胡文星, 蔡佳欣, 柯振宇,等. 基于注意力机制的 BiLSTM 动物声音情绪识别[J]. 智能计算机与应用,2024,14(7):57-63.
DOI:10.20169/j.issn.2095-2163.240708

# 基于注意力机制的 BiLSTM 动物声音情绪识别

胡文星, 蔡佳欣, 柯振宇, 彭烁钟, 胡 松, 赵小燕
(南京工程学院 信息与通信工程学院, 南京 211167)

**摘 要**: 声音是动物向外界表达情绪的一种重要方式,通过提取动物声音特征,建立特征值与动物情绪之间的映射关系,可以实现对动物情绪的感知和理解。为提高动物情绪识别性能,本文提出基于 Bahdanau 注意力机制的双向长短期记忆网络的动物声音情绪识别方法。该方法对动物声音进行特征提取,提取了的频谱质心、频谱带宽、频谱滚降点、过零率、均方根能量、频谱对比度、梅尔倒谱系数以及其一阶差分作为特征向量,输入双向长短期记忆网络,通过注意力机制对情绪特征进行通道方向的权重学习,最后由全连接层进行情感类型判别。本文以狗为例,对狗的声音进行了情绪识别实验,实验结果表明:相比于循环神经网络、双向长短期记忆网络,本文方法的识别准确度更高。

**关键词**: 动物情绪识别; Bahdanau 注意力机制; 双向长短期记忆网络; 特征提取
**中图分类号**: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2024)07-0057-07

## Emotion recognition of animal voices based on attention mechanism BiLSTM

HU Wenxing, CAI Jiaxin, KE Zhenyu, PENG Shuozhong, HU Song, ZHAO Xiaoyan
(School of Information and Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China)

**Abstract**: Sound serves as a crucial mean for animals to express their emotions to the outside world. By establishing the mapping relationship between animal's emotions and features extracted from animal sound, it becomes possible for computers to perceive and understand the emotional states of animals. In order to improve the performance of animal emotion recognition, a method for recognizing emotions in animal sounds based on the Bi-directional Long Short-Term Memory (BiLSTM) network with Bahdanau attention mechanism is presented in this paper. Feature extraction of the proposed method such as the spectral centroid, spectral bandwidth, spectral rolloff point, zero-crossing rate, root mean square energy, spectral contrast, Mel-frequency cepstral coefficients and their first-order differences, forming a feature vector from animal sound. The feature vector is treated as the input of BiLSTM network. Through the attention mechanism, the proposed method learns channel-wise weights for emotional features. Ultimately, a fully connected layer is utilized for the classification of emotional categories. Taking dogs as an example, experiments are conducted to recognize emotions in dog sounds. The experimental results demonstrate that the proposed method outperforms the methods based on Recurrent Neural Networks and BiLSTM networks with higher accuracy in emotion recognition.

**Key words**: animal emotion recognition; Bahdanau attention mechanism; bidirectional long short-term memory network; feature extraction

## 0 引 言

随着动物行为学研究的不断发展,越来越多的研究者开始关注动物情绪表达方面的问题。声音是动物向外界表达情绪的一种重要方式,通过提取动物声音特征,建立特征值与动物情绪之间的映射关系,可以实现对动物情绪的感知和理解,实时了解动物的情绪状态和情感需求,更好地维护动物的身心健康。

语音情感识别是指对人类语音进行情感分析,对语音信号提取丰富的情感特征,通过人类语音信号检测和识别如喜悦、愤怒、悲伤、惊讶、恐惧等多种情感类别[1]。20 世纪 80 年代中期,提出语音情感技术,开创了使用声学统计特征进行情感分类的先

**基金项目**: 江苏省大学生实践创新训练计划项目(202311276087Y); 南京工程学院引进人才科研启动基金项目(YKJ202019)。
**作者简介**: 胡文星(2004-),女,本科生,主要研究方向:信号处理,深度学习。
**通讯作者**: 赵小燕(1986-),女,博士,讲师,硕士生导师,主要研究方向:信号处理,深度学习。Email:xiaoyanzhao@njit.edu.cn
**收稿日期**: 2024-02-09

河[2]。近年来,随着深度学习的发展,越来越多的研究者基于深度神经网络识别人类语音情感。1999 年,Moriyama[3] 提出语音和情感之间的线性关联模型,将语音情感初步应用于电子商务中;乔冠楠等[4]提取声音特征参数与听觉参数,基于人工神经网络进行语音情感交叉识别。相较于传统的机器学习,研究者更加倾向于多种模型的融合,使深度学习模型达到更好的性能。Mirsamadi 等[5]提出将注意力机制应用于语音情感识别;Yoon 等[6]提出一种用于语音情感识别的多跳注意模型,使用 BiLSTM 从语音数据中提取隐藏信息,应用多跳注意模型产生最终的情感分类权重。彭祝亮等[7]提出一种结合双向长短记忆网络和方面注意力模块的情感分类方法。

动物声音与人类语音相似,可实现同物种之间的交流。利用动物声信号来识别动物情绪已成为一个新兴的研究方向。近年来,人工智能领域的机器学习和机器翻译技术为动物声音情绪识别带来新契机[8]。在《科学报告》上,一个国际科学家团队发表了一篇《Pig Grunts Reveal Their Emotions》,其中分析了猪从出生到死亡的数千个原声录音,最终将猪的呼噜声和其情绪联系在了一起,该研究发现利用人工智能识别动物情绪的正确率高达 92%。2008 年,MOLNÁR 等[9]应用机器学习的方法对狗的 6 种不同行为相对应的叫声进行了分析;2011 年,郭龙祥等[10]运用声学技术对鲸豚类动物声信号进行分析与识别以实现对海洋生物的保护;2018 年杨春勇等[11]公开了一种动物声音情绪识别系统及其方法;2020 年,黄玮等[12]基于 DWT (Discrete Wavelet Transformation)距离对动物声音展开研究,发现基于 DWT 距离算法的分类结果能反映动物的基本特征,且每种动物用于实验的声音的数量越多则分类的效果越理想;2022 年,石鑫鑫等[13]提出一种全连接算法与稀疏连接算法相结合的全卷积神经网络,用于蛙声识别;杨兴海等[14]把动物声音转换为数字信号,通过将标记有声音特征的动物声音输入到卷积神经网络中进行训练,从而得到训练好的卷积神经网络模型,实现了根据动物声音识别其情绪的效果。

为了解决当前动物声音情绪识别存在的维度不足和识别率低的问题,受人类语音情感识别的启发,本文提出了一种基于注意力机制的 BiLSTM 的动物声音情绪识别方法,该方法提取动物声音信号的多维特征参数,建立识别模型,实现对生气、受伤、伤心 3 种情绪的识别。

## 1 动物声音情绪识别模型

基于注意力机制的 BiLSTM 的动物情绪识别分为训练和测试两个阶段。在训练阶段,首先对采集的动物声音情绪训练样本进行预处理,然后提取训练样本的多种特征参数,融合多种特征来训练神经网络分类器的参数。在测试阶段,使用训练好的分类器进行情绪识别,最大后验概率对应的情绪即为识别结果。本文算法的系统框图如图 1 所示。

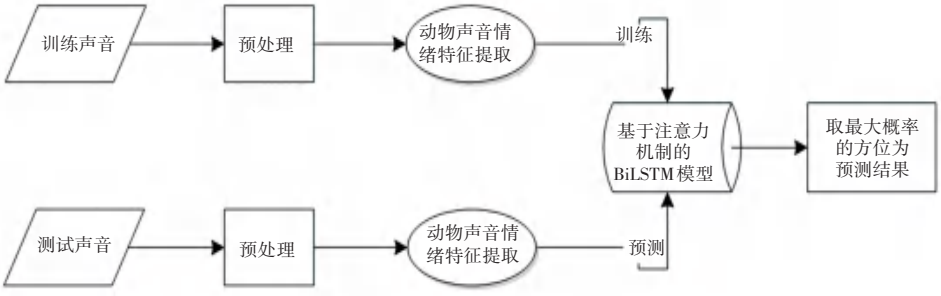


**图 1 基于注意力机制的 BiLSTM 动物情绪识别算法框图**

**Fig. 1 Block diagram of BiLSTM animal emotion recognition algorithm based on attention mechanism**

### 1.1 特征提取

从动物的声音信号中提取有效的情绪特征是情绪识别的关键,本文提取多种声音特征参数以提高算法的识别性能,特征提取过程如图 2 所示。首先,对声音信号分帧、加窗;其次,进行短时傅里叶变换得到声音信号频谱,从中计算频谱质心、频谱对比度、频谱带宽、频谱滚降;令信号的线性幅度通过美尔滤波器,对滤波器输出取对数,再进行 DCT (Discrete Cosine Transform)变换,得到美尔倒谱系数(Mel-Frequency Cepstral Coefficients, MFCC)。本算法提取的特征是一个 46 维的向量,其中包括 20 维的 MFCC 参数,20 维的 MFCC 一阶差分动态特征,以及频谱质心、频谱对比度、频谱带宽、频谱滚降、均方根能量和过零率。

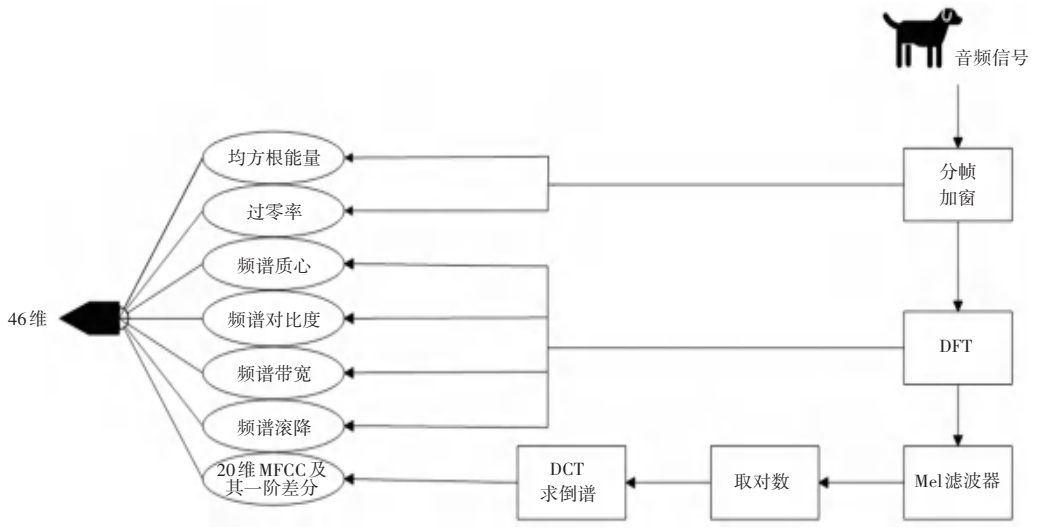


**图 2　动物声音信号情绪特征提取**

**Fig. 2　Emotional feature extraction of animal sound signals**

对动物声音信号做分帧、加窗处理。本文信号的采样率为 48 kHz，帧长为 512 个样本点，窗函数采用汉宁窗，则加窗后第 $i$ 帧声音信号可表示如下：

$$x_i(n) = w(n)x(iN + n) \tag{1}$$

$$w(n) = \begin{cases} 0.5(1 - \cos(2\pi n/(N-1))), & 0 \leqslant n \leqslant N-1 \\ 0, & \text{其他} \end{cases} \tag{2}$$

其中，$x(n)$ 为采样后的动物声音信号；$x_i(n)$ 为加窗后第 $i$ 帧声音信号；$N$ 表示帧长；$i$ 表示帧序号。

对各帧信号进行短时傅里叶变换，将时域数据转变为频域数据，得到信号的频谱，计算见式(3)：

$$X_i(k) = \sum_{n=1}^{N} x_i(n) e^{-j2\pi kn/N}, 1 \leqslant k \leqslant N \tag{3}$$

计算每帧信号的谱线能量，见式(4)：

$$P(i,k) = |X_i(k)|^2 \tag{4}$$

将求出的每帧谱线能量通过美尔滤波器组。美尔滤波器具有三角滤波特性，每个滤波器的传递函数，如式(5)：

$$Hm(k) = \begin{cases} \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \leqslant k \leqslant f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \leqslant k \leqslant f(m+1) \\ 0, & \text{其他} \end{cases} \tag{5}$$

式中：$f(m)$ 为心频率，$\sum_{m=0}^{M-1} H_m(k) = 1$，$M$ 为滤波器的个数，本文取 128。

对滤波输出求和并取对数，式(6)

$$G_i(m) = \log\left(\sum_{k=0}^{N-1} P(i,k) Hm(k)\right), 0 < m < M \tag{6}$$

通过美尔滤波器组后得到的系数高度相关，因此对美尔特征频谱做 DCT 变换去相关，使能量集中在 DCT 后的低频部分，DCT 变换公式(7)：

$$b_i(n) = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} G_i(m) \cos\left(\frac{\pi n}{M}(m - 0.5)\right) \tag{7}$$

其中，$b_i(n)$ 是 MFCC 特征参数。

本文 MFCC 系数的阶数为 20。MFCC 为声音信号的静态特征，但其动态特征更能反应情绪特征。因此本文采用 MFCC 的一阶差分来捕捉声音信号的动态变化。

除了 MFCC 参数，本文还提取了其他 6 种特征。在频域，分别提取了频谱质心、频谱对比度、频谱带宽和频谱滚降，通过计算声音频谱能量分布的平均点、信号最高频率与最低频率的差值，对比峰值能量与低谷能量等数值差异来识别不同的动物语音情绪。

频谱质心是描述声音属性的重要物理参数之一，是频率成分的重心，是在一定频率范围内通过能量加权平均的频率，频谱质心的计算公式(8)：

$$FC_i = \frac{\sum_{k=1}^{N} f(k) P(i,k)}{\sum_{k=1}^{N} P(i,k)} \tag{8}$$

其中，$f(k)$ 为信号频率。

动物声音信号是由不同频率的声波组成的，这些声波的强度会随着时间的变化而变化。谱对比度将频谱的每个帧都分为子带，对于每个子带通过将

峰值能量与谷值能量进行比较来估计能量对比,反映声音信号中不同频率成分的相对强度。频谱对比度可以帮助分析声音信号中不同频率成分的变化,从而提取声音信号中的有用信息,计算公式(9):

$$\text{频谱对比度}(\%) = \left(\frac{P_{\max} - P_{\min}}{P_{\max} + P_{\min}}\right) \tag{9}$$

其中,$P_{\max}$ 为峰值能量,$P_{\min}$ 为谷值能量。

声音信号的频谱能量集中在一定频率范围内。当低于某频率的所有频率分量的能量达到整个频带能量的某百分比(本文取 85%)时,该频率即为频谱滚降点。频谱滚降越大,说明信号的高频分量衰减得越快。

在时域,提取声音信号的均方根能量和过零率参数。均方根能量是指信号在一段时间内的平均功率,可以用来衡量声音的响度,计算公式(10):

$$E_i = \sqrt{\frac{1}{N}\sum_n x_i^2(n)} \tag{10}$$

过零率表示一帧声音信号波形穿过横轴(零电平)的次数[15],计算公式(11):

$$Z_i = \frac{1}{2}\sum_{n=0}^{N-1} \left|\operatorname{sgn}[x_i(n)] - \operatorname{sgn}[x_i(n-1)]\right| \tag{11}$$

$$\operatorname{sgn}[x] = \begin{cases} 1, & x \geqslant 0 \\ -1, & x < 0 \end{cases} \tag{12}$$

过零率实际是样本改变符号的次数,可以反应动物声音波动的激烈程度。

### 1.2 情绪识别模型

本文采用基于注意力机制的 BiLSTM 建立情绪识别模型。该模型建立在编码器-解码器的结构上,原始的时间序列数据首先在编码器中进行初步学习,将所得到的学习结果和编码器状态输入到注意力层中,注意力层利用 Bahdanau 注意力机制挖掘序列中的时域相关信息,并赋予相应的权重进行加权求和,将计算得到的上下文向量输入到解码器[16]。解码器对注意力层的输出进一步处理,并将解码器状态更新到注意力层中进行迭代,最终获得动物声音情绪的分类预测结果。基于注意力机制的 BiLSTM 结构如图 3 所示,主要由 3 部分组成:双向长短期记忆网络层(BiLSTM)、注意力层(Attention)和全连接层。

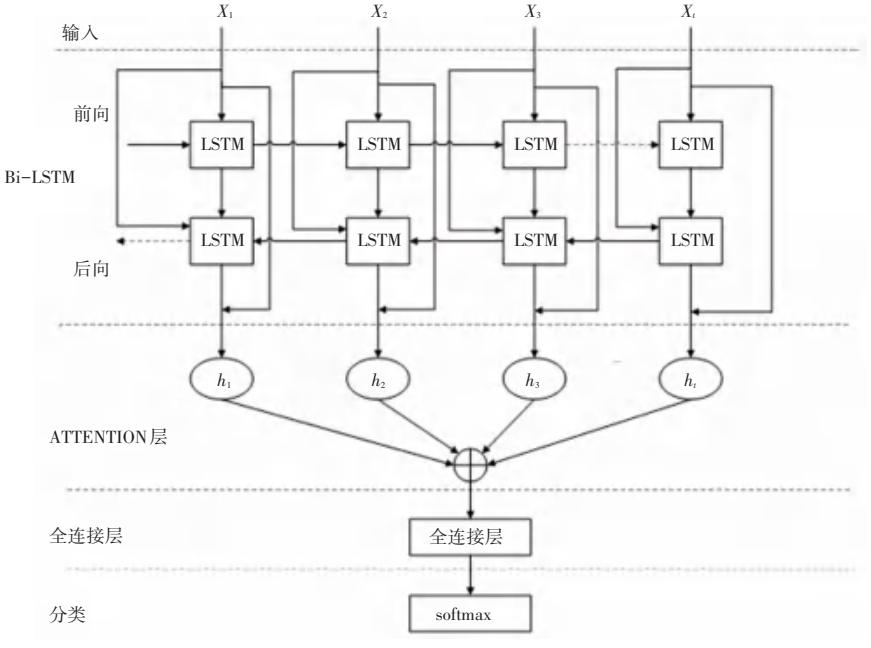


**图 3 基于注意力机制的 BiLSTM 结构图**

**Fig. 3 Structure diagram of BiLSTM based on attention mechanism**

#### 1.2.1 BiLSTM 模型

循环神经网络(Recurrent Neural Network,RNN)可用于处理可变长序数据,一个简单的循环神经网络由输入层、一个隐藏层和一个输出层组成。长短期记忆网络(Long-Short Term Memory, LSTM)是为了解决长依赖问题而衍生出的一种改良版本的循环神经网络,与传统循环神经网络 RNN 相比,能够有效克服 RNN 的记忆暂存、梯度弥散等问题,且兼具

长短期记忆功能，广泛应用于时间序列问题的预测[17]。LSTM 通过引入输入门、遗忘门、输出门 3 个门控单元以及专门的记忆细胞单元来控制信息的流动，允许网络选择性地记忆或遗忘信息，避免训练过程中的梯度爆炸和梯度消失问题，从而实现时间序列数据中长期依赖关系，提高数据预测的准确率。LSTM 共有 3 个 $\sigma$ 神经元和 1 个 tanh 神经元，其结构如图 4 所示。

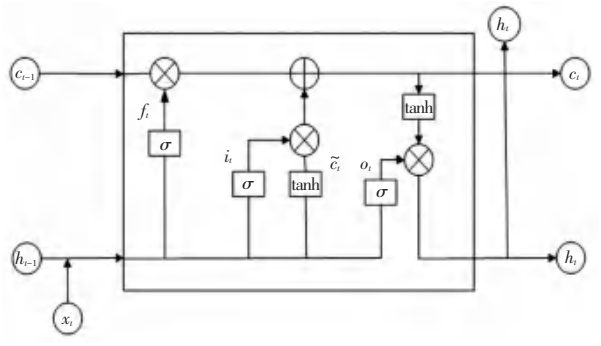


**图 4　LSTM 模型结构图**

**Fig. 4　Structure diagram of LSTM model**

$f_t$ 是遗忘门，用来选择前一时刻的记忆单元状态是否需要被遗忘。输出范围是[0,1]，表示保留多少前一时刻的记忆单元状态，计算公式(13)：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{13}$$

$i_t$ 是输入门，控制输入和当前计算的状态更新到记忆单元的程度，计算公式(14)和公式(15)：

$$i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, x_t] + b_i) \tag{14}$$

$$\tilde{C}_t = \tanh(\boldsymbol{W}_c \cdot [h_{t-1}, x_t] + b_c) \tag{15}$$

其中，$x_t$ 为当前输入状态；$h_{t-1}$ 为旧的隐藏层状态；tanh 函数作为生成候选记忆细胞 $\tilde{C}_t$ 的选项；$\boldsymbol{W}_i$ 代表输入矩阵；$\boldsymbol{W}_c$ 表示状态矩阵。

信息进入输入门和遗忘门，由公式(16)更新记忆状态：

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t} \tag{16}$$

其中，$C_{t-1}$ 为旧的记忆状态。

$o_t$ 是输出门，决定当前记忆单元需要保留输出的信息，以及能够传递下去的隐藏状态，用于控制当前时刻的输出。计算公式(17)：

$$o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, x_t] + \boldsymbol{b}_o) \tag{17}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{18}$$

其中，$\boldsymbol{W}_o$ 是输出矩阵，$\boldsymbol{b}_o$ 为偏置向量。

在处理动物声音情绪识别的问题时，当前时刻状态的输出不仅取决于过去的信息，还由未来的信息所决定。因此本文选择了双向长短期记忆(Bidirectional Long Short Term Memory，BiLSTM)神经网络模型进行动物声音情绪识别，由前向和后向两个反向的 LSTM 叠加而成，是对 LSTM 的扩展，具体表现为序列的双向学习。前向和后向 LSTM 在 $t$ 时刻的表达式为式(19)和式(20)：

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(x_t, \overrightarrow{h_{t-1}}) \tag{19}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(x_t, \overleftarrow{h_{t-1}}) \tag{20}$$

其中，$x_t$ 为 $t$ 时刻的输入序列。

分别以正逆两个方向输入至两个 LSTM 进行特征提取，得到前向 LSTM 隐藏层输出结果 $\overrightarrow{h_t}$ 和后向 LSTM 输出结果 $\overleftarrow{h_t}$，然后将 $\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$ 进行拼接形成新的向量作为此情感的特征表达。BiLSTM 通过前向和后向 LSTM 层同时提取上下文信息，对于多任务，尤其是自然语言处理中的标注和分类任务，是非常有效的。

### 1.2.2　Bahdanau 注意力机制

注意力机制是机器学习中常用的技术，通过对整体输入的计算快速聚焦所观察物体的学习重点，并给予更多注意力在重点信息上。由于输入的动物声音序列长短不一，本文采用 Bahdanau 注意力机制[18]，其是一种基于 seq2seq 模型注意力机制，能够对 BiLSTM 输出的特征向量给予不同的关注度。

注意力层的输入是 BiLSTM 隐藏层输出特征向量 $\boldsymbol{H} = \{h_1, h_2, \cdots, h_t\}$，注意力机制具体实现：将编码器的输出序列作为键向量，将编码器的隐藏状态作为查询向量，使用编码器中的 BiLSTM 网络的隐藏状态通过式(21)～式(23)计算每个时间步的注意力权重：

$$\mathrm{softmax}(\varepsilon_{tk}) = \tanh(W_j h_i + b_j) \tag{21}$$

$$a_{ij} = \frac{\exp(\varepsilon_{ij})}{\sum_{k=1}^{n} \exp(\varepsilon_{tk})} \tag{22}$$

$$h_i' = a_{ij} h_i \tag{23}$$

其中，$a_{ij}$ 代表解码器 $t$ 时刻对应编码器隐藏层状态的权重，将重点集中在情感突出时刻；$h_i'$ 为 $h_i$ 加权后的特征值；$W_j$ 和 $b_j$ 为可训练参数。

通过加权求和得到中间语义向量 $c_t$，公式(24)：

$$c_t = \sum_{i=1}^{t} a_{ij} h_i' \tag{24}$$

解码器隐藏层状态呈现为式(25)：

$$s_t = f(y_{t-1}, c_t, s_{t-1}), t = 1 \cdots T \tag{25}$$

通过全连接层将 46 维特征向量映射成一维，最

后由 Softmax 函数进行归一化,转化得到各情感的输出概率,式(26):

$$p(y_t) = g(y_{t-1}, s_t, c_t) \tag{26}$$

与传统的注意力机制不同,Bahdanau 注意力机制会将编码器的输出序列和解码器的隐藏状态进行非线性变换,并通过加法进行融合,从而更好地表达输入和输出之间的关系[19]。隐藏状态不仅要计算 BiLSTM 模型当前时刻输出的预测结果,还要与编码器生成的全部隐藏状态联合计算下一时刻的上下文信息。注意力的输出是上下文向量,包含了预测输出时源端所需要的信息[20]。通过与 BiLSTM 深度学习模型相结合,更好地处理输入数据,并取得更好的性能 。

## 2 实验结果与分析

### 2.1 实验设置

本文以狗为例,对狗的声音进行情绪识别。将狗的情绪分为“生气”、“受伤”和“伤心”。采集了柯基、萨摩耶、拉布拉多、泰迪、哈士奇、法斗、中华田园犬等各个品种在不同情绪下的声音,样本集中各情绪下音频数量见表 1。将不同的情绪声音数据按照 6∶4 的比例分为训练集和测试集。样本数据的采样率为 48 kHz。

**表 1 不同情绪的样本数量**

**Table 1 Sample size of different sentiments**

| 情绪类别 | 生气 | 受伤 | 伤心 |
|---|---|---|---|
| 总样本数量 | 147 | 102 | 124 |
| 测试集数量 | 59 | 40 | 49 |

狗“生气”、“受伤”和“伤心”3 种情绪时域波形分别如图 5~图 7 所示。

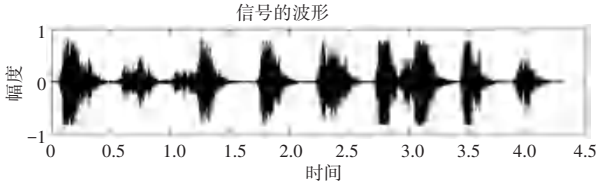


**图 5 狗“生气”情绪下的时域波形**

**Fig. 5 Time-domain waveform of the dog′s "angry" emotion**

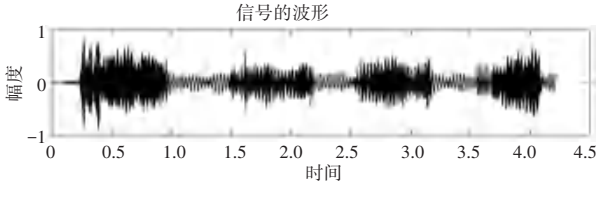


**图 6 狗“受伤”情绪下的时域波形**

**Fig. 6 Time-domain waveform of the dog′s "hurt" emotion**

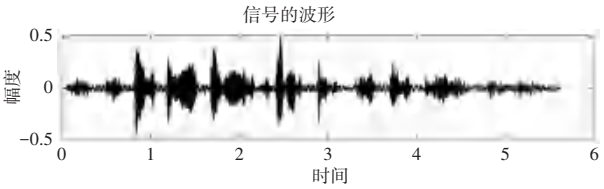


**图 7 狗“伤心”情绪下的时域波形**

**Fig. 7 Time-domain waveform of the dog′s "sad" emotion**

本文利用 Tensorflow 框架构建基于注意力机制的 BiLSTM 模型,对输入的狗声音特征进行处理。为使模型达到最佳性能,设置迭代次数(epoch)为 30,单次训练用的样本数(batch_size)为 32。采用 Adam 优化算法,通过计算梯度的一阶矩估计和二阶矩估计而为不同的参数设计独立的自适应性学习率,学习率(learning_rate)设置为 0.001。采用多分类任务中的交叉熵损失函数(categorical_crossentropy)作为损失函数,模型输出的激活函数为 Softmax 函数。

本文采用准确率(*Accuracy*)与精确率(*Precision*)作为评价指标。准确率关注全部类别的全部样本,指各类别正确预测的样本数占全部样本数的比例;精确率关注某一特定类别的预测样本,指某一类别正确预测的样本数占此类别所有预测样本的数量的比值。

对于本文所提出的基于注意力机制的 BiLSTM 的多分类模型,每一个特定的测试样本都有一个真实的分类,经过模型预测后,又会有一个预测分类。

### 2.2 实验结果

实验对比了基于 RNN 模型、基于 BiLSTM 模型及基于注意力机制的 BiLSTM 模型的狗情绪识别的性能,实验结果见表 2、表 3 和图 8 所示。

**表 2 3 种模型的分类精确率对比**

**Table 2 Comparison of classification accuracy of the three models %**

| 情绪类别 | 深度学习模型 | | |
|---|---|---|---|
| | RNN | BiLSTM | 基于注意力机制的 BiLSTM |
| 生气 | 81 | 85 | 92 |
| 受伤 | 83 | 93 | 95 |
| 伤心 | 89 | 89 | 93 |

**表 3 3 种模型的分类准确率对比**

**Table 3 Comparison of classification accuracy of the three models**

| 深度学习模型 | 准确率/% |
|---|---|
| RNN | 83.11 |
| BiLSTM | 88.51 |
| 基于注意力机制的 BiLSTM | 93.24 |

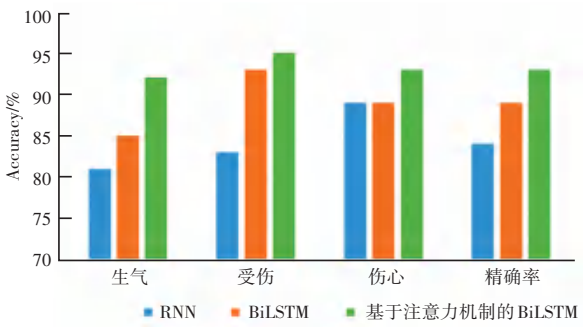


**图 8　3 种模型的动物情绪识别性能对比**

**Fig. 8　Comparison of animal emotion recognition performance of the three models**

由表 2、表 3 和图 8 可见，基于注意力机制的 BiLSTM 模型性能最佳，基于 BiLSTM 模型的性能次之，基于 RNN 的性能最低。基于注意力机制的 BiLSTM 模型的精确率比基于 BiLSTM 模型和基于 RNN 模型的精确率都要高，其中，“生气”的精确率分别提高了 7% 和 11%，“受伤”的精确率分别提高 2% 和 8%，“伤心”的精确率提高了 4%。基于注意力机制的 BiLSTM 模型的情绪准确率比基于 BiLSTM 模型和基于 RNN 模型的情绪识别准确率分别提高了约 4.74% 和 13.13%。此外，损失值指预测值与真实值之间的差值，基于注意力机制的 BiLSTM 模型、基于 BiLSTM 模型以及基于 RNN 模型的测试损失值分别为 0.19、0.36 和 0.51，基于注意力机制的 BiLSTM 模型相较于基于 RNN 和基于 BiLSTM 模型具有更好的情绪识别效果。

## 3　结束语

声音是动物向外界表达需求的一种方式，通过声音对动物情绪进行探究在诸多领域都有重要作用。本文提出一种基于注意力机制 BiLSTM 的动物声音情绪识别方法，BiLSTM 模型可以实现序列的双向学习，注意力机制可以对情绪特征进行通道方向的权重学习。情绪识别模型的输入为多种声音特征参数，包括 MFCC 及其一阶差分、频谱质心、频谱对比度、频谱带宽、频谱滚降、均方根能量和过零率。本文以狗为例，对狗的声音进行了 3 种情绪识别，实验结果表明，相比于 RNN、BiLSTM 模型，本文算法的情绪识别准确率更高。

虽然本文方法具有较好的识别效果，但仍存在一定的局限性。由于国内外对动物情绪研究较少，动物声音样本数量不足，且至今未发现能完全描述动物情绪的特征，使特征分析不完善，仍需展开进一步研究。

## 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[J]. 软件学报，2014，25(1)：37-50.

[2] VAN BEZOOIJEN R，OTTO S A，HEENAN T A. Recognition of vocal expressions of emotion：A three-nation study to identify universal characteristics[J]. Journal of Cross-Cultural Psychology，1983，14(4)：387-406.

[3] MORIYAMA T，OZAWA S. Emotion recognition and synthesis system on speech [C]//Proceedings of the 1999 IEEE on Multimedia Computing and Systems(ICMCS). Florence，Italy：IEEE，1999：840-844.

[4] 乔冠楠，胡剑凌，刘鹏. 声学参数和听觉参数结合的语音情感交叉识别[J]. 电声技术，2009，33(6)：52-56.

[5] MIRSAMADI S，BARSOUM E，ZHANG C. Automatic speech emotion recognition using recurrentneural networks with local attention[C]// Proceedings of 2017 IEEE International Conference on Acoustics，Speechand Signal Processing(ICASSP).IEEE，2017：2227-2231.

[6] YOON S，BYUN S，DEY S，et al. Speech emotion recognition using multi-hop attentionmechanism [C]//Proceedings of 2019 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2019：2822-2826.

[7] 彭祝亮，刘博文，范程岸，等. 基于 BLSTM 与方面注意力模块的情感分类方法[J]. 计算机工程，2020，46(3)：60-65，72.

[8] 刘恒，吴迪，苏家仪，等. 运用高斯混合模型识别动物声音情绪[J]. 国外电子测量技术，2016，35(11)：82-87.

[9] MOLNÁR C，KAPLAN F，ROY P，et al. Classification of dog barks：a machine learning approach[J]. Animal Cognition，2008，11(3)：389-400.

[10] 郭龙祥，梅继丹，张亮. 鲸豚类动物声信号的分析及识别[C]// 2011 年中国声学学会水声学学术会议. 2011：70-72.

[11] 杨春勇，侯金，陈少平，等. 动物声音情绪识别系统及其方法：CN 201510143593[P]. [2024-02-06].

[12] 黄玮，冉启斌. 基于 DTW 距离的动物声音分类研究[J]. 智能计算机与应用，2020，10(5)：6.

[13] 石鑫鑫，鱼昕，刘铭. FCNN 深度学习模型及其在动物语音识别中的应用[J]. 吉林大学学报(信息科学版)，2021，39(1)：60-65.

[14] 杨兴海，漆国强，杨兴荣，等. 基于卷积神经网络通过声音识别动物情绪的方法及系统：CN202111582306.9[P]. [2024-02-06].

[15] 李佳竞. 基于短时能量和短时过零率的远程广播信号监测研究[J]. 电声技术，2022，46(11)：100-102.

[16] 陈汀杨. 一种基于 Bahdanau 注意力机制的电价预测算法[J]. 中国科技信息，2023 (9)：65-68.

[17] BAHDANAU D，CHO K，BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv：1409.0473，2014.

[18] XU K，SHEN X，YAO T，et al. Greedy layer-wise training of long short term memory networks[C]//Proceedings of 2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE，2018：1-6.

[19] BAHDANAU D，CHO K，BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv：1409.0473，2014.

[20] 黄佳林. 基于深度学习的机场大气能见度及演化规律分析[D]. 郑州：华北水利水电大学，2023.