.46	63.26	27.37	54.41	65.75

BiDRN 的影响：为了研究 BiDRN 的效果，提出的 BiDRN 用 DialogueRNN 取代。-BiDRN 在两个数据集上的性能都有所下降，其中 MELD 上的下降更为显著，因为 MELD 中复杂的多人对话使得 DialogueRNN 难以准确捕获对话轮次间的模态信息。-BiDRN 在两个数据集上的较差性能证明了 BiDRN 的有效性。

Hi-Gated Fusion 的影响：-Hi-Gated Fusion 在两个数据集上所有情感类别中的表现都急剧下降，这证明了使用 Hi-Gated Fusion 有效捕捉文本、音频以及视觉跨模态的深层语义关联重要性和优越性。

CAMC 的影响：-CAMC 在两个数据集上的性能都有明显下降，尤其在少数和语义易混淆的情感类别上的下降更为显著。此外，MELD 的下降程度更为明显，因为 MELD 比 IEMOCAP 的情感类别明显更不平衡。这证明了 CAMC 在减轻少数和语义易混淆情绪类别分类困难方面的有效性。

4　结　　论

本文提出的多模态情感识别模型 FuseNet 用于 MERC 任务，解决了对话上下文表征单一以及长尾情感类别识别不足的问题。其中，提出了对话上下文编码器 BiDRN 可以同时关注历史与未来的对话信息，整合来自不同说话人的音频以及视觉线索，更准确识别对话中的情绪转变，并设计基于分层门控的多模态动态融合模型 Hi-

Gated Fusion有效捕捉跨模态间的语义关联,此外,提出的CAMC损失有效减轻了识别少数和语义相近的情感类别的困难。实验结果表明,在IEMOCAP数据集上,FuseNet的F1分数比最先进的基线模型CORECT提高了2.91%,尤其是在语义相近的情感类别"悲伤"和"沮丧"上分别达到了2.84%和3.39%的提升,并且在"中性"和"愤怒"情感上分别提升了1.58%和2.74%。在MELD数据集上,在F1分数上比最先进的基线模型CORECT提高了2.00%,并且在每个情感标签上的性能比其他模型更稳定。特别在少数情感类别"厌恶"和"愤怒"分别达到了7.37%和7.65%的提高,并且在"恐惧"情感上提升了15.56%。

然而,该模型还存在一些局限性,例如在实际场景中还会有其他模态信息对情绪转变造成影响,因此模态覆盖不足。未来研究将重点关注模型优化、多模态扩展等方面,进一步提升模型在复杂真实场景下的性能和实用性。

参考文献

[1] 范方朝,杜欣,谢城壁,等. 基于CNN-LSTM的脑电P300信号检测[J]. 电子测量技术,2022,45(23):159-165.
FAN F ZH,DU X,XIE CH B,et al. A P300 signal detection algorithm based on CNN and LSTM[J]. Electronic Measurement Technology,2022,45(23):159-165.

[2] 殷纪民. 对话场景下的端到端多模态情感识别研究[D]. 武汉:华中师范大学,2024.
YIN J M. End-to-end multimodal emotion recongnition in conversation[D]. Wuhan:Central China Normal University,2024.

[3] UJAH OGBUAGU B C, AMEEN A O, OKWUELEKA F N, et al. Taxonomy of opinion mining, approaches and domain applications: Future research direction[J]. SN Computer Science, 2025, 6(6):653.

[4] ZHAO L, GAO J, LI D, et al. The design and implementation of xiaoice,an empathetic social chatbot[J]. Computational Linguistics,2020,46(1):53-93.

[5] PORIA S, CAMBRIA E, HAZARIKA D, et al. Context-dependent sentiment analysis in user-generated videos[C]. 55th Annual Meeting of the Association for Computational Linguistics, 2017(1):873-883.

[6] JIAO W X, YANG H Q, LYU M R, et al. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition[J]. ArXiv preprint arXiv:1904.04446,2019.

[7] LI J, JI D, LI F, et al. HiTrans: A transformer based context- and speaker-sensitive model for emotion detection in conversations[C]. 28th International Conference on Computational Linguistics, 2020:4190-4200.

[8] MAO Y Z, SUN Q, LIU G, et al. Dialoguetrm: Exploring the intra-and inter-modal emotional behaviors in the conversation[J]. ArXiv preprint arXiv:2010.07637,2010.

[9] LIU Y H, ZHOU L J, XU R F, et al. Empathetic response generation with state management[J]. ArXiv preprint arXiv:2205.03676,2022.

[10] 童小钟,魏俊宇,苏绍璟,等. 融合注意力和多尺度特征的典型水面小目标检测[J]. 仪器仪表学报,2023,44(1):212-222.
TONG X ZH,WEI J Y,SU SH J,et al. Typical small surface target detection integrating attention and multi-scale features[J]. Chinese Journal of Scientific Instrument,2023,44(1):212-222.

[11] SHI T, SHAO L H. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations[C]. 61st Annual Meeting of the Association for Computational Linguistics,2023,1:14752-14766.

[12] MA H, WANG J, LIN H, et al. A transformer-based model with self-distillation for multimodal emotion recognition in conversations[J]. IEEE Transactions on Multimedia,2023,26:776-788.

[13] 考文君. 基于注意力机制的多模态的融合情感分析的研究[D]. 济南:山东交通学院,2024.
KAO W J. Research on multimodal fusion sentiment analysis based on attention mechanism[D]. Jinan: Shandong Jiaotong University,2024.

[14] YU X, WANG F, QIAO Z. SpikEmo: Enhancing emotion recognition with spiking temporal dynamics in conversations[J]. ArXiv preprint arXiv: 2411.13917,2024.

[15] PORIA S, HAZARIKA D, MAJUMDER N, et al. MELD: A multimodal multi-party dataset for emotion recognition in conversations[J]. ArXiv preprint arXiv: 1810.02508,2018.

[16] 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.

[17] 彭军强,张立坤,杨亚楠. 基于多模态轻量化混合模型的情绪识别[J]. 电子测量技术,2024,47(3):9-18.
PENG J Q, ZHANG L K, YANG Y N. Emotion recognition based on multimodal lightweight hybrid model[J]. Electronic Measurement Technology,2024,

47(3):9-18.

[18] LIU Y H, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach [J]. ArXiv preprint arXiv:1907.11692,2019.

[19] EYBEN F, WOLLMER M, SCHULLER B. Opensmile:The munich versatile and fast open-source audio feature extractor[C]. 18th ACM International Conference on Multimedia,2010:1459-1462.

[20] HUANG G, LIU Z, VAN D M L, et al. Densely connected convolutional networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[21] 李鹏. 基于改进 PSO-BP 算法的机器人目标位姿识别方法[J]. 国外电子测量技术,2023,42(1):7-12.
LI P. Robot target pose recognition method based on improved PSO-BP algorithm[J]. Foreign Electronic Measurement Technology,2023,42(1):7-12.

[22] WANG L, WU J, HUANG S L, et al. An efficient approach to informative feature extraction from multimodal data[C]. AAAI Conference on Artificial Intelligence,2019,33(1):5281-5288.

[23] MAJUMDER N, PORIA S, HAZARIKA D, et al. Dialoguernn:An attentive rnn for emotion detection in conversations [C]. AAAI Conference on Artificial Intelligence,2019,33(1):6818-6825.

[24] GHOSAL D, MAJUMDER N, PORIA S, et al. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation[J]. ArXiv preprint arXiv:1908.11540,2019.

[25] YANG H,GAO X,WU J,et al. Self-adaptive context and modal-interaction modeling for multimodal emotion recognition [C]. Association for Computational Linguistics:ACL 2023,2023:6267-6281.

[26] NGUYEN C V T, MAI A T, LE T S, et al. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction [J]. ArXiv preprint arXiv: 2311.04507,2023.

[27] AI W, ZHANG F C, MENG T, et al. A two-stage multimodal emotion recognition model based on graph contrastive learning[C]. 2023 IEEE 29th International Conference on Parallel and Distributed Systems(ICPADS), 2023:397-404.

[28] MENG T,ZHANG F CH,SHOU Y T,et al. Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing,2024,32:4298-4312.

[29] KINGMA P D,BA J. Adam: A method for stochastic optimization [J]. ArXiv preprint arXiv: 1412.6980,2014.

[30] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research,2014,15(1):1929-1958.

作者简介

赵亚芳,硕士研究生,主要研究方向为人工智能、自然语言处理。

E-mail:1508753785@qq.com

梁志剑(通信作者),博士,教授,主要研究方向为人工智能、自然语言处理。

E-mail:116585916@qq.com