·人工智能与模式识别·　　文章编号：1000-3428(2023)07-0094-08　　文献标志码：A　　中图分类号：TP391

基于动态卷积与残差门控的多模态情感识别

郭艳霞[1,2]，金　勇[1]，唐　宏[1,2]，彭金枝[1,2]

（1.重庆邮电大学 通信与信息工程学院，重庆 400065；2.重庆邮电大学 移动通信技术重庆市重点实验室，重庆 400065）

摘　要：为了防止一段话语中含有情感色彩的重要信息被无关信息淹没并实现多模态信息交互，通过挖掘高级局部特征以及设计有效的交互融合策略，提出一种基于动态卷积与残差门控的多模态情感识别模型。提取文本、音频和图像中的低级特征、高级局部特征以及上下文依赖关系，同时使用跨模态动态卷积对模态间和模态内交互信息进行建模，模拟长序列时域间的相互作用，捕捉不同模态的交互特征。设计一种残差门控融合方法来融合不同模态交互表征，自动学习每组交互表征对最终情感识别的影响权重，并将多模态融合特征输入分类器进行情感预测。在CMU-MOSEI和IEMOCAP数据集上的实验结果表明，该模型能够避免多模态中含有情感色彩的重要信息被无关信息淹没，情感分类准确率分别达到83.5%和83.9%，性能优于MulT、MFRM等基准模型。

关键词：自然语言处理；信息交互；多模态情感识别；动态卷积；门控机制

开放科学（资源服务）标志码（OSID）：

中文引用格式：郭艳霞，金勇，唐宏，等.基于动态卷积与残差门控的多模态情感识别[J].计算机工程，2023，49(7)：94-101.

英文引用格式：GUO Y X，JIN Y，TANG H，et al.Multi-modal emotion recognition based on dynamic convolution and residual gating[J].Computer Engineering，2023，49(7)：94-101.

Multi-modal Emotion Recognition Based on Dynamic Convolution and Residual Gating

GUO Yanxia[1,2]，JIN Yong[1]，TANG Hong[1,2]，PENG Jinzhi[1,2]

（1.School of Communication and Information Engineering，Chongqing University of Posts and Telecommunications，Chongqing 400065，China；2.Chongqing Key Laboratory of Mobile Communications Technology，Chongqing University of Posts and Telecommunications，Chongqing 400065，China）

【Abstract】To prevent important information containing emotional cues from being obscured by irrelevant information in discourse and to achieve multi-modal information interaction，a multi-modal emotion recognition model based on dynamic convolution and residual gating is proposed by mining advanced local features and designing effective interaction fusion strategies. Low-level features，high-level local features，and contextual dependencies from text，audio，and images are extracted. While using cross modal dynamic convolution to model inter-modal and intra-modal interactions，interactions are simulated between long sequences in time domain，and interaction features of different modalities are captured. A residual gated fusion method that fuses different modal interaction representations automatically learns the impact weight of each interaction feature on the final output，and inputs the multi-modal fusion feature into the classifier for emotion prediction. The experimental results show that this model prevents important information regarding emotional cues from being obscured by irrelevant information in multi-modal data. The accuracy of sentiment classification is 83.5% and 83.9% on the CMU-MOSEI and IEMOCAP datasets，respectively. The model outperforms benchmark models such as Multi-modal Transformer（MulT）and Multi-Fusion Residual Memory（MFRM）.

【Key words】natural language processing；information interaction；multi-modal emotion recognition；dynamic convolution；gating mechanism

DOI：10. 19678/j. issn. 1000-3428. 0064965

基金项目：长江学者和创新团队发展计划（IRT_16R72）。

作者简介：郭艳霞（1995—），女，硕士研究生，主研方向为情感识别；金　勇，高级工程师、硕士；唐　宏，教授、博士；彭金枝，硕士研究生。

收稿日期：2022-06-10　　**修回日期：**2022-08-22　**E-mail：**guoyx1228@163.com

0　概述

随着互联网的发展和智能手机的普及，社交媒体已成为人们日常生活中不可或缺的一部分，人们通过各种社交媒体来表达自己的观点和看法。情感识别技术通过分析和处理这些观点和看法获取人类所处情感状态，已成为人工智能和自然语言处理研究领域一项重要且备受关注的任务，广泛应用于服务型机器人、教育质量评估、人机交互[1]等任务。常见的情感识别方法按模态类型可分为基于非生理信号[2]、基于生理信号[3]和融合非生理信号和生理信号[4]的情感识别，按模态种类可分为基于单模态、双模态以及多模态情感识别。单模态情感识别一般不如双模态和多模态情感识别性能好。此外，由于有关生理信号的数据集较少，因此更多研究者倾向于融合文本、图像和音频进行多模态情感识别。

随着深度学习技术的快速发展，多模态情感识别取得了不错的应用效果。影响多模态情感分类最重要的两个因素是提取各个模态特征和融合多模态数据。对于文本、音频、图像的特征提取方法包括BERT[5]、HuBERT[6]、OpenFace[7]等。对于文本、音频、图像多模态数据的融合，传统多模态学习方法主要使用特征级融合和决策级融合，现有多数融合方法使用不同的数学公式对各个模态进行表征，如多模态层次融合、会话神经网络、不同特征的词级拼接等。

为了捕捉多模态情感识别中的不同模态的动态交互关系，ZADEH等[8]引入张量融合网络（Tensor Fusion Network，TFN）模型，3次使用笛卡尔积分别对文本、图像、语音中的单峰、双峰、三峰特征进行建模，但复杂度较高。AREVALO等[9]提出一种门控多模态单元（Gated Multi-modal Unit，GMU）模型，该模型使用乘法门决定模态如何影响单元的激活。PAN等[10]提出一种多模态注意力网络（Multi-Modal Attention Network，MMAN），使用多模态聚焦机制，选择性融合3个模态信息。上述方法都取得了较好的情感分类性能。

为了研究一个模态信息对另一个模态信息的影响，YANG等[11]提出一种跨模态BERT模型，通过使用语音信息动态调整文本相关词的权重，降低不相关词的权重，从而捕捉更丰富的情感信息，并降低噪声的影响。TSAI等[12]提出一种多模态变压器（Multi-modal Transformer，MulT）模型来融合不同模态特征，其在变压器的基础上引入模态强化单元，利用来自源模态的信息强化目标模态，实现异步序列的多模态融合。上述方法一般适用于不同模态信息的融合。

多模态数据包含丰富的情感信息，当人类通过多模态信息进行交流时，由于语言和面部表情发生时间不同步，因此情感信息在时间上存在不一致性。例如，当某人说："Today，I went to the talk show，I was so happy"，其面部表情先是微笑，再咧嘴大笑，最后保持平静。在整个过程中，与开心相关的情感词"happy"和视觉情感信息占整个文本序列长度的1/11和接近整个视觉序列长度的1/2，这会导致后续融合过程中含有情感色彩的重要信息容易被无关信息淹没，从而影响最终情感分类性能。虽然上述方法在情感分类上取得了较好的性能，但是无法解决含有情感色彩的重要信息易被无关信息淹没这一问题。

本文提出一种基于动态卷积与残差门控的多模态情感识别模型。首先，提取各个模态的低级特征、高级局部特征以及上下文依赖关系。然后，对文本、图像、音频模态内和模态间的交互信息进行建模。最后，通过残差门控自动学习每个交互信息在最终情感分类中的权重，并将多模态融合特征输入分类器进行情感预测。

1　模型描述

本文构建基于动态卷积和残差门控的多模态情感识别模型，如图1所示。

该模型主要包括：

1）特征提取模块，分别提取低级特征、高级局部特征以及各模态上下文依赖关系。

2）跨模态动态卷积（Cross-modal Dynamic Convolution，CDC）交互模块，对不同模态间的特征进行交互建模，得到交互表征。

3）残差门控（Residual Gating，RG）融合模块，设计一种残差门控融合方法，动态学习每组交互表征在最终情感分类中的权重。

基于此，将最终特征输入Softmax分类器得到情感分类结果。

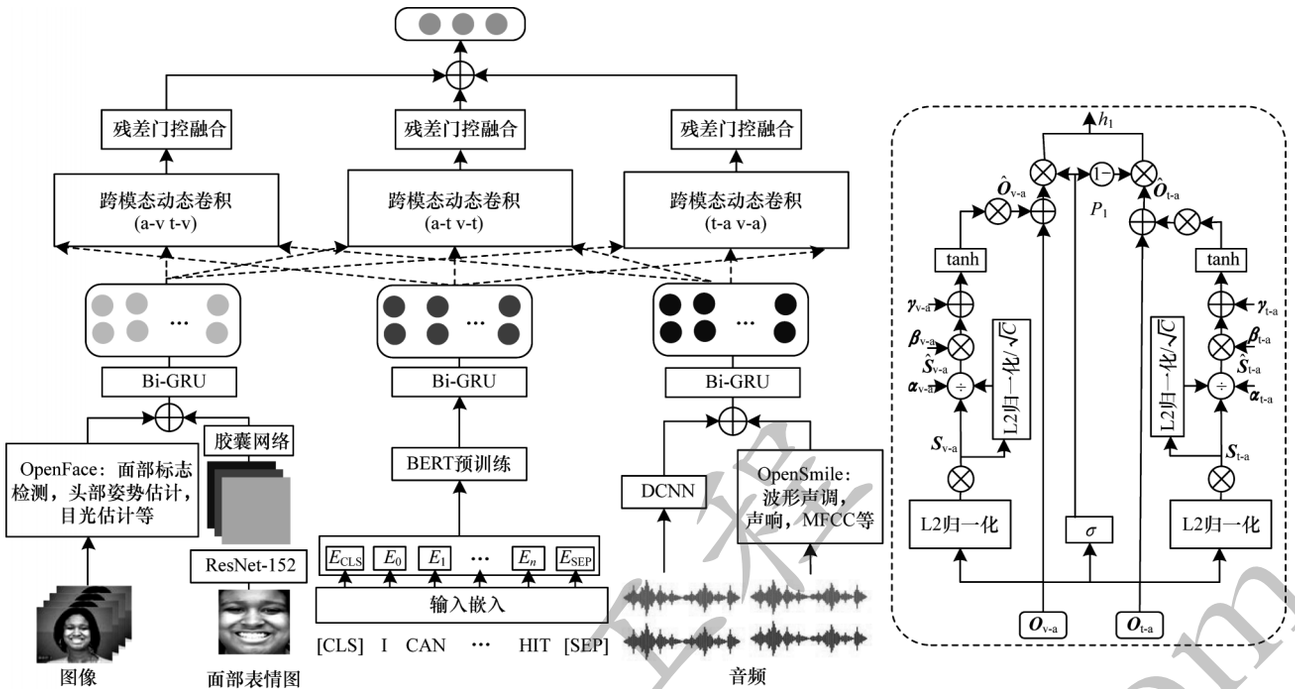

图1 基于动态卷积与残差门控的多模态情感识别模型结构

Fig.1 Structure of multi-modal emotion recognition model based on dynamic convolution and residual gating

1.1 特征提取模块

1.1.1 文本预处理

先将文本转录，再使用与BERT相同的标记方式对文本序列进行标记。例如，句子“Anyhow，It was a great day.”将被分为[‘Anyhow’，‘It’，‘was’，‘a’，‘great’，‘day’]，其中，‘Anyhow’又可分为‘Any’和‘##how’，‘##’代表了这两部分属于同一个单词。若给定一段经过标记的文本序列，输入BERT嵌入层后，将在文本序列开始位置和结束位置分别添加特殊标记[CLS]和[SEP]，表示为 $\boldsymbol{T}=[\mathrm{CLS},t_1,t_2,\cdots,t_n,\mathrm{SEP}]$，其中，[CLS]是用于分类的标识符，[SEP]是用于分割句子的标识符，$t_1,t_2,\cdots,t_n$ 表示文本单词标记序列。最终文本序列的嵌入层输出是相对应的令牌嵌入、分割嵌入和位置嵌入之和，即为 $E_{\mathrm{CLS}},E_1,E_2,\cdots,E_n,E_{\mathrm{SEP}}$。为了充分利用文本数据中的信息，本文对BERT-base模型进行微调。

1.1.2 音频特征提取

音频特征 $\boldsymbol{X}_{\mathrm{a}}$ 包含了低级特征 $\boldsymbol{X}_{\mathrm{OpenSmile}}$ 和高级局部特征 $\boldsymbol{X}_{\mathrm{RDCNN}}$。低级特征利用OpenSmile工具箱对不同说话者的韵律信息进行特征提取，主要包括声调、声响、梅尔频率倒谱系数(MFCC)等声学低级描述符。

相对于传统卷积神经网络(Convolutional Neural Network，CNN)，扩张卷积神经网络在保持参数不变的情况下增大了卷积核的感受野，可以捕捉更多的音频空间的显著局部情感特征[13]。本文采用一维残差扩张卷积神经网络(1D Residual Dilated Convolutional Neural Network，RDCNN)对音频中每个片段的高级局部特征进行提取。先使用一个DCNN从原始音频信号中提取局部特征，再输入局部特征增强块来增强局部情感特征，该块是由扩张率为3的扩张卷积神经网络、BN层和leaky_ReLU层组成，其中：BN层主要为了提高训练的性能和速度，从而避免梯度爆炸；leaky_ReLU层的作用是使模块中不存在线性关系。将该网络的扩张率设置为2，步幅设置为1。

1.1.3 图像特征提取

图像特征 $\boldsymbol{X}_{\mathrm{v}}$ 包含低级特征 $\boldsymbol{X}_{\mathrm{OpenFace}}$ 和高级局部特征 $\boldsymbol{X}_{\mathrm{Capsule}}$。低级特征是利用OpenFace工具箱从眼部区域、头部姿态等动作单元中提取面部标志特征。

为了捕捉更多面部情感特征，使用图像识别网络ResNet-152进行图像高级局部特征提取。首先将面部表情图进行预处理，调整为224×224像素的面部表情图 $\boldsymbol{V}'$；然后输入ResNet-152进一步特征提取，可表示为 $\boldsymbol{X}_{\mathrm{R}}=\mathrm{ResNet}(\boldsymbol{V}')$。由于ResNet-152无法处理图像中的位置信息，因此再将 $\boldsymbol{X}_{\mathrm{R}}$ 输入单层胶囊网络，可表示为 $\boldsymbol{X}_{\mathrm{C}}=\mathrm{Capsule}(\boldsymbol{X}_{\mathrm{R}})$。

1.1.4 单模态上下文依赖关系

双向门控循环单元(Bi-Gated Recurrent Unit，Bi-GRU)[14]是一种用于识别长序列上下文依赖关系的简化网络，广泛应用在时间序列数据中。使用Bi-GRU获取各个模态的上下文依赖关系，详细公式可参考文献[14]。文本、图像和音频特征经过Bi-GRU后可得到具有上下文依赖关系的文本特征 $\boldsymbol{z}_{\mathrm{t}}$、音频特征 $\boldsymbol{z}_{\mathrm{a}}$ 和图像特征 $\boldsymbol{z}_{\mathrm{v}}$。

1.2　跨模态动态卷积模块

注意力机制是将每个元素与其他元素进行比较来确定上下文元素的重要性，然而注意力对于远距离序列的关注较少[15]。一些研究表明[16]：动态卷积在每个时间步长时预测不同的卷积核，计算代价随输入目标序列长度的增加而线性增加，相比于注意力机制更加简单高效且易于叠加，可模拟长序列时域间的交互作用，交互也更加稳定。为了防止话语中含有情感色彩的重要信息被无关信息淹没以及实现多模态信息交互，使用跨模态动态卷积对音频-图像（a-v）、图像-文本（t-v）、音频-文本（a-t）、图像-文本（v-t）、文本-音频（t-a）、图像-音频（v-a）等6组模态间的局部信息进行建模，并利用多次多模态数据融合操作进行更好的表征。

首先，使用一维时间卷积将各个模态对应的Bi-GRU最后一层的输出转换到相同的维度，目的是使跨模态交互更加容易，计算公式可表示如下：

$$\boldsymbol{Z}_{\mathrm{m}}=1\mathrm{DConv}(z_{\mathrm{m}},\boldsymbol{W}_{\mathrm{m}}) \tag{1}$$

然后，跨模态动态卷积是利用不同模态间的互补性，从一个模态中提取信息，指导另一个模态的时间滤波，同时模拟跨模态时间交互。以图像v和音频a交互为例，可表示如下：

$$\begin{aligned}&\mathrm{CDC}_{\mathrm{v\text{-}a}}(\boldsymbol{Z}_{\mathrm{v}},\boldsymbol{Z}_{\mathrm{a}})=\mathrm{Softmax}\left(\boldsymbol{M}_{\mathrm{a,v}}\right)\boldsymbol{Z}_{\mathrm{v}}\boldsymbol{W}_{K_{\mathrm{v}}}=\\&\mathrm{Softmax}\left(\mathrm{TM}\left(\boldsymbol{K}_{\mathrm{a}}\right)\right)\boldsymbol{Z}_{\mathrm{v}}\boldsymbol{W}_{K_{\mathrm{v}}}=\\&\mathrm{Softmax}\left(\mathrm{TM}\left(\boldsymbol{Z}_{\mathrm{a}}\boldsymbol{W}_{K_{\mathrm{a}}}\right)\right)\boldsymbol{Z}_{\mathrm{v}}\boldsymbol{W}_{K_{\mathrm{v}}}\end{aligned} \tag{2}$$

其中：$\mathrm{CDC}_{\mathrm{v\text{-}a}}$是对应的跨模态动态卷积；TM是一个重新排列的函数，以逐行的方式将$\boldsymbol{K}_{\mathrm{a}}$重新排列为一个新的矩阵$\boldsymbol{M}_{\mathrm{a,v}}$；Softmax将输入在时间维度的权值进行归一化来控制权重的尺度；$\boldsymbol{K}_{\mathrm{a}}$是动态卷积核矩阵；$\boldsymbol{W}$是线性变换的权重矩阵。

最后，由于该模块具有方向性，因此利用跨模态动态卷积处理每一组模态对，共有6组模态对，计算过程可表示如下：

$$\boldsymbol{K}_{\mathrm{v}}^{(0)}=\hat{\boldsymbol{Z}}_{\mathrm{v}} \tag{3}$$

$$\boldsymbol{K}_{\mathrm{v\text{-}a}}^{(0)}=\boldsymbol{K}_{\mathrm{v}}^{(0)}=\hat{\boldsymbol{Z}}_{\mathrm{a}} \tag{4}$$

$$\hat{\boldsymbol{K}}_{\mathrm{v\text{-}a}}^{(i)}=\mathrm{MHCDC}_{\mathrm{v\text{-}a}}^{(i)}(\mathrm{LN}(\boldsymbol{K}_{\mathrm{v\text{-}a}}^{(i-1)}),\mathrm{LN}(\boldsymbol{K}_{\mathrm{v}}^{(0)})) \tag{5}$$

$$\boldsymbol{K}_{\mathrm{v\text{-}a}}^{(i)}=\boldsymbol{K}_{\mathrm{v\text{-}a}}^{(i-1)}+\mathrm{PWConv}^{(i)}(\mathrm{LN}(\hat{\boldsymbol{K}}_{\mathrm{v\text{-}a}}^{(i)})) \tag{6}$$

$$\boldsymbol{O}_{\mathrm{v\text{-}a}}=\boldsymbol{K}_{\mathrm{v\text{-}a}}^{(N)} \tag{7}$$

其中：$\mathrm{MHCDC}_{\mathrm{v\text{-}a}}^{(i)}$是第$i$层的多头跨模态动态卷积；LN是层标准化；$\mathrm{PWConv}^{(i)}$是第$i$层上的逐点卷积；$\boldsymbol{K}_{\mathrm{v\text{-}a}}^{(i)}$是第$i$层模态间交互层的输出；$\boldsymbol{O}_{\mathrm{v\text{-}a}}$是最终图像和音频的交互特征。

将对齐后的文本特征$\boldsymbol{Z}_{\mathrm{t}}$、音频特征$\boldsymbol{Z}_{\mathrm{a}}$和图像特征$\boldsymbol{Z}_{\mathrm{v}}$两两交互后得到6种交互特征，分别为$\boldsymbol{O}_{\mathrm{v\text{-}a}}$、$\boldsymbol{O}_{\mathrm{t\text{-}a}}$、$\boldsymbol{O}_{\mathrm{a\text{-}v}}$、$\boldsymbol{O}_{\mathrm{t\text{-}v}}$、$\boldsymbol{O}_{\mathrm{a\text{-}t}}$和$\boldsymbol{O}_{\mathrm{v\text{-}t}}$。

1.3　残差门控融合模块

为了充分利用6种交互特征，设计一个残差门控融合模块自动学习每组表征的权重。由于规范化操作时是无参数的，因此引入$\boldsymbol{\alpha}$、$\boldsymbol{\beta}$、$\boldsymbol{\gamma}$这3个可训练的参数使其具有自动学习能力，$\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_C]$主要用来控制每个通道的嵌入权值，$\boldsymbol{\beta}=[\beta_1,\beta_2,\cdots,\beta_C]$和$\boldsymbol{\gamma}=[\gamma_1,\gamma_2,\cdots,\gamma_C]$用来控制门通道的激活。使用L2归一化对通道关系进行建模，从而避免在特征输入时处于恒定的情况。将$\boldsymbol{O}_{\mathrm{v\text{-}a}}$和$\boldsymbol{O}_{\mathrm{t\text{-}a}}$作为第1组，$\boldsymbol{O}_{\mathrm{a\text{-}v}}$和$\boldsymbol{O}_{\mathrm{t\text{-}v}}$作为第2组，$\boldsymbol{O}_{\mathrm{a\text{-}t}}$和$\boldsymbol{O}_{\mathrm{v\text{-}t}}$作为第3组，对应每一组分别设计一个单独的门。首先，为了避免具有小感受野的信息引起的局部歧义，考虑使用较大感受野的信息，以$\boldsymbol{O}_{\mathrm{v\text{-}a}}$、$\boldsymbol{O}_{\mathrm{t\text{-}a}}$为例，可表示如下：

$$\boldsymbol{S}_{\mathrm{v\text{-}a}}^{(c)}=\boldsymbol{\alpha}_{\mathrm{v\text{-}a}}^{(c)}\left\|\boldsymbol{O}_{\mathrm{v\text{-}a}}^{(c)}\right\|_2=\boldsymbol{\alpha}_{\mathrm{v\text{-}a}}^{(c)}\left[\sum_{i=1}\left(\boldsymbol{O}_{\mathrm{v\text{-}a}}^{(c_i)}\right)^2+\varepsilon\right]^{\frac{1}{2}} \tag{8}$$

$$\boldsymbol{S}_{\mathrm{t\text{-}a}}^{(c)}=\boldsymbol{\alpha}_{\mathrm{t\text{-}a}}^{(c)}\left\|\boldsymbol{O}_{\mathrm{t\text{-}a}}^{(c)}\right\|_2=\boldsymbol{\alpha}_{\mathrm{t\text{-}a}}^{(c)}\left[\sum_{i=1}\left(\boldsymbol{O}_{\mathrm{t\text{-}a}}^{(c_i)}\right)^2+\varepsilon\right]^{\frac{1}{2}} \tag{9}$$

其中：ε是一个数值较小的常数，避免在取0时无法求导；$\boldsymbol{\alpha}_{\mathrm{v\text{-}a}}^{(c)}$控制每个通道的权重，当$\boldsymbol{\alpha}_{\mathrm{v\text{-}a}}^{(c)}=0$时表示该通道不参与最终情感预测。

其次，为了创建$\boldsymbol{O}_{\mathrm{v\text{-}a}}$和$\boldsymbol{O}_{\mathrm{t\text{-}a}}$通道间的竞争与合作关系，引入L2归一化对通道关系进行操作，可表示如下：

$$\hat{\boldsymbol{S}}_{\mathrm{v\text{-}a}}^{(c)}=\frac{\sqrt{C}\,\boldsymbol{S}_{\mathrm{v\text{-}a}}^{(c)}}{\left\|\boldsymbol{S}_{\mathrm{v\text{-}a}}\right\|}=\frac{\sqrt{C}\,\boldsymbol{S}_{\mathrm{v\text{-}a}}^{(c)}}{\left[\sum_{C=1}^{C}\left(\boldsymbol{S}_{\mathrm{v\text{-}a}}^{(c)}\right)^2+\varepsilon\right]^{\frac{1}{2}}} \tag{10}$$

$$\hat{\boldsymbol{S}}_{\mathrm{t\text{-}a}}^{(c)}=\frac{\sqrt{C}\,\boldsymbol{S}_{\mathrm{t\text{-}a}}^{(c)}}{\left\|\boldsymbol{S}_{\mathrm{t\text{-}a}}\right\|}=\frac{\sqrt{C}\,\boldsymbol{S}_{\mathrm{t\text{-}a}}^{(c)}}{\left[\sum_{C\neq1}^{C}\left(\boldsymbol{S}_{\mathrm{t\text{-}a}}^{(c)}\right)^2+\varepsilon\right]^{\frac{1}{2}}} \tag{11}$$

其中：$\boldsymbol{S}_{\mathrm{v\text{-}a}}=\left[\boldsymbol{S}_{\mathrm{v\text{-}a}}^{(1)},\boldsymbol{S}_{\mathrm{v\text{-}a}}^{(2)},\cdots,\boldsymbol{S}_{\mathrm{v\text{-}a}}^{(c)}\right]$；$C$是通道数，$C\in\{1,2,\cdots,c\}$，$\sqrt{C}$是用来规范化$\boldsymbol{S}_{\mathrm{v\text{-}a}}$和$\boldsymbol{S}_{\mathrm{t\text{-}a}}$的大小，防止当$C$过大时$\hat{\boldsymbol{S}}_{\mathrm{v\text{-}a}}^{(c)}$和$\hat{\boldsymbol{S}}_{\mathrm{t\text{-}a}}^{(c)}$过小。

最后，引入门控机制来自适应输入特征，通过对输入特征的调整来促进通道的竞争和合作关系，可表示如下：

$$\hat{\boldsymbol{O}}_{\mathrm{v\text{-}a}}^{(c)}=\boldsymbol{O}_{\mathrm{v\text{-}a}}^{(c)}\left[1+\tanh\left(\boldsymbol{\gamma}_{\mathrm{v\text{-}a}}^{(c)}\hat{\boldsymbol{S}}_{\mathrm{v\text{-}a}}^{(c)}+\boldsymbol{\beta}_{\mathrm{v\text{-}a}}^{(c)}\right)\right] \tag{12}$$

$$\hat{\boldsymbol{O}}_{\mathrm{t\text{-}a}}^{(c)}=\boldsymbol{O}_{\mathrm{t\text{-}a}}^{(c)}\left[1+\tanh\left(\boldsymbol{\gamma}_{\mathrm{t\text{-}a}}^{(c)}\hat{\boldsymbol{S}}_{\mathrm{t\text{-}a}}^{(c)}+\boldsymbol{\beta}_{\mathrm{t\text{-}a}}^{(c)}\right)\right] \tag{13}$$

其中：$\boldsymbol{\gamma}=\left[\gamma^{(1)},\gamma^{(2)},\cdots,\gamma^{(c)}\right]$、$\boldsymbol{\beta}=\left[\beta^{(1)},\beta^{(2)},\cdots,\beta^{(c)}\right]$是控制门通道激活的权重和偏置。当通道$\boldsymbol{\gamma}_{\mathrm{v\text{-}a}}^{(c)}$正向激活时，该通道与其他通道处于竞争状态，反之为合作状态。

$$\boldsymbol{P}_1=\sigma\left(\boldsymbol{W}_{P_1}\left[\boldsymbol{O}_{\mathrm{v\text{-}a}},\boldsymbol{O}_{\mathrm{t\text{-}a}}\right]\right) \tag{14}$$

$$\boldsymbol{h}_1=\boldsymbol{P}_1\odot\hat{\boldsymbol{O}}_{\mathrm{v\text{-}a}}+\left(1-\boldsymbol{P}_1\right)\odot\hat{\boldsymbol{O}}_{\mathrm{t\text{-}a}} \tag{15}$$

其中：$\boldsymbol{P}_1$是第1组在情感分类中贡献的权重；$\odot$是逐元素相乘；σ是Sigmoid函数；$\boldsymbol{h}_1$是第1个残差门控的最终输出。

同理可得，第2组、第3组的门控输出为$\boldsymbol{h}_2$、$\boldsymbol{h}_3$。将最终门控输出$\boldsymbol{h}=\boldsymbol{h}_1+\boldsymbol{h}_2+\boldsymbol{h}_3$通过全连接层和Softmax层得到情感分类结果。

2　实验与结果分析

2.1　数据集

使用CMU-MOSEI[17]和IEMOCAP[18]数据集进

行实验仿真。

1）CMU-MOSEI数据集包含22 777个语句和3 228个视频，被分为不同的情感强度和6种情绪（快乐、悲伤、愤怒、恐惧、厌恶和惊讶）。使用16 326、1 871、4 659个样本分别作为训练集、验证集和测试集。

2）IEMOCAP数据集包含10名演员在12 h内的302个视频数据，记录愤怒、厌恶、恐惧、悲伤、中性、快乐、兴奋等情绪。选择并评估快乐、悲伤、愤怒、中性等4种常用的情绪类别。使用4 290、2 124、1 208个样本分别作为训练集、验证集和测试集。

2.2 模型参数设置及评价指标

实验基于PyTorch深度学习架构，在Windows 10操作系统上运行，配置为NVIDIA Quadro RTX 6000 GPU。实验参数设置如表1所示。使用分类准确率（ACC-2、ACC-7、ACC）、F1值、平均绝对误差（Mean Absolute Error，MAE）、Corr等评价指标对模型性能进行评估。

表1 实验参数设置

Table 1 Experiment parameter setting

网络	参数	CMU-MOSEI	IEMOCAP
BERT-base	层数	12	12
	隐藏状态大小	768	768
	注意力头数	12	12
GRU	单元状态维度	200	200
	隐藏状态维度	200	200
训练网络	优化器	Adam	Adam
	Dropout	0.5	0.3
	学习率	0.000 1	0.000 5
	Batch Size	16	32
	Epoch	100	50
动态卷积	核大小	19	9
	通道数/个	40	40
时间卷积	核大小(t/v/a)	3/3/3	3/3/5

2.3 结果分析

2.3.1 与基准模型的对比

为了验证所提模型的有效性，将其与LF-LSTM[12]、EF-LSTM[12]、MulT[12]、MCTN[19]、RAVEN[20]、MFRM[21]、PMR[22]、Multilogue-Net[23]、MFN[24]、MAG-BERT[25]等模型在CMU-MOSEI数据集和IEMOCAP数据集上进行对比实验。表2给出了各模型在CMU-MOSEI数据集上的实验结果，该结果是进行多次实验取得的最佳结果，其中：ΔSOTA是所提模型相对于当前最佳模型在各指标上的变化值，↑表示提高，↓表示下降；最优结果用加粗字体表示；MAE值越小，模型性能越好。

表2 CMU-MOSEI数据集上的实验结果

Table 2 Experimental results on the CMU-MOSEI dataset %

模型	ACC-2	ACC-7	F1值	MAE	Corr
LF-LSTM	80.6	48.8	80.6	61.9	65.9
MCTN	79.8	49.6	80.6	60.9	67.0
MulT	82.5	51.8	82.3	58.0	70.3
RAVEN	79.1	51.0	79.5	61.4	66.2
MFRM	82.4	50.9	82.6	59.8	69.0
Multilogue-Net	82.1		80.3	59.0	
MAG-BERT	82.2		82.6	**54.3**	**76.4**
PMR	83.3	52.5	82.6		
所提模型	**83.5**	**52.8**	**83.1**	56.8	71.2
ΔSOTA	0.2↑	0.3↑	0.5↑	2.5↑	5.2↓

由表2可以看出：

1）MCTN、RAVEN和LF-LSTM相对于其他模型表现较差。MCTN将一种模态转换为另一种模态进行情感分类，LF-LSTM使用晚期融合方式，RAVEN通过在文本词语中出现的图像和音频来建立非语言表征，这3个模型性能较差的原因可能为未充分利用重要情感信息或融合过程中存在信息缺失的情况。

2）MulT、MFRM、MAG-BERT和Multilogue-Net相对于前3个模型准确率有所提高。MFRM不断保留之前的信息，MulT捕获不同模态间的相关信息，Multilogue-Net捕捉了对话上下文信息以及倾听者和说话者的情感状态。这些因素都可以提升情感分类性能。

3）所提模型在分类指标上优于当前最优模型PMR，ACC-2提高了0.2个百分点，ACC-7提高了0.3个百分点，F1值提高了0.5个百分点，MAE比MAG-BERT降低了2.5个百分点，Corr提高了5.2个百分点，达到当前已知的最好结果。PMR是使用基于跨模态交叉注意力的渐进模态增强的方式来交换彼此的特征，所提模型是通过跨模态动态卷积对不同模态间的交互进行建模，相对于跨模态交叉注意力更轻量化，着重于避免模态中的重要情感特征被无关特征淹没。上述结果证明了所提模型的有效性。

表3给出了各模型在IEMOCAP数据集上的实验结果，该结果是进行多次实验取得的最佳结果。由表3可以看出：1）所提模型的10个评价指标值中有5个评价指标值优于对比模型，悲伤和中性两种情感上的ACC和F1值优于当前最佳对比模型MulT，ACC分别提高0.1和0.8个百分点，F1值分别提高0.4、1.7个百分点；2）所提模型在快乐和愤怒两种情感上的ACC和F1值略低于最佳对比模型，这说明所提模型可以有效提高悲伤和中性两种情感的分类准确率，但对于快乐和愤怒的情感分类性能没有MulT和MFRM性能好。

表3　IEMOCAP数据集上的实验结果

Table 3　Experimental results on the IEMOCAPI dataset　　%

模型	快乐		悲伤		愤怒		中性		整体	
	ACC	F1值	ACC	F1值	ACC	F1值	ACC	F1值	ACC	F1值
EF-LSTM	86.0	84.2	80.2	80.5	85.2	84.5	67.8	67.1	79.8	79.0
MFN	86.5	84.0	83.5	82.1	85.0	83.7	69.6	69.2	81.1	79.7
MulT	**90.7**	**88.6**	86.7	86.0	87.4	87.0	72.4	70.7	**84.3**	83.0
MFRM	87.6	85.9	86.1	85.4	**89.4**	**89.4**	70.7	69.8	83.4	82.6
RAVEN	87.3	85.8	83.4	83.1	87.3	86.7	69.7	69.3	81.9	81.2
所提模型	87.1	86.9	**86.8**	**86.8**	89.0	88.9	**72.8**	**72.4**	83.9	**83.7**
ΔSOTA	3.6↓	1.7↓	0.1↑	0.8↑	0.4↓	0.5↓	0.4↑	1.7↑	0.4↓	0.7↑

2.3.2　权重分析

$\boldsymbol{P}_1$、$\boldsymbol{P}_2$和$\boldsymbol{P}_3$分别对应第1组、第2组和第3组的权重。通过分析这些权重进一步研究所提模型的情感分类性能。图2和图3给出了所提模型在五折交叉验证(5-fold cross validation)实验中每组交互表征的平均权重,可以看出第1组的$\boldsymbol{O}_{t\text{-}a}$、第2组的$\boldsymbol{O}_{t\text{-}v}$以及第3组$\boldsymbol{O}_{a\text{-}t}$和$\boldsymbol{O}_{v\text{-}t}$在所有交互特征上的所占权重始终较高,并且都含有文本模态,这表明文本模态在情感分类中的贡献最大,可能的原因为文本模态中含有更加直接和丰富的情感特征。

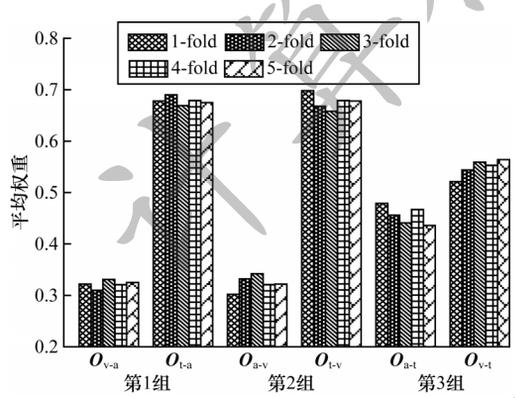

图2　在CMU-MOSEI数据集上每组表征的平均权重

Fig.2　Average weight of each representation on the CMU-MOSEI dataset

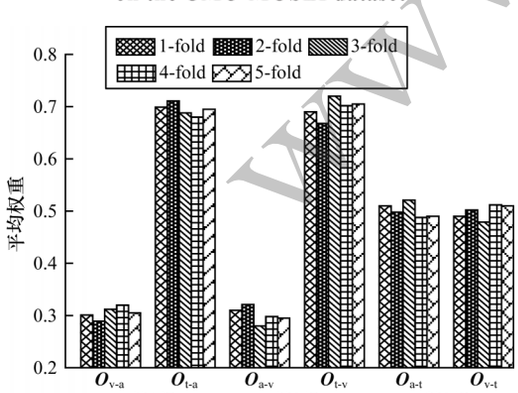

图3　在IEMOCAP数据集上每组表征的平均权重

Fig.3　Average weight of each representation on the IEMOCAP dataset

2.4　消融实验

2.4.1　不同输入模态对模型性能的影响

为了评估不同输入模态对模型性能的影响,在CMU-MOSEI数据集上进行消融实验,如图4所示。由图4可以看出:当输入单模态时,文本比音频和图像的情感分类性能都要好;当输入双模态时,相比于单模态时情感分类性能有所提高,并且t+a和t+v比a+v情感分类性能好;当输入三模态t+a+v时,情感分类性能优于单模态和双模态。由此可见,文本模态在情感分类中具有最为显著的作用,同时证明了融合文本、图像和音频3个模态有助于提升情感识别的性能。

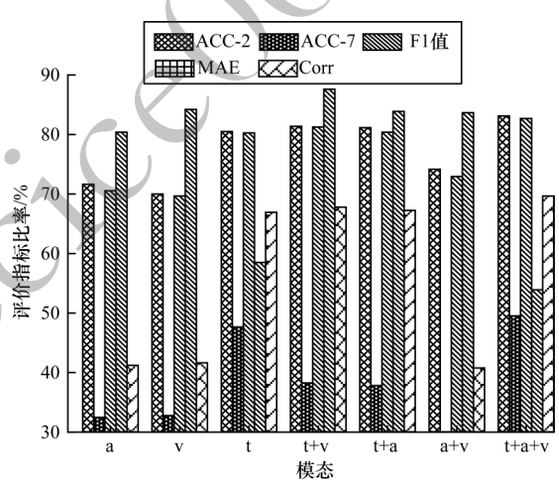

图4　不同输入模态在CMU-MOSEI数据集上的消融实验结果

Fig.4　Ablation experimental results of different input modals on the CMU-MOSEI dataset

2.4.2　跨模态动态卷积交互模块和残差门控融合模块对模型性能的影响

为了研究跨模态动态卷积交互模块和残差门控融合模块对情感分类性能的影响,在CMU-MOSEI数据集上进行消融实验,如图5所示。将Bi-GRU最后一层输出特征分别输入NCR、CDC和所提模型中,其中,NCR删除了跨模态动态卷积交互模块和残差门控融合模块,CDC只删除了残差门控融合模块。

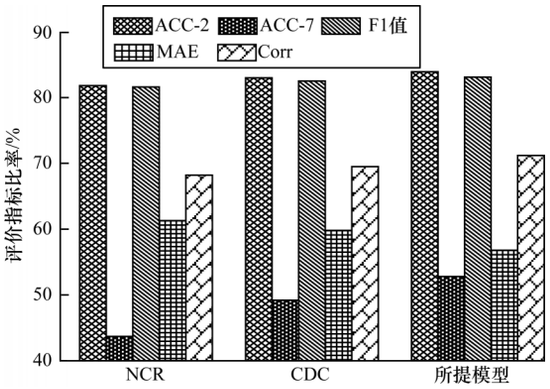

图5 在CMU-MOSEI数据集上跨模态动态卷积交互模块和残差门控融合模块的消融实验结果

Fig.5 Ablation experimental results of cross-modal dynamic convolutional interaction module and residual gating fusion module on the CMU-MOSEI dataset

由图5可以看出，与所提模型相比，不使用跨模态动态卷积交互模块和残差门控融合模块的模型情感分类性能指标大幅度下降，ACC-2、ACC-7、F1值、MAE和Corr分别为81.8%、43.7%、81.6%、61.3%、68.2%，当使用跨模态动态卷积交互模块时，ACC-2、ACC-7、F1值和Corr分别提高了1.2、5.5、0.9、1.3个百分点，MAE降低了1.5个百分点，证明了跨模态动态卷积交互模块的有效性。使用CDC模型进行情感分类时，ACC-2、ACC-7、F1值、MAE和Corr分别为83.0%、49.2%、82.5%、59.8%、69.5%，使用所提模型进行情感分类时，ACC-2、ACC-7、F1值和Corr分别提高了0.9、3.6、0.6、1.7个百分点，MAE降低了3.0个百分点，证明了残差门控融合模块的有效性。

2.4.3 不同跨模态动态卷积核大小和多模态间交互层数对模型性能的影响

为了研究跨模态动态卷积核大小和多模态间交互层数的变化对ACC和F1值的影响，在IEMOCAP数据集上进行消融实验，如表4所示，其中，跨模态动态卷积核大小的变化范围为{3,5,7,9,11,13,15,17,19}，多模间交互层数的变化范围为{2,3,4,5,6,7,8}。由表4可以看出，较深的模型相对于较浅的模型性能略好，对于相同的交互层，跨模态动态卷积核大小为9的模型性能略优于核大小为3的模型性能。但是，跨模态动态卷积核过大并不会有更好的性能，例如跨模态动态卷积核大小为19的模型性能比核大小为9的模型性能略差，可能的原因为当动态卷积核过大时会提取到更多无关特征，而多模态间交互层数对模型性能的影响不大。

表4 在IEMOCAP数据集上跨模态动态卷积核大小和多模态间交互层数的消融实验结果

Table 4 Ablation experimental results on the kernel size of cross-modal dynamic convolution and the number of interaction layers between multi-modals on the IEMOCAP dataset %

核大小	2层		3层		4层		5层		6层		7层		8层	
	ACC	F1值	ACC	F1值	ACC	F1值	ACC	F1值	ACC	F1值	ACC	F1值	ACC	F1值
3	83.0	81.3	83.6	81.3	82.9	81.6	83.0	81.4	83.5	81.2	83.5	81.1	83.0	81.1
5	83.5	81.8	82.8	80.7	83.5	81.3	83.3	81.7	83.4	81.5	83.2	81.4	83.0	81.0
7	82.9	81.6	83.5	81.2	83.2	81.2	83.5	81.4	83.4	81.7	83.5	82.5	83.6	80.6
9	83.5	81.0	83.5	81.9	83.1	81.5	83.5	81.0	83.5	81.7	83.6	81.1	83.6	82.0
11	82.8	82.5	83.5	81.1	83.2	81.0	83.2	81.7	83.5	81.6	83.2	81.2	83.5	82.0
13	83.5	81.5	82.9	81.2	83.5	81.1	83.5	81.4	83.5	81.5	83.5	82.5	83.2	81.2
15	83.0	81.0	83.0	81.5	83.5	81.1	83.3	81.8	83.5	81.0	83.5	80.9	83.2	81.3
17	83.1	81.9	83.2	81.2	82.9	81.2	83.5	81.6	83.4	81.2	83.3	81.5	83.5	81.9
19	83.0	81.5	83.3	81.5	82.8	81.5	83.2	81.3	83.4	80.3	83.0	81.1	83.2	81.0

3 结束语

本文提出一种基于动态卷积与残差门控的多模态情感识别模型，先分别提取文本、图像、音频3个模态中的低级特征、高级局部特征和上下文依赖关系，再通过跨模态动态卷积对不同模态时间维度上的相互作用关系进行建模，识别与情感相关联的线索，同时避免重要信息被无关信息淹没。为了自动学习每组交互表征在最终情感分类中的权重，设计残差门控融合方法，通过竞争或合作的方式有效融合多个交互表征。在CMU-MOSEI和IEMOCAP数据集上的实验结果表明，所提模型在情感分类任务上的性能优于基准模型，并且验证了跨模态动态卷积交互模块和残差门控融合模块的有效性。后续考虑将所提模型拓展至脑电波信号、皮肤电信号等其他模态数据的情感识别，进一步提升其适用范围。

参考文献

[1] CHEN C H. Research on multi-modal mandarin speech emotion recognition based on SVM[C]//Proceedings of IEEE International Conference on Power, Intelligent Computing and Systems. Washington D. C. , USA: IEEE Press, 2019: 173-176.

[2] 乔栋，陈章进，邓良，等. 基于改进语音处理的卷积神经网络中文语音情感识别方法[J]. 计算机工程，2022，48(2)：281-290.
QIAO D, CHEN Z J, DENG L, et al. Method for Chinese

speech emotion recognition based on improved speech-processing convolutional neural network[J]. Computer Engineering, 2022, 48(2): 281-290. (in Chinese)

[3] 柳素红，孙晓，李春彬. 基于位置信息重建与时频域信息融合的脑电信号情感识别[J]. 计算机工程，2021，47(12)：95-102.
LIU S H, SUN X, LI C B. Emotion recognition using EEG signals based on location information reconstruction and time-frequency information fusion[J]. Computer Engineering, 2021, 47(12): 95-102. (in Chinese)

[4] TAN Y, SUN Z, DUAN F, et al. A multimodal emotion recognition method based on facial expressions and electroencephalography[J]. Biomedical Signal Processing and Control, 2021, 70: 103029.

[5] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding[EB/OL]. [2022-05-17]. https://arxiv.org/abs/1810.04805.

[6] HSU W N, BOLTE B, TSAI Y H H, et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2021, 29: 3451-3460.

[7] BALTRUŠAITIS T, ROBINSON P, MORENCY L P. OpenFace: an open source facial behavior analysis toolkit[C]//Proceedings of IEEE Winter Conference on Applications of Computer Vision. Washington D. C., USA: IEEE Press, 2016: 1-10.

[8] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[EB/OL]. [2022-05-17]. https://arxiv.org/abs/1707.07250.

[9] AREVALO J, SOLORIO T, MONTES-Y-GÓMEZ M, et al. Gated multimodal units for information fusion[EB/OL]. [2022-05-17]. https://arxiv.org/abs/1702.01992.

[10] PAN Z, LUO Z, YANG J, et al. Multi-modal attention for speech emotion recognition[EB/OL]. [2022-05-17]. https://arxiv.org/abs/2009.04107.

[11] YANG K C, XU H, GAO K. CM-BERT: cross-modal BERT for text-audio sentiment analysis[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York, USA: ACM Press, 2020: 521-528.

[12] TSAI Y H, BAI S J, LIANG P P, et al. Multimodal Transformer for unaligned multimodal language sequences[C]// Proceedings of Association for Computational Linguistics Meeting. Philadelphia, USA: ACL Press, 2019: 6558-6569.

[13] KWON S. MLT-DNet: speech emotion recognition using 1D dilated CNN based on multi-learning trick approach[J]. Expert Systems with Applications, 2021, 167: 114177.

[14] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. [2022-05-17]. https://arxiv.org/abs/1412.3555.

[15] TANG G, MÜLLER M, RIOS A, et al. Why self-attention? A targeted evaluation of neural machine translation architectures[EB/OL]. [2022-05-17]. https://arxiv.org/abs/1808.08946.

[16] WU F, FAN A, BAEVSKI A, et al. Pay less attention with lightweight and dynamic convolutions[EB/OL]. [2022-05-17]. https://arxiv.org/abs/1901.10430.

[17] ZADEH A, PU P. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: ACL Press, 2018: 1-8.

[18] 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.

[19] PHAM H, LIANG P P, MANZINI T, et al. Found in translation: learning robust joint representations by cyclic translations between modalities[C]//Proceedings of AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2019: 6892-6899.

[20] WANG Y, SHEN Y, LIU Z, et al. Words can shift: dynamically adjusting word representations using nonverbal behaviors[C]//Proceedings of 2019 AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2019: 7216-7223.

[21] MAI S J, HU H F, XU J, et al. Multi-fusion residual memory network for multimodal human sentiment comprehension[J]. IEEE Transactions on Affective Computing, 2022, 13(1): 320-334.

[22] LÜ F M, CHEN X, HUANG Y Y, et al. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2021: 2554-2562.

[23] SHENOY A, SARDANA A. Multilogue-Net: a context aware RNN for multi-modal emotion detection and sentiment analysis in conversation[EB/OL]. [2022-05-17]. https://arxiv.org/abs/2002.08267.

[24] ZADEH A, LIANG P P, MAZUMDER N, et al. Memory fusion network for multi-view sequential learning[EB/OL]. [2022-05-17]. https://arxiv.org/abs/1802.00927.

[25] RAHMAN W, HASAN M K, LEE S W, et al. Integrating multimodal information in large pretrained Transformers[C]// Proceedings of the Conference Association for Computational Linguistics Meeting. Philadelphia, USA: ACL Press, 2020: 2359-2369.

编辑　陆燕菲