DOI:10.11918/j.issn.0367-6234.201905193

嵌入注意力机制并结合层级上下文的语音情感识别

程艳芬[1],陈垚鑫[2],陈逸灵[1],杨 益[1]

(1.武汉理工大学 计算机科学与技术学院,武汉 430063;2.湖北工业大学 计算机学院,武汉 430068)

摘 要: 由于情感语料问题、情感与声学特征之间关联问题、语音情感识别建模问题等因素,语音情感识别一直充满挑战性.针对传统基于上下文的语音情感识别系统仅局限于特征层造成标签层上下文细节丢失以及两层级差异性被忽略的缺陷,本文提出嵌入注意力机制并结合层级上下文学习的双向长短时记忆(BLSTM)网络模型.模型分3个阶段完成语音情感识别任务,第1阶段提取情感语音特征全集后采用SVM-RFE特征排序算法降维得到最优特征子集,并对其进行注意力加权;第2阶段将加权后的特征子集输入BLSTM网络学习特征层上下文获得最初情感预测结果;第3阶段利用情感标签值对另一独立BLSTM网络训练学习标签层上下文信息并据此在第2阶段输出结果基础上完成最终预测.模型嵌入注意力机制使其自动学习调整对输入特征子集的关注度,引入标签层上下文使其联合特征层上下文实现层级上下文信息融合提高鲁棒性,提升了模型对情感语音的建模能力,在SEMAINE和RECOLA数据集上实验结果表明:与基线模型相比RMSE和CCC均得到较好改善.

关键词: 语音情感识别;注意力机制;上下文;双向长短时记忆网络

中图分类号: TN912.34 **文献标志码:** A **文章编号:** 0367-6234(2019)11-0100-08

Speech emotion recognition with embedded attention mechanism and hierarchical context

CHENG Yanfen[1], CHEN Yaoxin[2], CHEN Yiling[1], YANG Yi[1]

(1. School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China;
2. School of Computer, Hubei University of Technology, Wuhan 430068, China)

Abstract: A challenging task remains with regarding to speech emotion recognition due to issues such as emotional corpus problems, association between emotion and acoustic features, and speech emotion recognition modeling. Conventional context-based speech emotion recognition system risks of losing the context details of the label layer and neglecting the difference of the two-level due to solely limited to the feature layer. This paper proposed a Bidirectional Long Short-Term Memory (BLSTM) network with embedded attention mechanism combined with hierarchical context learning model. The model completed the speech emotion recognition task in three phases. The first phase extracted the feature set from the emotional speech, then used the SVM-RFE feature-sorting algorithm to reduce the feature in order to obtain the optimal feature subset and assigned attention weights. The second phase, the weighted feature subset was input into the BLSTM network learning feature layer context to obtain the initial emotional prediction result. The third phase used the emotional value to train another independent BLSTM network for learning label layer context information. According to the information, the final prediction was completed based on the output result of the second phase. The model embedded the attention mechanism to automatically learn to adjust the attention to the input feature subset, introduced the label layer context to associate the feature layer context so as to achieve the hierarchical context information fusion and improve the robustness, and improved the model's ability to model the emotional speech. The experimental results on the SEMAINE and RECOLA datasets showed that both RMSE and CCC were significantly improved than the baseline model.

Keywords: speech emotion recognition; attention mechanism; context; BLSTM

为使计算机更顺畅和谐地与人交流,作为情感计算重要分支的语音情感识别正被作为人工智能热门研究课题之一[1].目前常用描述情感状态的方法分为离散情感描述模型和维度情感描述模型,前者将情感状态划分为人们日常交互中常见的几种基本情感,研究者多数采用美国心理学家Ekman提出的6类(高兴、悲伤、生气、厌恶、恐惧、惊讶)情感划分方法[2];后者将情感描述为连续情感空间中的不同坐标点,情感空间中不同维度代表不同心理学属性.由于人们日常生活中的情感体验丰富多彩,离散模型包含的情感种类较少,难以传达沮丧、低沉、欣喜若狂等非基本情感状态,因此研究人员开始基于维

收稿日期: 2019-05-27
基金项目: 国家自然科学基金(51179146)
作者简介: 程艳芬(1970—),女,博士,副教授
通信作者: 陈垚鑫,littlebabyball@163.com

度模型展开情感计算相关工作，本文选择当前常用二维情感空间 Valence-Arousal（效价度–唤醒度）进行维度语音情感识别研究.

针对孤立情感语句的语音情感识别已取得不错成果，但日常生活中人们说话时前后语句存在一定联系，情感表达是连续流畅的，因此基于上下文的语音情感识别方法受到越来越多研究者的关注[3]，分析当前情感语句的前后一句或若干句情感信息能有效提升连续语句的语音情感识别效果. 当前基于上下文的语音情感识别方法能够学习情感语音序列的时间上下文信息[4]，但一般简单地将上下文多帧特征直接作为输入而忽略了每帧各自的特点，并且多局限于特征层上下文[5]. Sariyanidi 等[6]通过在时间窗口内进行特征提取利用特征序列的上下文信息，讨论了多种采用这种方式进行特征处理以学习特征层上下文信息的识别方法. Shao L 等[7]发现从语音数据中提取的语音情感特征和语音所蕴含的情感状态有较大差异. 例如在自然型语音数据库中，语音情感特征可能变化十分迅速，而说话人的情感状态却变化缓慢[8]，这种差异性可能来自于录制情感语音时说话人受到了与情感无关因素的影响，从而导致语音内容变化剧烈，在语音情感识别中必须考虑这种差异性. 受上述工作启发，本文首先计算各帧特征注意力权值用于对特征序列加权，然后将加权后的特征序列输入 BLSTM 网络进行特征层上下文学习，最后在上述输出基础上增加标签层上下文学习，综合考虑两层级间差异性加强模型鲁棒性，进一步提升语音情感识别准确率.

1　BLSTM 语音情感识别模型

随着深度学习的不断发展，神经网络结构逐渐变得复杂，与过去简单前馈神经网络相比，循环神经网络的关键在于其隐层之间既有前馈连接，又有内部反馈连接. BLSTM 网络结合了长短时记忆（LSTM）网络和双向循环神经网络（BRNN）的优点，能学习语音序列数据的时间上下文信息[9-13]. LSTM 网络将传统循环神经网络（RNN）中的节点替换为 LSTM 节点，解决了 RNN 输出误差逐渐消失导致记忆衰退的问题；BRNN 的网络结构使其能充分利用过去和未来时间信息，BLSTM 网络将 BRNN 网络中两个方向上的神经元节点都替换为 LSTM 节点[14-15]，因此 BLSTM 可以更好地学习前后时刻的时间上下文信息. 本文基线模型使用基于 BLSTM 的语音情感识别模型，BLSTM 网络包含输入层、BLSTM 层和输出层，按时间展开示意图如图 1 所示.

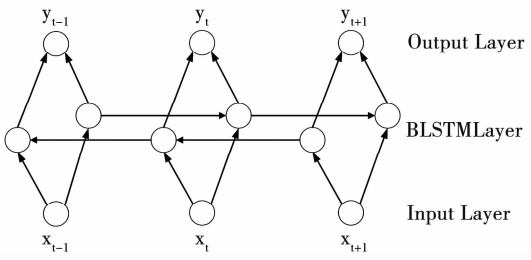

图 1　BLSTM 网络示意

Fig. 1　BLSTM network diagram

图中 $\boldsymbol{x}_t$ 表示时刻 t 的输入数据，$\boldsymbol{y}_t$ 表示时刻 t 的目标输出数据. 其中由若干记忆模块构成的 BLSTM 层中每个模块包含一个或多个自连接的记忆单元以及控制信息流动的输入门、输出门和遗忘门，本文使用的记忆模块结构如图 2 所示.

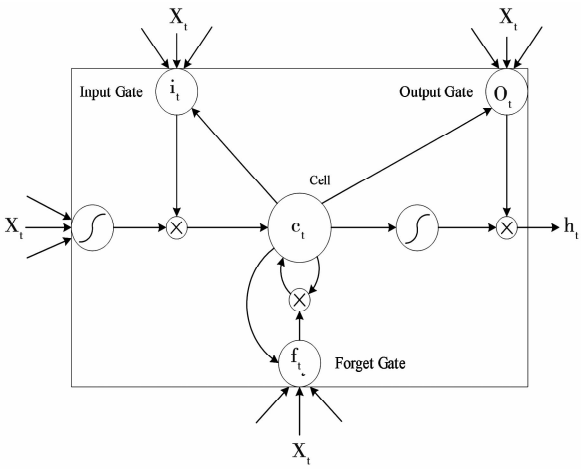

图 2　LSTM 记忆模块结构

Fig. 2　LSTM unit structure diagram

将特征序列表示为 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_{T-1},\boldsymbol{x}_T)$，其中 T 表示序列时间索引值，BLSTM 层按照时刻 $t=1\sim T$ 依次计算 3 个门和记忆单元的激活值，t 时刻的计算公式如下.

输入门：

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t+\boldsymbol{W}_{hi}\boldsymbol{h}_{t-1}+\boldsymbol{b}_i). \tag{1}$$

遗忘门：

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t+\boldsymbol{W}_{hf}\boldsymbol{h}_{t-1}+\boldsymbol{b}_f). \tag{2}$$

记忆单元：

$$\boldsymbol{c}_t=\boldsymbol{f}_t*\boldsymbol{c}_{t-1}+\boldsymbol{i}_t*\tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t+\boldsymbol{W}_{hc}\boldsymbol{h}_{t-1}+\boldsymbol{b}_c). \tag{3}$$

输出门：

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t+\boldsymbol{W}_{ho}\boldsymbol{h}_{t-1}+\boldsymbol{b}_o). \tag{4}$$

隐层输出：

$$\boldsymbol{h}_t=\boldsymbol{o}_t*\tanh(\boldsymbol{c}_t). \tag{5}$$

式中：$\boldsymbol{W}_{x\#}$（代表 $\boldsymbol{W}_{xi}$，$\boldsymbol{W}_{xf}$，$\boldsymbol{W}_{xc}$，$\boldsymbol{W}_{xo}$）是输入 $\boldsymbol{x}_t$ 与记忆模块间连接矩阵，$\boldsymbol{W}_{h\#}$（代表 $\boldsymbol{W}_{hi}$，$\boldsymbol{W}_{hf}$，$\boldsymbol{W}_{hc}$，$\boldsymbol{W}_{ho}$）为隐含层上一时刻输出 $\boldsymbol{h}_{t-1}$ 与记忆模块间连接矩阵，$\boldsymbol{b}_{\#}$

(代表$\boldsymbol{b}_i$,$\boldsymbol{b}_f$,$\boldsymbol{b}_c$,$\boldsymbol{b}_o$)为偏置向量,σ 是 sigmoid 函数,tanh 是双曲正切函数,* 为向量间逐个元素相乘计算符号,当前时刻隐含层输出$\boldsymbol{h}_t$ 将作为下一时刻输入.

2 嵌入注意力机制并结合层级上下文的语音情感识别

上述基于 BLSTM 的语音情感识别基线模型忽略了特征序列中各帧的特点,并存在仅局限于特征层上下文而导致标签层上下文细节丢失的缺点,因此本文提出嵌入注意力机制并结合层级上下文的语音情感识别模型,如图 3 所示.

该模型识别过程主要包括语音情感特征注意力加权和层级上下文学习 2 个步骤. 注意力加权对来自不同时刻的帧特征给予不同关注;层级上下文学习包括特征层上下文学习和标签层上下文学习,主要学习语音数据的时间上下文信息并据此做情感识别. 识别前需对语音情感特征全集进行特征降维,获取最优特征子集.

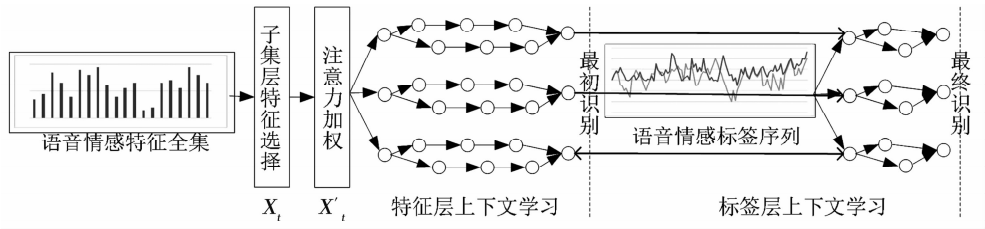

图 3 嵌入注意力机制并结合层级上下文的语音情感识别模型

Fig. 3 BLSTM model based on hierarchical context and attention mechanism

2.1 子集层特征选择

由于本文改进主要针对语音情感识别建模过程,不涉及找到影响识别效果的最佳特征全集,因此语音情感特征提取参考“The Interspeech 2009 Emotion Challenge”竞赛中的基准特征集[16],包含过零率、能量、基频、谐波噪声比、1 ~ 12 阶梅尔频率倒谱系数等 16 个低层语音情感特征,在这些低层特征基础上应用最大值、最小值、均值、标准差等 12 个全局统计函数共得到 384 维语音情感特征,如表 1 所示. 本文利用开源软件 OpenSMILE 提取这些语音特征参数并进行归一化.

表 1 声学特征及统计特征

Tab. 1 Acoustic features and statistical features

声学特征	数量	统计特征	数量
Energy	1	max, min, range, amean, stddev	5
MFCC1 ~ 12	12	maxPos, minPos	2
ZCR	1	skewness, kurtosis	2
HNR	1	linregc1, linregc2, linregerrQ	3
F0	1		
合计	16	合计	12

在特征全集上使用 SVM-RFE 特征选择算法进行子集层特征选择,SVM-RFE 算法根据支持向量机建模过程中的特征权重不断迭代淘汰排名靠后的特征从而实现特征排序,算法流程如图 4 所示,其中 k 代表特征维数.

SVM 分类器常用排序系数是各特征对目标函数具有的判别信息量,即特征权重向量 $\boldsymbol{w}$ 对分类面 $y=\boldsymbol{w}\cdot\boldsymbol{x}+b$ 的贡献值,权重 w_i 越大表明该特征含有判别信息越多,对决策函数影响越大. 因此每次迭代剔除一个 $\|\boldsymbol{w}\|^2$ 最小的特征,同时更新特征排序表进行递归训练直至得到特征全集最终排序结果.

其中 SVM 目标函数为

$$\begin{cases} J=\min\dfrac{1}{2}\|\boldsymbol{w}\|^2, \\ \text{s.t. } \boldsymbol{y}_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\geqslant 1, i=1,2,\cdots,k. \end{cases} \tag{6}$$

当剔除第 i 个特征后 J 的变化为

$$\Delta J(i)=\frac{\partial J}{\partial w_i}\Delta w_i+\frac{\partial^2 J}{\partial w_i^2}(\Delta w_i)^2. \tag{7}$$

式中 w_i 为第 i 个特征的权值. 求 J 的最优解得

$$\Delta J(i)\approx(\Delta w_i)^2. \tag{8}$$

其中:$\Delta w_i=w_i$,因此 SVM-RFE 以 $\|\boldsymbol{w}\|^2$ 为排序准则可以保证特征排序过程中优先保留信息量大的特征子集,从而实现特征降维减小后续识别计算复杂度.

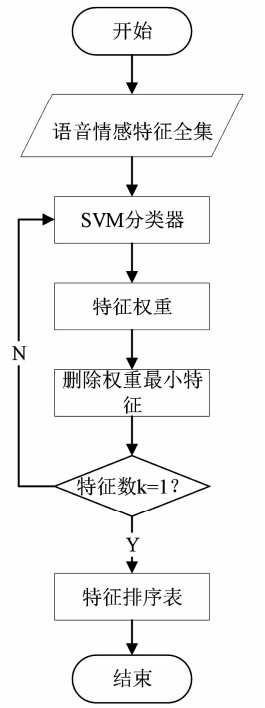

图 4　SVM-RFE 算法流程图

Fig. 4　SVM-RFE algorithm flow chart

2.2　注意力机制

注意力机制是一种思想，不同任务中它的实现方式可以完全不同[17]，本文针对 BLSTM 网络模型输入——语音情感特征，实现注意力机制，使模型学习调整对来自不同时刻的帧特征给予不同的关注度.

传统基于特征层上下文学习的语音情感识别模型在 t 时刻的输入 $\boldsymbol{x}_t$ 由 P 帧语音情感上下文特征扩展而成，经过若干网络层计算最终输出语音情感某一维度预测值. 由于每帧特征对当前时刻情感预测贡献并不一定相同，因此上述做法存在输入 P 帧语音情感特征内部时间信息被忽略的缺点. 注意力机制通过神经网络计算输入特征 $\boldsymbol{x}_t$ 的注意力权值 $\boldsymbol{\alpha}_t$ 对 P 帧特征分别加权，加权后的特征 $\boldsymbol{x}'_t$ 替换 $\boldsymbol{x}_t$ 作为 BLSTM 网络的输入. 具体实现如式(9) ~ (12).

$$\boldsymbol{e}_t = \mathrm{Attend}(\boldsymbol{x}_t, \boldsymbol{s}_{t-1}, \boldsymbol{\alpha}_{t-1}). \tag{9}$$

式中 Attend(·)为计算注意力得分 $\boldsymbol{e}_t$ 的神经网络，由上式可知注意力得分取决于当前时刻输入 $\boldsymbol{x}_t$、上一时刻预测值 $\boldsymbol{s}_{t-1}$ 以及上一时刻注意力权值 $\boldsymbol{\alpha}_{t-1}$.

$$a_{tp} = \frac{\exp(e_{tp})}{\sum_{p=1}^{P}\exp(e_{tp})}. \tag{10}$$

式中 e_{tp} 为第 p 帧注意力得分，上式表明注意力权值 a_{tp} 通过注意力得分经 Softmax 函数归一化至 0 ~ 1 之间.

$$\boldsymbol{x}'_{tp} = a_{tp}x_{tp}. \tag{11}$$

式中 $\boldsymbol{x}_{tp}$ 为第 p 帧特征向量，上式利用第 p 帧注意力权值 a_{tp} 对 x_{tp} 加权得到考虑 P 帧特征各自贡献程度信息后的第 p 帧特征表示 x'_{tp}.

$$\boldsymbol{y}_t = \mathrm{BLSTM}(\boldsymbol{x}'_t). \tag{12}$$

$\boldsymbol{x}'_t$ 送入式(12) BLSTM 网络模型即可得到语音情感最初预测值 $\boldsymbol{y}_t$.

2.3　层级上下文学习过程

在 BLSTM 网络模型部分已阐述其结构和计算过程，本文使用的多层 BLSTM 网络计算方法. 假设 BLSTM 网络一共有 N 层，第一层是输入层，第二层到第 N − 1 层是 BLSTM 层，第 N 层是输出层，特征层上下文学习阶段计算公式如下.

$$(\overrightarrow{\boldsymbol{h}}_t^n, \overrightarrow{\boldsymbol{c}}_t^n) = \overrightarrow{H}^n(\overrightarrow{\boldsymbol{h}}_t^{n-1}, \overrightarrow{\boldsymbol{h}}_{t-1}^n, \overrightarrow{\boldsymbol{c}}_{t-1}^n), \tag{13}$$

$$(\overleftarrow{\boldsymbol{h}}_t^n, \overleftarrow{\boldsymbol{c}}_t^n) = \overleftarrow{H}^n(\overleftarrow{\boldsymbol{h}}_t^{n-1}, \overleftarrow{\boldsymbol{h}}_{t-1}^n, \overleftarrow{\boldsymbol{c}}_{t-1}^n), \tag{14}$$

$$\boldsymbol{y}'_t = \boldsymbol{W}_{\overrightarrow{h}_t^{N-1}y}\overrightarrow{\boldsymbol{h}}_t^{N-1} + \boldsymbol{W}_{\overleftarrow{h}_t^{N-1}y}\overleftarrow{\boldsymbol{h}}_t^{N-1} + \boldsymbol{b}_y. \tag{15}$$

式中：$1 \leqslant t \leqslant T, 2 \leqslant n \leqslant N-1, \overrightarrow{\boldsymbol{h}}_t^1 = \boldsymbol{x}'_t, \overleftarrow{\boldsymbol{h}}_t^1 = \boldsymbol{x}'_t$，$\boldsymbol{y}'$ 是网络的输出. 每一 BLSTM 网络层中参数 $\overrightarrow{\boldsymbol{h}}_0^n, \overleftarrow{\boldsymbol{h}}_0^n, \overrightarrow{\boldsymbol{c}}_0^n, \overleftarrow{\boldsymbol{c}}_0^n$ 的值均随机产生. 式(13)、(14)是多层 BLSTM 网络的计算方法，函数 H(·)接收当前时刻上一隐藏层节点的输出 $\overrightarrow{\boldsymbol{h}}_t^{n-1}$、当前隐藏层节点上一时刻的输出 $\overrightarrow{\boldsymbol{h}}_{t-1}^n$ 和当前隐藏层节点上一时刻记忆单元的输出 $\overrightarrow{\boldsymbol{c}}_{t-1}^n$ 3 个数据作为输入，运算后输出当前时刻当前隐藏层节点的输出 $\overrightarrow{\boldsymbol{h}}_t^n$ 和当前时刻当前隐藏层节点记忆单元的输出 $\overrightarrow{\boldsymbol{c}}_t^n$，$\overrightarrow{H}$(·)和 $\overleftarrow{H}$(·)是 H(·)计算函数在 2 个方向上的运用. 式(15)是输出层的计算方式，同时考虑来自 2 个方向上的数据. 所有计算都需要从 $t = 1$ 时刻逐步计算至 $t = T$ 时刻.

用于特征层上下文学习的 BLSTM1 与用于标签层上下文学习的 BLSTM2 网络参数训练均采用 BPTT 算法. 设 t 时刻网络输入层向量为 $\boldsymbol{x}(t)$，隐层向量为 $\boldsymbol{h}(t)$，输出层向量为 $\boldsymbol{o}(t)$，输入层与隐层间连接矩阵为 $\boldsymbol{V}$，隐层与隐层间连接矩阵为 $\boldsymbol{U}$，隐层与输出层间连接矩阵为 $\boldsymbol{W}$，隐层与输出层的偏置分别为 b 和 a. $h^p(t)$ 和 $o^p(t)$ 分别为第 p 帧在 t 时刻的隐层变量和输出变量，$\boldsymbol{\delta}^p(v^p(t))$ 和 $\boldsymbol{\delta}^p(u^p(t))$ 分别为第 p 帧在 t 时刻输出层误差反向信号变量和隐层误差反向信号向量，L_p 为模型总损失，$\frac{\partial L_p}{\partial \boldsymbol{W}}$、$\frac{\partial L_p}{\partial \boldsymbol{V}}$、$\frac{\partial L_p}{\partial \boldsymbol{U}}$ 分别为对权值 $\boldsymbol{W}$、$\boldsymbol{V}$、$\boldsymbol{U}$ 的偏导，$\frac{\partial L_p}{\partial a}$、$\frac{\partial L_p}{\partial b}$ 分别为对偏置 a 和 b 的偏导.

首先随机初始化所有权值和偏置，初始化 $\frac{\partial L_p}{\partial \boldsymbol{W}}$，

$\frac{\partial L_p}{\partial \boldsymbol{V}},\frac{\partial L_p}{\partial \boldsymbol{U}},\frac{\partial L_p}{\partial a},\frac{\partial L_p}{\partial b}=0$. $t=0$ 时定义隐层变量为 0,随着时间从 $t=1$ 到 T 正向传播,更新第 p 帧在 t 时刻的隐层变量和输出层变量:

$$h^p(t)=f(u^p(t))=f(\boldsymbol{V}x^p(t)+\boldsymbol{U}h^p(t-1)+b), \tag{16}$$

$$o^p(t)=g(v^p(t))=f(\boldsymbol{W}h^p(t)+a). \tag{17}$$

随着时间从 $t=T$ 到 1 反向传播,计算第 p 帧在 t 时刻输出层和隐层的误差反向信号变量:

$$\delta^p(v^p(t))=o^p(t)-y^p(t)\cdot g'(v^p(t)), \tag{18}$$

$$\delta^p(u^p(t))=[\boldsymbol{W}^T\delta^p(v^p(t))]\cdot f'(u^p(t)). \tag{19}$$

更新权值 $\boldsymbol{W}$、$\boldsymbol{V}$、$\boldsymbol{U}$ 和偏置 a 和 b 的偏导:

$$\frac{\partial L_p}{\partial \boldsymbol{W}}=\frac{\partial L_p}{\partial \boldsymbol{W}}+\sum_{p=1}^{P}\delta^p(v^p(t))\,(h^p(t))^T, \tag{20}$$

$$\frac{\partial L_p}{\partial \boldsymbol{V}}=\frac{\partial L_p}{\partial \boldsymbol{V}}+\sum_{p=1}^{P}\delta^p(u^p(t))\,(x^p(t))^T, \tag{21}$$

$$\frac{\partial L_p}{\partial \boldsymbol{U}}=\frac{\partial L_p}{\partial \boldsymbol{U}}+\sum_{p=1}^{P}\delta^p(u^p(t))\,(x^p(t))^T, \tag{22}$$

$$\frac{\partial L_p}{\partial a}=\frac{\partial L_p}{\partial a}+\sum_{p=1}^{P}\delta^p(v^p(t)), \tag{23}$$

$$\frac{\partial L_p}{\partial b}=\frac{\partial L_p}{\partial b}+\sum_{p=1}^{P}\delta^p(u^p(t)). \tag{24}$$

特征层上下文学习阶段将注意力加权后的特征序列 $\boldsymbol{x}'_t(\boldsymbol{x}'_1,\boldsymbol{x}'_2,\boldsymbol{x}'_3,\cdots,\boldsymbol{x}'_{T-1},\boldsymbol{x}'_T)$ 输入 BLSTM1,计算输出与标签序列$(\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{y}_3,\cdots,\boldsymbol{y}_{T-1},\boldsymbol{y}_T)$之间的均方根误差(RMSE)作为训练时的损失函数,计算方法如式(25)所示.

$$R_{\mathrm{RMSE}}=\sqrt{\frac{1}{T}\sum_1^T(\boldsymbol{y}_t-\boldsymbol{y}'_t)^2}. \tag{25}$$

由于语音情感特征与语音情感状态之间存在差异,因此需增加标签层上下文学习更好地利用标签序列的上下文信息,使最终识别结果更趋近标签值,从而达到提升识别准确率的目的. 标签层上下文学习阶段将标签序列$(\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{y}_3,\cdots,\boldsymbol{y}_{T-1},\boldsymbol{y}_T)$输入 BLSTM2,计算输出与标签序列之间的 RMSE 值作为损失函数.

$(\boldsymbol{y}'_1,\boldsymbol{y}'_2,\boldsymbol{y}'_3,\cdots,\boldsymbol{y}'_{T-1},\boldsymbol{y}'_T)$为 BLSTM1 输出所得初步识别结果,将其作为训练好的 BLSTM2 的输入即可获得最终识别结果$(\hat{\boldsymbol{y}}_1,\hat{\boldsymbol{y}}_2,\hat{\boldsymbol{y}}_3,\cdots,\hat{\boldsymbol{y}}_{T-1},\hat{\boldsymbol{y}}_T)$,该结果综合考虑了特征层和标签层上下文信息,准确率得到提升.

3 实验结果与分析

3.1 语音情感数据库

本文在 SEMAINE 和 RECOLA 数据库上评估所提方法有效性. SEMAINE 可供研究者无偿使用(http://semaine-db.eu/),音频采样率为 48 kHz,量化精度为 24 bit,数据时长约为 7 小时[18-19]. RECOLA 录制过程中参与者通过视频会议完成指定任务,记录视频会议内容得到音频、视频、心电图和皮肤电活动 4 种模态数据[20]. 实验过程中,将语音数据库中语料随机划分成三等份,其中两份作为训练集,一份用作测试集.

3.2 评价指标

本文采用 RMSE 和一致性相关系数(CCC)作为模型性能评估指标,RMSE 计算方法在式(25)中已经给出,可看出 RMSE 值越小性能越好. CCC 计算方法为

$$C_{\mathrm{CCC}}=\frac{2\rho\delta_y\delta_{y'}}{\delta_y^2+\delta_{y'}^2+(\mu_y-\mu_{y'})^2}. \tag{26}$$

式中:$\boldsymbol{y}$ 为维度情感标注值序列数据,$\boldsymbol{y}'$为模型实际输出序列数据,ρ 是两个序列数据的 Pearson 相关系数,μ_y和$\mu_{y'}$分别代表两个序列数据的均值,δ_y和$\delta_{y'}$分别表示两个序列数据的方差. 从计算公式可以看出 CCC 值越大模型性能越好.

3.3 参数选择

本文使用 SVM-RFE 算法进行子集层特征选择,SVM 采用线性核函数,依照 SVM-RFE 算法得到特征排序表后,从 1 ~ 100 对特征维数取值,实验结果表明选择前 72 维特征子集的性能最优. 在开始特征层和标签层上下文信息学习之前,首先在训练集上进行网络层数和隐层节点数两个重要网络结构参数的选定,利用性能较好的参数作为测试集最终参数. 确定模型所用 BLSTM 网络层数时固定隐层节点数,通过从 1 ~5 设置层数比较相应模型性能寻找合适网络层数. 同理根据选定的网络层数,以步长 $0.25\times M$ 在$[0.25\times M,1.25\times M]$($M$ 是特征维度)范围内选择不同隐层节点数,比较各情况下模型性能,最终确定特征层上下文信息学习阶段 BLSTM 网络层数为 3,节点数为(64,64,64);标签层上下文信息学习阶段 BLSTM 网络层数为 1,节点数为 1. 输入层节点数为输入数据的大小,特征层和标签层上下文信息学习阶段都采用线性回归作为输出层,输出层节点数为 1. 训练最大迭代次数为 200 次,采用 early-stop 方式,当模型在训练集上一致性相关系数值连续 10 次都不再提高时停止训练,将训练集上情感识别效果最好的模型权重作为测试用最终权重,权重的更新方法参考 Adadelta 方法[21].

训练时对语音数据所提取的特征序列按确定帧数作为一个时序长度进行切分构成若干时序段,然后将每一个时序段作为一个样本进行 BLSTM 网络

训练. 实验过程中发现随着时序长度的改变，BLSTM 网络的收敛程度呈现明显差别，图 5 展示了 SEMAINE 数据库中序列 ID 为 13 的情感语音 Valence 维度 CCC 值随时序段长度的变化规律.

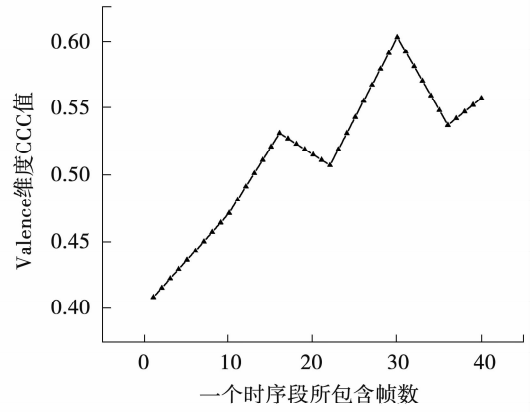

图 5　不同时序段长度性能对比

Fig. 5　Comparison of performance of different time segments

图中横坐标表示一个时序段包含的帧数，纵坐标表示识别结果在 Valence 维度的 CCC 值，总体来看在一定区间内时序长度越长，BLSTM 网络收敛效果越好，这是因为更长的时序长度包含更多的时序变化信息，能为语音情感预测任务提供更大信息量. 数据库中语音长度各不相同，折中选择 30 帧为一个时序长度，可使绝大多数语音数据包含至少 3 个时序段，同时能令 BLSTM 保持较好收敛效果.

3.4　实验结果分析

为检验本文改进效果，将特征层上下文学习阶段的最初识别结果与增加标签层上下文学习阶段的最终识别结果以及同时引入标签层上下文学习、注意力机制的识别结果进行比较，对比数据如表 2、表 3所示.

表 2 显示，SEMAINE 数据库中 Valence 维度 RMSE 值在增加标签层上下文学习后减小 3.36%，嵌入注意力机制相较上述结果上再获得 0.34% 的缩小值，同时其 CCC 值在逐步改进过程中依次提高 25.06% 和 28.04%. 表 3 表明 RECOLA 数据库中两个维度 RMSE 值平均约减小 4.6%，CCC 值平均约增大 27%. 更好的模型对应更低的 RMSE 和更高的 CCC，以上数据说明同时引入标签层上下文学习以及注意力机制优于增加标签层的 BLSTM 网络模型，更优于仅基于特征层上下文学习的语音情感识别模型，其在 2 个数据集上与基线特征层上下文学习模型相比均取得了性能提升效果.

表 2　SEMAINE 数据库改进前后结果比较

Tab. 2　Comparison of results on SEMAINE database

情感维度	特征层		标签层		标签层 + 注意力机制	
	R_{RMSE}	C_{CCC}	R_{RMSE}	C_{CCC}	R_{RMSE}	C_{CCC}
Valence	0.137 2	0.316 1	0.103 6	0.566 7	0.100 2	0.596 5
Activation	0.196 0	0.312 8	0.162 8	0.562 9	0.155 4	0.598 7
Power	0.176 6	0.226 7	0.144 7	0.476 4	0.132 5	0.502 6
Expectation	0.217 5	0.262 8	0.184 8	0.513 7	0.172 5	0.538 4
Intensity	0.187 7	0.247 8	0.156 4	0.498 9	0.135 2	0.511 6

表 3　RECOLA 数据库改进前后结果比较

Tab. 3　Comparison of results on RECOLA database

情感维度	特征层		标签层		标签层 + 注意力机制	
	R_{RMSE}	C_{CCC}	R_{RMSE}	C_{CCC}	R_{RMSE}	C_{CCC}
Arousal	0.213 0	0.401 3	0.180 3	0.642 2	0.166 2	0.650 2
Valence	0.166 8	0.290 5	0.133 4	0.540 2	0.122 5	0.580 5

为更清晰地对比分析各模型执行效率，将 3 组模型分别在 2 个数据集上进行多次实验并记录其平均运行时间，具体结果如表 4 所示.

表 4　模型训练时间对比

Tab. 4　Comparison of model training time

数据集	特征层	标签层	标签层 + 注意力机制
SEMAINE	3 h 48 m	3 h 58 m	2 h 34 m
RECOLA	4 h 36 m	4 h 39 m	3 h 29 m

由表 4 可知，数据集越大模型训练时间越长. 由于特征层上下文学习网络 BLSTM1 与标签层上下文学习网络 BLSTM2 相互独立，且 BLSTM2 输入数据维度远小于 BLSTM1 的输入，故增加标签层上下文学习过程在提升识别准确率的同时对模型执行效率并未造成较大影响. 而嵌入注意力机制后相对于基线特征层上下文学习模型训练时间平均减少了1 h 15 min，表明注意力机制能聚焦于对当前任务更关键的信息，降低对其他信息关注度甚至过滤无关信

息,从而较大程度加快模型训练速度,提高模型执行效率.

进一步地,对 SEMAINE 数据库中序列 ID 为 13 的语音数据 Valence 维度预测结果进行可视化如图 6所示.

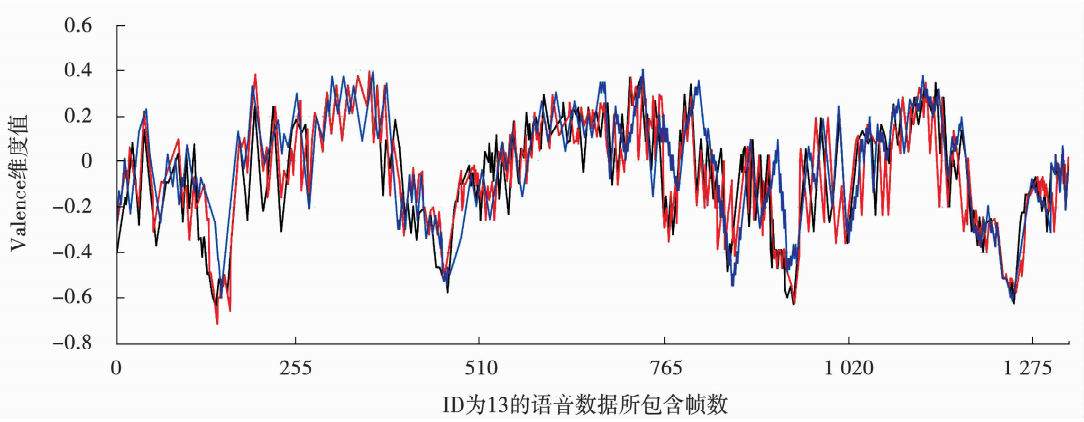

图 6 本文模型与基线模型 Valence 维度预测结果比较

Fig. 6 Comparison of the Valence dimension prediction results between the model and the baseline model

图 6 提供了本文模型与基线模型在 Valence 维度情感预测结果的比较. 黑色曲线代表数据库提供的情感标注值,红色曲线代表本文模型预测结果,蓝色曲线代表基线模型预测结果. 从图中可以看出前后帧变化不大的各帧两个预测结果比较相似,但是前后帧 Valence 维度值变化大的一帧例如第 150 帧,黑色曲线与灰色曲线的吻合程度明显优于蓝色曲线. 分别计算模型改进前后预测结果的 RMSE 值见表 5,对比结果显示本文模型在 Valence 维度的 RMSE 值相较基线模型降低了 12%,说明引入标签层上下文学习和注意力机制的预测结果更贴近情感标注值.

表 5 RMSE 值对比

Tab. 5 RMSE value comparison

评估指标	基线模型	本文模型
RMSE	0.28	0.16

最后将本文模型与现有模型方法进行对比.

模型一:Valstar M 等[22]介绍的 AVEC 2016 中用于音频与视频情感识别的机器学习方法.

模型二:张雪英等[23]提出的一种基于 SVR 的语音维度情感预测模型.

模型三:Ehab A. Albadawy 等[24]提出的集合和端到端的联合建模方法.

模型四:本文改进前仅进行特征层上下文学习的 BLSTM 网络模型.

模型五:本文设计的基于层级上下文和注意力机制的 BLSTM 网络模型.

表中 R 代表 RMSE 值,C 代表 CCC 值. 由表可知,模型 5 的 RMSE 值相较模型 1、2、4 在 Valence、Activation、Power、Expectation、Intensity 这五个维度上均有不同程度减小,平均降低 1.5%,CCC 值平均提高 15.1%. 通过与现有其他模型对比,本文所提方法具有更好的情感分辨力,在不同语音情感数据库也拥有更好的稳定性.

表 6 与现有其他模型方法对比

Tab. 6 Compared with other existing model methods

维度	模型 1		模型 2		模型 3		模型 4		模型 5	
	R	C	R	C	R	C	R	C	R	C
V	–	–	0.122	0.338	–	–	0.104	0.567	0.100	0.597
A	–	–	0.171	0.361	–	–	0.163	0.563	0.155	0.599
P	–	–	0.159	0.262	–	–	0.145	0.476	0.133	0.503
E	–	–	0.187	0.311	–	–	0.185	0.514	0.173	0.538
I	–	–	0.162	0.283	–	–	0.156	0.499	0.135	0.512
Arousal	–	0.396	–	–	–	0.528	0.180	0.642	0.166	0.650
Valence	–	0.401	–	–	–	0.519	0.133	0.540	0.123	0.581

4 结　论

本文针对语音情感识别提出基于层级上下文和注意力机制的 BLSTM 模型，首先原始特征全集经过特征选择得到最优特征子集，消除高维数低层次特征的冗余性和不稳定性；然后对特征子集进行注意力加权，充分考虑输入各帧特征中的时间信息，使模型对输入层中每帧特征给予不同关注度；其次学习加权后的特征序列上下文信息得到初步情感预测结果，最后在上述初步结果基础上增加标签层上下文学习做最终识别. 该方法抓住语音情感在表达过程中的连续性特点，利用 BLSTM 网络学习语音情感特征序列以及语音情感标签值序列两层级的上下文信息，综合考虑其差异性，实验结果表明，与基线模型相比本文模型提升了对情感语音信号的建模能力，并有效提高了语音情感识别准确率. 本文采用的 BLSTM 模型虽然在上下文学习方面表现优秀，但是结构复杂不易训练，因此后续将对优化 BLSTM 进行研究进一步提升识别性能.

参考文献

[1] Björn W. Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends[J]. Communications of the Acm, 2018, 61(5):90. DOI:10.1145/3129340

[2] JING Shaoling, MAO Xia, CHEN Lijiang. Prominence features: Effective emotional features for speech emotion recognition[J]. Digital Signal Processing, 2018, 72:216. DOI:10.1016/j.dsp.2017.10.016

[3] Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: A review[J]. International Journal of Speech Technology, 2018, 21(1):93. DOI:10.1007/s10772-018-9491-z

[4] Tzinis E, Potamianos A. Segment-based speech emotion recognition using recurrent neural networks[C]// Seventh International Conference on Affective Computing & Intelligent Interaction. 2018. DOI:10.1109/ACII.2017.8273599

[5] Mustafa M B, Yusoof M A M, Don Z M, et al. Speech emotion recognition research: An analysis of research focus[J]. International Journal of Speech Technology, 2018, 21(1):137. DOI:10.1007/s10772-018-9493-x

[6] Sariyanidi E, Gunes H, Cavallaro A. Automatic analysis of facial affect: A survey of registration, representation, and recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(6): 1113. DOI:10.1109/TPAMI.2014.2366127

[7] SHAO Ling, ZHEN Xiantong, TAO Dacheng, et al. Spatio-temporal laplacian pyramid coding for action recognition[J]. IEEE Transactions on Cybernetics, 2014, 44(6): 817. DOI:10.1109/TCYB.2013.2273174

[8] Meng H, Bianchiberthouze N. Affective state level recognition in naturalistic facial and vocal expressions[J]. IEEE Transactions on Cybernetics, 2013, 44(3): 315. DOI: 10.1109/tcyb.2013.2253768

[9] Rodriguez P, Cucurull G, Gonalez J, et al. Deep pain: Exploiting long short-term memory metworks for facial expression classification[J]. IEEE Transactions on Cybernetics, 2017. DOI: 10.1109/TCYB.2017.2662199

[10] Ringeval F, Eyben F, Kroupi E, et al. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data[J]. Pattern Recognition Letters, 2014: S0167865514003572. DOI:10.1016/j.patrec.2014.11.007

[11] Nicolaou M A, Gunes H, Pantic M. Audio-visual classification and fusion of spontaneous affective data in likelihood space[C]// International Conference on Pattern Recognition. IEEE, 2010. DOI: 10.1109/ICPR.2010.900

[12] Ringeval F, Valstar M, Jaiswal S, et al. AV + EC 2015-The first affect recognition challenge bridging across audio, video, and physiological data[C]//5th International Workshop on Audio/Visual Emotion Challenge (AVEC). ACM, 2015. DOI: 10.1145/2808196.2811642

[13] Metallinou A, Martin Wöllmer, Katsamanis A, et al. Context-sensitive learning for enhanced audio visual emotion classification[J]. IEEE Transactions on Affective Computing, 2012, 3(2): 184. DOI:10.1109/T-AFFC.2011.40

[14] HE Lang, JIANG Dongmei, YANG Le, et al. Multimodal Affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks[C]//International Workshop on Audio/visual Emotion Challenge. ACM, 2015. DOI:10.1145/2808196.2811641

[15] Khorrami P, Le Paine T, Brady K, et al. How deep neural networks can improve emotion recognition on video data[J]. IEEE International Conference on Image Processing (ICIP) 2016:619

[16] Tzinis E, Potamianos A. Segment-based speech emotion recognition using recurrent neural networks[C]//Seventh International Conference on Affective Computing & Intelligent Interaction. IEEE, 2018. DOI:10.1109/ACII.2017.8273599

[17] 张宇，张鹏远，颜永红. 基于注意力 LSTM 和多任务学习的远场语音识别[J]. 清华大学学报(自然科学版)，2018(3)
ZHANG Yu, ZHANG Pengyuan, YAN Yonghong. Long short-term memory with attention and multitask learning for distant speech recognition[J]. Journal of Tsinghua University (Science and Technology), 2018(3)

[18] Mckeown G, Valstar M, Cowie R, et al. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent[J]. IEEE Transactions on Affective Computing, 2012, 3(1): 5. DOI:10.1109/t-affc.2011.20

[19] Mckeown G, Valstar M F, Cowie R, et al. The SEMAINE corpus of emotionally coloured character interactions[C]//IEEE International Conference on Multimedia & Expo. 2010. DOI:10.1109/ICME.2010.5583006

[20] Ringeval F, Sonderegger A, Sauer J, et al. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions[C]//IEEE International Conference & Workshops on Automatic Face & Gesture Recognition. 2013. DOI:10.1109/FG.2013.6553805

[21] Zeiler M D. ADADELTA: An adaptive learning rate method[J]. Computer Science, 2012

[22] Valstar M, Gratch J, Schuller B, et al. AVEC 2016-depression, mood, and emotion recognition workshop and challenge[J]. 2016. DOI:10.1145/2988257.2988258

[23] 张雪英，张婷，孙颖，等. 基于 PAD 模型的级联分类情感语音识别[J]. 太原理工大学学报，2018，49(5). DOI:10.16355/j.cnki.issn1007-9432tyut.2018.05.013
ZHANG Xueying, ZHANG Ting, SUN Ying, et al. Cascaded classification emotional speech recognition base on PAD model[J]. Taiyuan University of Technology, 2018, 49(5). DOI:10.16355/j.cnki.issn1007-9432tyut.2018.05.013

[24] Ehab A. Albadawy, Yelin Kim (2018). Joint discrete and continuous emotion prediction using ensemble and end-to-end approaches. 366. 10.1145/3242969.3242972

（编辑　苗秀芝）