计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(5):269−275 [doi: 10.15888/j.cnki.csa.007917]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合数据平衡和注意力机制的 CNN+LSTM 的自然语音情感识别①

陈　港[1], 张石清[2], 赵小明[1,2]

[1](浙江理工大学 机械与自动控制学院, 杭州 310018)
[2](台州学院 智能信息处理研究所, 台州 318000)
通讯作者: 赵小明, E-mail: tzxyzxm@163.com

摘　要: 为了解决语音情感识别中数据集样本分布不平衡的问题, 提出一种结合数据平衡和注意力机制的卷积神经网络 (CNN) 和长短时记忆单元 (LSTM) 的语音情感识别方法. 该方法首先对语音情感数据集中的语音样本提取对数梅尔频谱图, 并根据样本分布特点对进行分段处理, 以便实现数据平衡处理, 通过在分段的梅尔频谱数据集中微调预训练好的 CNN 模型, 用于学习高层次的片段语音特征. 随后, 考虑到语音中不同片段区域在情感识别作用的差异性, 将学习到的分段 CNN 特征输入到带有注意力机制的 LSTM 中, 用于学习判别性特征, 并结合 LSTM 和 Softmax 层从而实现语音情感的分类. 在 BAUM-1s 和 CHEAVD2.0 数据集中的实验结果表明, 本文提出的语音情感识别方法能有效地提高语音情感识别性能.

关键词: 卷积神经网络; 长短时记忆单元; 注意力机制; 语音情感识别

引用格式: 陈港,张石清,赵小明.结合数据平衡和注意力机制的 CNN+LSTM 的自然语音情感识别.计算机系统应用,2021,30(5):269–275. http://www.c-s-a.org.cn/1003-3254/7917.html

Natural Speech Emotion Recognition by Integrating Data Balance and Attention Mechanism Based on CNN+LSTM

CHEN Gang[1], ZHANG Shi-Qing[2], ZHAO Xiao-Ming[1,2]

[1](Faculty of Mechanical Engineering & Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China)
[2](Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, China)

Abstract: In order to solve the problem of unbalanced sample distribution in a dataset in Speech Emotion Recognition (SER), this study proposes a SER method combining a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) units with data balance and an attention mechanism. This method first extracts the log-Mel spectrogram from the samples in a speech emotion dataset and devides the sample distribution into segments according to sample distribution for balance. Then, this method fine-tunes the pre-trained CNN model in the segmented Mel-spectrum dataset to learn high-level speech segments. Next, given the differences in the emotion recognition of different segments in speech, the learned segmented CNN features are input into the LSTM with an attention mechanism for learning discriminative features, and speech emotions are classified with LSTM and Softmax layers. The experimental results in the BAUM-1s and CHEAVD2.0 datasets show that the method proposed in this study has much better performance than conventional methods.

Key words: Convolutional Neural Network (CNN); Long Short-Term Memory (LSTM) unit; attention mechanism; speech emotion recognition

① 基金项目: 国家自然科学基金 (61976149); 浙江省自然科学基金 (LZ20F020002)
Foundation item: National Natural Science Foundation of China (61976149); Natural Science Foundation of Zhejiang Province (LZ20F020002)
收稿时间: 2020-09-23; 修改时间: 2020-10-21; 采用时间: 2020-10-28; csa 在线出版时间: 2021-04-28

人类的语言不仅包含了丰富的文本信息, 同时也携带着包含人们情绪表达的音频信息, 如语音的高低、强弱、抑扬顿挫等变化. 如何让计算机从语音信号中自动识别出说话人的情感状态, 即所谓的“语音情感识别”方面的研究, 已成为人工智能、模式识别、情感计算等领域中的一个热点研究课题. 该研究旨在让计算机通过分析说话人的语音信号对用户的情感信息进行获取、识别和响应, 从而实现用户与计算机之间的交互更加和谐与自然. 该研究在智能人机交互、电话客服中心、机器人等方面具有重要的应用价值.

目前, 在语音情感识别领域中, 大量的前期工作[1–4]主要是针对模拟情感而进行的, 因为这种模拟情感数据库的建立相对自然情感而言, 要容易得多. 近年来, 针对实际环境下的自然语音情感识别方面的研究备受研究者的关注, 因为它更接近实际, 而且比模拟情感的识别要困难得多.

语音情感特征提取, 是语音情感识别中的一个关键步骤, 其目的是从情感语音信号中提取能够反映说话人情感表达信息的特征参数. 目前, 大量语音情感识别文献[5–10]采用手工设计的特征用于情感识别, 如韵律特征 (基频、振幅、发音持续时间)、音质特征 (共振峰、频谱能量分布、谐波噪声比), 谱特征 (梅尔频率倒谱系数 (Mel Frequency Cepstrum Coefficient, MFCC)) 等. 近年来, 也出现了一些代表性的包含几千个手工设计特征的声学特征集, 如 Interspeech-2010 特征集[11], ComParE 特征集[12], AVEC-2013 特征集[13] 以及 GeMAPS 特征集[14]. 尽管这些手工设计的特征参数已经取得了较好的语音情感识别性能, 但它们是低层次的, 对于情感的判别力还不够高, 与人类理解的情感标签还存在“语义鸿沟”问题.

为了解决这个问题, 近年来新出现的深度学习技术[15]可能提供了线索. 近年来, 一些代表性的深度学习方法, 如深度信念网络 (Deep Belief Network, DBN)[15]、卷积神经网络 (Convolutional Neural Networks, CNN)[16]和长短时记忆单元 (Long Short-Term Memory, LSTM)[17]都已经用于语音情感识别. 当使用深度学习方法时, 其输入一般为手工设计的声学特征参数, 或者原始的语音频谱. 例如, 文献 [18] 采用 DBN 直接从提取的 MFCC 等声学特征参数中提取高层次的属性特征, 然后使用极限学习机 (Extreme Learning Machine, ELM) 实现情感的分类任务. 文献 [19–22] 也开始成功使用 CNN 从原始的语音频谱中提取出合适的特征参数用于语音情感识别. 例如, 文献 [19] 采用稀疏自动编码器和 1 层 CNN 结构的方法从原始的语音频谱中学习情感语音特征.

值得指出的是, 文献 [19–21] 采用样本数量非常有限的情感语音数据集来训练自己的浅层 CNN 模型 (1 或 2 个卷积层). 然而, 在计算机视觉领域, 利用已训练好的深度 CNN 模型, 如 AlexNet[16], 在目标图像数据集进行迁移学习往往取得比浅层 CNN模型更好的性能. 主要原因是, 深度 CNN 模型可以通过采用多层的卷积和池化操作来捕获图像的高层属性特征. 为了充分发挥深度 CNN 模型的优势, 我们之前的一个工作[22]提出将一维的情感语音信号转换成类似于 RGB 图像的三通道语音频谱片段作为深度 CNN 模型的输入, 然后将在 ImageNet 图像数据集已训练好的 AlexNet 模型在目标语音情感数据集进行跨模态的迁移学习, 取得了比浅层 CNN 模型更好的语音情感识别性能. 此外, 文献 [23] 提出一种多尺度的 CNN+LSTM的混合深度学习模型, 获得了较好的语音情感性能.

然而, 现有基于深度学习方法的语音情感识别研究存在一些问题:

(1) 没有考虑语音数据集客观存在的情感类别不平衡问题, 即数据集中各种情感类型的样本数量极不均衡. 在这种数据类别不平衡的情况下, 深度学习方法训练时对样本少的情感类型容易出现过拟合的现象.

(2) 语音信号是一种时间序列信号, 但不同时间上的片段区域对情感识别的作用大小是不一样的. 现有文献大都忽略了一句语音中不同片段区域在情感识别方面作用的差异性, 从而限制了深度学习方法的特征表征能力. 因此, 有必要将人类视觉注意力机制 (attention mechanism) 与深度学习方法相结合[24,25], 用于语音情感识别的特征学习.

针对上述问题, 本文提出一种结合数据平衡和注意力机制的 CNN+LSTM 的语音情感识别方法, 并且用于自然语音情感类型的识别. 首先, 采用欠采样和过采样的方式实现情感语音数据集中的样本片段数量的类别平衡. 所谓欠采样是在多数类中的语音样本进行部分采样, 而过采样是在少数类中的语音样本进行部分重复采样. 其次, 采用在大规模音频数据集上已训练好的 VGGish[26] 模型在目标语音情感数据集上进行迁移学习. 最后, 通过对 VGGish 学习到的一句语音中不

同的片段特征区域, 生成时间分布上的权重, 然后和特征图进行加权求和运算, 从而监督双向 LSTM 网络 (Bi-LSTM) 的学习, 以便给一句语音中不同片段特征区域分配不同的注意力权重值. 这样, 该方法能够聚焦于一句语音中对情感识别起作用的目标片段区域的特征学习, 从而改善深度神经网络的特征表征能力, 进一步提高语音情感识别性能.

在自然情感语音数据库 BAUM-1s[27] 和 CHEAVD-2.0[28] 的试验结果表明, 本文方法取得的语音情感识别性能优于其它方法.

1 提出的方法

图 1 给出了结合数据平衡和注意力机制的 CNN+LSTM 的语音情感识别方法的系统流程图. 由图 1 所示, 该方法包括 4 个步骤: (1) 对数梅尔频谱 (Log Mel-spectrogram) 的创建和数据平衡 (data balance); (2) 基于 CNN 的深度片段特征学习; (3) 基于注意力机制的 Bi-LSTM 的情感分类. 图 1 中每个步骤的实现, 具体如下所述.

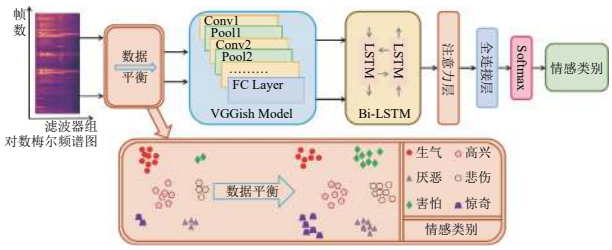

图 1 结合数据平衡和注意力机制的 CNN+LSTM 的语音情感识别方法

1.1 对数梅尔频谱的创建和数据平衡

对原始的一维情感语音信号重采样为 16 kHz 的单声道格式, 然后采用帧移为 10 ms, 时长为 25 ms 的汉宁窗进行短时傅里叶变换, 计算出整句语音信号的声谱图. 将声谱图映射到 64 阶 Mel 滤波器组 (Filter banks) 中计算出 Mel 声谱并取对数, 得到稳定的对数 Mel 频谱.

考虑到用于后续特征学习的 VGGish[26] 模型输入的片段大小是 96×64, 其时长为 0.96 s (含 96 帧, 每帧 0.01 s), 因此需要通过控制时长进行欠采样或重采样, 以便实现训练数据的平衡. 图 2 和图 3 分别给出了 BAUM-1s 和 CHEAVD2.0 上的数据平衡示意图.

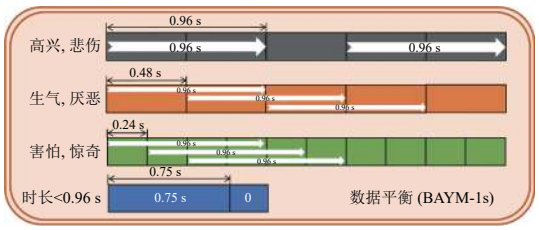

图 2 BAUM-1s 数据集上的数据平衡

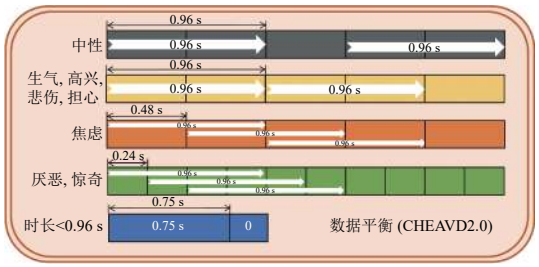

图 3 CHEAVD2.0 数据集上的数据平衡

如图 2, 图 3 所示, 由于数据集中的每个语音样本的时长是不同的, 然而模型的输入大小是固定的, 因此使用数据平衡方法对输入的语音样本进行规范化处理, 数据平衡的实现分为 3 种情况: (1) 对于时长小于 0.96 s (时长<0.96 s) 的语音样本, 在其后面进行重复补值为零的帧, 从而得到一个时长为 0.96 s 的片段, 即含有 96 帧的对数梅尔频谱图; (2) 对于时长大于 0.96 s 的语音样本, 会根据其所属类别的样本总数占数据集总样本数的比例的多少进行欠采样或重采样. 例如, 对于 BAUM-1s 上类别样本数量较多的高兴 (Joy) 和悲伤 (Sadness) 的语音样本采用欠采样方式减少片段数,即只在它们的首尾段处各取一个 0.96 s 的片段, 而对于样本数量不多的语音样本采用重采样方式进行片段数的放大. 对于样本数最少的害怕 (Fear) 和惊奇 (Surprise) 语音样本, 片段重叠长度为 0.24 s, 而样本数较少的生气 (Anger) 和厌恶 (Disgust), 其片段重叠长度为 0.48 s. 数据平衡中产生的片段样本的类别等于其所在整句语音的情感类别. 表 1 和表 2 分别列出了采用数据平衡前后在 BAUM-1s 和 CHEAVD2.0 训练数据集上的各种情感类型的样本片段数量的对比. 由表 1 和表 2 可知, 在数据平衡之前, 每个类别的样本数量差异较大, 如害怕 (Fear) 和惊奇 (Surprise) 类别的语音样本数量, 与其他类别的样本数量相差较大, 在数据平衡之后, 数据集中的每个类别的样本数量达到一种近似的数据平衡状态, 从而能训练出更具鲁棒性的网络模型. 作为举例, 由于 BAUM-1s

包含 30 人, 实验时采用 5 次交叉验证方法, 即所有数据平均分成 5 组, 使用其中一组用于测试, 剩下的 4 组用于训练, 共循环 5 次, 最后取 5 次的平均结果作为最终的结果. 因此, 表 1 只列出了第一次交叉验证时的数据平衡前后结果, 其中测试数据包含随机选择的 6 人 (S03、S19、S20、S24、S28、S29).

表 1 BAUM-1s 数据平衡前后的训练数据样本片段数量比较

数据平衡	生气	厌恶	害怕	高兴	悲伤	惊奇
平衡前	144	268	93	491	487	83
平衡后	314	522	326	405	345	309

表 2 CHEAVD2.0 数据平衡前后的训练数据样本片段数量比较

数据平衡	生气	焦虑	厌恶	高兴	中性	悲伤	惊奇	担心
平衡前	3096	1798	470	2597	5233	2433	468	2345
平衡后	3096	3352	1663	2597	2777	2433	1580	2345

1.2 基于 CNN 的深度片段特征学习

目前, 在计算机视觉领域, 文献 [29,30] 已经证明在目标数据集上对预先训练好的 CNN 模型进行微调 (Fine-tuning), 是一种有效的减轻数据不足的方法. 具体的实现方式为首先通过一个大型的数据集来训练网络模型, 由于网络模型的参数初始化对于模型的训练是至关重要的, 因此由大型数据集训练完成保存的模型参数可用于初始化目标任务的模型参数. 如文献 [30] 首先使用 Imagenet 数据集训练网络模型, 完成训练后, 将模型的最后一个全连接层之前的所有参数迁移到目标任务中的模型, 以此进行目标任务模型的参数初始化. 为此, 采用在大规模音频数据集上已训练好的 VGGish[26] 模型在目标语音情感数据集上进行迁移学习.

VGGish 是由 Google 的语音理解技术团队于 2017 年 3 月发布的一个在大规模音频数据集训练得到类似于 VGG 的模型, 旨在为音频事件检测提供常见的大规模评估任务. 该模型包括 6 个卷积层 (Conv1, Conv2,⋯, Conv6)、4 个最大池化层 (Pool1, Pool2, Pool2, Pool3, Pool4) 和 4 个全连接层 (FC1, FC2, FC3, FC4). VGGish 的训练数据来自于包括 600 多个音频事件类的本体的 200 万个人标记的 10 s YouTube 视频音轨组成的数据集. VGGish 能够从原始的语音信号片段 96×64 中提取高层次的 128-D 大小的特征向量.

给定第 i 个输入片段数据a_i及其对应的情感类别 y_i, 对预训练好的 VGGish 网络 (V) 进行微调, 则相当于求解下面的最优化问题:

$$\min_{W^V,\theta^V}\sum_{i=1}^{K}L\left(softmax\left(W^V\Upsilon^V\left(a_i;\theta^V\right)\right),y_i\right) \tag{1}$$

式中, $\Upsilon^V(a_i;\theta^V)$表示网络V的最后一个全连接层的输出特征, θ^V表示网络 V 的网络参数, W^V 表示网络 V 的 Softmax 层的权重参数. 损失函数 L 则表示为:

$$L(V,y)=-\sum_{j=1}^{l}y_j\log\left(y_j^V\right) \tag{2}$$

式中, y_j表示第 j 个语音片段样本真实的类别号, y_j^V表示网络 V 的 Softmax 层所预测的第 j 个样本类别号, t 表示情感类别数目.

1.3 基于注意力机制的 Bi-LSTM 的情感分类

为了获取一句语音的长时间动态信息, 采用基于注意力机制的双向长短时记忆网络 (Bi-LSTM) 对 CNN 在不同语音频谱片段上学习到的 128-D 特征序列进行时间上的动态信息建模, 并输出整句语音样本的情感识别结果.

给定一个时间长度为 T 的输入序列 (x_1, x_2, $\cdots x_T$), Bi-LSTM 旨在通过计算网络节点激活函数的输出, 将输入序列 (x_1, x_2, $\cdots x_T$) 映射到一个输出序列(p_1, p_2,$\cdots$,p_T), 如下所示:

$$i_t=\sigma(W_{xi}x_t+W_{hi}h_{t-1}+W_{ci}c_{t-1}+b_i) \tag{3}$$

$$f_t=\sigma\left(W_{xf}+W_{hf}h_{t-1}+W_{cf}c_{t-1}+b_f\right) \tag{4}$$

$$c_t=f_tc_{t-1}+i_t\tanh(W_{xc}x_t+W_{hc}h_{t-1}+b_c) \tag{5}$$

$$\sigma_t=\sigma(W_{xo}x_t+W_{ho}h_{t-1}+W_{co}c_{t-1}+b_o) \tag{6}$$

$$h_t=\sigma_t\tanh(c_t) \tag{7}$$

式中, i_t、f_t、c_t、σ_t、h_t分别是 LSTM 模型中的输入门、忘记门、细胞存储单元和输出门的激活输出向量. x_t和h_t分别表示第t个时间步长的输入向量和隐层向量. $W_{\alpha\beta}$表示α和β之间的权重矩阵. 例如, W_{xi}是从输入x_t到输入门i_t的权重矩阵. b_α是的偏置值, σ表示 Sigmoid 激活函数$\sigma(x)=1/(1+e^{-x})$.

为了采用视觉注意力机制获取需要聚焦的重点语音片段区域, 并抑制其他无用的信息, 拟在 Bi-LSTM 的基础上添加一个注意力层 (attention layer). 该注意力层首先采用式 (8), 计算出不同时间序列片段特征的权重参数, 然后采用式 (9) 对不同片段特征进行加权求

和, 得到整句语音的特征表示μ, 紧接着就可以采用 Softmax 分类器来预测整句语音的情感类别.

$$\mu_t = \frac{\exp(Wh_t)}{\sum_{t=1}^{T}\exp(Wh_t)} \tag{8}$$

$$\mu = \sum_{t=1}^{T}\mu_t h_t \tag{9}$$

2 实验与结果分析

为了检验所提出方法的自然语音情感识别性能, 采用自然情感语音数据集 BAUM-1s[27] 和 CHEAVD2.0[28] 进行自然语音情感识别的实验测试.

2.1 数据集

BAUM-1s[27] 数据集是一个在 2016 年建立的音视频情感数据集. 它包括 1222 个视频样本, 来自于 30 个土耳其人. 实验中拟采用 6 种基本情感进行测试, 即生气 (Anger)、厌恶 (Disgust)、害怕 (Fear)、高兴 (Joy)、悲伤 (Sadness) 和惊奇 (Surprise). 这样, 最终收集到 520 个情感语音样本用于实验测试.

CHEAVD2.0[28] 是中科院自动化所在 2017 年为多模态情感识别竞赛时提供的音视频情感数据集. 它包含 8 种情感类型: 生气 (Anger)、厌恶 (Disgust)、害怕 (Fear)、高兴 (Joy)、悲伤 (Sadness)、惊奇 (Surprise)、担心 (Worry) 和焦虑 (Anxiety). 该数据集共有 7030 个样本, 分为 3 部分: 训练集 (4917 个样本)、验证集 (707 个样本) 和测试集 (1406 个样本). 本文使用训练集和验证集来检验所提出方法的自然语音情感识别性能, 因为测试集只对参加竞赛者开放和获取.

2.2 结果分析

表 3 和表 4 分别列出了本文方法采用数据平衡前后的试验结果的比较. 从表 3 和表 4 可见:

(1) 不管是否采用数据平衡策略, 本文采用的带注意力机制方法 (VGGish+LSTM+Attention) 取得的语音情感识别性能, 都要优于未带注意力机制方法 (VGGish+LSTM). 这表明采用注意力机制能够改善 LSTM 的特征表征能力, 因为注意力机制能够聚焦于一句语音中对情感识别起作用的目标片段区域的特征学习.

(2) 做数据平衡之后, 所有方法取得的识别性能都要明显高于数据平衡之前的结果. 在数据平衡之前, 本文方法 (VGGish+LSTM+Attention) 在 BAUM-1s 和 CHEAVD2.0 数据集分别取得了 44.09% 和 41.73%. 然而, 在数据平衡之后, 本文方法则分别在 BAUM-1s 和 CHEAVD2.0 数据集上提高了 3.18% 和 2.05%. 这说明采用数据平衡策略能够进一步改善深度学习方法的特征学习能力.

表 3 数据平衡之前的语音情感识别性能 (%)

方法	BAUM-1s	CHEAVD2.0
VGGish+LSTM	43.14	39.89
VGGish+LSTM+注意力	44.09	41.37

表 4 数据平衡之后的语音情感识别性能 (%)

方法	BAUM-1s	CHEAVD2.0
VGGish+LSTM	45.78	42.29
VGGish+LSTM+注意力	47.27	43.42

为了进一步给出每种情感类型的具体正确识别率, 图 4 和图 5 分别列出了本文方法 (VGGish+LSTM+Attention) 在数据平衡之后的两个数据集上识别结果的模糊矩阵. 由图 4 可知, 在数据平衡之后的 BAUM-1s 数据集上, 高兴 (Joy) 和悲伤 (Sadness) 的正确识别率分别达到了 64.53% 和 61.19%, 而其它 4 种情感类别的识别性能较差, 不足 40%. 由图 5 可知, 在数据平衡之后的 CHEAVD2.0 数据集上, 生气 (Anger) 和中性 (Neutral) 的正确识别率分别达到了 65.62% 和 64.5%, 而其它 6 种情感类型的正确识别率都不足 45%. 主要原因可能是这些情感类型学习到的特征表征区分度不高, 导致它们相互之间容易混淆.

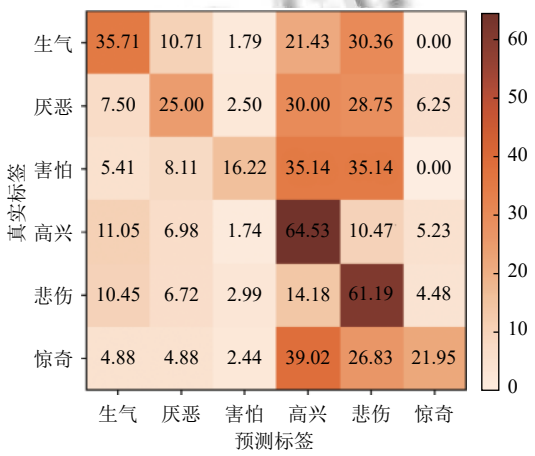

图 4 本文方法在数据平衡之后的 BAUM-1s 数据集上识别结果的模糊矩阵

为了进一步说明本文方法的优越性, 表 5 列出了本文方法取得的识别结果与其它现有文献报道结果之间的比较. 从表 5 可知, 本文方法不仅优于一些采用手工特征的方法, 如 MFCC[27], GeMAPS ADDIN EN.

CITE.DATA [28−31], 也优于一些基于深度卷积神经网络 CNN 的方法, 如采用 AlexNet 的跨模态迁移方法[22,32]. 以及与含有 20 层的 ResNet 网络的自训练网络方法[33] 获得的准确率相近.

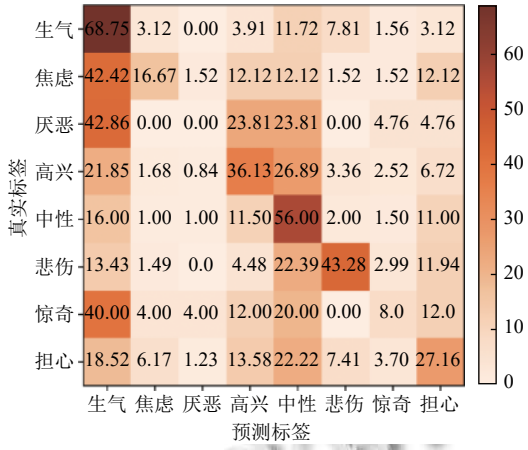

图 5 本文方法在数据平衡之后的 CHEAVD2.0 数据集上识别结果的模糊矩阵

表 5 与现有文献报道的结果比较

数据集	文献	特征	识别率(%)
BAUM-1s	Zhalehpour S[27]	MFCC	29.41
	Zhang S[22]	CNN	42.46
	Ma Y[32]	CNN	42.38
	本文方法	CNN+LSTM+注意力+数据平衡	**47.27**
CHEAVD2.0	Li Y [28]	GeMAPS	39.90
	Xi YX[33]	CNN(ResNet)	**43.96**
	Miao HT[31]	GeMAPS	40.31
	本文方法	CNN+LSTM+注意力+数据平衡	43.42

3 结论与展望

本文提出了一种结合数据平衡和注意力机制的卷积神经网络 (CNN) 和长短时记忆单元 (LSTM) 的语音情感识别方法, 采用数据平衡方法对语音情感数据集中每个类别的样本进行相应的预处理, 再采用预训练好的 VGGish 网络在目标数据集上进行微调, 从而学习出分段语音特征, 再通过带有注意力机制的 LSTM 从分段语音特征中学习对应的高阶判别性特征, 最后通过 Softmax 层进行情感分类. 本文在两个自然语音情感数据集 BAUM-1s 和 CHEAVD2.0 中的实验证明了提出的语音情感识别方法有较好的识别性能. 由于本文提出的网络模型的训练方式并不是端到端的训练方式, 因此, 在未来的工作中, 希望能以端到端的方式训练网络模型; 由于高阶注意力生成的全局性特征更具表达力, 并能辅助网络模型对语音信息长范围的相关性进行建模, 因此, 在未来的工作中, 希望能将二阶注意力机制或更高阶的注意力机制与现有的网络模型进行结合, 以学习出更具判别性的语音情感特征.

参考文献

1 韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述. 软件学报, 2014, 25(1): 37–50. [doi: 10.13328/j.cnki.jos.004497]

2 张石清, 李乐民, 赵知劲. 人机交互中的语音情感识别研究进展. 电路与系统学报, 2013, 18(2): 440–451, 434.

3 Schuller BW. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM, 2018, 61(5): 90–99. [doi: 10.1145/3129340]

4 Akçay MB, Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 2020, 116: 56–76. [doi: 10.1016/j.specom.2019.12.001]

5 Song P. Transfer linear subspace learning for cross-corpus speech emotion recognition. IEEE Transactions on Affective Computing, 2019, 10(2): 265–275. [doi: 10.1109/TAFFC.2017.2705696]

6 Demircan S, Kahramanli H. Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech. Neural Computing and Applications, 2018, 29(8): 59–66. [doi: 10.1007/s00521-016-2712-y]

7 Zhao XM, Zhang SQ. Spoken emotion recognition via locality-constrained kernel sparse representation. Neural Computing and Applications, 2015, 26(3): 735–744. [doi: 10.1007/s00521-014-1755-1]

8 Gharavian D, Sheikhan M, Nazerieh A, *et al*. Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Computing and Applications, 2012, 21(8): 2115–2126.

9 Zhang ZX, Coutinho E, Deng J, *et al*. Cooperative learning and its application to emotion recognition from speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(1): 115–126.

10 朱菊霞, 吴小培, 吕钊. 基于 SVM 的语音情感识别算法. 计算机系统应用, 2011, 20(5): 87–91. [doi: 10.3969/j.issn.1003-3254.2011.05.019]

11 Kayaoglu M, Eroglu Erdem C. Affect recognition using key frame selection based on minimum sparse reconstruction. Proceedings of the 2015 ACM on International Conference

on Multimodal Interaction. Seattle, WA, USA. 2015. 519–524.
12 Schuller B, Steidl S, Batliner A, *et al*. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association. Lyon, France. 2013. 148–152.
13 Valstar M, Schuller B, Smith K, *et al*. AVEC 2013: The continuous audio/visual emotion and depression recognition challenge. Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge. Barcelona, Spain. 2013. 3–10.
14 Eyben F, Scherer KR, Schuller BW, *et al*. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 2016, 7(2): 190–202. [doi: 10.1109/TAFFC.2015.2457417]
15 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507. [doi: 10.1126/science.1127647]
16 Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Siem Reap, Cambodia. 2012. 1097–1105.
17 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
18 Han K, Yu D, Tashev I. Speech emotion recognition using deep neural network and extreme learning machine. 15th Annual Conference of the International Speech Communication Association. Singapore. 2014. 223–227.
19 Mao QR, Dong M, Huang ZW, *et al*. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia, 2014, 16(8): 2203–2213. [doi: 10.1109/TMM.2014.2360798]
20 Trigeorgis G, Ringeval F, Brueckner R, *et al*. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China. 2016. 5200–5204.
21 Huang CW, Narayanan SS. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. 2017 IEEE International Conference on Multimedia and Expo (ICME). Hong Kong, China. 2017. 583–588.
22 Zhang SQ, Zhang SL, Huang TJ, *et al*. Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(10): 3030–3043. [doi: 10.1109/TCSVT.2017.2719043]
23 Zhang SQ, Zhao XM, Tian Q. Spontaneous speech emotion recognition using multiscale deep convolutional LSTM. IEEE Transactions on Affective Computing, 2019. [doi: 10.1109/TAFFC.2019.2947464]
24 黎万义, 王鹏, 乔红. 引入视觉注意机制的目标跟踪方法综述. 自动化学报, 2014, 40(4): 561–576.
25 孙小婉, 王英, 王鑫, 等. 面向双注意力网络的特定方面情感分析模型. 计算机研究与发展, 2019, 56(11): 2384–2395. [doi: 10.7544/issn1000-1239.2019.20180823]
26 Hershey S, Chaudhuri S, Ellis DP, *et al*. CNN architectures for large-scale audio classification. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA. 2017. 131–135.
27 Zhalehpour S, Onder O, Akhtar Z, *et al*. BAUM-1: A spontaneous audio-visual face database of affective and mental states. IEEE Transactions on Affective Computing, 2017, 8(3): 300–313. [doi: 10.1109/TAFFC.2016.2553038]
28 Li Y, Tao JH, Schuller B, *et al*. MEC 2017: Multimodal emotion recognition challenge. 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia). Beijing, China. 2018. 1–5.
29 Campos V, Jou B, Giró-i-Nieto X. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. Image and Vision Computing, 2017, 65: 15–22. [doi: 10.1016/j.imavis.2017.01.011]
30 Oquab M, Bottou L, Laptev I, *et al*. Learning and transferring mid-level image representations using convolutional neural networks. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 1717–1724.
31 Miao HT, Zhang YF, Li WP, *et al*. Chinese multimodal emotion recognition in deep and traditional machine leaming approaches. 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia). Beijing, China. 2018. 1–6.
32 Ma YX, Hao YX, Chen M, *et al*. Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. Information Fusion, 2019, 46: 184–192. [doi: 10.1016/j.inffus.2018.06.003]
33 Xi YX, Li PC, Song Y, *et al*. Speaker to emotion: Domain adaptation for speech emotion recognition with residual adapters. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Lanzhou, China. 2019. 513–518.