西北大学学报(自然科学版)
2024 年 4 月,第 54 卷第 2 期,Apr. ,2024,Vol. 54,No. 2
Journal of Northwest University (Natural Science Edition)

·人工智能情感计算·

一种基于语音、文本和表情的多模态情感识别算法

吴　晓[1],牟　璇[1],刘银华[2,3],刘晓瑞[1,2]

(1. 青岛大学 自动化学院,山东 青岛　266071;2. 青岛大学 未来研究院,山东 青岛　266071;
3. 山东省工业控制技术重点实验室,山东 青岛　266071)

摘要　针对当前多模态情感识别算法在模态特征提取、模态间信息融合等方面存在识别准确率偏低、泛化能力较差的问题,提出了一种基于语音、文本和表情的多模态情感识别算法。首先,设计了一种浅层特征提取网络(Sfen)和并行卷积模块(Pconv)提取语音和文本中的情感特征,通过改进的 Inception-ResnetV2 模型提取视频序列中的表情情感特征;其次,为强化模态间的关联性,设计了一种用于优化语音和文本特征融合的交叉注意力模块;最后,利用基于注意力的双向长短期记忆(BiLSTM based on attention mechanism,BiLSTM-Attention)模块关注重点信息,保持模态信息之间的时序相关性。实验通过对比 3 种模态不同的组合方式,发现预先对语音和文本进行特征融合可以显著提高识别精度。在公开情感数据集 CH-SIMS 和 CMU-MOSI 上的实验结果表明,所提出的模型取得了比基线模型更高的识别准确率,三分类和二分类准确率分别达到 97. 82% 和 98. 18%,证明了该模型的有效性。

关键词　多模态;情感识别;并行卷积;交叉注意力
中图分类号:TP391　　**DOI**:10. 16152/j. cnki. xdxbzr. 2024-02-004

A multimodal emotion recognition algorithm based on speech, text and facial expression

WU Xiao[1], MOU Xuan[1], LIU Yinhua[2,3], LIU Xiaorui[1,2]

(1. Automation School, Qingdao University, Qingdao 266071, China;
2. Institute of Future, Qingdao University, Qingdao 266071, China;
3. Shandong Key Laboratory of Industrial Control Technology, Qingdao 266071, China)

Abstract　Aiming at the problems of low recognition accuracy and poor generalization ability of current multimodal emotion recognition algorithms in modal feature extraction and information fusion between modalities, a multimodal emotion recognition algorithm based on speech, text and expression is proposed. Firstly, a shallow feature extraction network (Sfen) combined with parallel convolution module (Pconv) is designed to extract the emotional features in speech and text. A modified Inception-ResnetV2 model is adopted to capture the emotional features of expression in video stream. Secondly, in order to strengthen the correlation among modal-

收稿日期:2023-10-13
基金项目:国家重点研发计划"智能机器人"专项资助项目(2020YFB1313600);青岛市自然科学基金资助项目(23-2-1-126-zyyd-jch);山东省高等学校优秀青年创新团队支持计划项目(2022KJ142)。
第一作者:吴晓,男,从事多模态情感识别研究,3208602731@qq. com。
通信作者:刘晓瑞,男,博士,从事机器人学、人机社会交互研究,liuxiaorui@qdu. edu. cn。

ities, a cross attention module is designed to optimize the fusion between speech and text modalities. Finally, a bidirectional long and short-term memory module based on attention mechanism (BiLSTM-Attention) is used to focus on key information and maintain the temporal correlation between modalities. By comparing the different combinations of the three modalities, it is found that the hierarchical fusion strategy that processes speech and text in advance can obviously improve the accuracy of the model. Experimental results on the public emotion datasets CH-SIMS and CMU-MOSI show that the proposed model achieves higher recognition accuracy than the baseline model, with three-class and two-class accuracy reaching 97.82% and 98.18% respectively, which proves the effectiveness of the model.

Keywords multimodal; emotion recognition; parallel convolution; cross attention

近年来,随着人工智能技术的快速发展,人机交互逐渐成为了当前科研人员研究的热点。情感分析作为人机交互的重要组成部分,也呈现出了模态多元化的趋势[1],比如使用语音、文本、表情,甚至脑电等生理信号来进行情感分析。因此,如何处理和融合这些异构信息,实现对其准确的分析与判断,成为了当前需要解决的重点问题。

在情感识别领域中,传统的机器学习方法如朴素贝叶斯(naive Bayes, NB)、支持向量机(support vector machine, SVM)等[2-3]被广泛应用。但随着深度学习技术的发展,以卷积神经网络(convolutional neural network, CNN)、循环神经网络(recurrent neural network, RNN)、深度卷积神经网络(deep convolutional neural network, DCNN)[4-6]为代表的数据驱动方法逐渐成为情感分析的主流。目前,研究人员已经在单模态情感识别领域取得了一定进展。在文本情感识别方面,Xu 等人提出一种基于 CNN 的微博情绪分类模型 CNN_Text_Word2vec,使模型的整体准确率比主流方法提高了 7.0%[7];在图像情感识别方面,郑剑等人提出了一种基于 DCNN 的 FLF-TAWL 网络,该网络能够自适应捕捉人脸重要区域,提高人脸识别的有效性[8];在语音情感识别方面,部分研究将声学特征和 RNN 进行结合,如 Dutta 等人提出一种语音识别模型,利用 RNN 提取线性预测编码(linear predictive coding, LPC)和 Mel 频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)特征,并在识别阿萨姆语上取得了一定效果[9]。

近期的研究表明,多模态情感模型能够将来自不同感知模态的信息有效融合。由于充分利用了数据的多样性,多模态模型表现出比单模态模型更大的优势。针对多模态情感识别,国内外学者已经开展了深入的研究工作。如 HOU 等人提出一种早期融合模型 EF-LSTM,通过拼接语音、文本和表情 3 种模态的特征并利用 LSTM 进行编码,有效提取了模态间的交互信息[10]。Zadeh 等人设计一种张量融合网络(TFN),通过采用多维张量的外积操作,较好地捕获了不同模态间的交互信息[11]。Liu 等人设计一种低秩多模态融合算法(LMF),在 TFN 的基础上进行低秩多模态张量融合,使网络效果得到一定的提升[12]。Zadeh 等人提出一种记忆融合网络(MFN),通过利用注意力机制和多视图门控网络,同步捕捉了时序序列和模态间的交互信息[13]。Tsai 等人提出一种跨模态网络 Transformer(MulT),通过扩展多式 Transformer 结构,成功解决了不同模态数据的长期依赖性问题,进一步提高了模型性能[14]。Yu 等人提出一种自监督多任务学习网络 Self-MM,通过设计基于自监督学习策略的标签生成模块,并引入权重自调整策略,较好地实现了对情感的预测分类[15]。虽然研究者不断探索新的情感识别模型以提升多模态情感识别的准确率,但仍存在一些不足。在情感特征提取方面,上述多模态情感模型主要通过预训练模型实现对情感特征提取。但预训练模型往往需要进行微调或迁移学习来达到适应特定任务的目的,可能会导致在小样本数据集或特定应用中出现泛化性能力不足的问题。在特征融合方面,上述多模态模型虽然采用了一些改进型的融合方法,但在融合过程中没有很好地考虑模态特征间的相关性及模态的选择性问题,导致最终的识别准确率偏低。

针对上述问题,本文在现有研究的基础上提出了一种基于语音、文本和表情的多模态情感识别算法。该算法利用 Sfen 网络和 Pconv 模块充分提取语音和文本情感特征;采用改进的 Inception-ResnetV2 网络[16]提取表情情感特征;通过交叉注意力融合(cross attention fusion, CAF)模块强化语音和文本特征的相关性;最后,利用 BiLSTM-At-

tention 模块获取关键信息,保持信息在时间上的连续性。

1　多模态情感识别模型

构建多模态情感识别模型通常包括以下几个方面:多模态信息预处理、情感特征提取、情感识别模型的设计与选择、特征融合方案[17]。如何确定有效的模态组合方案,并实现有效的特征融合是本文需要研究的重点问题。本文利用语音(A)、文本(T)与表情(V)3 种模态构建多模态情感识别模型,该模型主要是由 Sfen 网络、Pconv 模块、BiLSTM-Attention 模块和交叉注意力融合(CAF)模块组成,整体框架如图 1 所示。

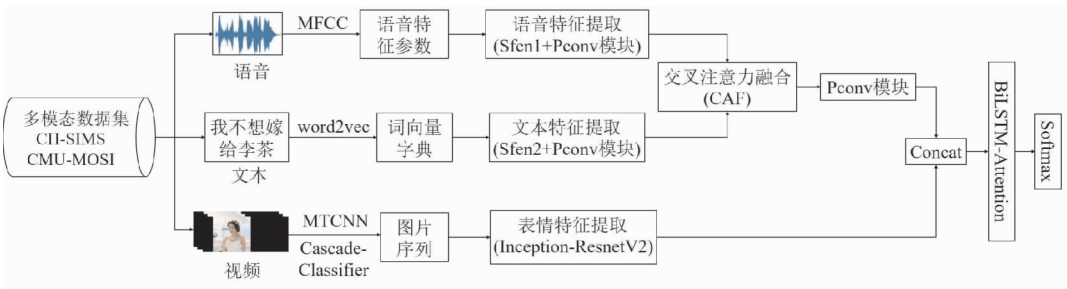

图 1　三模态情感模型框架图

Fig. 1　Framework diagram of the trimodal emotion model

在图 1 所示的模型框架中,首先利用 Sfen 网络和 Pconv 模块提取语音和文本的情感特征,并通过 CAF 模块实现 2 个模态间的信息互补,优化模态间的信息融合。对于基于视频的表情信息,该情感识别方法以图片识别分类常用的 Inception-ResnetV2 模型为基础进行改进,以提高在多种环境背景下的鲁棒性。在融合策略上,本文将语音-文本特征与表情特征进行特征级[18]融合,并通过 BiLSTM-Attention 模块后,利用 Softmax 实现对情感的识别分类。

1.1　数据预处理

目前语音信号预处理的方法主要有傅里叶变换、神经网络、动态时间规划和梅尔频率倒谱系数(MFCC)[19]等,其中,梅尔频率倒谱系数提取到的特征参数更接近人耳感知的特点。本文利用 MFCC 对视频中的原始语音信号进行预处理,通过对提取到的语音数据进行预加重、分帧和加窗等操作,将原始语音信号转换为语音特征参数。针对原始文本数据,首先,采用文本分类中常用的 jieba 分词工具[20]对文本中的分词进行分类;然后,利用停止词数据库去除文本信息中的停止词,避免无用信息的干扰;最后,通过 word2vec[7]模型将文本转换成词向量形式,构建词向量字典。针对研究中使用的文本数据量,使用了 word2vec 中的 CBOW[21]作为本文的神经网络语言模型。

数据集中原始视频片段的背景、光线和环境等因素[22]可能会导致从视频中提取到的连续帧无法被准确地识别为人脸。因此,本文首先将每个视频片段逐帧处理成连续的图片,利用 MTCNN[23]模型和 OpenCV 库中的 CascadeClassifier[24]人脸级联检测器实现对人脸的检测,提高对人脸的检测精度;然后,将检测到的人脸图像裁剪成 149 × 149 的统一尺寸大小;最后,经过归一化、灰度化后,输出处理后的图片序列。

1.2　语音文本特征提取

在情感识别的过程中,浅层特征提取主要从输入的文本、语音或图像中提取有关情感的表层信息,是数据预处理后的一项关键步骤。针对语音和文本模态,本文设计了一种 Sfen 网络实现对 2 种模态浅层特征的提取,Sfen 网络结构如图 2 所示。

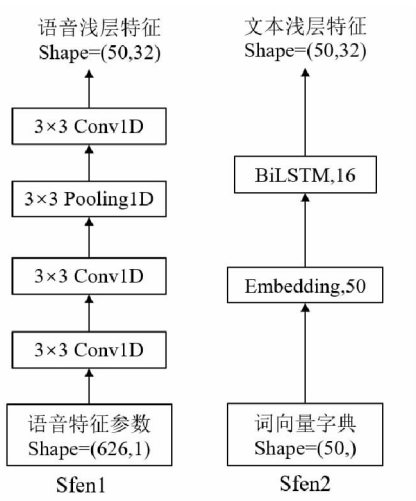

图 2　Sfen 网络结构图

Fig. 2　Sfen network structure diagram

对于音频输入,语音特征参数经过 Sfen1 网络中一维卷积层和池化层的处理后得到语音浅层特征(卷积核大小为 3 ×3)。类似地,对于文本输入,词向量字典通过 Sfen2 网络中的 Embedding 和 BiLSTM 层后得到文本浅层特征。其中,Embedding 层增强了文本特征之间的相关性,在 Embedding 层之后引入 BiLSTM 能够捕获更丰富的上下文信息,同时保持文本间的序列关系。语音特征参数和词向量字典经过各自的 Sfen 网络处理后,其输出特征维度保持相同,确保了后续语音和文本特征融合的可行性。

为获取深层次的情感特征,本文利用残差网络[25](residual network,ResNet)的思想将最大池化层与卷积层进行拼接,针对语音和文本 2 个模态设计了一种 Pconv 模块,其结构如图 3 所示。

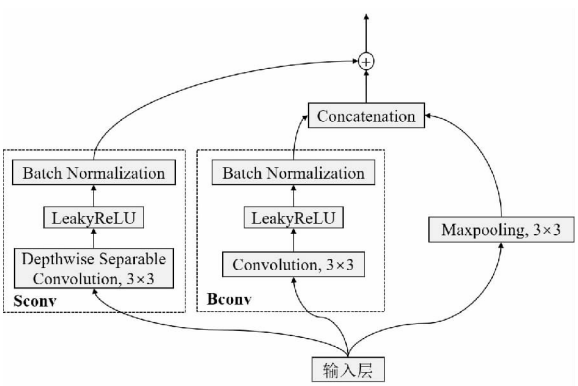

图 3 Pconv 模块结构

Fig. 3 Pconv module structure

在图 3 中,Pconv 模块由最大池化层、Bconv 单元和 Sconv 单元 3 部分组成。其中,Bconv 单元由 3 层组成:传统卷积层、LeakyReLU 激活函数、批标准化(Batch Normalization)。Sconv 单元与 Bconv 单元类似,但在输入环节使用了卷积核大小为 3 ×3 的深度可分离卷积层[26](Depth Separable Convolution),进一步减少运算参数的数量,提高运算效率。在次级输出环节,本文将最大池化层的输出和 Bconv 单元的输出进行拼接,其输出再与 Sconv 单元的输出特征进行叠加。上述设计中的拼接环节可以增加最终输出特征的多样性,而叠加环节又可以在输出前对每个维度的特征进行增强和补充。该 Pconv 模块采用的残差连接的方法,避免了神经网络中的信息冗余和梯度爆炸[27]问题,使得网络能够更有效地学习到数据的特征表示,保证了特征提取的充分性。

1.3 表情特征提取

目前处理视频序列中面部表情信息的方法主要是 3D 卷积和 2D 卷积,其中,3D 卷积能够在时间维度上捕捉连续视频帧之间的动态信息,2D 卷积能够在每个视频帧中提取空间特征。本文将 3D 卷积与 2D 卷积相结合,先利用 2D 卷积提取图像帧的空间特征,再使用 3D 卷积捕捉时间维度的特征,不仅可以形成更深层次的特征表示,还能够有效地提高面部表情的识别效率。

Inception-ResnetV2 神经网络模型具有良好的特征提取能力和泛化性能,常用于图像分类、目标检测等任务。本研究采用的表情情感识别模型是在 Inception-ResnetV2 模型的基础上进行的改进,利用 3D 卷积与 2D 卷积相结合的多尺度卷积核[28]处理表情数据信息。改进后的模型结构如图 4 所示。在传统的 Inception-ResnetV2 模型的基础上,将其前半部分的特征提取层由 2D 转换为 3D,利用三维卷积核滑动提取相应特征。由于时间维度较小,当时间维度卷积为 1 时,再次通过压缩方式(squeeze)将 3D 卷积转换为 2D 卷积,减少训练参数的产生,降低运算难度。

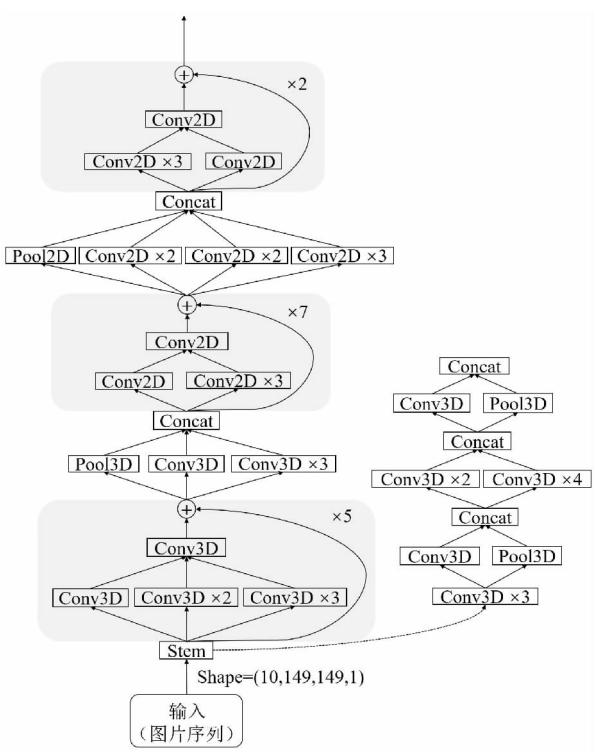

图 4 表情特征提取网络结构图

Fig. 4 Expression feature extraction network structure diagram

1.4 交叉注意力融合模块

模态特征的融合需要考虑不同模态间的耦合关系。目前的研究表明 T(文本)和 A(语言)2 种模态之间存在紧密的时序与特征耦合关系[29]。

本文改变了传统的特征融合方式,设计了一种基于交叉注意力的融合模块,在保留模态内特征的同时,有效地编码 T 和 A 模态间的信息。该融合模块结构如图 5 所示。

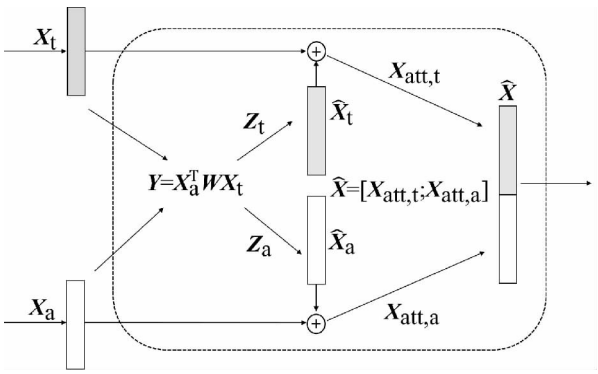

图 5　交叉注意力模块结构示意图

Fig. 5　Schematic diagram of the cross attention module structure

在图 5 所示的交叉注意力融合模块中,$\boldsymbol{X}_t$ 和 $\boldsymbol{X}_a$ 分别代表数据集中的视频序列 $\boldsymbol{X}$ 经过 Pconv 模块后提取出的 T 和 A 的深层特征。为使模态间的异质性最小化,设置了一个可学习的权重矩阵 $\boldsymbol{W} \in \mathbf{R}^{k\times k}$,相互计算的关系如式(1)所示,

$$\boldsymbol{Y} = \boldsymbol{X}_a^{\mathrm{T}} \boldsymbol{W} \boldsymbol{X}_t \tag{1}$$

式中:$\boldsymbol{Y} \in \mathbf{R}^{l\times l}$;$\boldsymbol{W}$ 代表文本和语音的相互关系权重;k 代表文本和语音的特征维度。相关矩阵 $\boldsymbol{Y}$ 给出了 T 和 A 特征之间的相关性度量,较高的相关系数说明子序列对应的 T 和 A 特征之间具有较强的相关性。基于以上思路,分别利用 $\boldsymbol{Y}^{\mathrm{T}}$ 和 $\boldsymbol{Y}$ 的 softmax 函数进一步计算 T 和 A 特征的交叉注意力权重 $\boldsymbol{Z}_t$ 和 $\boldsymbol{Z}_a$。计算如式(2)和(3)所示。

$$\boldsymbol{Z}_{t_{i,j}} = \frac{\mathrm{e}^{\boldsymbol{Y}^{\mathrm{T}}_{i,j/T_s}}}{\sum_{k=1}^{K} \mathrm{e}^{\boldsymbol{Y}^{\mathrm{T}}_{i,k/T_s}}} \tag{2}$$

$$\boldsymbol{Z}_{a_{i,j}} = \frac{\mathrm{e}^{\boldsymbol{Y}_{i,j/T_s}}}{\sum_{k=1}^{K} \mathrm{e}^{\boldsymbol{Y}_{k,j/T_s}}} \tag{3}$$

式中:i 和 j 表示矩阵 $\boldsymbol{Y}$ 的第 i 行和第 j 列元素;T_s 表示 softmax 系数。

在上述计算中,权重矩阵 $\boldsymbol{W}$ 是基于 T 和 A 特征的相互关系学习的,即一种模式的注意力权重是由另一种模式确定的,从而有效地利用了 T 和 A 这 2 个模态的互补特性。在得到交叉注意力权重后,利用交叉注意力权重计算获得 T 和 A 的特征注意力图 $\hat{\boldsymbol{X}}_t$ 和 $\hat{\boldsymbol{X}}_a$,如式(4)和(5)所示。

$$\hat{\boldsymbol{X}}_t = \boldsymbol{X}_t \cdot \boldsymbol{Z}_t \tag{4}$$

$$\hat{\boldsymbol{X}}_a = \boldsymbol{X}_a \cdot \boldsymbol{Z}_a \tag{5}$$

式中:$\boldsymbol{Z}_t$ 和 $\boldsymbol{Z}_a$ 分别代表 T 和 A 特征的交叉注意力权重。通过将重加权的注意力图添加到相应的特征上,可获得 2 种模态的深层特征表征 $\boldsymbol{X}_{att,t}$ 与 $\boldsymbol{X}_{att,a}$,如式(6)和(7)所示。

$$\boldsymbol{X}_{att,t} = \tanh(\boldsymbol{X}_t + \hat{\boldsymbol{X}}_t) \tag{6}$$

$$\boldsymbol{X}_{att,a} = \tanh(\boldsymbol{X}_a + \hat{\boldsymbol{X}}_a) \tag{7}$$

将 $\boldsymbol{X}_{att,t}$ 和 $\boldsymbol{X}_{att,a}$ 拼接起来,得到 T 和 A 的特征表示,即 $\hat{\boldsymbol{X}} = [\boldsymbol{X}_{att,t}, \boldsymbol{X}_{att,a}]$。经过交叉注意力模块融合后的特征将再次输入到下一级 Pconv 模块中,通过其并行结构充分提取融合后的信息。

1.5　BiLSTM-Attention 模块

长短时记忆网络[30](long short term memory, LSTM)利用 3 个不同门结构,有效解决了序列数据的依赖性和语序问题,其结构如图 6 所示。

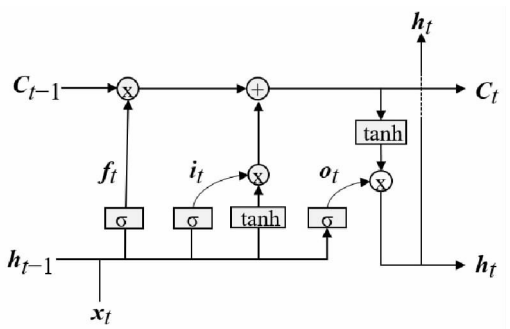

图 6　LSTM“门”结构

Fig. 6　LSTM “gate” structure

在 t 时刻,将当前隐层状态记为 h_t,各门状态更新如下:

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_f) \tag{8}$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t * \boldsymbol{C}_{t-1} + \boldsymbol{i}_t * \tanh(\boldsymbol{W}_c \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_c) \tag{9}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_i) \tag{10}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + b_o) \tag{11}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * \tanh(\boldsymbol{C}_t) \tag{12}$$

式中:$\boldsymbol{x}_t$ 表示当前输入单元状态;$\boldsymbol{f}_t$、$\boldsymbol{C}_t$、$\boldsymbol{i}_t$、$\boldsymbol{o}_t$ 分别表示当前遗忘门、存储单元、输入门、输出门;b_* 表示偏置项;$\boldsymbol{W}_*$ 表示权重矩阵;σ 是激活函数。

LSTM 只能获取输出时刻前的信息,不能利用反向信息,本文利用了 2 个单向 LSTM 构成双向长短时记忆网络(BiLSTM),同时处理前向与后向信息。此外,注意力机制[31](attention)能够在训练过程中根据特征序列信息的重要程度赋予权重值,选择性忽略非重要信息,最大化相关向量的贡献。为使模型更好获取输入序列中不同位置的重要性,在 BiLSTM 层的基础上添加注意力层提高网络对关键信息的感知和利用能力。BiLSTM-Attention 模块结构如图 7 所示。

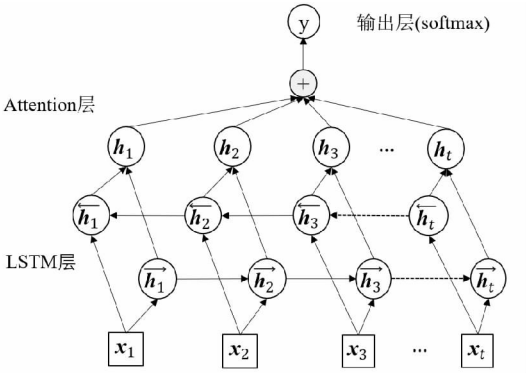

图 7 BiLSTM-Attention 模块结构图

Fig. 7 BiLSTM-Attention module structure diagram

2 多模态情感识别实验

2.1 数据集

实验数据集选用公开的多模态情感数据集 CH-SIMS[32] 和 CMU-MOSI[33]。CH-SIMS 数据集取材自 60 部电影、电视剧与综艺节目,包括2 281 个视频片段。每个视频片段中的情感状态由 5 个人给予标注,以平均标注结果作为该片段的情绪状态。CMU-MOSI 数据集包含 YouTube 上收集的 90 个视频,并将其人工划分为 2 199 个视频片段。其中,CH-SIMS 数据集的情绪状态分为消极、中性和积极 3 种(对应标签 0、1、2),CMU-MOSI 数据集的情绪状态分为消极和积极 2 种(对应标签 0、1)。同时,将数据集划分训练集、验证集和测试集。数据集信息如表 1 所示。

表 1 数据集信息

Tab. 1 Datasets Information

数据集	CH-SIMS	CMU-MOSI
训练集	1 596	1 539
验证集	456	440
测试集	229	220
总 计	2 281	2 199

2.2 参数设置与评估指标

实验基于 TensorFlow 深度学习框架进行模型搭建,在 NF5468 型 24 * GPU 服务器上进行模型训练。训练中采用 SGD 作为网络优化函数,LeakRelu 作为激活函数。训练时的 Batch size 设置为 32,Epoch = 1 000,学习率为 1e-4,LSTM 层的隐藏层单元数量为 128。为防止网络在训练中出现过拟合现象,在 BiLSTM-Attention 层后使用 P = 0.5 的 Dropout 作为补偿。

本文采用了准确率(Accuracy,式中简记 R_{Acc})和 F1 值(F1-score,式中简记 F_1)作为模型整体性能的评估指标。具体计算如式(13)和(14)所示。

$$R_{\mathrm{Acc}}=\frac{N_{\mathrm{TP}}+N_{\mathrm{TN}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}+N_{\mathrm{TN}}} \tag{13}$$

$$F_1=\frac{2N_{\mathrm{TP}}}{2N_{\mathrm{TP}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}} \tag{14}$$

式中:N_{TP}表示实际与预测均为正的样本数;N_{FP}表示实际为负但预测为正的样本数;N_{TN}表示实际与预测均为负的样本数;N_{FN}表示实际为正但预测为负的样本数。

2.3 组合方案讨论

为验证提出的多模态情感框架中采用的模态组合方式的有效性,本文共讨论了 4 种(AT-V、AV-T、TV-A、A-T-V)模态组合方案,如图 8 所示。

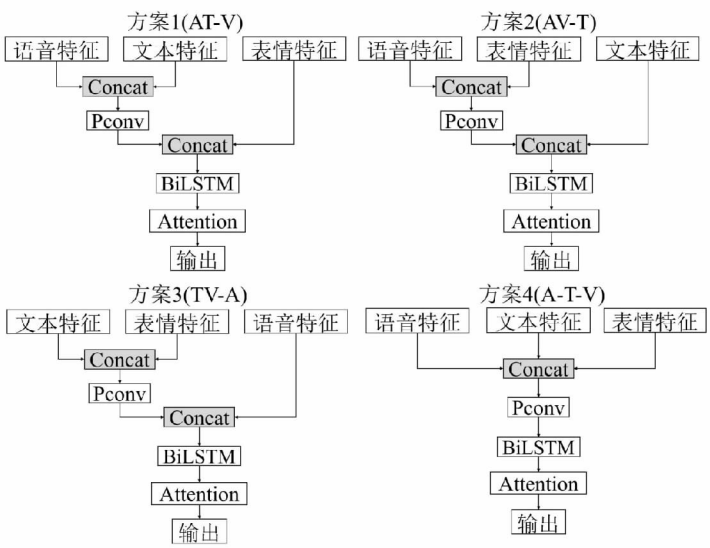

图 8 三模态组合方案

Fig. 8 Three modals combination schemes

为保证实验的可靠度,4 组实验均在 CH-SIMS 和 CMU-MOSI 数据集上进行验证且训练超参数保持一致,实验结果如表 2 所示。其中,Acc-2 和 Acc-3 分别表示二分类和三分类的准确率。通过表 2 可以看出,方案 1 中的模态组合 AT-V 在 2 类数据集上都取得比另外 3 种方案更好的识别效果。其中,方案 1 在 CH-SIMS 上的 Acc-3、F1 分别达到了 96.94%、96.67%;在 CMU-MOSI 上分别达到 97.73% 和 97.52%。表明本文采用的语音和文本先进行特征融合是最优的三模态组合方式。

表 2　三模态组合方案结果比较

Tab. 2　Comparison of results of three modals combination schemes　　单位:%

方案	模态组合	CH-SIMS		CMU-MOSI	
		Acc-3	F1	Acc-2	F1
方案 1	AT-V	96.94	96.67	97.73	97.52
方案 2	AV-T	96.51	96.63	97.70	97.28
方案 3	TV-A	96.07	95.67	96.36	96.41
方案 4	A-V-T	96.51	95.98	96.82	96.86

2.4　消融实验

2.4.1　融合方式消融实验

在确定 2.3 节中方案 1 为最优的三模态组合(AT-V)后,为验证本文提出的交叉注意力融合模块(CAF)的优势,进一步将方案 1 中的语音和文本特征融合的方式由 Concat 分别替换为 Self-Attention[34] 和 CAF 并进行消融实验。其中,Concat 表示不添加注意力的简单特征拼接,Self-Attention 表示自注意力融合方式,其强调相关特征的组成部分。实验结果如表 3 所示。

表 3　融合方式消融结果比较

Tab. 3　Comparison of ablation results of fusion methods　　单位:%

融合方式	CH-SIMS		CMU-MOSI	
	Acc-3	F1	Acc-2	F1
Concat	96.94	96.67	97.73	97.52
Self-Attention	97.20	96.97	97.95	97.76
CAF(Proposed)	97.82	97.33	98.18	97.87

通过表 3 可以看出,在引入了交叉注意力后,该模型在 2 类数据集上的评估指标均得到了显著的提升。在 CH-SIMS 数据集上,Acc-3 和 F1 值分别达到 97.82% 和 97.33%;在 CMU-MOSI 数据集上,Acc-2 和 F1 值分别达到 98.18% 和 97.87%。相对于简单的特征拼接(Concat)的融合方式,自注意力(Self-Attention)融合方法虽在一定程度提高了系统的性能,凸显了相关的特征组成部分,但是其计算方式较为复杂,增加了模型的复杂性。相对于自注意力融合,由于交叉注意力融合(CAF)机制通过利用 A-T 特征之间的相互关联性,且计算方式更为简便,有效地捕获了 2 种模态的互补性,进一步提高了模型性能。以上结果符合本文的预期设想,证明了提出的交叉注意力能够更好地利用语音和文本间的特征互补关系,进一步提高特征融合的效果。

2.4.2　BiLSTM-Attention 模块消融实验

为验证本文利用的 BiLSTM-Attention 模块的作用,做了 3 组对比实验。①FC:语音、文本与表情 3 种模态进行特征融合后输入到全连接层输出;②LSTM:在特征融合后通过 LSTM 网络输出;③BiLSTM:特征融合后通过双向 LSTM 输出。实验结果如表 4 所示。

表 4　BiLSTM-Attention 消融结果比较

Tab. 4　Comparison of BiLSTM-Attention ablation results　　单位:%

模型	CH-SIMS		CMU-MOSI	
	Acc-3	F1	Acc-2	F1
FC	96.07	96.15	96.36	96.32
LSTM	96.94	96.90	96.82	96.78
BiLSTM	97.37	97.18	97.73	97.65
BiLSTM-Att(Ours)	97.82	97.33	98.18	97.87

从表 4 可以看出,在以上 4 种模型对比实验结果中,本文的 BiLSTM-Attention 模块在 Acc 和 F1 值上均取得了最优。在 CH-SIMS 数据集上较其他 3 种模型至少高出了 0.004 5 和 0.001 5;在 CMU-MOSI 数据集上至少高出了 0.004 5 和 0.002 2。通过以上不同模型的对比实验结果可知,本文采用的 BiLSTM 与 Attention 相结合的方法有助于更好地实现对多模态情感的分析和预测,进一步表明了该网络模块对多模态情感模型的重要性。

2.4.3　模态消融实验

为验证本文提出的网络模型的适用性,在 CH-SIMS 数据集分别进行了单模态、双模态及三

模态 7 种组合的消融实验。具体的消融实验结果如表 5 所示。

表 5 CH-SIMS 模态消融实验结果

Tab. 5 Results of the CH-SIMS modal ablation experiments

单位:%

模态组合	Acc-3	F1
T	77. 13	75. 67
A	72. 05	71. 83
V	87. 81	87. 26
T + A(CAF)	93. 38	93. 33
T + V	94. 76	94. 28
A + V	95. 20	94. 64
A + T + V(Proposed)	97. 82	97. 33

通过表 5 可以观察到三模态的 Acc-3 和 F1 指标均优于单模态和双模态,效果最好。在单模态情感识别实验中,表情模态信息预测真实情感能力最强,Acc-3 达到 87. 81%,F1 达到 87. 26%。在双模态情感识别实验中,A + V 组合效果最好,Acc-3、F1 分别达到 95. 20%、94. 64%,T + V 和 T + A 次之。心理学家 Mehrabian 的研究发现,人们在日常生活中的情感信息主要是通过表情与语言传达的[35],这也与消融实验中 A + V 模态组合的实验结果相符。以上的消融实验不仅验证了利用语音、文本和表情进行多模态情感识别的必要性,也证明了本文提出的引入 CAF 思想的多模态情感融合方法的可行性和有效性。

2.5 对比实验

本节将提出的多模态模型与目前多种经典的情感模型进行对比,基线模型介绍如下。

EF-LSTM[10]:早期融合的 LSTM 模型。首先拼接 3 种模态的特征向量,然后利用 LSTM 对拼接后的特征进行编码。

LF-LSTM[10]:晚期融合的 LSTM 模型。首先 LSTM 编码 3 个模态特征向量,然后结合 LSTM 最后一层的隐层向量构成多模态的特征表示。

MAG-BERT[36]:多模态自适应门模型。通过提出一种多模态自适应门机制(MAG),使 BERT 和 XLNet 能够在微调过程中接受多模态数据的输入。

MuIT[14]:多模态 Transformer 模型。通过考虑不同模态之间的时序依赖关系,实现在非对齐数据集上的跨模态交互。

MMIM[37]:多模态分层互信息最大化框架。在多模态分析任务中引入互信息理论,最大化输入级和融合级特征表征的互信息。

MISA[38]:模态不变和模态特定表征框架。针对不同模态学习模态不变和模态特定的特征表示,对不同种类的表示向量提出分布相似性损失、重建损失、正交损失及任务预测损失。

Self-MM[15]:自监督多任务学习网络。通过一种基于自监督策略的标签生成模块获取单模态表征,并在训练阶段设计一种平衡不同任务损失的权重调整策略。

CMFIB[39]:跨模态融合与信息瓶颈模型。利用互信息估计模块优化多模态表示向量与真实标签之间的互信息下限,最小化输入数据与多模态表示向量间的互信息。

经过多次对比实验,在 2 类数据集上和其他基线模型的评估指标对比结果如表 6 所示。

表 6 各模型性能对比结果

Tab. 6 Performance comparison of each model

单位:%

融合方式	CH-SIMS		CMU-MOSI	
	Acc-3	F1	Acc-2	F1
EF-LSTM	57. 38	56. 89	76. 65	76. 69
LF-LSTM	70. 20	65. 29	76. 73	76. 82
MuIT	68. 27	64. 23	77. 41	77. 20
MAG-BERT	76. 37	76. 42	84. 19	84. 16
MMIM	77. 90	77. 92	84. 14	84. 00
MISA	79. 21	78. 53	84. 20	84. 24
Self-MM	79. 87	79. 87	84. 33	84. 35
CMFIB	80. 28	80. 27	86. 56	86. 50
Ours	97. 82	97. 33	98. 18	97. 87

由表 6 可知,本文提出的模型在 Acc 和 F1 值 2 类评估指标上要优于对比的基线模型,尤其在 CMU-MOSI 数据集上表现更好,Acc-2 和 F1 指标比最优基线模型分别提升了 0. 116 2 和 0. 113 7;在 CH-SIMS 数据集上,Acc-3 和 F1 值比最优基线模型分别提升了 0. 175 4 和 0. 170 6。该结果表明,本文设计的特征提取网络以及交叉注意力机制等组件能够有效地挖掘模态间的特征关系,增强模态间的相互依赖性。这对于多模态数据的融合和各项评估指标的提升产生了显著效果。

在上述基线模型中,EF-LSTM 和 LF-LSTM 效果表现最差。这是因为 2 种模型直接拼接 3 种特

征,保留了大量噪声,无法筛选出重要信息。本文的注意力机制能够对关键信息进行加权处理,增强其显著性,进而提升模型的性能。与 MuIT 和 MAG-BERT 相比,本文的模型的 Acc 指标在 CH-SIMS 上至少提升了约 21 个百分点,在 CMU-MOSI 上至少提升了约 14 个百分点。MuIT 在计算模态间的依赖关系时,未考虑上下文信息,且网络结构较为复杂。MAG-BERT 虽较 MuIT 有一定的提升,但在预训练或微调过程中需要大量的多模态数据,可能会导致模型计算困难。本文模型在情感计算时通过利用多尺度卷积核和 BiLSTM 网络,降低了计算量并保持了上下文时序相关性,提高了计算效率。与 MMIM 和 MISA 相比,本文模型采用的交叉注意力融合机制更加适用于多模态识别任务,在有效利用不同模态互补特性的同时增强了模态间的相关性。与 Self-MM 和 CMFIB 相比,所提出的方法在 2 类数据集的评估指标上表现出色,取得了较好的效果。Self-MM 在任务间特征共享方面容易过拟合某些任务,可能导致其性能的下降。CMFIB 在情感分析时只能捕捉到变量之间的关联性,难以充分捕捉模态的深层情感特征。本文设计的 Pconv 模块利用并行架构和特定网络层降低了过拟合的风险,并有效提取了深层次的特征。

3　结语

针对当前多模态情感模型存在识别精度低等问题,本文提出了一种基于语音、文本和表情的多模态情感识别算法。该模型由 Sfen 网络、Pconv 模块和改进的 Inception-ResnetV2 网络提取多模态特征,利用交叉注意力融合机制强化语音-文本双模态的关联性,并通过 BiLSTM-Attention 模块实现对情感的预测和分类。在 CH-SIMS 和 CMU-MOSI 数据集上的实验表明,该模型可以更好地提取模态特征并进行特征融合,显著提高情感识别的精度。接下来本研究将进一步细化情感类别,并探讨在细粒度识别任务下的多模态融合算法的架构设计。

参考文献

[1] 李霞, 卢官明, 闫静杰, 等. 多模态维度情感预测综述[J]. 自动化学报, 2018, 44(12): 2142-2159.
LI X, LU G M, YAN J J, et al. A review of multimodal dime-nsional sentiment prediction[J]. Journal of Auctomatica Sinica, 2018, 44(12): 2142-2159.

[2] RISH I. An empirical study of the naive Bayes classifier [J]. Journal of Universal Computer Science, 2001, 1(2):127.

[3] 赵健, 周莉芸, 武孟青, 等. 基于人工智能的抑郁症辅助诊断方法[J]. 西北大学学报(自然科学版), 2023, 53(3): 325-335.
ZHAO J, ZHOU L Y, WU M Q, et al. Assistant diagnosis method of depression based on artificial intelligence [J]. Journal of Northwest University (Natural Science Edition), 2023, 53(3): 325-335.

[4] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[5] ELMAN J L. Finding structure in time[J]. Cognitive Science, 1990, 14(2): 179-211.

[6] MAJUMDER N, HAZARIKA D, GELBUKH A, et al. Multimodal sentiment analysis using hierarchical fusion with context modeling[J]. Knowledge-Based Systems, 2018, 161: 124-133.

[7] XU D L, TIAN Z H, LAI R F, et al. Deep learning based emotion analysis of microblog texts[J]. Information Fusion, 2020, 64: 1-11.

[8] 郑剑, 郑炽, 刘豪, 等. 融合局部特征与两阶段注意力权重学习的面部表情识别[J]. 计算机应用研究, 2022, 39(3): 889-894.
ZHENG J, ZHENG C, LIU H, et al. Deep convolutional neural network fusing local feature and two-stage attention weight learning for facial expression recognition[J]. Application Research of Computers, 2022, 39(3): 889-894.

[9] DUTTA K, SARMA K K. Multiple feature extraction for RNN-based Assamese speech recognition for speech to text conversion application[C]//2012 International Conference on Communications, Devices and Intelligent Systems. Kolkata: IEEE, 2012: 600-603.

[10] HOU M, TANG J J, ZHANG J H, et al. Deep multimodal multilinear fusion with high-order polynomial pooling[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019: 12156-12166.

[11] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 1103-1114.

[12] LIU Z, SHEN Y, LAKSHMINARASIMHAN V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 2247-2256.

[13] ZADEH A, LIANG P P, MAZUMDER N, et al. Memory fusion network for multi-view sequential learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 5634-5641.

[14] TSAI Y H H, BAI S J, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[J]. Proceedings of the Conference Association for Computational Linguistics Meeting, 2019, 2019: 6558-6569.

[15] YU W M, XU H, YUAN Z Q, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(12): 10790-10797.

[16] ZHAO J, ZHANG M, HE C, et al. A novel facial attractiveness evaluation system based on face shape, facial structure features and skin[J]. Cognitive Neurodynamics, 2020, 14(5): 643-656.

[17] 贾宁, 郑纯军. 融合音频、文本、表情动作的多模态情感识别[J]. 应用科学学报, 2023, 41(1): 55-70.
JIA N, ZHENG C J. Multimodal emotion recognition by fusing audio, text, and expression-action[J]. Journal of Applied Sciences, 2023, 41(1): 55-70.

[18] WANG Y Y, GU Y, YIN Y F, et al. Multimodal transformer augmented fusion for speech emotion recognition[J]. Frontiers in Neurorobotics, 2023, 17: 1181598.

[19] 焦亚萌, 周成智, 李文萍, 等. 融合多头注意力的 VGGNet 语音情感识别研究[J]. 国外电子测量技术, 2022, 41(1): 63-69.
JIAO Y M, ZHOU C Z, LI W P, et al. Research on speech emotion recognition with VGGNet incorporating multi-headed attention [J]. Foreign Electronic Measurement Technology, 2022, 41(1): 63-69.

[20] ZHANG Y M, SUN M H, REN Y, et al. Sentiment analysis of sina weibo users under the impact of super typhoon lekima using natural language processing tools: A multi-tags case study[J]. Procedia Computer Science, 2020, 174: 478-490.

[21] 刘亚姝, 侯跃然, 严寒冰. 基于异质信息网络的恶意代码检测[J]. 北京航空航天大学学报, 2022, 48(2): 258-265.
LIU Y S, HOU Y R, YAN H B. Malicious code detection based on heterogeneous information networks [J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 258-265.

[22] 邱世振, 白靖文, 张晋行, 等. 基于六轴机械臂驱动的微波球面扫描成像系统[J]. 电子测量与仪器学报, 2023, 37(4): 98-106.
QIU S Z, BAI J W, ZHANG J X, et al. Microwave spherical scanning imaging system driven by six-axis manipulator [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(4): 98-106.

[23] KU H C, DONG W. Face recognition based on MTCNN and convolutional neural network[J]. Frontiers in Signal Processing, 2020, 4(1): 37-42.

[24] 付而康, 周佳玟, 姚智, 等. 基于机器视觉识别的户外环境情绪感受测度研究[J]. 景观设计学(中英文), 2021, 9(5): 46-59.
FU E K, ZHOU J C, YAO Z, et al. A study on the measurement of emotional feelings in outdoor environments based on machine vision recognition[J]. Landscape Architecture Frontiers, 2021, 9(5): 46-59.

[25] ZHANG K, SUN M, HAN T X, et al. Residual networks of residual networks: Multilevel residual networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(6): 1303-1314.

[26] DING W, HUANG Z Y, HUANG Z K, et al. Designing efficient accelerator of depthwise separable convolutional neural network on FPGA[J]. Journal of Systems Architecture, 2019, 97(C): 278-286.

[27] 梁宏涛, 刘硕, 杜军威, 等. 深度学习应用于时序预测研究综述[J]. 计算机科学与探索, 2023, 17(6): 1285-1300.
LIANG H T, LIU S, DU J W, et al. Research review on application of deep learning to time series prediction [J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(6): 1285-1300.

[28] 焦义, 徐华兴, 毛晓波, 等. 融合多尺度特征的脑电情感识别研究[J]. 计算机工程, 2023, 49(5): 81-89.
JIAO Y, XU H X, MAO X B, et al. Research on EEG emotion recognition by fusing multi-scale features [J]. Computer Engineering, 2023, 49(5): 81-89.

[29] XU Y R, SU H, MA G J, et al. A novel dual-modal emotion recognition algorithm with fusing hybrid features of audio signal and speech context[J]. Complex & Intelligent Systems, 2023, 9(1): 951-963.

[30] 王兰馨, 王卫亚, 程鑫. 结合 Bi-LSTM-CNN 的语音文本双模态情感识别模型[J]. 计算机工程与应

用, 2022, 58(4): 192-197.
WANG L X, WANG W Y, CHENG X. Combined Bi-LSTM-CNN for speech-text bimodal emotion recognition model[J]. Computer Engineering and Applications, 2022, 58(4): 192-197.

[31] 祁宣豪, 智敏. 图像处理中注意力机制综述[J]. 计算机科学与探索,2024,18(2):345-362.
QI X H, ZHI M. A review of attention mechanisms in image processing [J]. Journal of Frontiers of Computer Science and Technology, 2024,18(2):345-362.

[32] YU W M, XU H, MENG F P, et al. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 3718-3727.

[33] ZADEH A, ZELLERS R, PINCUS E, et al. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[EB/OL]. (2016-08-12)[2023-09-25]. http://arxiv.org/abs/1606.06259.

[34] ZHANG X C, QIU X P, PANG J M, et al. Dual-axial self-attention network for text classification[J]. Science China Information Sciences, 2021, 64(12): 80-90.

[35] WANG Y, SONG W, TAO W, et al. A systematic review on affective computing: Emotion models, databases, and recent advances [J]. Information Fusion, 2022, 83/84: 19-52.

[36] RAHMAN W, HASAN M K, LEE S W, et al. Integrating multimodal information in large pretrained transformers[J]. Proceedings of the Conference Association for Computational Linguistics Meeting, 2020, 2020: 2359-2369.

[37] HAN W, CHEN H, PORIA S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis [EB/OL]. (2021-09-16)[2023-09-25]. http://arxiv.org/abs/2109.00412.

[38] HAZARIKA D, ZIMMERMANN R, PORIA S. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis[C]// Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020: 1122-1131.

[39] 程子晨, 李彦, 葛江炜, 等. 利用信息瓶颈的多模态情感分析[J]. 计算机工程与应用, 2024, 60(2):137-146.
CHENG Z C, LI Y, GE J W, et al. Multi-modal sentiment analysis using information bottleneck [J]. Computer Engineering and Applications, 2024, 60(2): 137-146.

(编　辑　李　静)