PAPER

Combining CNN and Broad Learning for Music Classification

Huan TANG[†], ***Nonmember and*** **Ning CHEN**[†a)], ***Member***

SUMMARY Music classification has been inspired by the remarkable success of deep learning. To enhance efficiency and ensure high performance at the same time, a hybrid architecture that combines deep learning and Broad Learning (BL) is proposed for music classification tasks. At the feature extraction stage, the Random CNN (RCNN) is adopted to analyze the Mel-spectrogram of the input music sound. Compared with conventional CNN, RCNN has more flexible structure to adapt to the variance contained in different types of music. At the prediction stage, the BL technique is introduced to enhance the prediction accuracy and reduce the training time as well. Experimental results on three benchmark datasets (GTZAN, Ballroom, and Emotion) demonstrate that: i) The proposed scheme achieves higher classification accuracy than the deep learning based one, which combines CNN and LSTM, on all three benchmark datasets. ii) Both RCNN and BL contribute to the performance improvement of the proposed scheme. iii) The introduction of BL also helps to enhance the prediction efficiency of the proposed scheme.

key words: *deep learning, broad learning, random convolutional neural network (RCNN), music classification*

1. Introduction

Digital music and online streaming have become very popular these days due to the increase in the number of users. As a result, how to help the users find valuable data, such as trends, popular genres and artists, has become a great challenge. Automatic music classification technique can help to solve this problem. It has become a hot research topic within the field of Music Information Retrieval (MIR) in the past two decades.

It has been verified in [1]–[3], [5] that compared with the conventional hand-crafted feature extraction strategies, deep learning-based ones have their superiorities: i) The nonlinear mapping strategy in deep architecture helps to describe the prominent time-varying nonlinear property of music, precisely. ii) The hierarchical architecture of the deep architecture can represent the time (onset, rhythm) and frequency (note,chord)-based hierarchical nature contained in the music. iii) The Recurrent Neural Network (RNN)-based deep learning architecture (e.g. LSTM) can grasp the prominent long-term dependency based properties, such as recurrent harmonics and music structure contained in the music. These are the possible reasons why deep learning architecture based schemes have achieved tremendous success in various MIR tasks, such as onset detection [6], emotion recognition [7], chord estimation [8], rhythm stimuli recognition [9], source separation [10], music recommendation [11] and auto-tagging [4], [12], [14], [15].

For music classification tasks, CNN and RNN are the two most adopted deep learning architectures. In most cases, CNN is adopted to analyze the spectrogram image of the music sound to learn the high-level descriptors of it. In [4], [5], [17], the influence of the filter shape of CNN on the performance of music feature extraction was studied. It was verified that different shapes of filters may be fit for extracting different features of music sound. For example, the wider and higher filters may learn longer temporal dependency and more spreader timbral features, respectively [17]. In [16], the problem that if randomly weighted CNNs can obtain equivalent classification accuracy as trained CNN was studied. It was believed that the former is close to matching the accuracies obtained by the latter. However, as shown in [17], CNNs may be good at modeling the local context (such as instrument timbre or musical units), but not the long-term dependencies (such as music structure or recurrent harmonies). To solve this problem, RNN-based schemes are studied for music classification tasks. For example, in [18], LSTM was adopted to analyze the MFCC feature sequences for classification. However, since the temporal modeling is performed on the linear feature, it is difficult to disentangle underlying factors of variation with the input. Also, since there is no intermediate nonlinear hidden layer in LSTM, the history of previous inputs may not be summarized, efficiently.

To take full advantage of the complementarity between CNN and RNN in representing different aspects of music sound, some researchers proposed to construct hybrid architectures of CNN and RNN for music classification [2], [4], [12], [13]. In [13], a hybrid architecture consisting of the paralleling CNN and Bi-RNN blocks was proposed. CNN and Bi-RNN were adopted to extract spatial feature and temporal frame orders, respectively. Finally, the outputs of CNN and Bi-RNN are fused to obtain the whole feature, which was then used for classification. It was verified that the Bi-RNN block was an excellent complement to CNNs. In [4], three channels of CNNs with different shapes of filters were applied on the spectrogram image of the music sound, respectively, to extract its pitch-, tempo- and bass-relevant descriptors, respectively. Then, the outputs of each CNN channel were concatenated and put into

Manuscript received June 21, 2019.
Manuscript revised October 19, 2019.
Manuscript publicized December 5, 2019.

[†]The authors are with the School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China.
a) E-mail: nchen@ecust.edu.cn

DOI: 10.1587/transinf.2019EDP7175

an LSTM to extract the long-term dependency-based property contained in the concatenated feature sequence. It was shown that the adding of the LSTM layer helps to enhance classification accuracy, greatly, especially for the emotion classification task. The possible reason is that the detection of emotion is more dependent on the long-term dependency-based property, which can be grasped by LSTM, precisely. Despite the high performances achieved by the hybrid architecture proposed in [4], it has some shortcomings: i) Since the probability distributions of the features (such as pitch, tempo, or bass) may vary, greatly, among different music datasets (GTZAN, Ballroom, or Emotion), it may be impossible to find a fixed size of the filter for each CNN channel to make them fit for different music datasets. ii) For each CNN channel, only one hidden layer is included, thus the hierarchical feature of the sound may not be grasped, precisely. In addition, the shapes and the number of the filters in the hidden layer should be set manually in advance, which reduces its flexibility, greatly. iii) The training of the LSTM architecture takes too much time, which is not desired in real applications.

To solve these problems, a hybrid architecture-based music classification scheme, which takes full advantage of the merits of RCNN [19] in representing the nonlinear property and inherent hierarchical nature of music, and that of Broad Learning (BL) in ensuring training efficiency [20], is proposed in this paper. On the one hand, since RCNN can set the number of hidden layers and that of filters in each layer, flexibly, according to the property of the input, it may be robust to the variance contained in different music datasets well. On the other hand, since sparse autoencoder is adopted in the training procedure of BL, it can overcome the randomness nature of the input feature matrix. Also, due to the high training efficiency of BL, the proposed scheme may be fit for real applications. Experimental results on three benchmark datasets demonstrate that: i) The proposed model, denoted as RCNNBL, performs better than the hybrid deep learning architecture-based one [4]. ii) Compared with Multi-Channel CNN (MCC)-based feature extraction scheme [4], RCNN-based one achieves higher classification accuracy. iii) Compared with the LSTM-based prediction strategy in [4], the BL-based one achieves much lower training time and a little bit higher classification accuracy.

2. Proposed Scheme

As shown in Fig. 1, the proposed scheme is composed of three RCNNBL architectures, which are represented as red, green, and purple, respectively. The final predicted tag is obtained by applying the majority voting on the predicted ones achieved by the above three architectures. Each architecture comprises 4 steps: preprocessing, RCNN-based feature extraction, BL-based tag prediction, and majority voting.

2.1 Preprocessing

First, the original music sound, whose sampling rate is 44.1 kHz, is segmented into frames of 2048 samples (50% overlap) with the Blackman-harris window. Then, Discrete-Time Fourier Transform (DTFT) is applied on each frame to obtain the spectrum, which is further filtered by a Mel filter bank that is composed of k_1 filters to generate the Mel-spectrogram. Finally, the whole Mel-spectrogram sequence is split into L chunks of k_2 frames, which are denoted as $I_q, q = 1, \ldots, L$, along the time axis. Thus, the size of each chunk is $k_1 \times k_2$.

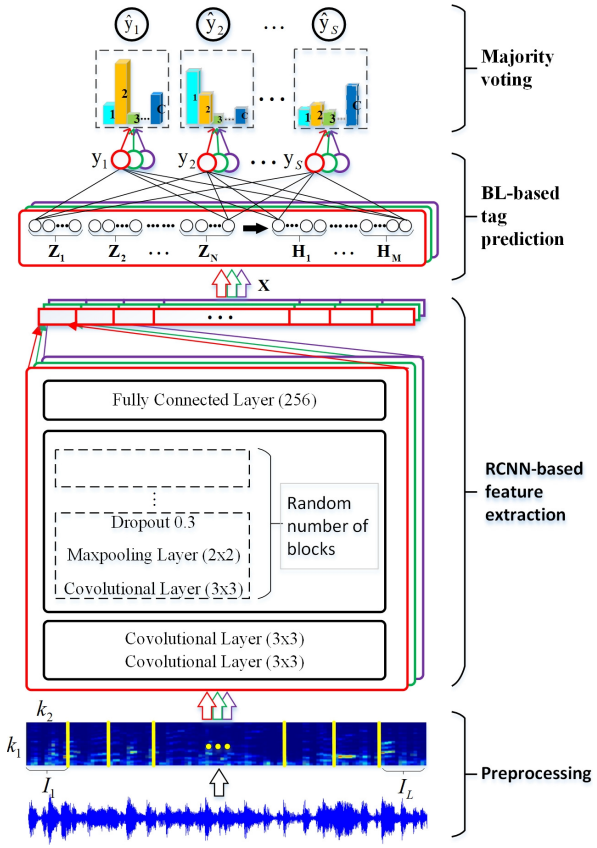

Fig. 1 The block diagram of the proposed RCNNBL model.

2.2 RCNN-Based Feature Extraction

In the hybrid deep architecture-based classification scheme proposed in [4], three CNN channels, which are composed of different shapes of (vertical bar, horizontal bar, and rectangular) of filters, were adopted to extract pitch-, tempo- and bass-based feature from the Mel-spectrogram, respectively. The outputs of each CNN channel are concatenated to obtain the combined feature. However, as shown in Figs. 2 (a), (b) and (c), it has been verified that for each CNN channel, different sizes of filters are set to obtain the best performance on different datasets (GTZAN, Ballroom, and Emotion). As a result, it is quite difficult to obtain a filter size combination that makes all three CNN channels perform the best on different datasets. In addition, in each CNN channel of in [4], there is only one hidden layer, so the hierarchical nature of the input sound may not be grasped,

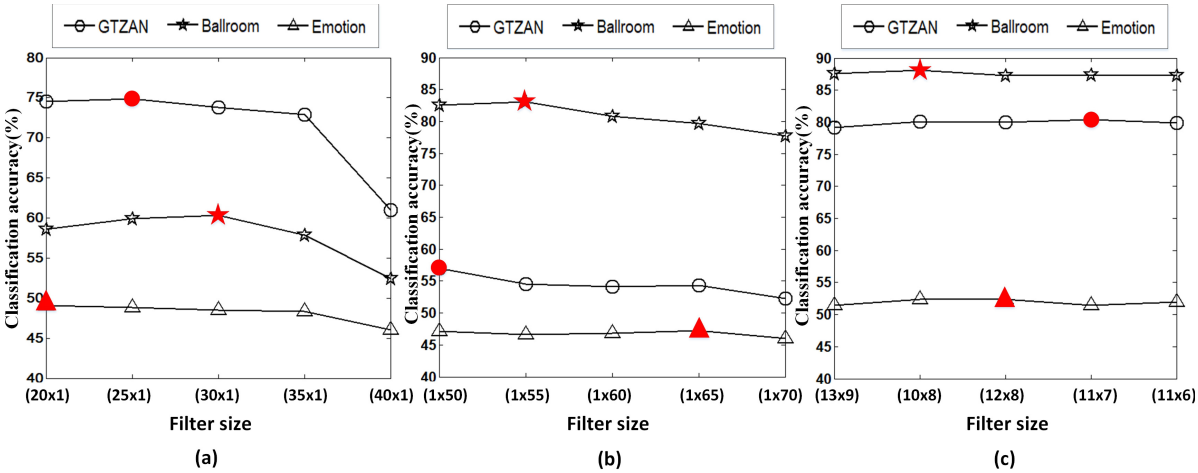

Fig. 2 Classification accuracy obtained by CNN channels with different filter shapes (a) vertical bar, (b) horizontal bar, and (c) rectangle at different sizes on three benchmark datasets. The red mark corresponds to the highest classification accuracy.

precisely. Last and most importantly, in [4], the number of filters in the hidden layer needs to be set in advance, which is inflexible and unreasonable.

To solve the above problems contained in the scheme proposed in [4], the RCNN is adopted in the proposed model to extract features from the Mel-spectrogram. As shown in Fig. 1, the RCNN-based feature extraction stage of the proposed scheme is composed of two stacks of RCNN. First, each of the Mel-spectrogram chunk, $I_q, q = 1, \ldots, L$, is passed through the first stack of RCNN, which is composed of two convolutional layers. Each layer comprises a random number ($128 \sim 384$) of 3×3 filters. 3×3 filter shape is chosen because it is the smallest size to capture the notion of left/right or up/down center [22]. Second, the output of the first stack of RCNN is put into the second stack of RCNN, which is composed of a random number ($3 \sim 5$) of CNN blocks. Each CNN block is made up of the convolutional layer, the max-pooling layer, and the dropout layer. Specifically, each convolutional layer includes a random number ($128 \sim 384$) of 3×3 filters, and the convolution stride is fixed to 2 pixels. Third, the output of the second stack of RCNN is passed through a Fully Connected (FC) layer, which is composed of P nodes. Finally, the outputs of the FC layer for all Mel-spectrogram chunks are concatenated to obtain the whole feature of the input music sound. It should be noted that to train the RCNN architecture, for each convolutional layer the ReLU activation function and 30% dropout is chosen, and for each stack, the batch size of 128 and 80 epoches are used. In addition, the Adam [23] with the learning rate of 1e-3 is adopted for learning rate control, and the cross-entropy is used as the loss function.

2.3 BL-Based Tag Prediction

Assuming that the feature of the s-th song is $\boldsymbol{x}_s, s = 1, \ldots, S$, where S is the number of music songs contained in the music collection, the feature matrix of the whole music dataset can be denoted as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_s]$.

First, the original feature matrix $\boldsymbol{X}$ is linearly projected to N mapped feature matrices, which are denoted as $\boldsymbol{Z}_i, i = 1, \ldots, N$, with Eq. (1)

$$\boldsymbol{Z}_i = \phi(\boldsymbol{X}\boldsymbol{W}_{e_i} + \boldsymbol{\beta}_{e_i}), i = 1, \ldots, N \tag{1}$$

where, ϕ is a linear function, $\boldsymbol{W}_{e_i}$ and $\boldsymbol{\beta}_{e_i}$ are the random weights with proper dimension.

Next, all the obtained mapped features are concatenated with Eq. (2) to obtain the linear-combined feature, denoted as $\boldsymbol{Z}$.

$$\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N] \tag{2}$$

Then, M nonlinear mappings (see Eq. (3)) are performed on $\boldsymbol{Z}$ to obtain the nonlinear feature matrices, denoted as $\boldsymbol{H}_j, j = 1, \ldots, M$

$$\boldsymbol{H}_j = \xi(\boldsymbol{Z}\boldsymbol{W}_{h_j} + \boldsymbol{\beta}_{h_j}), j = 1, \ldots, M \tag{3}$$

where ξ is tansig function, $\boldsymbol{W}_{h_j}$ and $\boldsymbol{\beta}_{h_j}$ are the random weights with proper dimension.

Finally, all the obtained nonlinear feature matrices $\boldsymbol{H}_j, j = 1, \ldots, M$ are concatenated to obtain the non-linear combined feature matrix, denoted as $\boldsymbol{H}$, with Eq. (4). The number of enhancement nodes is Q.

$$\boldsymbol{H} = [\boldsymbol{H}_1, \boldsymbol{H}_2, \ldots, \boldsymbol{H}_M] \tag{4}$$

Assuming that the true labels of the songs in the whole training set are $\boldsymbol{Y} = \{y_s | s = 1, \ldots, S\}$, then, the broad learning can be represented as Eq. (5)

$$\boldsymbol{Y} = [\boldsymbol{Z}|\boldsymbol{H}]\,\boldsymbol{W} \tag{5}$$

where $\boldsymbol{W}$ is the connecting weights for the broad structure and it can be obtained through the ridge regression approximation shown in Eq. (6)

$$\boldsymbol{W} = [\boldsymbol{Z}|\boldsymbol{H}]^{+}\,\boldsymbol{Y} \tag{6}$$

So, the training of the BL structure is quite efficient. Then, at the testing stage, the predicted tags of the songs in the whole testing set $\widehat{\boldsymbol{Y}}$ can be calculated with Eq. (7)

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{X}}\boldsymbol{W} \tag{7}$$

where $\widehat{\boldsymbol{X}}$ is the feature matrix of the testing set.

2.4 Majority Voting

To ensure the robustness and generalization, three RCNNBL architectures, which are represented as red, green and purple, respectively, in Fig. 1, are combined in the proposed model. Then, the majority voting is applied to the outputs of the above three RCNNBL architectures to obtain the final prediction of the tag. Specifically, assuming that the output obtained with the t-th architecture for the s-th song is $\widehat{y}_{st}, t = 1, \ldots, T, s = 1, \ldots, S$, the predicted label of the s-th song, denoted as $\widehat{y}_s$, can be obtained with Eq. (8)

$$\widehat{y}_s = \arg\max_t \left[softmax(\widehat{y}_{st}) \right] \tag{8}$$

3. Experiments

To evaluate the effectiveness and efficiency of the proposed classification model in comparison with the deep learning-based ones [4], [17], three benchmark datasets that are composed of different types of music are included in the experiments. To make a fair comparison, 10-fold cross validation is performed for each scheme to obtain the classification accuracy. Each dataset is randomly split into training, validation, and testing sets with the proportion of 8:1:1. The parameters of the proposed scheme are listed in Table 1, and those of the 3-channel (MCC-3)-based scheme and MCCLSTM scheme are set as shown in [17] and [4], respectively. All the experiments are carried out on NVIDIA TITAN Xp GPU with 12 GB memory.

3.1 Datasets

- GTZAN dataset [24]: This dataset and Ballroom dataset are adopted to test the performance of the proposed scheme in genre classification. The 1000 songs in this dataset are classified into 10 genres (classical, country, disco, hiphop, jazz, rock, blues, reggae, pop, and metal). For each genre, there are 100 songs. The length, sampling rate, and quantization precision are 30 seconds, 22050 Hz, and 16-bits/sample, respectively.
- Ballroom dataset [25]: This dataset consists of 698 songs of ballroom dance music. All the songs (each lasts 30 seconds) are divided into 8 genres, such as cha-cha-cha (111), jive (60), quickstep (82), rumba (98), samba (86), tango (86), viennese valtz (65), and slow waltz (110).
- Emotion dataset [26]: This dataset, which is composed of 2906 songs, is included to test the performance of the proposed scheme in emotion classification. Compared with GTZAN and Ballroom, the size of this dataset is much larger. The tracks of this dataset are classified into 4 classes (angry 639, happy 753, relax 750, and sad 764). The length of the tracks varies from 30 seconds to 60 seconds. In the experiment, only the first 30 seconds of each track are used.

Table 1 Parameters for the proposed model.

Steps	Parameters
Preprocessing	$k_1 = 40$ $k_2 = 40$ (GTZAN), 80 (Ballroom, Emotion)
RCNN-based feature extraction	$P = 256$
BL-based tag prediction	$N\epsilon\{10, 11\}$ $Q\epsilon\{100, 110, 120, 130, 140\}$
MV	$T = 3$

3.2 Effectiveness of RCNN-Based Feature Extraction

In this experiment, to verify the superiority of RCNN in feature extraction over MCC-3, the output of RCNN in the proposed model and that of MCC-3 in [4] are directly used for classification, which is performed by softmax function and majority voting. The comparison results shown in Table 2 demonstrate that the RCNN-based feature extraction method performs much better than the MCC-3-based one on all three datasets. The RCNN-based one even achieves higher performance than the whole scheme proposed in [4], which is based on the combination of CNN and LSTM and is denoted as MCCLSTM in this paper. In addition, unlike the MCC-3-based one, which needs to set the size and the number of filters in advance manually, in RCNN, the shape of the filter is fixed as 3×3 and the number of the filters is randomly chosen between 128 and 384 according to the input, automatically. Thus, the RCNN-based feature extraction method is more convenient and flexible. Considering that the block size of the RCNN architecture in the proposed model is between 3 and 5, and the number of convolutional layers is between 128 and 384, the performances of RCNN and those of the Fixed CNN (FCNN) architectures with 3 blocks and 128 convolutional layers, or 5 blocks and 384 convolutional layers, are compared on Ballroom dataset (see Fig. 3). It can be seen that when compared with FCNN, the accuracy of RCNN is a little bit lower. But, when BL is combined, the proposed scheme performs better than the combination of FCNN and BL.

Table 2 Effectiveness of RCNN in feature extraction in comparison with deep learning-based schemes [4].

Schemes	Classification accuracy: mean(%)±std(%)		
	GTZAN	Ballroom	Emotion
MCC-3 [4]	82.90 ± 2.11	89.45 ± 2.18	52.96 ± 2.78
MCCLSTM [4]	84.69 ± 1.76	91.90 ± 2.33	56.33 ± 2.15
RCNN	**88.20 ± 2.67**	**92.27 ± 2.22**	**56.59 ± 2.53**

3.3 Efficiency of BL-Based Prediction

To verify the training efficiency of the BL-based prediction model adopted in the proposed scheme in comparison with the LSTM-based one adopted in [4], BL and LSTM are adopted to make prediction based on the feature extracted by MCC-3 and RCNN schemes, respectively. For the same feature (MCC-3 or RCNN), the training time, which is obtained as the mean value of the training time for the 10-fold cross validation, needed by the BL-based model and LSTM-based one is compared in Table 3. It is obvious that: i) The training efficiency of the BL-based model is much higher than that of LSTM-based one, especially on the Emotion dataset. ii) Unlike the LSTM-based model, whose training time changes greatly with the size of the datasets, the training time of BL-based one remains stable among the three datasets. iii) When RCNN-based feature is considered, on the largest dataset included, Emotion, the BL-based prediction model will save about 9 minutes than LSTM-based one. While, in real application, the music collection may include more than 200,000 songs. In such case, the BL-based model will save more than 10 hours' training time. In addition, the training speed of BL and that of three-layer MLP when equivalent classification accuracies (95.71% and 94.56%, respectively) are achieved is compared on Ballroom. The result is that the training time of three-layer MLP, which is 135.39s, is much larger than that of BL, which is 49.48s.

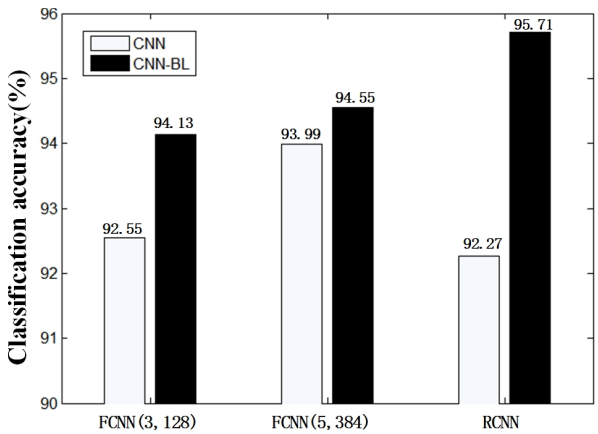

Fig. 3 Classification accuracy comparison between the FCNN-based and RCNN-based schemes on the Ballroom dataset.

3.4 Effectiveness of BL-Based Prediction

To verify the effectiveness of the BL-based prediction scheme in comparison with LSTM-based one [4], BL and LSTM are adopted to make tag prediction based on the same feature (MCC-3 or RCNN), respectively. From the experimental results shown in Table 4, it can be seen that, in most cases, for the same input feature (MCC-3 or RCNN), BL-based prediction model can achieve a little bit higher classification accuracy than LSTM-based one, which means that

Table 3 Training efficiency comparison between BL-based prediction model and LSTM-based one [4].

Schemes	Training time (seconds)		
	GTZAN	Ballroom	Emotion
MCCLSTM [4]	95.63	73.84	247.96
MCCBL	**32.96**	**32.94**	**40.66**
RCNNLSTM	260.28	179.18	615.57
RCNNBL	**76.84**	**49.48**	**55.74**

Table 4 Effectiveness of BL-based prediction model in comparison with LSTM-based one [4].

Schemes	Classification accuracy: mean(%)±std(%)		
	GTZAN	Ballroom	Emotion
MCCLSTM [4]	84.69 ± 1.76	91.90 ± 2.33	**56.33 ± 2.15**
MCCBL	**85.24 ± 1.42**	**93.21 ± 2.12**	55.81 ± 2.64
RCNNLSTM	89.50 ± 1.96	93.84 ± 2.21	57.33 ± 3.03
RCNNBL	**90.50 ± 1.69**	**95.71 ± 2.03**	**59.56 ± 2.16**

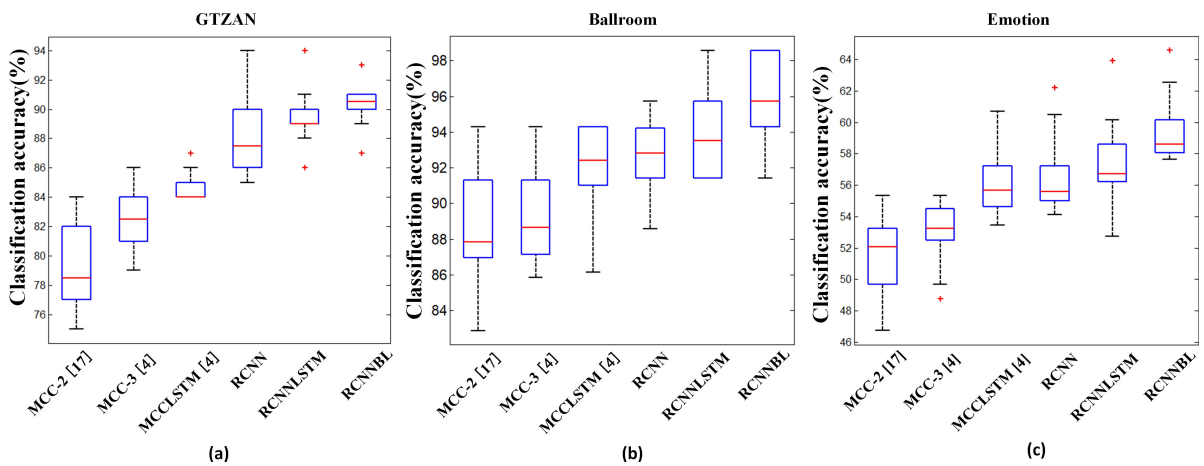

Fig. 4 Comparison of the classification accuracy distribution obtained with 10-cross-validation by different schemes on (a) GTZAN, (b) ballroom, and (c) emotion dataset.

the BL-based prediction strategy may help to enhance classification accuracy as well. The only abnormal case occurs on the Emotion dataset when the MCC-based feature is considered.

3.5 Performance Comparison with State-of-the-Art Schemes

In Fig. 4, the distributions of the classification accuracy (obtained by 10-cross-validation) achieved by different models (MCC-2 [17], MCC-3 [4], MCCLSTM [4], RCNN, RCNNLSTM, and RCNNBL) are compared on three datasets. The horizontal red line in each box is the median of the classification accuracies obtained by 10-cross-validation. The lower and upper horizontal block lines of the box indicate the 25- and 75-percentiles, respectively. The horizontal black lines further above or below represent the furthest points not considered outlines. Points beyond this range are depicted as red plus signs.

It can be observed in Fig. 4 that: i) The proposed model achieves higher median classification accuracy than the other 5 schemes on all three datasets. ii) The lowest classification accuracy obtained by the proposed scheme is equal to or higher than those of the other 5 schemes on all three datasets. iii) When compared with the hybrid deep architecture-based scheme MCCLSTM [4], the proposed scheme enhances the classification performance, obviously, on all three datasets. iv) By comparing the performances obtained by MCCLSTM [4], RCNNLSTM and RCNNBL on each dataset, we may draw the conclusion that both RCNN and BL contribute to the performance enhancement of the proposed scheme.

4. Conclusions

In this article, we present a hybrid architecture, which combines RCNN and BL, for music classification tasks. The proposed model aims to take advantage of the effectiveness and flexibility of RCNN in representing the music's nonlinear and hierarchical features, and the model training efficiency of BL to enhance both the effectiveness and the efficiency of music classification tasks. Extensive experimental results on three benchmark datasets (GTZAN, Ballroom and Emotion) demonstrate that the proposed hybrid architecture can enhance both the classification accuracy and prediction efficiency, greatly.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 61771196, 61671156, 61872143).

References

[1] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," 17th International Society of Music Information Retrieval (ISMIR), pp.805–811, Aug. 2016.

[2] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," 16th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.2392–2396, March 2017.

[3] E.J. Humphrey, J.P. Bello, and Y. LeCun, "Feature learning and deep architectures: new directions for music informatics," Journal of Intelligent Information Systems, vol.41, no.3, pp.461–481, 2013.

[4] N. Chen and S. Wang, "High-level music descriptor extraction algorithm based on combination of multi-channel CNNs and LSTM," 18th International Society of Music Information Retrieval (ISMIR), pp.509–514, Oct. 2017.

[5] J. Pons, O. Nieto, M. Prockup, E.M. Schmidt, A.F. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," 19th International Society for Music Information Retrieval Conference (ISMIR), pp.637–644, Sept. 2018.

[6] J. Schluter and S. Bock, "Improved musical onset detection with convolutional neural networks," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6979–6983, May 2014.

[7] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked convolutional and recurrent neural networks for music emotion recognition," Sound and Music Computing (SMC) Conference, pp.208–214, Oct. 2017.

[8] J. Deng and Y.-K. Kwok, "Large vocabulary automatic chord estimation using bidirectional long short-term memory recurrent neural network with even chance training," Journal of New Music Research, vol.47, no.1, pp.53–67, 2017.

[9] S. Stober, D.J. Cameron, and J.A. Grahn, "Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings," Advances in Neural Information Processing Systems (NIPS), pp.1449–1457, Dec. 2014.

[10] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: a multi-scale neural network for end-to-end audio source separation," 19th International Society for Music Information Retrieval Conference (ISMIR), pp.334–340, Sept. 2018.

[11] A. Abdul, J. Chen, H.Y. Liao, and S.H. Chang, "An emotion-aware personalized music recommendation system using a convolutional neural networks approach," Applied Sciences, vol.8, no.7, pp.1–16, 2018.

[12] J. Dai, S. Liang, W. Xue, C.J. Ni, and W.J. Liu, "Long short-term memory recurrent neural network based segment features for music genre classification," International Symposium on Chinese Spoken Language Processing (ISCSLP), pp.1–5, IEEE, 2016.

[13] L. Feng, S. Liu, and J. Yao, "Music genre classification with paralleling recurrent convolutional neural network," arXiv preprint arXiv:1712.08370, 2017.

[14] R.L. Aguiar, Y.M.G. Costa, and C.N. Silla, "Exploring data augmentation to improve music genre classification with ConvNets," 2018 International Joint Conference on Neural Networks (IJCNN), pp.1–8, IEEE, July 2018.

[15] T. Raissi, A. Tibo, and P. Bientinesi, "Extended pipeline for content-based feature engineering in music genre recognition," 43rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.2661–2665, IEEE, April 2018.

[16] J. Pons and X. Serra, "Randomly weighted CNNs for (music) audio classification," 44th International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.336–340, IEEE, May 2019.

[17] J. Pons, T. Lidy, and X. Serra, "Experimenting with musically motivated convolutional neural networks," 14th IEEE International Workshop on Content-Based Multimedia Indexing (CBMI), pp.1–6, June 2016.

[18] K.H. Wong, C.P. Tang, K.L. Chui, Y.K. Yu, Z.L. Zeng, X. Jiang, G. Chen, and Z. Chen, "Music genre classification using a hierarchical long short term memory (LSTM) model," 3rd International Workshop on Pattern Recognition (IWPR), vol.10828, pp.108281B, May 2018.

[19] K. Kowsari, M. Heidarysafa, D.E. Brown, K.J. Meimandi, and L.E. Barnes, "RMDL: random multimodel deep learning for classification," International Conference on Information System and Data Mining (ICISDM), pp.19–28, April 2018.
[20] C.L.P. Chen and Z.L. Liu, "Broad learning system: an effective and efficient incremental learning system without the need for deep architecture," IEEE Trans. Neural Netw. Learn. Syst., vol.29, no.1, pp.10–24, 2018.
[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol.15, no.1, pp.1929–1958, 2014.
[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 3rd International Conference on Learning Representations (ICLR), May 2015.
[23] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," The 3rd International Conference for Learning Representations (ICLR), pp.1–15, May 2015.
[24] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Transactions on Speech and Audio Processing, vol.10, no.5, pp.293–302, 2002.
[25] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," 25th International Conference on Audio Engineering Society (AES), pp.196–204, 2004.
[26] V. Kirandziska and N. Ackovska, "Finding important sound features for emotion evaluation classification," European Conference on Electronics (Eurocon), IEEE, pp.1637–1644, July 2013.

Huan Tang received the B.S. degree in Photoelectric Information Engineering from the Changshu Institute of Technology, Suzhou, China, in 2017. He is currently pursuing the M.S. degree with the School of Information Science and Engineering of East China University of Science and Technology. His research interests include audio signal processing and music information retrieval.

Ning Chen received the Ph.D. degree in Electronic Engineering from Shanghai Jiaotong University in 2008. From 2008 to 2010, she worked as a postdoctoral in Shanghai University. In 2010, she joined the School of Information Science and Engineering, East China University of Science and Technology, Shanghai, P.R. China. Now, she is a professor of signal processing. Her main research interest include the analysis and processing of audio and music signals using techniques, such as matrix factorization, probability theory, and deep learning.