ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 6,Nov. 2022,pp. 1353－1362
DOI:10. 16337/j. 1004-9037. 2022. 06. 016

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于声学和文本特征的多模态情感识别

顾　煜[1]，金　赟[1,2]，马　勇[3]，姜芳艽[1]，俞佳佳[1]

(1. 江苏师范大学物理与电子工程学院，徐州 221116;2. 江苏师范大学科文学院，徐州 221116;3. 江苏师范大学语言科学与艺术学院,徐州 221116)

摘　要：在语音模态中,利用OpenSMILE工具箱可以从语音信号中提取浅层声学特征,通过Transformer Encoder网络从浅层声学特征中挖掘深层特征,并将深浅层特征融合,从而获取更丰富的情感表征。在文本模态中,考虑到停顿因素与情感之间的关联性,将语音和文本对齐以获得说话停顿信息,采用停顿编码的方式将停顿信息添加到转录文本中,再通过DC-BERT模型获取话语级文本特征。将获得的声学与文本特征进行融合,利用基于注意力机制的双向长短时记忆(Bi-directional long short-term memory-attention,BiLSTM-ATT)神经网络进行情感分类。最后,本文对比了3种不同注意力机制融入BiLSTM网络后对情感识别的影响,即局部注意力、自注意力和多头自注意力,发现局部注意力的效果最优。实验表明,本文提出的方法在IEMOCAP数据集上的4类情感分类的加权准确率达到了78.7%,优于基线系统。

关键词：多模态情感识别;深浅特征融合;DC-BERT模型;注意机制;停顿编码

中图分类号：TN912.34　　**文献标志码**：A

Multimodal Emotion Recognition Based on Acoustic and Lexcial Features

GU Yu[1], JIN Yun[1,2], MA Yong[3], JIANG Fangjiao[1], YU Jiajia[1]

(1. School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou 221116, China;2. Kewen College, Jiangsu Normal University, Xuzhou 221116, China;3. School of Linguistic Sciences and Arts, Jiangsu Normal University, Xuzhou 221116, China)

Abstract: In the speech mode, the OpenSMILE toolbox is used to extract low-level acoustic features from the speech signal. Transformer Encoder is richer to excavate deep features from low level acoustic features and fuses them so as to obtain more useful emotional representation. In the text mode, considering the association between pause and emotion, the speech and text are aligned to obtain the pause information and the pause information is added to the transcript text by pause encoding. The utterance-level lexical features are obtained by the improved DC-BERT model. Then, acoustic features and lexical features are fused and the bi-directional long short-term memory based on attention neural network (BiLSTM-ATT) is used for emotion classification. Finally, this paper compares the effects of three different attention mechanisms integrated into BiLSTM on emotion recognition (local attention, self-attention and multi-headed attention), and local attention is found to be the most effective. In the experiments on IEMOCAP dataset, the method proposed in this paper achieves 78.7% in weighted accuracy for four emotion categories, which

基金项目：国家自然科学基金青年基金项目(52005267);江苏省高校自然科学基金(18KJB510013,17KJB510018);校创新项目(2021XKT1250)。
收稿日期：2022-01-04;**修订日期**：2022-11-07

is better than the baseline system.

Key words: multimodal emotion recognition; deep and shallow feature fusion; DC-BERT model; attention mechanism; pause encoding

引　言

尽管语音情感识别(Speech emotion recognition, SER)和自然语言处理(Natural language processing, NLP)已经取得了很大的进展,但人类仍然无法与机器进行自然地交流。因此,建立一套能够在人机交互中检测情感的系统至关重要。但由于人类情感的多变性和复杂性,这仍然是一项具有挑战性的任务。传统的情感识别主要针对于单个模态,如文本、语音和图像等,在识别性能上存在一定的局限性[1]。如在早期的语音情感识别任务中,研究人员主要利用的是语音中的声学特征和一些相关的韵律学特征,往往忽视了语音中所包含的具体语义信息(文本信息)。但在日常会话和社交媒体中,声音往往是对一段文本内容的复述,二者密切相关。考虑到语音和文本模态之间的同一性、互补性和强相关联性,不少研究人员从单模态转向了多模态的情感识别研究。其中,融合语音和文本这两种不同模态信息来进行情感识别也成为了热点研究方向。实验表明,与单个模态相比,同时考虑多种模态信息可以更加准确地捕捉情感[2]。在多模态融合方面,主要采用3种融合策略:特征层融合、决策层融合以及混合融合。Kim 等[1]利用深度神经网络(Deep neural network, DNN)提取话语级声学瓶颈特征和以分布表征和情感词汇为基础的文本特征,将这些声学和文本特征进行早期融合后输入至另一个DNN网络进行分类,并取得了良好的效果。文献[3]使用OpenSMILE工具箱提取的特征和原始的倒谱特征作为语音的话语级声学特征,而在文本特征方面利用N-gram语言模型进行捕获,并将两个模态先分别训练识别,再进行决策融合。也有研究人员另辟蹊径,将侧重点放在两个模态信息融合上,文献[4]提出一种新颖的多模态交叉的自注意力网络(Multimodal cross and self-attention network, MCSAN),该网络主要利用交叉注意力机制来引导一个模态关注另一个模态,从而实现特征的更新。

随着技术的发展,许多研究机构也在不断探索新的语言模型。2019谷歌研究所[5]首次提出一种新型语言表征模型BERT,该模型可以生成深层次的语言双向表征,对自然语言处理各项任务的结果都有很大的提升。文献[6]利用BERT获得上下文词嵌入来表征转录文本中所包含的信息,但没有考虑到因BERT复杂网络结构与情感语料库数据量不足而不匹配的问题。BERT虽然可以用来生成文本信息的表征,但无法弥补转录文本自身忽视一些潜在情感信息的不足。在转录文本时并不会体现出说话过程中的停顿信息。文献[7]调研了说话停顿信息与情感之间的联系,发现与快乐、积极相比,在悲伤、害怕的情感状态下,沉默停顿的平均时长占整段语音的比例增加了,且注意到处于不同情感状态时,说话停顿的频率、持续时间以及停顿发生的位置也会有所区别。另一方面,基于注意力机制的深度网络在解码阶段显示了优越的性能,在自然语言处理和语音识别领域中得到了广泛的应用。而在语音情感识别中,由于情感特征在语句中分布并不均匀,因此不少研究人员在情感识别任务中增加了注意力机制,如文献[8-10],使得网络对包含情感信息较多的部分具有指导性机制,重点突出局部最具情感的信息。

针对提高情感识别性能,本文提出了一种基于声学和文本特征的多模态识别方法。在文本模态上,原始的转录文本缺失了情感相关的说话人停顿信息,因而利用语音和转录文本的强制对齐,将停顿信息编码后添加至文本。为解决传统BERT复杂的网络结构与情感数据量少的不匹配问题,将文本输入分层密集连接BERT模型(Densely connected bi-directional encoder representation from transformers, DC-BERT)提取话语级文本特征。在语音模态上,利用OpenSMILE提取语音情感的浅层特征,并与Transformer Encoder学习浅层特征后得到的深层特征进行融合生成多层次的声学特征。本文专注于特

征提取的质量与有效性,利用早期特征层融合技术来补充声学和文本特征之间相互缺失的信息,并采用了基于注意力机制的双向长短时记忆神经网络(BiLSTM-ATT)作为分类器。其中BiLSTM网络的优势是能够充分利用先验知识,获取有效的上下文信息,而注意力机制有助于抽取特征中突显情感信息的部分,避免信息冗余。最后,本文对比了目前使用较为广泛的3种注意力机制,即局部注意力机制[11]、自注意力机制[12]、多头自注意力机制[12]对情感信息的捕获能力。最终,本文方法在IEMOCAP数据集[13]上4类情感分类中加权准确率达到78.7%。与基线系统相比,展示了良好的性能。

1　多模态情感识别模型

本节主要描述了系统的整体框架及其所涉及的技术。该系统由声学特征提取模块、文本特征提取模块和BiLSTM-ATT网络模型组成,系统整体框架如图1所示。

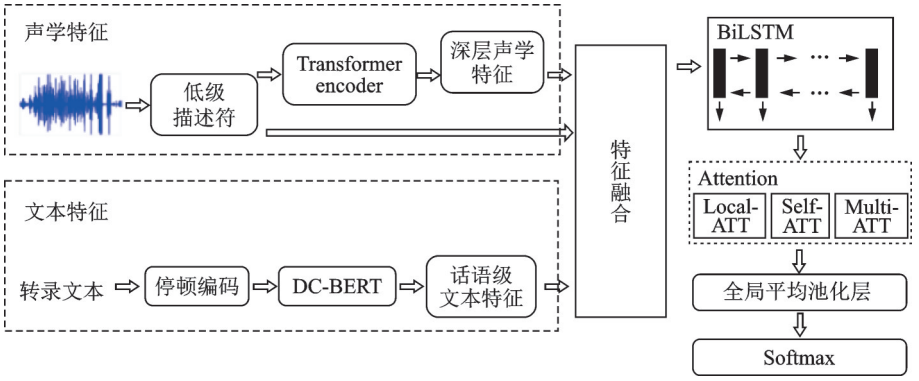

图1　多模态情感识别模型的系统框架

Fig.1　System framework of the proposed model for multimodal emotion recognition

1.1　声学特征提取

本文使用OpenSMILE工具箱[14]中的Emobase特征集提取了988维浅层声学特征。它们由低级描述符(Low-level descriptors, LLDs)组成,如强度、响度、梅尔频率倒谱系数(Mel-frequency cepstral coefficients,MFCC)、音调以及它们在话语级上每个短帧的统计值,如最大值、最小值、平均值和标准偏差等。但是,低级描述符只包含全局浅层信息,仅仅使用其表达情感是不够的,需要从中挖掘出更细节的情感描述特征。

受自然语言处理领域Transformer模型[12]的启发,采用Transformer Encoder网络结构对低级描述符进行2次学习提取深层特征。Transformer模型最早用于机器翻译任务,可以很好地解决序列到序列(Sequence to sequence,Seq2seq)的问题,从而广泛应用于自然语言处理领域。该模型主要包括编码器、解码器。其中,在Seq2seq模型中,编码器主要将输入单词序列映射为高维的连续表征序列,而解码器则是在给定高维连续表征序列的情况下,生成一个单词序列作为输出。

但在语音情感分类任务中,一句话对应一个情感标签,且数据量不如机器翻译任务,因而本文仅采用Transformer的编码器结构,其强大的特征学习能力受益于内部的自注意机制,可以有效地从浅层声学特征中挖掘到与情感状态高度相关的深层表征。

1.2　文本特征提取

考虑到说话停顿对情感表达的影响,本文通过宾夕法尼亚大学语音标签强制对齐工具(Penn phonetics lab forced aligner, P2FA)对预处理后的转录文本和音频进行强制对齐,从而确定停顿的位置和持

续时间。文本预处理包含删除转录文本的标点符号,以及单词统一转换为小写。根据文献[15]的经验,将停顿时长分为6个区间:0.05～0.1 s,0.1～0.3 s,0.3～0.6 s,0.6～1.0 s,1.0～2.0 s和大于2.0 s。对这6个区间分别进行编码:“..”“…”“….”“……”“……”“…….”;最后添加“.”在每个说话人的句尾作为结束的标志。该流程如图2所示。

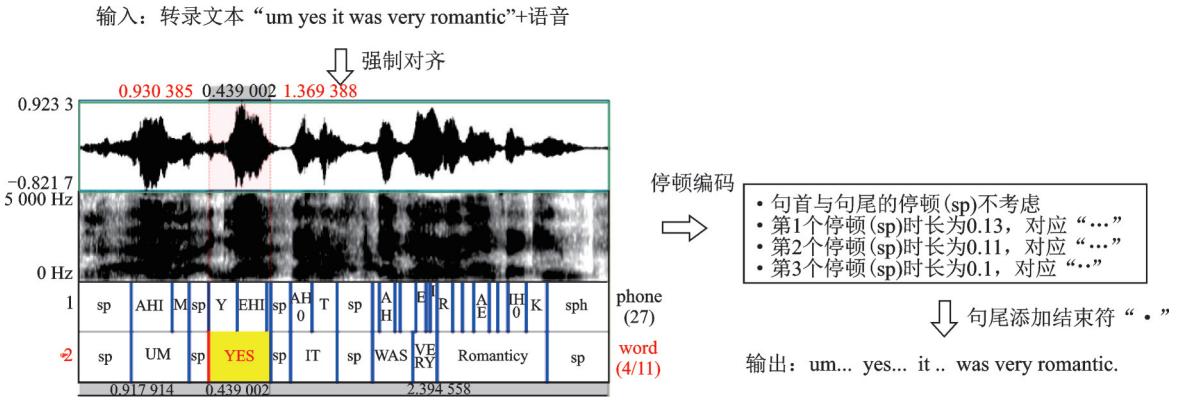

图2 停顿编码流程图

Fig.2 Procedure for pause encoding

针对上文所述,BERT复杂网络结构与数据量不匹配,本文采用了一种改进的BERT模型作为文本特征提取器,即分层密集连接BERT模型,它保留了Transformer中每个多头自注意层[16]内部的残差连接,在层与层之间新增了密集连接,即每一个多头自注意力层的输入额外增加了前两层的特征信息,目的是加快模型的收敛速度,使网络的损失函数更加平滑,而每一层提取的特征也可以在不同的注意层之间被重复使用,提高了特征的利用率。DC-BERT网络结构如图3所示,其中Trm为Transformer。假设给定一个输入特征序列X,X中元素如式(1)所示。

$$x_i = H(x_{i-1}) + \alpha x_{i-1} + \beta x_{i-2} \tag{1}$$

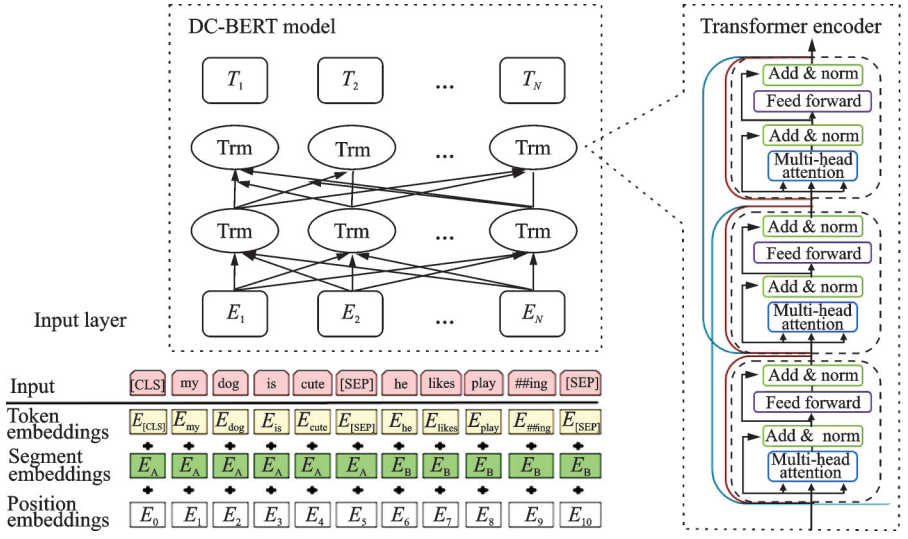

图3 DC-BERT模型结构

Fig.3 DC-BERT model structure

式中:x_i为输入特征序列X的第i个元素;H为非线性函数;α和β为保留前两层信息的权重系数,使得每一层都能得到前两层处理的结果,却又不占主导地位。

DC-BERT模型由12层Transformer组成,每一层的输出理论上都可以作为话语级的文本特征。根据之前的实验经验,本文选择DC-BERT倒数第2层的768维输出序列作为话语级文本特征。

1.3　模型结构

LSTM网络可以解决长距离信息依赖问题,以及在训练过程中避免梯度消失或爆炸。BiLSTM网络是由前向LSTM和反向LSTM组成,相较于单向的LSTM网络,BiLSTM网络能够充分利用先验知识,更好地捕捉和考虑上下文信息。

本文在BiLSTM网络中引入注意力机制来关注话语中包含强烈情感特征的特定部分,即BiLSTM-ATT模型,同时对比了3种注意力机制,即局部注意力机制[11]、自注意力机制[12]、多头自注意力机制[12]。

1.3.1　局部注意力机制

为了解决计算开销问题,本文采用了一种局部注意力机制,该机制只关注一部分编码隐藏层。局部注意力首先在时间t上,为当前节点生成一个对齐位置p_t,然后选择性地设置1个固定大小为$2D+1$的上下文窗口。

$$c_t=\sum_{i=P_t-D}^{P_t+D}\alpha_{t,i}\bar{h}_i \tag{2}$$

式中:D根据经验选择;P_t为窗口中心,由当前隐藏状态的h_t决定,是一个实数;编码器的全部隐藏状态为$\bar{h}_i$;对齐权重的计算过程和传统attention相似,即

$$\alpha_{t,i}=\frac{\exp\left(\text{score}\left(h_t,\bar{h}_i\right)\right)}{\sum_{i}^{T_x}\exp\left(\text{score}\left(h_t,h_i\right)\right)}\exp\left(-\frac{i-P_t^2}{2\sigma^2}\right) \tag{3}$$

式中标准偏差σ根据经验设定。

1.3.2　自注意力机制

自注意力机制利用了输入特征序列元素之间的加权相关性。具体来说,输入序列的每个元素都可以通过一个线性函数投影成3种不同的表示形式:查询(query)、键(key)、值(value)[17],即

$$\boldsymbol{q}_i=\boldsymbol{w}_q^{\mathrm{T}}\boldsymbol{u}_i,\ \boldsymbol{v}_i=\boldsymbol{w}_v^{\mathrm{T}}\boldsymbol{u}_i,\boldsymbol{k}_i=\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{u}_i \tag{4}$$

式中:$\boldsymbol{w}_q$、$\boldsymbol{w}_v$、$\boldsymbol{w}_k$分别为查询、键、值的权重矩阵;$\boldsymbol{u}_i$为输入的第i个词向量。

最终注意矩阵为

$$z=\text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{5}$$

式中:$\boldsymbol{Q}$为查询矩阵;$\boldsymbol{K}$为键矩阵;V为句子的值矩阵;d_k为比例因子。

1.3.3　多头自注意力机制

为了扩展模型对不同位置的关注能力,本文在自注意力机制的基础上对比了多头自注意力机制对语音情感识别任务的影响。多头是指输入特征序列的每个变量(query、key和value)的投影数不止一组。也就是说,在参数不共享的前提下,将$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$通过参数矩阵映射后,做单层的自注意力,然后将自注意力层层叠加。多头自注意力计算公式为

$$\text{Multihead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\text{head}_1,\cdots,\text{head}_n) \tag{6}$$

$$\text{head}_i=\text{attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}\right) \tag{7}$$

式中$\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$、$\boldsymbol{W}_i^{\boldsymbol{V}}$为第$i$个权重矩阵。

2 实验验证

2.1 数据准备

为了验证所提方法的有效性，本文在IEMOCAP情感数据集[13]上进行了多组实验。该数据集包含5组二元互动的会话，共包括12 h的视听数据（音频、转录文本、视频和面部动作捕捉）。本文仅使用了音频和转录文本，一些多模态情感识别利用自动语音识别（Automatic speech recognition，ASR）系统将语音翻译成文本，本文并没有针对该语音训练一个专门的ASR系统，而是直接使用IEMOCAP数据库所包含的转录文本，减少了因ASR系统识别错误带来的消极影响，Li等做了相应的实验验证了直接使用转录文本能够提升情感识别的准确率[18]。

IEMOCAP数据库共有10类情感（愤怒、高兴、悲伤、中立、沮丧、兴奋、恐惧、惊讶、厌恶、其他），每句话都由3位注释员进行情感判定。为了与先前的研究结果具有对比性，选取了4种情感进行分类，其中将高兴与兴奋划分为一类，以平衡数据在不同类别之间的分布。最终实验数据共计5 531句话语，类别占比分别为：愤怒19.9%，快乐29.5%，中立30.8%，悲伤19.5%。

2.2 参数设置

本文采用特定人的十折交叉验证作为最终实验结果。模型的参数主要根据交叉验证的结果进行调整。为了增加模型的泛化能力，在交叉验证中，把训练数据分成10份，其中训练集9份和验证集1份，通过十折的交叉验证求取平均值来获得模型的参数。此外，设置了Dropout防止模型过拟合，在全连接层加入Dropout可以随机地将某些输出置0，相当于增加了噪声，从而防止模型过拟合。实验结果也表明，本文提出的方法具有较好的泛化能力。最终模型的参数为：BiLSTM网络的神经元数设置为200（100个正向节点和100个反向节点），训练批次大小设置为64，迭代次数设置为20，Dropout设置为0.5；采用IEMOCAP数据集最常用的评价指标：加权准确率WA和未加权准确率UA来评估模型性能的优劣。WA是整个测试数据的总体准确率，UA是每个情感类别的平均准确率。采用交叉熵损失函数作为模型的损失函数，其公式如下

$$C=-\frac{1}{n}\sum_{x}\left[y\ln a+(1-y)\ln(1-a)\right] \tag{8}$$

式中：n为样本个数；y为期望输出；a为神经元实际输出。

2.3 实验结果

为了分析验证所提的多模态融合方法以及该模型的优越性，本文分3个步骤进行验证分析。首先针对单语音模态验证深浅特征融合的有效性，本文做了如下几组对比实验：(1) IS09+ BiLSTM：使用384维的IS09特征集作为声学特征，并采用BiLSTM网络进行分类；(2) emobase+BiLSTM：使用988维的emobase特征集作为声学特征，并采用BiLSTM网络进行分类；(3) emobase+deep features (proposed)：使用988维的emobase特征集作为浅层特征，将其输入Transformer Encoder提取深层特征，再将深浅特征融合，送入BiLSTM网络进行分类。对比实验结果如表1所示。由表1可以看出，在BiLSTM网络参数与上述设置一致的前提下，对于单语音模态而言，Emobase+deep features (proposed)的WA和UA分别可以达到67.55%和66.39%。深浅融合特征明显优于仅有低级描述

表1 仅语音模态的实验对比结果

Table 1 Experimental comparison results for only speech modal

模型	WA/%	UA/%
IS09+BiLSTM	65.66	64.67
emobase+BiLSTM	66.81	65.97
emobase+deep features (proposed)	67.55	66.39

符的浅层特征。同时验证了利用Transformer Encoder是可以从浅层特征中提取更显著的局部情感信息。

其次,针对单文本模态,本文做了如下几组对比实验:(1) Word2vec+ BiLSTM:使用传统词嵌入模型word2vec提取文本特征,并采用BiLSTM网络进行分类;(2) BERT+BiLSTM:直接采用转录文本,将其输入BERT预训练模型后,提取倒数第2层的768维输出序列作为文本特征,并采用BiLSTM网络进行分类;(3) DC-BERT+BiLSTM:直接采用转录文本,将其输入DC-BERT预训练模型后,提取倒数第2层的768维输出序列作为文本特征,并采用BiLSTM网络进行分类;(4) Pause+BERT+BiLSTM:使用经过停顿编码后的转录文本,将其输入BERT预训练模型后,提取倒数第2层的768维输出序列作为文本特征,并采用BiLSTM网络进行分类;(5) Pause+DC-BERT+BiLSTM:使用经过停顿编码后的转录文本,将其输入DC-BERT预训练模型后,提取倒数第2层的768维输出序列作为文本特征,并采用BiLSTM网络进行分类。对比结果如表2所示。由表2可知,在BiLSTM网络参数与上述设置一致的前提下,对于单文本模态而言,DC-BERT+BiLSTM的WA可以达到69.01%,UA达到了68.93%;而BERT+BiLSTM的WA为68.78%,UA为68.69%, Word2vec+ BiLSTM的WA仅为65.21%。由此DC-BERT的性能要优于BERT模型和word2vec。除此之外,不难发现相较于直接使用转录文本,采用经过停顿编码后的文本新增了语义与停顿信息的联结,在一定程度上是对语义信息无声的补充,可以有效地提高情感识别的准确性,而DC-BERT与停顿编码的组合也进一步提升了识别的准确性,其中WA和UA分别达到了70.13%和70.34%。

表2　仅文本模态的实验对比结果

Table 2　Experimental comparison results for only text (transcribed text) modal

模型	WA/%	UA/%
Word2vec+ BiLSTM	65.21	—
BERT+BiLSTM	68.78	68.69
DC-BERT+BiLSTM	69.01	68.93
Pause+ BERT+BiLSTM	69.66	69.64
Pause+DC-BERT+BiLSTM	70.13	70.34

最后将语音和文本模态融合的结果(本文采用基于特征层融合的策略)与最近的一些实验结果比较,其中这些引用皆使用了相同的情感语料库,同时在此基础上,本实验对比了3种不同注意力机制,如表3所示。

表3　多模态模型在IEMOCAP数据集上的对比结果

Table 3　Comparison results on IEMOCAP dataset using multimodal models

模型	WA/%	UA/%
Concat (Yoon et al., 2018)[19]	71.80	—
Concat(Gu et al., 2018)[20]	72.70	—
Concat(Xu et al., 2019)[21]	72.50	70.90
BN+v_{w2v}^{utt} + v_{w2v}^{utt} (Kim et al. ,2019)[1]	—	75.50
Concat(Pepino et al., 2020)[6]	—	65.10
Concat(Patamia et al., 2021)[2]	70.18	—
LLDs+word2vec+BiLSTM	71.10	—
BiLSTM-LocalAtt(proposed)	78.70	79.51
BiLSTM-SelfAtt(proposed)	77.99	78.77
BiLSTM-MultiAtt(proposed)	76.39	75.97

(1) Concat (Yoon et al., 2018)[19]:提出一种多模态双循环编码器模型,使用双向RNN对语音和文本序列进行编码,再使用前馈神经网络将编码序列组合从而完成情感类别预测,最终在IEMOCAP数据集上获得了71.8%的识别率。

(2) Concat(Gu et al., 2018)[20]:提出一种多模态分层注意力结构(Multimodal hierarchical attention structure),该结构主要包括文本注意力模块、语音注意力模块和融合模块,在预处理阶段,将文本和语音进行强制对齐。然后,文本注意模块和语音注意模块从相应的输入中提取特征,并通过融合后的特征进行情感预测,最终在IEMOCAP数据集上获得了72.7%的识别率。

(3) Concat(Xu et al., 2019)[21]:使用注意力机制来学习语音帧和文本词之间的对齐,再将对齐的

多模态特征输入至序列模型中进行情感识别，最终在IEMOCAP数据集上的WA和UA分别为72.50%和70.90%。

(4) BN+v_{w2v}^{utt}+v_{w2v}^{utt} (Kim et al. ,2019)[1]：利用DNN提取话语级声学瓶颈特征，以及以分布表征和情感词汇为基础的文本特征，将声学和文本特征进行早期融合，然后输入至另一个DNN分类，最终在IEMOCAP数据集上的UA为75.5%。

(5) Concat(Pepino et al., 2020)[6]：通过BERT获得的上下文词嵌入作为转录文本的特征，利用OpenSMILE工具包提取36维的声学特征，采用模型融合的方式将两个模态的信息整合，最终在IEMOCAP数据集上的UA为65.10%。

(6) Concat(Patamia et al., 2021)[2]：利用librosa获取34维声学特征，通过BERT获得的上下文词嵌入作为文本的特征，并将两个模态的特征输入神经网络获取更深层的特征，采用特征层融合的方式整合两个模态的信息，最终在IEMOCAP数据集上的WA为70.18%。

(7) LLDs+word2vec+BiLSTM：将语音模态的988维LLDs和文本模态中使用word2vec提取的词嵌入进行简单的特征拼接，再送入与上述参数设置一致的BiLSTM网络中进行情感识别，最终WA为71.10%。

本实验在多模态的基础上，将注意力机制引入BiLSTM来引导网络关注特征中情感浓烈的地方，并对比了3种不同注意机制(LoaclAtt、SelfAtt和MultiAtt)，其WA分别是78.70%、77.99%和76.39%，UA为79.51%、78.77%和75.97%。显然，与其他先进的方法进行比较，本文所提模型的性能优于上述模型。本模型相较于上述模型识别效果有所提高主要在于两个模态特征提取的创新，在语音模态，本文对浅层声学特征进行2次学习，从浅层声学特征中挖掘深层声学特征，并将深浅层特征融合，得到的新特征包含更丰富的信息，可以多层次的去识别情感；在文本模态，本文将语音中的停顿时长信息以编码的形式添加至转录文本中，这是把语音模态中的特定信息与文本模态信息融合，使得文本所带的语义信息中加入了停顿信息，让文本内容变得更加丰富。最终将两个模态的特征进行融合，并采用注意力机制去关注情感信息突出的部分，获得了较好的实验结果。

3种不同注意力机制下的分类混淆矩阵如图4所示，发现基于局部注意力机制的BiLSTM网络要比基于自注意力机制或多头自注意力机制的BiLSTM网络表现更好。可以看出，除中立类别外，其他类的识别率几乎都在75%以上。文献[22]曾表述高兴是一种正效价和唤醒值情感，仅靠浅层特征信息是无法很好预测的。在本文实验中，高兴的识别率在80%左右，远高于文献[22]，证明了利用Transformer从浅层特征中学习深层特征的方法是有效的。

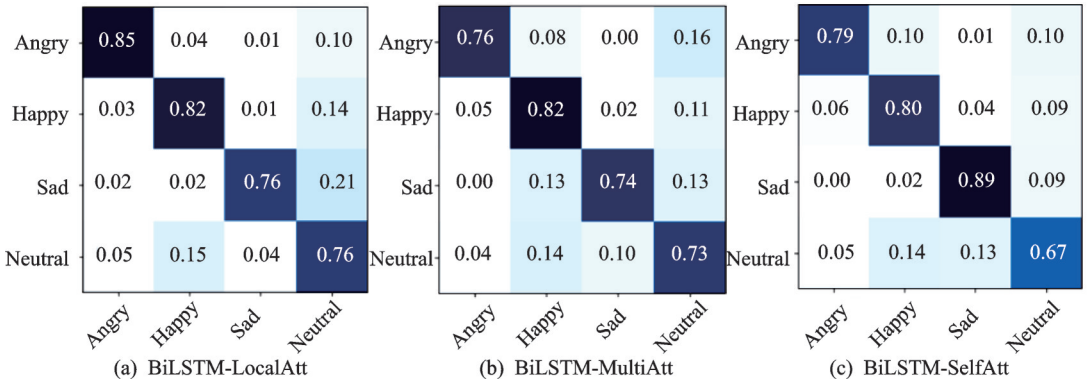

图4 在3种不同注意力机制下4类情感识别结果的混淆矩阵

Fig.4 Confusion matrices of four categories of emotion recognition results under three different attention mechanisms

为了进一步验证BiLSTM-LocalAtt模型在语音情感识别方面的优势,本文在IEMOCAP数据库上进行了参数量(Params)和计算复杂度(FLOPs)对比实验。如表4所示,BiLSTM-MultiAtt模型的网络参数量最多,计算复杂度最大,但其识别准确率最低,可见对于小数据量的情感识别任务,较为庞大的网络结构未必能取得预期效果。BiLSTM-LocalAtt模型和BiLSTM-SelfAtt模型的网络参数量和计算复杂度是一样的,但局部注意力机制的效果要优于自注意力机制,可见对于整句语音而言,情感并不是平均分布的,而是相对集中在某几个地方,因此局部注意力机制会更适合情感识别任务。

表4　网络复杂度对比实验结果

Table 4　Comparison of experimental results for network complexity

模型	Params/10^6	FLOPs/10^6
BiLSTM-LocalAtt	1.713	8.360
BiLSTM-SelfAtt	1.713	8.360
BiLSTM-MultiAtt	1.861	8.660

3　结束语

本文提出了一种有效的从语音和转录文本中识别情感的方法。通过Transformer Encoder模型从OpenSMILE工具箱提取的浅层特征中2次学习获得深层特征,再把深浅层特征融合以补全信息的完整性。利用两个模态的对齐获取语音中的停顿信息,并以停顿编码的方式将说话停顿添加到转录文本中,补充了文本模态除语义信息外的其他从属信息,使得文本信息更加多元化。最终结果表明,与直接使用转录文本相比,具有停顿信息的转录文本可以提高情感识别的准确性;再使用DC-BERT模型提取的话语级文本特征,以弥补因BERT复杂网络结构与数据量不足而不匹配的问题。本文将两种改进后的模态特征融合并输入到BiLSTM-ATT网络中进行情感分类。实验结果表明,该方法在情感识别效果上优于其他方法。同时本文对比了3种注意力机制在情感识别任务中的影响,发现在本实验数据情况下,局部注意力机制的效果要优于另外两个注意力机制。

参考文献:

[1] KIM E, SHIN J W. DNN-based emotion recognition based on bottleneck acoustic features and lexical features[C]//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 6720-6724.

[2] PATAMIA R A, JIN W, ACHEAMPONG K N, et al. Transformer based multimodal speech emotion recognition with improved neural networks[C]//Proceedings of 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML). Chengdu, China: IEEE, 2021: 195-203.

[3] LI B, DIMITRIADIS D, STOLCKE A. Acoustic and lexical sentiment analysis for customer service calls[C]//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 5876-5880.

[4] SUN Licai, LIU Bin, TAO Jianhua, et al. Multimodal cross and self-attention network for speech emotion recognition[C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021: 4275-4279.

[5] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.

[6] PEPINO L, RIERA P, FERRER L, et al. Fusion approaches for emotion recognition from speech using acoustic and text-based features[C]//Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 6484-6488.

[7] ESZTER T, CSABA P. Ascribing emotions depending on pause length in native and foreign language speech[J]. Speech Communication, 2014, 56(1): 35-48.

[8] LI Yuanchao, ZHAO Tianyu, TATSUYA K. Improved end-to-end speech emotion recognition using self attention mechanism

and multitask learning[C]//Proceedings of Interspeech 2019. Graz, Austria: ISCA, 2019: 2803-2807.

[9] CHEN Mingyi, HE Xuanji, YANG Jing, et al. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition[J]. IEEE Signal Processing Letters, 2018, 25(10): 1440-1444.

[10] YEH S, LIN Y, LEE C. An interaction-aware attention network for speech emotion recognition in spoken dialogs[C]//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 6685-6689.

[11] LUONG M, PHAM H, MANNING C D, et al. Effective approaches to attention-based neural machine translation[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP). Lisbon, Portugal: Association for Computational Linguistics, 2015: 1412-1421.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Advances in Deural Information Processing Systems. Doha, Qatar: Association for Computational Linguistics, 2017: 5998-6008.

[13] 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.

[14] EYBEN F, WÖLLMER M, SCHULLER B. Opensmile: The munich versatile and fast open-source audio feature extractor [C]//Proceedings of International Conference on Multimedia. Firenze, Italy: ACM, 2010: 1459-1462.

[15] YUAN Jiahong, CAI Xingyu, CHURCH K. Pause-encoded language models for recognition of alzheimer's disease and emotion[C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021: 7293-7297.

[16] 王楠禔. 基于BERT改进的文本表示模型研究[D]. 重庆:西南大学,2019.
WANG Nanzhi.Study on text representation model based on improved BERT[D].Chongqing: Southwest University, 2019.

[17] TARANTINO L, GARNER P N, LAZARIDIS A, et al. Self-attention for speech emotion recognition[C]//Proceedings of Interspeech 2019. Graz, Austria: ISCA, 2019: 2578-2582.

[18] LI B, DIMITRIADIS D, STOLCKE A. Acoustic and lexical sentiment analysis for customer service calls[C]//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 5876-5880.

[19] YOON S, BYUN S, JUNG K. Multimodal speech emotion recognition using audio and text[C]//Proceedings of the 2018 IEEE Spoken Language Technology (SLT). Greece: IEEE,2018: 112-118.

[20] GU Yue, YANG Kangning, FU Shiyu, et al. Multimodal affective analysis using hierarchical attention strategy with word-level alignment[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(ACL). Melbourne, Australia: Association for Computational Linguistics, 2018: 2225-2235.

[21] XU Haiyang, ZHANG Hui, HAN Kun, et al. Learning alignment for multimodal emotion recognition from speech[C]//Proceedings of Interspeech 2019. Graz, Austria: ISCA, 2019: 3569-3573.

[22] LIU J, LIU Z, WANG L, et al. Speech emotion recognition with local-global aware deep representation learning[C]//Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 7174-7178.

作者简介:

顾煜(1997-),男,硕士研究生,研究方向:语音信号处理,E-mail:guyuluck666@163.com。

金赟(1979-),**通信作者**,男,副教授,研究方向:语音信号处理、人工智能、机器学习等,E-mail:jiny@jsnu.edu.cn。

马勇(1977-),男,讲师,博士研究生,研究方向:语音与音频信号处理,E-mail:may@jsnu.edu.cn。

姜芳艽(1971-),女,副教授,硕士生导师,研究方向:Web数据管理、数据挖掘和大数据融合,E-mail:jiangfj@jsnu.edu.cn。

俞佳佳(1997-),女,硕士研究生,研究方向:语音情感识别,E-mail:1137218386@qq.com。

(编辑:刘彦东)