基于深度学习的减压音乐重构研究

李 哲 陈 宇 张 博 陈 亮 郭 滨
(长春理工大学电子信息工程学院 吉林 长春 130022)

摘 要 音乐在旋律与和弦之间有复杂的匹配关系,音乐重构是长时间序列生成的算法研究。通过计算多轨道音乐序列的音乐频谱质心,使用栈式自编码器(SAE)对频谱质心较高的音乐进行音符特征提取,将音乐特征输入长短期记忆循环神经网络(LSTM),构建多轨道音乐重构模型。分析重构音乐的和谐度和音符分布均方误差,结果表明该方法好于单独 LSTM 网络重构方法。设计受试者焦虑状态测评实验,分析播放重构音乐前后受试者的焦虑程度,从而验证生成重构的音乐可以有效减压。

关键词 深度学习 音乐重构 频谱质心 栈式自编码器 长短期记忆循环神经网络

中图分类号 TP391 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2022.08.024

RESEARCH ON DECOMPRESSION MUSIC RECONSTRUCTION BASED ON DEEP LEARNING

Li Zhe Chen Yu Zhang Bo Chen Liang Guo Bin
(*School of Electronics and Information Engineering, Changchun University of Science and Technology, Changchun 130022, Jilin, China*)

Abstract Music has a complex matching relationship between melody and chord. Music reconstruction is a study of algorithms for long-term sequence generation. Multi-track music was constructed by calculating the music spectrum centroid of the multi-track music sequence, the stack autoencoder(SAE) was used to extract note features from music with a high spectral centroid, and by inputting the music features into long-term and short-term memory recurrent neural networks(LSTM) to construct a multi-track music reconstruction model. After analysis, the harmony degree of the music and the mean square error of the note distribution were analyzed. The results show that the method is better than the LSTM network reconstruction method alone. An anxiety assessment test was designed for subjects to analyze the subject's anxiety before and after playing reconstructed music to verify that the generated reconstructed music can effectively reduce stress.

Keywords Deep learning Music reconstruction Spectrum centroid Stacked auto encoder Long and short-term memory recurrent neural network

0 引 言

随着社会现代化的发展,生活节奏加快给人们带来焦虑情绪。对于影响情绪的方式而言,音乐是非入侵的有效方法,其作用在临床应用上也得到了认可,研究表明音乐能够有效影响人的情绪[1]。

传统生成音乐的方法有概率模型[2],由于音符被认为是无序的,重构音乐时忽略上下文的连接。现已有更多算法用于重构音乐,包括马尔可夫模型和遗传算法[3]。马尔可夫模型重构音乐时处理长时间序列会随着时间增加而衰减,遗传算法每次迭代延迟较高。在神经网络的基础上,音乐重构在深度学习上有更好的发展。重构音乐序列过程可以映射到深度学习算法上[4]。文献[5]为了解决循环神经网络(RNN)生成长时间序列时的梯度消失问题,利用长短期记忆循环神

收稿日期:2020-03-26。吉林省自然科学基金项目(20150101013JC)。**李哲**,硕士,主研领域:信号与信息处理技术。**陈宇**,副教授。**张博**,硕士。**陈亮**,硕士。**郭滨**,教授。

经网络(LSTM)生成鼓的节奏序列。通过训练 LSTM 网络可以生成指定风格的音乐,如 Hutchings 等[6]通过训练 LSTM 网络生成爵士乐,文献[7]基于 BiLSTM 和 NN 网络生成巴赫风格的曲子。但以上研究没有提出基于情感背景重构影响情绪的音乐。

临床上测评焦虑程度多采用汉密尔顿焦虑量表(HAMA)和焦虑自评量表(SAS)。文献[8]采用 HAMA 对 66 名重度抑郁焦虑合并症患者进行心理评估,以分析焦虑情绪与症状的因果关系。SAS 应用于焦虑情绪疗效评估的效度信度高,适用于广泛人群测评焦虑程度[9]。

基于以上研究,本文提出重构减压音乐模型。对多轨道音乐进行减压特征提取,通过训练 LSTM 网络重构音乐,并分析重构的音乐质量,结合 HAMA 和 SAS 设计合理化减压实验,分析重构音乐的减压效果。

1　原理分析

1.1　频谱质心

频谱质心是音乐在频域上的特征,通过快速傅里叶变换(FFT)由音频转化到频域上分析音乐信号频谱包络的质心。频谱质心可以用来衡量乐曲中所含高频分量和低频分量的比重。频谱质心较低时,乐曲有较多低频内容,乐曲呈现的低沉阴郁品质,频谱质心较高时,乐曲有更多明亮舒缓的高频内容。频谱质心计算公式如下:

$$P_i = \frac{\sum_{k=0}^{N-1} k\,|S_i(k)|}{\sum_{k=0}^{N-1} |S_i(k)|} \tag{1}$$

式中:N 为帧长;k 表示频率下标;$S_i(k)$ 代表在第 i 帧信号位置的快速傅里叶在 k 处的幅度值。

1.2　栈式自编码器

自动编码器 AE(Auto Enconer)是无监督学习的神经网络。AE 的构建是基于反向传播算法,有输入层,隐藏层和输出层。输入层与输出层的神经元数量相同,隐藏层的神经元数少于输入层的神经元数。如图 1 所示,AE 分为编码器(Encoder)和解码器(Decoder)两部分,$\boldsymbol{x}$ 代表原始输入数据,$\boldsymbol{h}$ 代表表征,$\boldsymbol{r}$ 代表输出数据。编码器是输入层到隐藏层的数据处理过程,使原始数据 $\boldsymbol{x}$ 被迫降维,压缩成潜在的维度表征 $\boldsymbol{h}$,并能学习到样本数据的特征。解码器的过程是隐藏层到输出层,由于隐藏层的维数少于输出层的维数,输出数据 $\boldsymbol{r}$ 通过压缩数据的表征重新构造得到。此过程处理多维训练数据,提取更高的潜在特征。

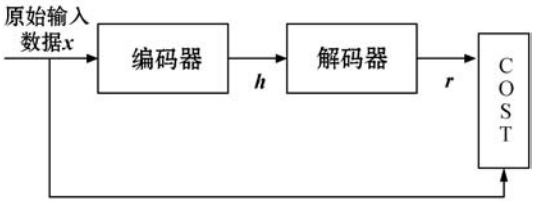

图 1　自编码器结构图

编码器的函数表达式如下:

$$\boldsymbol{h} = f(\boldsymbol{x}) \tag{2}$$

编码过程激活函数使用 sigmoid 函数,表达式如下:

$$f(\boldsymbol{x}) = \mathrm{sigmoid}(\boldsymbol{x}) = \frac{1}{1+\mathrm{e}^{-x}} \tag{3}$$

解码过程函数表达式如下:

$$\boldsymbol{r} = g(\boldsymbol{h}) = g(f(\boldsymbol{x})) \tag{4}$$

用于数据重构的误差函数为均方误差函数(MSE),表达式如下:

$$L = \|\boldsymbol{x} - g(f(\boldsymbol{x}))\|^2 \tag{5}$$

代价函数表达式如下:

$$J = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\|\boldsymbol{x}^i - g(f(\boldsymbol{x}^i))\|^2\right] = \frac{1}{2N}\sum_{i=1}^{N}\|\boldsymbol{x}^i - \boldsymbol{r}^i\|^2 \tag{6}$$

式中:N 为训练样本的数量;$\boldsymbol{x}^i$ 为输入向量;$\boldsymbol{r}^i$ 为自动编码器网络重构的输出结果。训练自动编码器网络的目的是学习输入向量 $\boldsymbol{x}^i$ 与输出向量 $\boldsymbol{r}^i$ 相似的关系,为了减小输出与原始数据的误差,训练代价函数使其值减小。

栈式自编码器(Stacked Auto Encoder,SAE)是堆叠多个自编码器构成[10],又称深度自编码器。本研究使用的栈式自编码器结构是由两个自编码器嵌套组成,如图 2 所示。隐藏层的神经元数目是逐层减少的,前一层的输出是下一层的输入,通过深层压缩,提取潜在特征,二阶特征为此网络学习到和弦与旋律的更高维度的特征。

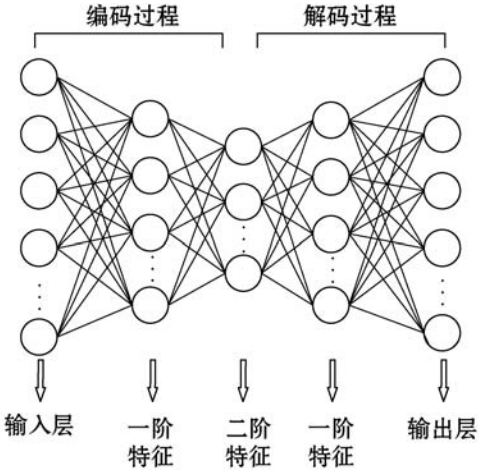

图 2　栈式自编码器网络结构图

1.3 LSTM 原理分析

循环神经网络(RNN)是闭合反馈神经网络,利用历史信息通过隐藏层的网络结构,影响当前处理的数据。但训练长序列的重构比较困难,主要原因在于向前传播和反向传播都会乘上多次隐藏层的参数。隐藏层参数小于1,导致前向传播中小于1的值乘上多次时减少对输出的影响,反向传播时会导致梯度弥散问题。在传统 RNN 的基础上,Hochreiter 等[11]提出了长短期记忆循环神经网络(LSTM)解决 RNN 的梯度消失,改善长序列生成问题。

LSTM 设计目的是通过使用常数误差流(CEC)来获得长时间的恒定误差流。LSTM 通过"门"的结构遗忘和增强信息到神经元的能力,来记忆长期的信息,"门"是一种让信息选择性通过的方法。LSTM 中历史信息存储在存储单元,更新和处理数据通过输入门、遗忘门和输出门控制。LSTM 单元结构如图3所示。

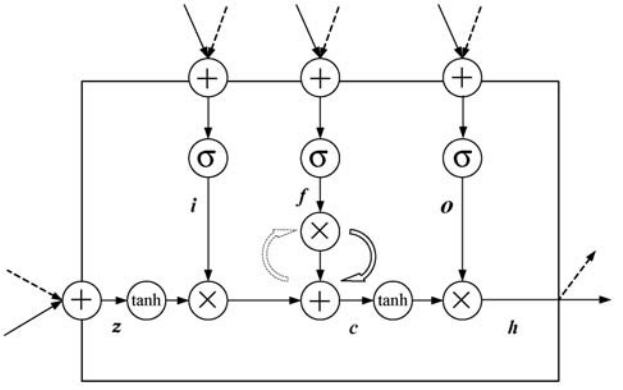

图3 LSTM 单元结构图

时间步长 t 下的输入值为 $\boldsymbol{x}_t$,隐藏状态值为 $\boldsymbol{h}_t$,$\boldsymbol{w}$ 为加权矩阵,$\boldsymbol{b}$ 为偏置向量,σ 为 sigmoid 激活函数。LSTM 向前传播具体计算过程如下。

遗忘门的输入通过 sigmoid 函数,使其输出的值在0~1之间,1表示信息完全保留,0表示信息完全遗忘。遗忘门可以选择性遗忘神经元状态中无意义信息,控制历史信息对当前状态的影响。遗忘门计算公式如下:

$$\boldsymbol{f}_t=\sigma(\boldsymbol{w}_{xf}\boldsymbol{x}_t+\boldsymbol{w}_{hf}\boldsymbol{h}_{t-1}+\boldsymbol{b}_f) \tag{7}$$

输入门控制对当前神经元状态的更新,计算公式如下:

$$\boldsymbol{i}_t=\sigma(\boldsymbol{w}_{xi}\boldsymbol{x}_t+\boldsymbol{w}_{hi}\boldsymbol{h}_{t-1}+\boldsymbol{b}_i) \tag{8}$$

输出门控制储存单元的状态值的输出,计算公式如下:

$$\boldsymbol{o}_t=\sigma(\boldsymbol{w}_{xo}\boldsymbol{x}_t+\boldsymbol{w}_{ho}\boldsymbol{h}_{t-1}+\boldsymbol{b}_o) \tag{9}$$

t 时刻候选记忆值:

$$\boldsymbol{z}_t=\sigma(\boldsymbol{w}_{xz}\boldsymbol{x}_t+\boldsymbol{w}_{hz}\boldsymbol{h}_{t-1}+\boldsymbol{b}_z) \tag{10}$$

t 时刻记忆单元值:

$$\boldsymbol{c}_t=\boldsymbol{i}_t\times\boldsymbol{z}_t+\boldsymbol{f}_t\times\boldsymbol{c}_{t-1} \tag{11}$$

t 时刻输出值:

$$\boldsymbol{h}_t=\boldsymbol{o}_t\times\tanh(\boldsymbol{c}_t) \tag{12}$$

3 研究方法

3.1 音乐序列预处理

下载480首舒缓的 MIDI(Musical Instrument Digital Interface)格式音乐,MIDI 格式的音乐文件携带数字化音乐信息,每个轨道表示一种乐器,准确记录每个乐器的演奏过程。由于节奏和调式影响音乐对情绪的引导,筛选出4/4拍慢节奏420首音乐,并转化成 C 大调音乐。将 MIDI 格式音乐转化为 WAV 格式的音频,音频通过快速傅里叶变换转换到频域,计算每首音乐的频谱质心。筛选出频谱质心高于平均水平的音乐,共360首音乐作为训练样本。

将频谱质心较高的360首音乐转化 MIDI 格式,每首音乐截取成90 s的音乐序列。在90 s的 MIDI 格式音乐中只保留钢琴、吉他、小提琴、短笛四个轨道,删除其他乐器所占的轨道。MIDI 为各轨道的乐器定义出128个音符,编号为0~127,中央 C 编号为60。对每个音符演奏的力度定义编号为0~127。采用 Python 的 music21 包读取 MIDI 音乐文件的音符与和弦信息保存文本格式,通过 One-hot 编码生成128维向量作为训练数据。

3.2 减压特征提取

栈式自编码器用于提取频谱质心较高的音乐样本,提取各个乐器的和弦与旋律的潜在特征称为减压特征。前一层的输出即为下一层的输入,逐层训练之后再进行反向训练。通过已有数据对栈式自编码器进行预训练,在 Tensorflow 框架中构建栈式自编码网络,设置栈式自编码器输入与输出神经元数量相等,输出数据的维度与输入数据相同为128维,一阶特征层设置88维,二阶特征层设置64维。由栈式自编码器提取二阶特征层的数据即为潜在的减压音乐特征向量,用 $\boldsymbol{x}<n>$ 表示。

3.3 音乐序列重构模型

将音乐特征向量 $\boldsymbol{x}<n>$ 向左平移一个单位为 $\boldsymbol{y}<n>$,即 $\boldsymbol{x}<n+1>=\boldsymbol{y}<n>$。音乐特征向量输入 LSTM,LSTM 是用先前值预测下一个值,给出向量序列 $\boldsymbol{x}<1>$,$\boldsymbol{x}<2>$,…,$\boldsymbol{x}<i>$ 构建模型预测 $\boldsymbol{y}<i>$。通过不断迭代,得到的误差值传入 LSTM 反向传播中,更新加权参数 $\boldsymbol{w}$。输入向量是栈式自编码器的二阶特征层的特征向量,LSTM 网络输入维度设置64维,隐藏层设置128

维。为了防止过拟合,节点被丢弃概率(Dropout 值)为 30%,全连接层(Dense)神经元数等于输出音符不同的个数,通过 Softmax 预测 LSTM 单元的输出,输出的序列转换成一首 5 min 时长的 MIDI 音乐,完整音乐重构模型如图 4 所示。

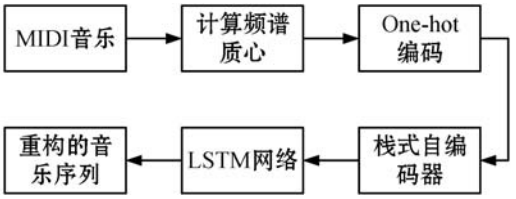

图 4　完整音乐重构模型图

LSTM 网络的隐藏层层数分别选择 2、3、4 层进行训练,LSTM 网络层数等于隐藏层层数,迭代次数对训练结果的影响如图 5 所示,其中横坐标为迭代次数,纵坐标为损失函数的值。

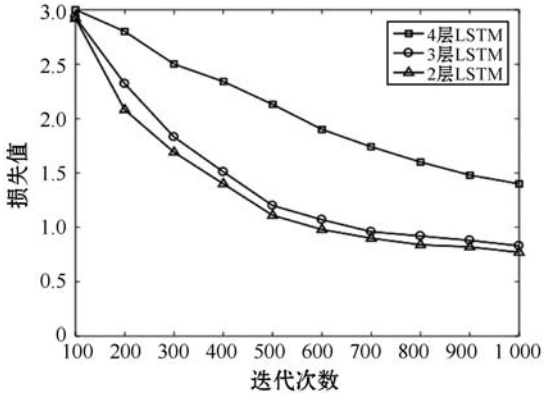

图 5　不同层 LSTM 网络的损失值

LSTM 网络层数设置过多时,梯度下降明显速率缓慢,3 层 LSTM 网络收敛效果较好。重构音乐模型的生成音乐序列部分设置 3 层的 LSTM 网络。

3.4　重构音乐结果分析

3.4.1　重构音乐的和谐度

重构音乐采用单独 LSTM 网络和加入 SAE 结构的 LSTM 网络(以下称 SAE-LSTM 网络)分别训练,设置产生相同的随机种子用于两种模型重构音乐,分别重构 100 首音乐作为音乐质量分析。

为了评定重构音乐的多轨道之间的和旋配合,对生成的音乐进行和谐度分析,轨道之间演奏相似的和旋则说明音乐和谐。和谐度计算公式为:

$$H = \sum_{i=1}^{i} \Delta\left(\bigcap_{k=1}^{k} C_i^k\right) \tag{12}$$

式中:i 为重构音乐序列的乐段数;K 为轨道数;C_k^i 表示第 k 轨道的第 k 和旋。以钢琴演奏主旋律为基准,计算两种模型各重构 100 首音乐中的其他轨道和弦和谐度并取均值,如表 1 所示。

表 1　重构音乐和谐度平均值

乐器	LSTM	SAE-LSTM
小提琴	0.425	0.642
吉他	0.551	0.583
短笛	0.372	0.419

SAE-LSTM 网络重构音乐较单独 LSTM 网络重构音乐的和谐度更高,结果说明 SAE 对音乐特征的提取可使 LSTM 网络在多轨道音乐上学习效果更好,重构的乐曲中各乐器之间搭配更和谐。

3.4.2　重构音乐的音符分布均方误差

评估重构音乐模型学习乐器特性的效果,通过音符分布均方误差来衡量。乐器特性含义为每个乐器有独自分布的音域,例如钢琴有 88 个琴键,音域为 $A_2 - c^5$,而小提琴多演奏高音,音域为 $g - c^4$。音符分布均方误差越小说明重构音乐的模型学习效果好,计算公式为:

$$Notes_{\mathrm{MSE}} = \frac{\sum_{i=1}^{U}\sum_{j=1}^{V}\left(\frac{a_i}{V} - \frac{\hat{y}_i}{V}\right)^2}{UV} \tag{13}$$

式中:U 表示音乐的数量;V 表示音符不同的数量;a_i 代表原始音符;$\hat{y}_i$ 代表重构后的音符。计算 LSTM 网络和 SAE-LSTM 网络分别重构音乐的各个轨道音符分布均方误差,结果如表 2 所示。

表 2　重构音乐音符分布均方误差　($\times 10^{-4}$)

轨道	LSTM	SAE-LSTM
钢琴	4.976	3.732
小提琴	7.454	6.218
吉他	3.438	3.126
短笛	3.227	2.853

结果表明,SAE-LSTM 网络重构音乐的音符分布均方误差较小,SAE 对音乐特征的提取有效提高 LSTM 网络学习不同乐器的特性,从而提高重构音乐的质量。

3.5　重构音乐减压实验设计

3.5.1　实验条件及对象

招募本校由自感由压力引发焦虑状态的志愿者 30 人,年龄 20 ~ 24 岁,均听力正常且受过音乐训练,告知实验内容均同意作为受试者参加减压实验。减压实验在心理治疗室中进行,保持室内安静和整洁并设置音响、耳机设备。

3.5.2　测评指标

由于压力引发焦虑症状,通过受试者测评焦虑自评量表(SAS)和汉密尔顿焦虑量表(HAMA)的方式判断受试者的减压情况。SAS 含有 20 个项目测评,每个项目有 4 个程度的选项,选项分值为 1 分、2 分、3 分、

4 分。计算测评总分数,总分在 50 以下表示没有焦虑症状,50 ~ 59 为轻度焦虑,60 ~ 69 为中度焦虑,69 分以上为重度焦虑,分数越高代表焦虑程度越显著。HAMA 包含 14 个项目,每个项目为 0 ~ 4 分五级评分法,总得分 7 分以下表示无焦虑,大于 7 分代表有明显焦虑,大于 21 分为严重焦虑。情绪越焦虑,测评出的分数越高。

3.5.3 实验过程

通过计算频谱质心,由 SAE 提取减压特征输入的 LSTM 网络重构 100 首音乐,音乐总播放时长为 342 分钟 52 秒。30 名受试者依次单独进入心理治疗室中,测评 SAS 与 HAMA 并记录得分情况。受试者被引导闭眼,坐于沙发以舒适姿势随机聆听重构的减压音乐,播放时长为十分钟。间歇三分钟,再次随机播放减压音乐十分钟。受试者聆听结束后,在无人干预的情况下再次测评 SAS 和 HAMA,并记录得分情况。

3.5.4 重构音乐减压效果分析

30 名受试者依次参与完成上述实验过程,为了提高实验结果的效信度,记录减压音乐干预焦虑情绪前后的 SAS 与 HAMA 得分情况并分析,如表 3 所示。其中 N 代表参加实验的受试者人数。受试者在实验开始之前焦虑情绪均为显著,在聆听减压音乐之后焦虑情绪得到舒缓,根据每个人对音乐感知不同,减压效果也有所差异。得分情况表明该方法重构的音乐有效调节焦虑情绪,达到减压效果。

表 3 减压音乐干预前后的 SAS 与 HAMA 得分

类别	N	干预前 (均值 ± 标准差)	干预后 (均值 ± 标准差)
SAS	30	57.935 ± 1.346	49.663 ± 1.218
HAMA	30	15.638 ± 2.015	11.638 ± 1.379

4 结 语

本研究提出多轨道减压音乐特征的提取方法,构建重构减压音乐模型。计算样本音乐的频谱质心,通过栈式自编码器对频谱质心较高的音乐进行特征高维度压缩,提取的音乐特征为减压特征,将特征输入 LSTM 网络训练重构减压音乐序列。在研究中发现栈式自编码器提高了 LSTM 网络对音乐特性与和弦搭配的学习,重构的多轨道音乐和谐度更高。设计减压实验,实验结果表明本文方法重构的音乐序列有减压效果。

参 考 文 献

[1] Eich E, Ng J T W, Macaulay D, et al. Combining music with thought to change mood[M]//Handbook of Emotion Elicitation and Assessment. Oxford University Press, 2007: 124 - 136.

[2] Yu L G, Bu J J, Chen C. Rhythm generation based on inside-outside algorithm[J]. Journal of Zhejiang University (Engineering Science), 2005, 39(12): 1969 - 1972

[3] Fernandez J D, Vico F. AI methods in algorithmic composition: A comprehensive survey[J]. Journal of Artifificial Intelligence Research, 2013, 48: 513 - 582.

[4] 李星达. 钢琴多音估计问题和音乐生成问题的深度学习方法[D]. 吉林:吉林大学,2019.

[5] Makris D, Kaliakatsos-Papakostas M, Karydis I, et al. Conditional neural sequence learners for generating drums' rhythms[J]. Neural Computing & Applications, 2019, 6: 1793 - 1804.

[6] Hutchings P, Mccormack J. Using autonomous agents to improvise music compositions in real-time[J]. Computational Intelligence in Music, Sound, Art and Design, Evomusart 2017, 10198: 114 - 127.

[7] 叶文豪. 基于 BiLSTM 和 GANs 算法的自动作曲[D]. 吉林:吉林大学,2019.

[8] Huang Y C, Lee Y, Lin P Y, et al. Anxiety comorbidities in patients with major depressive disorder: The role of attachment[J]. International Journal of Psychiatry in Clinical Practice, 2019, 23(4): 286 - 292.

[9] 杨红棉. 大学生自我能力感知、焦虑、神经质及压力的关系研究[D]. 重庆:西南大学,2019.

[10] 朱成. 栈式自编码器特征表达能力研究[J]. 电信快报, 2019(3): 32 - 37.

[11] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735 - 1780.

[12] 王程,周婉,何军. 面向自动音乐生成的深度递归神经网络方法[J]. 小型微型计算机系统,2017,38(10):2412 - 2416.

[13] Veire L V, Bie T D. From raw audio to a seamless mix: Creating an automated DJ system for drum and bass[J]. EURASIP Journal on Audio Speech and Music Processing, 2018, 134: 1 - 21.

[14] Han X, Li B Y, Wang Z R. An attention-based neural framework for uncertainty identification on social media texts [J]. 清华大学学报自然科学版(英文版),2020,25(1): 117 - 126.

[15] 鲁亚平. 面向深度网络的自编码器研究[D]. 苏州:苏州大学,2016.

[16] 王雅思,姚鸿勋,孙晓帅,等. 深度学习中的自编码器的表达能力研究[C]//第十届和谐人机环境联合学术会议论文集,2014:1 - 8.

[17] 杨帅,王鹃. 基于堆栈降噪自编码器改进的混合推荐算法[J]. 计算机应用,2018,38(7):1866 - 1871.

[18] 姜力,詹国华,李志华. 基于递归神经网络的散文诗自动生成方法[J]. 计算机系统应用,2018,27(8):263 - 268.