计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(1):296−301 [doi: 10.15888/j.cnki.csa.008769]

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

基于 Attention-CGRU 网络的中文语音情感识别①

王茂林, 郝　刚

(天津理工大学 计算机科学与工程学院, 天津 300010)
通信作者: 郝　刚, E-mail: ganghao@email.tjut.edu.cn

摘　要: 正确识别语音中包含的情感信息可以大幅提高人机交互的效率. 目前, 语音情感识别系统主要由语音特征抽取和语音特征分类两步组成. 为了提高语音情感识别准确率, 选用语谱图而非传统声学特征作为模型输入, 采用基于 attention 机制的 CGRU 网络提取语谱图中包含的频域信息和时域信息. 实验结果表明: 在模型中引入注意力机制有利于减少冗余信息的干扰, 并且相较于基于 LSTM 网络的模型, 采用 GRU 网络的模型预测精确度更高, 且在训练时收敛更快, 与基于 LSTM 的基线模型相比, 基于 GRU 网络的模型训练时长只有前者的 60%.

关键词: 语音情感识别; 注意力机制; 门控循环单元; 语谱图; 深度学习

引用格式: 王茂林,郝刚.基于 Attention-CGRU 网络的中文语音情感识别.计算机系统应用,2023,32(1):296–301. http://www.c-s-a.org.cn/1003-3254/8769.html

Chinese Speech Emotion Recognition Based on Attention-CGRU Network

WANG Mao-Lin, HAO Gang

(School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300010, China)

Abstract: Accurate recognition of speech emotion information can help to greatly improve the efficiency of human-computer interaction. At present, the speech emotion recognition system mainly consists of two steps: speech feature extraction and speech feature classification. In order to improve the accuracy of speech emotion recognition, the spectrogram is used as the model input instead of traditional acoustic features, and the CGRU network based on the attention mechanism is adopted to extract the frequency domain and time domain information in the spectrogram. The experimental results show that the introduction of the attention mechanism in the model is beneficial to reduce the interference of redundant information, and compared with the model based on the LSTM network, the model using the GRU network can fast converge during training and has higher prediction accuracy. In addition, the training time of the GRU-based model is only 60% of that of the LSTM-based baseline model.

Key words: speech emotion recognition; attention mechanism; gate recurrent unit (GRU); spectrogram; deep learning

语音是人们生活中常见的交流方式之一, 其中除了语言信息外还包含大量非语言信息. 如果想建立一个真正智能的语音交互系统, 除了需要识别语音中的语言信息, 还需要理解这些非语言信息. 语音情感识别的研究, 是实现这一目标的强大助推力. 因此, 语音情感识别 (speech emotion recognition, SER) 被众多研究者关注.

典型的语音情感识别系统, 主要有 3 部分内容[1], 分别是合适的语音情感数据库、提取有效的语音情感特征和特征分类算法. 选择合适的声学特征作为语音情感识别系统的输入, 有助于提高系统的识别准确率. 传统声学特征有振幅、MFCC、基音频率等. 宋春晓使用 MFCC 特征在 EMO-DB 数据库上得到了 82.47% 的准确率[2]. 目前的研究表明, 语音中包含的时域信息

① 收稿时间: 2022-02-10; 修改时间: 2022-03-03, 2022-03-18; 采用时间: 2022-03-28; csa 在线出版时间: 2022-10-28
CNKI 网络首发时间: 2022-11-15

和频域信息同样重要, 基于短时傅里叶方法生成的语谱图, 既包含语音的频域信息, 又包含传统声学特征没有的时域信息, 因此语谱图也被广泛用于语音情感识别研究.

在对声学特征做分类时, 传统的方法有支持向量机和隐马尔科夫模型[3,4]. 随着深度学习技术的快速发展, 越来越多的研究人员将深度学习应用于语音情感识别. Kim 等[5] 通过深度神经网络来提取语音中包含的情感特征. Lee 等[6] 利用双向长短期记忆 (BiLSTM) 网络强大的上下文学习能力, 创建了一个准确率达到 63.89% 的语音情感识别系统. Satt 等[7] 基于 LSTM 网络提出一种高效的语音情感特征分类算法, 并使用语谱图作为模型输入, 提高了模型的识别精度. 薛艳飞等[8] 将多任务学习用于语音情感识别, 在离散情感语料库上系统识别准确率达到了 75.83%.

虽然相较于传统方法, 将深度学习应用在语音情感识别领域, 可以大幅提高识别准确率, 但深度学习网络结构复杂, 参数量大, 尤其是基于 LSTM 的语音情感识别模型, 在训练时模型收敛慢, 训练时间长. 本文在前人研究的基础上, 提出一种 Attention-CGRU 模型, 模型采用从原始语音信号提取到的语谱图作为输入, 通过卷积神经网络提取语谱图中包含的高阶情感特征. 模型通过引入注意力机制, 减少冗余信息对识别准确率的影响, 并使用门控循环单元[9,10] 来代替常用于语音情感识别系统中的长短期记忆网络[11,12], 使得模型更高效, 减小模型训练成本, 提高特征识别率.

1 语音情感识别

1.1 GRU

GRU 属于循环神经网络[13] 的一种, 与 LSTM 功能类似. 在很多应用场景下, GRU 和 LSTM 的表现相差无几, 但二者控制输出值的方式不同. LSTM 通过输入门、遗忘门和输出门这 3 个门函数来控制输入值、记忆值和输出值, 而 GRU 通过重置门和更新门来控制前一时刻的状态对现在所处状态的影响力大小. LSTM 和 GRU 都是通过各种门函数来保留上一时刻的输出特征, 但与 LSTM 相比, GRU 参数更少, 模型训练过程中迭代速度更快, 更易于计算, 很大程度上提高了模型的训练效率. 图 1 是 GRU 的基本结构.

重置门和更新使用当前时间步输入 X_t 与上一时间步隐藏状态 H_{t-1} 作为输入, 并通过激活函数为 Sigmoid 函数的全连接层计算输出. 重置门 R_t 和更新门 Z_t 的计算方式如下:

$$R_t = \sigma(X_t W_{xr} + H_{t-1} W_{hr} + b_r) \tag{1}$$

$$Z_t = \sigma(X_t W_{xz} + H_{t-1} W_{hz} + b_z) \tag{2}$$

其中, W_{xr}, W_{hr} 和 W_{xz}, W_{hz} 是模型的权重参数, 重置门和更新门都是通过输入 X_t 和上一个时刻的隐藏状态 H_{t-1} 得到, b_r 和 b_z 则是偏置参数.

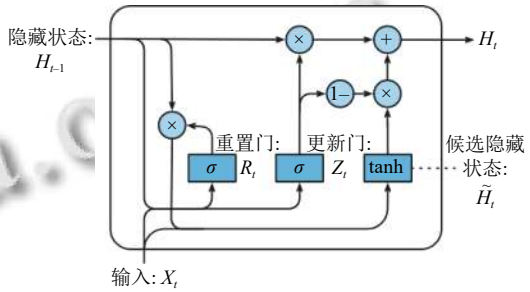

图 1 GRU 的基本结构

在计算辅助隐藏状态时需要用到候选隐藏状态$\tilde{H}_t$, 候选隐藏状态是通过将当前时刻重置门的输出 R_t 和上一时刻的隐藏状态 H_{t-1} 做元素乘法, 其计算方式如下:

$$\tilde{H}_t = \tanh(X_t W_{xh} + (R_t \cdot H_{t-1}) W_{hh} + b_h) \tag{3}$$

当前时刻 t 的隐藏状态的计算需要用到当前时刻的更新门输出 Z_t 和上一时刻的隐藏状态输出 H_{t-1}, 以及当前时刻计算得到的候选隐藏状态输出$\tilde{H}_t$, 计算方式如下:

$$H_t = Z_t \cdot H_{t-1} + (1 - Z_t) \cdot \tilde{H}_t \tag{4}$$

1.2 注意力机制

注意力机制[14,15] 就是在信息处理的过程中, 对不同的内容分配不同的注意力权重, 即分配更多的注意力给信息的关键部分. 在输入给模型原始信息中, 含有大量对情感识别没有帮助无效信息, 将注意力机制引入语音情感识别的主要目的是让模型对语音中的关键信息分配更大的注意力权重, 剔除语音中的冗余信息的影响. 图 2 为注意力机制结构示例图.

在本文的基于注意力机制的 CGRU 模型中, 将经过 GRU 的输出 $H_t = \{h_1, h_2, h_3, \cdots, h_L\}$作为 self-attention 层的输入, 其中 L 是输入信号的长度, 模型会在计算前初始化一个参数向量 W, $W = (w_1, w_2, w_3, \cdots, w_L)$, 其中每个 w_i 代表每一帧输入信号的权重大小, 每一帧权重的计算公式为:

$$\alpha_i = \frac{\exp\left(w^{\mathrm{T}} y_t\right)}{\sum \exp\left(w^{\mathrm{T}} y_t\right)} \tag{5}$$

Attention 层会给 GRU 的输出计算出一个注意力分布, 在最后的情感分类阶段, 根据这些注意力分布能够更有选择性提取信息. 在模型计算时的表现为根据注意力权重矩阵, 对输入信息进行加权求和, 最终得到模型的分类结果. 在引入 attention 层后, 可以有效过滤掉 GRU 层的输出中包含的冗余信息.

1.3 基于 attention 机制的 CGRU 网络

基于注意力机制的 CGRU 中文语音情感分类模型整体框架如图 3 所示, 模型以从原始语音数据中提取的二维语谱图作为输入信号. 通过卷积神经网络完成初步的特征提取, 再经过由 GRU 和 attention 组成的网络主体部分, 最后经过一个全连接层得到分类结果. 初步的特征提取通过两层 CNN 网络完成, 第 1 个卷积层 (3 维 conv) 的过滤器大小设为 (3, 5, 2), 步长设为 1. 输入信号经过第 1 层卷积后再经过一次最大池化操作, 数据沿通道方向的维度从最初的 3 变为 1, 再经过过滤器大小为 (3, 5) 的第 2 个卷积层, 接着再经过一层池化层. 经过卷积神经网络完成初步的特征提取后, 将提取到的高级特征作为输入送到 GRU 和 attention 组成的分类网络中, 完成特征分类.

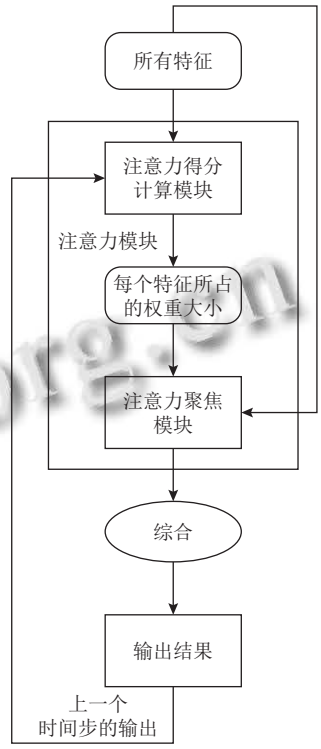

图 2 Attention 机制结构示意图

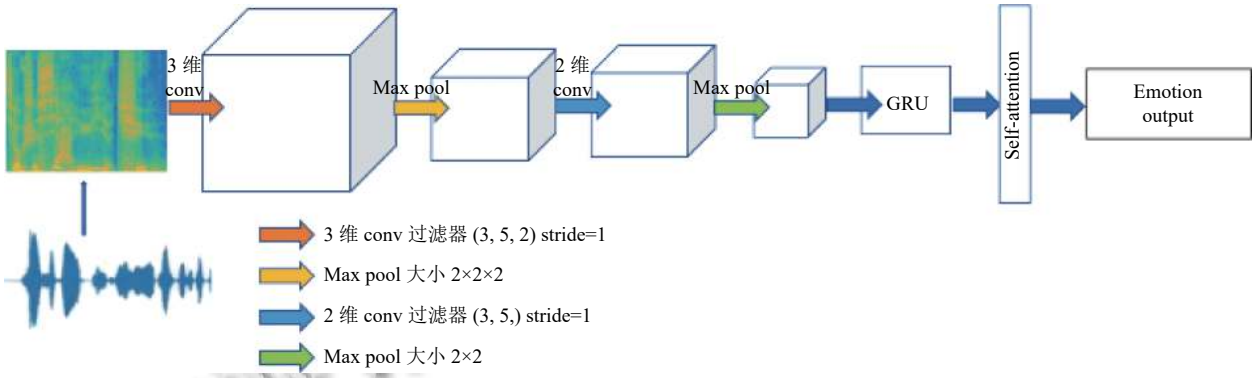

图 3 基于 attention 机制的 CGRU 网络结构图

2 语音数据库的选取与语谱图生成

2.1 语音情感数据库

选择一个合适的情感语音数据库直接关系到后续语音特征的提取以及语音情感识别系统的正确率. 本文选用中国科学院自动化研究所录制的 CASIA 汉语情感语料库进行实验. CASIA 中所有的语音样本是由 4 个专业发音人录制的, 共包含 6 种情绪, 分别为生气、高兴、害怕、悲伤、惊讶和中性, 语音样本总量为 9 600 句.

2.2 语音特征提取

语谱图与只包含语音频域信息的传统声学特征不同, 可以反映原始声音信号频谱随时间的变化. 图 4 是从原始语音数据中获取语谱图[16] 的流程图.

首先获取不同语音信号在整个时域内的频域信息, 再将整个时间段分为一个个小的时间段 (frames), 通过短时傅里叶变换输出语音信号的频谱矩阵, 再对频谱矩阵进行逐位平方, 将幅度转化为功率, 最终得到语音信号对应的语谱图. 式 (6) 为频谱矩阵计算式:

$$F(\omega)=\tau\cdot\frac{Sa(\tau-\omega)}{2} \tag{6}$$

下面使用 CASIA 数据集中“就算下雨也去”这句话举例, 图 5 是这句话分别在在 sad、fear、surprise 和 angry 下的语谱图结果. 语谱图的横轴表示时间, 纵轴表示语音频率, 语谱图中每个点的坐标值表示该语音信号的数据能量, 由于语谱图是在二维平面表示三维信息, 因此该能量值是通过每个坐标点的颜色表示, 坐标点颜色越深则能量越大.

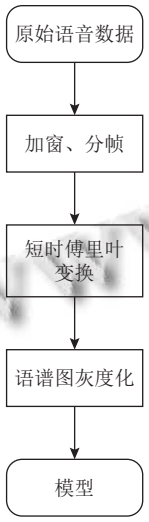

图 4 语谱图生成流程

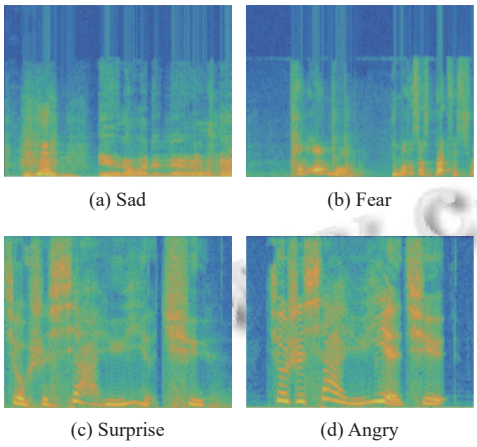

图 5 语谱图示例

3 实验分析

在本次实验中, 模型训练时学习率[17] 设为 0.005, 损失函数选用 Cross Empty Loss, 优化器选用 Adam[18], batchsize 设为 64, epoch 设为 100. 为了避免由于训练数据和测试数据划分引起的模型准确率变化, 采用 K 折交叉验证 (K-fold cross-validation)[19] 的方式评估模型性能. 模型性能评估指标使用准确率 (accuracy)、召回率 (recall)、精确率 (precision) 和 $F1$ 值. 使用混淆矩阵来描述实验结果, 即将模型预测出来的每一条语音信息在实际语音情感样本中所占的比例用一个二维向量矩阵表示. 将混淆矩阵列向量中被语音样本正确识别出来的数目表示为 TN, 将混淆矩阵行向量中被语音样本错误分类的数目表示为 FP, 将混淆矩阵列向量中被语音样本错误分类的数目表示为 FN, 将所有混淆矩阵行向量中语音样本数表示为 P, 将所有混淆矩阵列向量中的语音样本数表示为 N.

为了验证本文提出的基于 attention 机制 CGRU 模型的实际效果, 一共进行了 3 组对比实验, 第 1 组实验将本文提出的基于 attention 机制的 CGRU 模型与未采用 attention 机制的模型对比. 第 2 组实验将本文提出的模型与基线模型对比. 第 3 组实验验证了不同超参数 (学习率和优化器) 的选取对模型性能的影响. 图 6 所示的混淆矩阵是本文提出的基于 attention 机制的 CGRU 模型的实验结果, 图 7 是模型训练时的损失变化曲线.

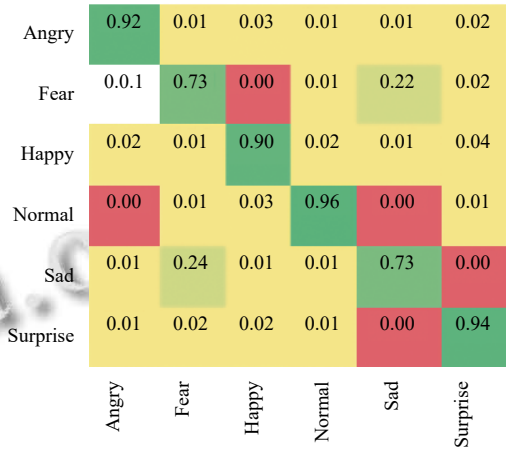

图 6 基于 attention 机制的 CGRU 模型实验结果

3.1 注意力机制有效性验证实验

为了验证 attention 机制在模型学习语音情感特征时的有效性, 将未加入 attention 机制的 CGRU 模型与基于 attention 机制的 CGRU 模型做对比. 从而验证注意力机制在语音情感识别中是否可以提高语音情感识别系统的准确率. 图 8 是未加入 attention 机制的 CGRU 模型的实验结果. 图 9 是基于 attention 机制 CGRU 模型与未加入 attention 机制的 CGRU 模型在 CASIA 中各类情绪样本上模型预测的准确率对比图.

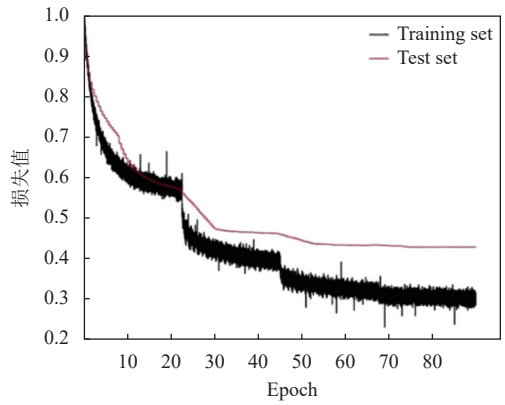

图 7 基于 attention 机制的 CGRU 模型训练时损失值变化

图 8 未加入 attention 机制的 CGRU 模型实验结果

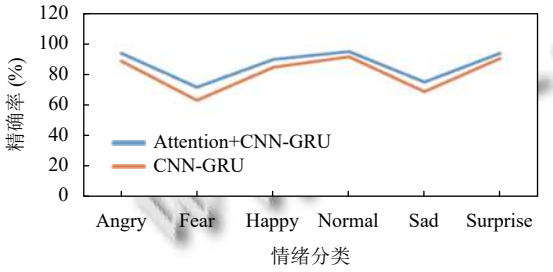

图 9 注意力机制有效性验证实验结果对比图

从对比结果可以看出, 在加入 attention 机制后, 模型在 6 类情绪的测试样本下, 预测精确率都有所提升, 其中模型在害怕 (fear) 这类测试样本中, 精确度提升最为明显, 加入注意力机制后模型精确率提升了 8 个百分点. 表明加入 attention 机制可以更有效地提取语音信号内的高阶情感特征, 并消除冗余特征的影响, 提高模型的检测精度.

3.2 与基线模型实验

本文选用基于 LSTM 的语音情感识别模型作为基线模型, 该模型也是现在语音情感识别领域的主流模型. 将本文提出的基于 attention 机制的 CGRU 模型与基线模型在各个情绪分类下的 $F1$ 值对比, 横轴表示各个情绪分类, 纵轴标识 $F1$ 值. 从图 10 中可以看到, 本文提出的基于 attention 机制的 CGRU 模型与其他基线模型相比, 在各个情绪分类上的 $F1$ 值都高于平均水平, 且通过 Attention-LSTM 模型[20] 与 DCNN-LSTM 模型的对比, 再一次说明了将 attention 机制引入语音情感识别上后, 对模型效果带来了显著提升, 可以更好地提取语音中的情感特征. 此外, 在相同的训练 epoch 和 batchsize 下, 本文提出的基于 attention 机制的 CGRU 模型训练所花费的时间只有基于 LSTM 的基线模型的 60%, 可以看到使用 GRU 有利于提高模型的训练效率, 降低模型训练成本.

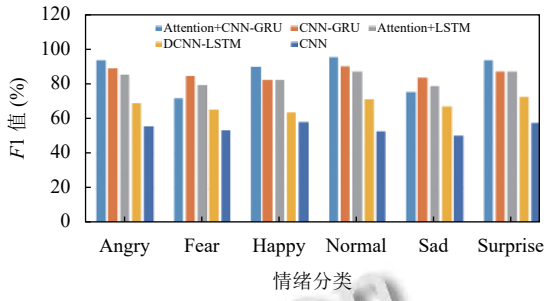

图 10 与基线模型的 $F1$ 值对比图

3.3 超参数对模型性能影响实验

为了测试不同超参数 (优化器和学习率) 对模型的影响, 模型训练时的 batchsize 和 epoch 都保持一致, batchsize 设为 64, epoch 设为 100, 调整模型训练时的学习率大小以及使用的优化其类型. 图 11 是模型在不同优化器和学习率 (lr) 下得到的平均精确率和平均召回率对比.

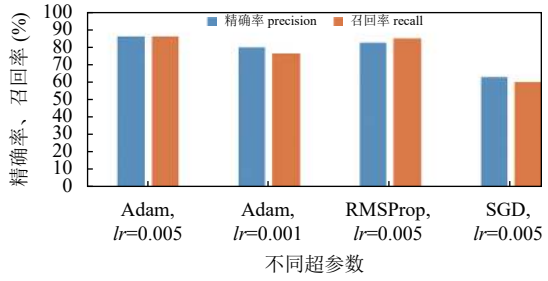

图 11 不同超参数下模型精确率对比图

通过图 11 可以看出, 在其他参数确定的情况下, 不同超参数的选择对模型性能有着不同的影响, 通过

精确率和召回率的综合考虑, 模型优化器选用 Adam, 学习率选择 0.005 时模型效果最佳, 而优化器 SGD[21,22] 不适合本文提出的模型结构.

4 结论与展望

本文提出了一种基于 attention 机制的 CGRU 的语音情感识别模型, 模型使用原始语音中提取的既包含频域信息又包含时域信息的语谱图作为模型输入, 首先通过卷积神经网络提取语谱图中深层次的语音特征, 再通过 GRU 网络完成高阶情感特征分类, 通过对比实验研究发现, 将 attention 机制加入到模型中, 使得模型更加关注语音中的关键信息, 消除冗余信息的影响, 提高模型的精确率. 本文提出的模型与其他基于 LSTM 的模型相比, 在训练时模型收敛更快, 在 CASIA 汉语情感语料库上, 本文提出的模型识别的精确率更高. 但仍存在跨语料库识别时精度降低的问题, 后续研究工作是提高模型泛化能力, 以及跨语料库的语音情感识别研究.

参考文献

1 张雪英, 孙颖, 张卫, 等. 语音情感识别的关键技术. 太原理工大学学报, 2015, 46(6): 629–636, 643.

2 宋春晓. 情感语音的非线性特征提取及特征优化的研究 [硕士学位论文]. 太原: 太原理工大学, 2018.

3 Hu H, Xu MX, Wu W. GMM supervector based SVM with spectral features for speech emotion recognition. Proceedings of 2007 IEEE International Conference on Acoustics, Speech and Signal Processing. Honolulu: IEEE, 2007. IV-413–IV-416.

4 Lin ZH, Feng MW, Santos CND, *et al*. A structured self-attentive sentence embedding. Proceedings of the 5th International Conference on Learning Representations. Toulon: ICLR, 2017. 1–15.

5 Kim Y, Lee H, Provost EM. Deep learning for robust feature generation in audiovisual emotion recognition. Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013. 3687–3691.

6 Lee J, Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition. Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden: ISCA, 2015. 1537–1540.

7 Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms. Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm: ISCA, 2017. 1089–1093.

8 薛艳飞, 毛启容, 张建明. 基于多任务学习的多语言语音情感识别方法. 计算机应用研究, 2021, 38(4): 1069–1073.

9 张小川, 刘连喜, 戴旭尧, 等. 基于词性特征的 CNN_BiGRU 文本分类模型. 计算机应用与软件, 2021, 38(11): 155–161. [doi: 10.3969/j.issn.1000-386x.2021.11.024]

10 朱星浩, 胥备. 基于 GRU 算法的音乐和词语的情感语义匹配算法. 计算机技术与发展, 2021, 31(11): 46–51. [doi: 10.3969/j.issn.1673-629X.2021.11.008]

11 翟社平, 杨媛媛, 邱程, 等. 基于注意力机制 Bi-LSTM 算法的双语文本情感分析. 计算机应用与软件, 2019, 36(12): 251–255. [doi: 10.3969/j.issn.1000-386x.2019.12.040]

12 李金宇, 王晓晔, 彭宪, 等. 基于双向 LSTM 的文本情感倾向分类. 计算机科学与应用, 2021, 11(5): 1401–1410.

13 Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 2020, 404: 132306. [doi: 10.1016/j.physd.2019.132306]

14 陈海涵, 吴国栋, 李景霞, 等. 基于注意力机制的深度学习推荐研究进展. 计算机工程与科学, 2021, 43(2): 370–380.

15 任欢, 王旭光. 注意力机制综述. 计算机应用, 2021, 41(S1): 1–6. [doi: 10.11772/j.issn.1001-9081.2020101634]

16 陶华伟, 查诚, 梁瑞宇, 等. 面向语音情感识别的语谱图特征提取算法. 东南大学学报 (自然科学版), 2015, 45(5): 817–821. [doi: 10.3969/j.issn.1001-0505.2015.05.001]

17 贺昱曜, 李宝奇. 一种组合型的深度学习模型学习率策略. 自动化学报, 2016, 42(6): 953–958. [doi: 10.16383/j.aas.2016.c150681]

18 Kingma DP, Ba J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2014.

19 Srinivasan K, Cherukuri AK, Vincent DR, *et al*. An efficient implementation of artificial neural networks with K-fold cross-validation for process optimization. Journal of Internet Technology, 2019, 20(4): 1213–1225.

20 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. [doi: 10.15888/j.cnki.csa.007917]

21 Ruder S. An overview of gradient descent optimization algorithms. arXiv:1609.04747, 2016.

22 仝卫国, 李敏霞, 张一可. 深度学习优化算法研究. 计算机科学, 2018, 45(S2): 155–159. [doi: 10.11896/j.issn.1002-137X.2018.11A.029]

(校对责编: 孙君艳)