Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09643-6

Bottom-up broadcast neural network for music genre classification

Caifeng Liu[1] · Lin Feng[2] · Guochao Liu[3] · Huibing Wang[4] · Shenglan Liu[2]

Received: 23 October 2019 / Revised: 30 June 2020 / Accepted: 18 August 2020 /
Published online: 27 October 2020

Abstract

Music genre classification based on visual representation has been successfully explored over the last years. Recently, there has been increasing interest in attempting convolutional neural networks (CNNs) to achieve the task. However, most of the existing methods employ the mature CNN structures proposed in image recognition without any modification, which results in the learning features that are not adequate for music genre classification. Faced with the challenge of this issue, we fully exploit the low-level information from spectrograms of audio and develop a novel CNN architecture in this paper. The proposed CNN architecture takes the multi-scale time-frequency information into considerations, which transfers more suitable semantic features for the decision-making layer to discriminate the genre of the unknown music clip. The experiments are evaluated on the benchmark datasets including GTZAN, Ballroom, and Extended Ballroom. The experimental results show that the proposed method can achieve 93.9%, 96.7%, 97.2% classification accuracies respectively, which to the best of our knowledge, are the best results on these public datasets so far. It is notable that the trained model by our proposed network possesses tiny size, only 0.18M, which can be applied in mobile phones or other devices with limited computational resources. Codes and model will be available at https://github.com/CaifengLiu/music-genre-classification.

Keywords Music genre classification · CNN · Spectrogram

This study was funded by National Natural Science Foundation of People's Republic of China (No.61672130, No.61602082, No.91648205, No.61627808, No.61972064), the National Key Scientific Instrument and Equipment Development Project (No. 61627808), the LiaoNing Revitalization Talents Program (No. XLYC1806006).

✉ Lin Feng
fenglin@dlut.edu.cn

Caifeng Liu
liucaifeng12345@mail.dlut.edu.cn

[1] Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, China

[2] School of Innovation and Entrepreneurship, Dalian University of Technology, Dalian, China

[3] Department of Functional RD, JD, Beijing, China

[4] College of Information Science and Technology, Dalian Maritime University, Dalian, China

1 Introduction

With the rapid development of multimedia technology, a tremendous number of digital audio are uploaded on the Internet. Except for the benefits brought by these audio tasks, the explosive growth of these audio causes fatal effects on various aspects. Therefore, managing these audio appropriately is a burdensome task crying out for reliable solutions. Researchers all over the world have devoted plenty of efforts to deal with various audio. Music information retrieval (MIR) is one of those studies, which aims to retrieve information from music, e.g. genre, emotion, mood. Due to the urgent need for many applications, such as recommender systems, music generation, automatic categorization music search, MIR has attracted attention widely. Classification as a basic understanding of the music field has become an essential tool for MIR to analyze and process the music information. As a core issue of MIR, genre classification focuses on assigning a specific genre (classical, rock, jazz, etc.) to an unknown music clip. Expert annotation is notoriously expensive and intractable for large catalogs, Therefore, automatic content-based genre recognition of music will be the crucial service for content distribution vendors.

Even though researchers have proposed various algorithms from different perspectives, most of them rely on excellent hand-crafted features or construct appropriate classifiers for music data. Spectrograms of music data have been proved to be one effective tool to describe audio signals. Similar to the images, spectrograms are also visual representations that can adequately maintain the information of time-frequency from music data. It builds a bridge between the algorithms for both image data and audio signals. Some mature algorithms for image processing can directly be adopted by related methods for audio signals. For the stage of feature extraction, the spectrograms are able to store most time-frequency information in their texture, which is of vital importance for the representation of various audio. Furthermore, there are many descriptors which can exploit the texture information to some extent, including Local Binary Patterns (LBP), Gabor Filters, Local Phase Quantization, etc.. After feature extraction, classification on the extracted features can directly determine the performances of music genre recognition. Some of those classification algorithms are common choices in this field, such as Support Vector Machine (SVM), Gaussian Mixture Models (GMM), and Music Classifier Systems (MCS) with different fusion strategies, etc. [8]. Even though the feature representations and classification algorithms for music collections seem to be maturing, it's difficult for those traditional hand-crafted methods to design appropriate features for a specific task automatically. Therefore, it is far more beneficial to adopt a data-driven method rather than designing those hand-crafted ones.

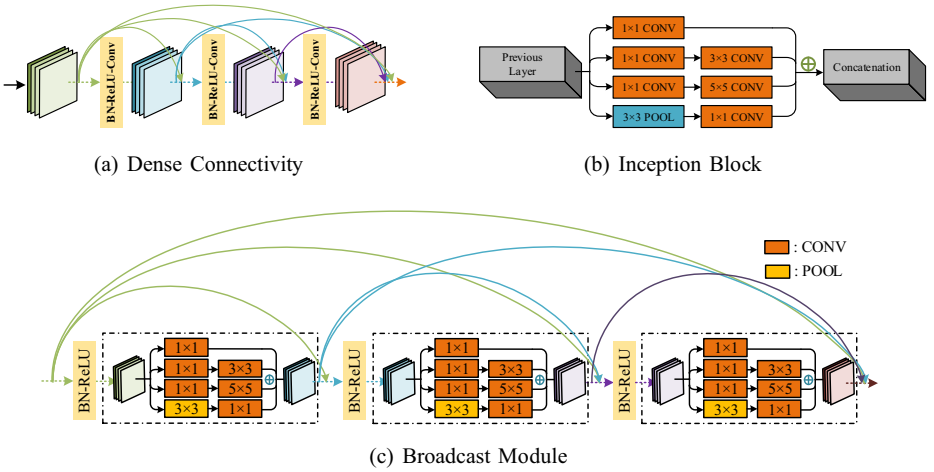
Over the last decade, we have witnessed a surge of Convolutional Neural Network (CNN) architectures which have achieved satisfying performances in many fields like image recognition [42] and natural language processing. Meanwhile, Music genre classification has also been inspired by the remarkable successes of CNN. It is well known that CNNs can extract enough information from images due to their hierarchical structures. Low-level features, such as underlying texture, etc., are constructed into high-level semantic information through all layers of CNN. Similar to images, music also consists of hierarchical structures, which inspires us to develop an appropriate CNN model to deal with the problem of music classification. For instance, pitch and loudness combine over time to form chords, melodies, and rhythms. And all these elements above can form the whole music layer by layer. Furthermore, it has been verified that CNNs are very sensitive to the textural information [9] of images. The special ability of CNNs can help the task of music classification to exploit abundant information from the spectrogram which contains rich texture information from music signals.

Up to now, most of the CNN-based music classification models are constructed by directly utilizing those mature architectures to deal with this problem. For example, Choi et al. [3] introduced a musical transfer learning system, in which a CNN model was trained on a large music dataset [1] as a feature extractor and an SVM classifier was stacked on it and Jakubik et al. [15] introduced two Recurrent Neural Network (RNN) architectures from image domain with a different mechanism of gating: Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). They reported the accuracies of 91% and 92% respectively on GTZAN, which showed their potential for music content analysis.

Even though these architectures have achieved excellent performances in the music domain, the results have not been nearly as convincing as they have been in the visual realm. Most CNN-based models directly fed the visual spectrogram representations into these CNNs without any modification. Because traditional CNNs are constructed just for image processing tasks, it cannot achieve good performances without awareness of the difference between spectrogram of audio and images. Stimulated by the problems above, two main motivations of our paper are listed as follows:

- Even though the genres are positioned in different levels or time-scales in each hierarchy, previous deep learning-based works predict the genre from the same scale of time and level frequency, which is similar with the task of image classification. However, sound events are accumulated by frequency over time domain which causes the individual genre has different performance sensitivity to different time scales and levels of features. Therefore, it is necessary to design a specific CNN structure which can comprehensively handle multi-scale of audio features.
- Previous network structures of music genre classification mainly focus on abstracting high-level semantic features layer by layer. It leads to a massive loss of lower-level features including a large amount of critical information for making the decision. However, the low-level features tend to be more contributed for improving the genre classification performance [3]. Therefore, how to construct an appropriate CNN structure to maximally abstract high-level information and preserve the lower-level features simultaneously, which is just for the task of music classification is of vital importance but challenging.

Based on the analysis above, the direct way to deal with these problems is to exploit an appropriate CNN model that can make full use of both the high-level semantic information and the low-level features from various music. In this paper, we surveyed the problems in the field of music classification and proposed a novel architecture named Bottom-up Broadcast Neural Network (BBNN) which adopts a relatively wide and shallow structure. The main idea of the BBNN architecture is to develop an effective block and connection manner between different blocks to fully exploit and preserve the low-level information to the higher layers. So, low-level information of the spectrogram is able to participate in the decision-making layer throughout the network, which is very important for the task of music classification. Therefore, BBNN is equipped with a novel Broadcast Module (BM) which consists of Inception blocks and connects them by dense connectivity. We have shown the architecture of BM in Fig. 1c. Because the Inception block (IB) can perceive the feature maps with different scales, it is able to extract information embedded in the time-frequency of the audio signal from different scales simultaneously. Moreover, BBNN densely connects those basic blocks and transforms the low-level information to the decision-making layers, which ensures that the low-level information can be maintained as much as possible. Most Deep CNN (DCNN) models have to adopt various data-augmentation pre-processings [36] to enlarge the size of the training dataset. Compared with those traditional DCNN models,



(a) Dense Connectivity (b) Inception Block

(c) Broadcast Module

Fig. 1 Comparison of prior network structures (**a**, **b**) and proposed module Module(**c**)

our proposed BBNN has few parameters to be learned. Therefore, a smaller dataset without any data-augmentation techniques, such as Ballroom, is enough for the training stage of BBNN.

The rest of the paper is organized as follows: Section 2 are the review of recently related works and our motivation. Section 3 describes the details of proposed method. Section 4 shows the experiments. Section 5 illustrates conclusion and possible directions of future work.

2 Related work

Music genre classification based on visual representation (e.g. STFT spectrogram, mel-spectrogram, MFCC, etc.) has been successfully explored over the last years. The spectrogram has a strong ability to describe the temporal change of energy distribution over frequency bins. Since they are similar to the images and include sufficient texture information, different types of traditional texture descriptors in the computer visual field have been utilized to describe the content of these spectrograms further. These descriptors include Local Binary Patterns (LBP), Gabor Filters, Local Phase Quantization, etc.. Then, the classifier, such as Support Vector Machine (SVM), Gaussian Mixture Models (GMM), can be trained with these texture descriptors. Although these traditional ways of classification have achieved the improving results compared with human-level (70%) on several music datasets, they are still lying exactly on the feature engineering [5].

Deep neural network architectures can alleviate the need of the task-depend prior knowledge. It has been achieved dramatic advantages against hand-crafted methods in computer vision [13, 20]. In parallel, efforts were considered to apply these successful architectures for musical genre classification [18, 21]. Lee et al. [22] are the pioneers of introducing a deep learning framework to audio classification. They trained a Convolutional Deep Belief Network (CDBN) with two hidden layers to learn hierarchical features from a pre-processed spectrogram input. To some extent, the paper [22] inspired later researches on deep learning approaches applied to audio recognition tasks as feature extractors. Later, Hamel et al.

[11] stacked three hidden layers of 50 units each over frames of a spectrogram to extract abstract features, then these features were fed into a Support Vector Machine(SVM) to predict genre label for an unknown music clip. Sigtia et al. [38] further explored modifications to this system [11], in particular, using Rectified Linear Units(ReLUs) instead of standard sigmoid units and replacing the SVM classifier with Random Forest. They validated their framework on the GTZAN database and achieved 83% accuracy. Li et al. [23] employed a CNN with three convolutional layers as a feature extractor to learn musical pattern features for training a variety of decision tree classifiers available in the WEAK machine learning system [10] and obtained a classification accuracy of 84% on the GTZAN dataset. Choi et al. [3] introduced a musical transfer learning system, in which a CNN model was trained on a large music dataset [1] as a feature extractor and an SVM classifier was stacked on it. While these groups delivered better performance than that using hand-engineered features in many genre classification tasks, they are still limited to feature learning without the supervision of classifiers. This may lead to unsatisfactory prediction abilities of the trained classification models and the adopted framework still has two stages.

Different from the above architectures, most recent methods based on CNN integrated feature learning process and appropriate classifier at one stage. Jakubik et al. [15] introduced two Recurrent Neural Network(RNN) architectures from image domain with a different mechanism of gating: Long-Short Term Memory(LSTM) and Gated Recurrent Unit(GRU). They reported the accuracies of 91% and 92% respectively on GTZAN, which showed their potential for music content analysis. The work NNet2 [46] achieved an accuracy of 87% on the same dataset by proposing a new CNN structure introduced shortcut connection [12] to all layers in their network. In addition, to improve learning capacity, max- and average-pooling were combined in this network to offer more statistical information for subsequent layers. Considering that different temporal intervals of a song have different decision-making importance, Yu et al. [45] incorporate attention mechanism with Bidirectional Recurrent Neural Network. Moreover, the attention mechanism is also used in the work [32] and integrated stacking attention modules.

There are great differences as discussed in the previous section between pictorial classification and musical classification. Most of the previous methods based on CNN directly introduced the various mature network structures from the visual domain, which tend to neglect these differences. This is the reason that CNN-based methods have not achieved satisfactory classification accuracies in the musical genre recognition domain. In this article, a specialized network structure for music genre classification will be proposed.

3 Proposed design and approach

3.1 Broadcast module

The widely accepted consensus is that individual genre has different performance sensitivity to various frequency bands and time intervals. Inspired by [39], we combine convolutions with different kernel size to form an Inception block (Fig. 1b). The Inception blocks are stacked on top of each other as the basic extraction units to sufficiently learn features from multiple reception fields. It can decrease the network susceptibility to frequency-shifts in a spectrogram. To further strengthen feature propagation in the BM, we utilize dense connection paths to connect all Inception blocks in BM, which can bottom-up transmit the extracted feature maps to all subsequent blocks in BM. One of the main beneficial aspects of BM architecture is that it maximally transmit and preserve all extracted feature maps to

higher-layers so that the decision layers make a prediction based on all feature-maps in the network. Another practically useful aspect of BM design is that it aligns with the intuition that audio information should be perceived at various time-frequency scales simultaneously.

As shown in Fig. 2, BM consists of L identical Inception blocks connected to each other by dense connectivity, which allows each block to receive inputs directly from all its previous blocks. We denote $\mathbf{X}_{SL}$ outputted by shallow layers as the input of the BM, L as the number of Inception blocks. In the music genre recognition task, we fix the $L = 3$. Thus, the input of l-th block, $l = 1, ..., L$, can be represented as:

$$\mathbf{X}_l = f_I\left(\left[\mathbf{X}_{SL}, \mathbf{X}_1, ..., \mathbf{X}_{l-1}\right]\right), \tag{1}$$

where $\left[\mathbf{X}_{SL}, \mathbf{X}_1, ..., \mathbf{X}_{l-1}\right]$ refers to the concatenation of the feature maps produced by blocks $0, ..., l-1$ and f_I is a composite function of all operations in the Inception block. In each Inception block, the filter sizes of convolutions are mainly adopted $1 \times 1, 3 \times 3, 5 \times 5$ with stride 2 and then, 1×1 convolutions are utilized to compute reductions before them. Before each convolution, the BN and rectified linear activation operations are implemented. An Inception block consists of layers of the above types stacked upon each other, with occasional max-pooling layers restricted to the size 3×3 with stride 2 to halve the resolution

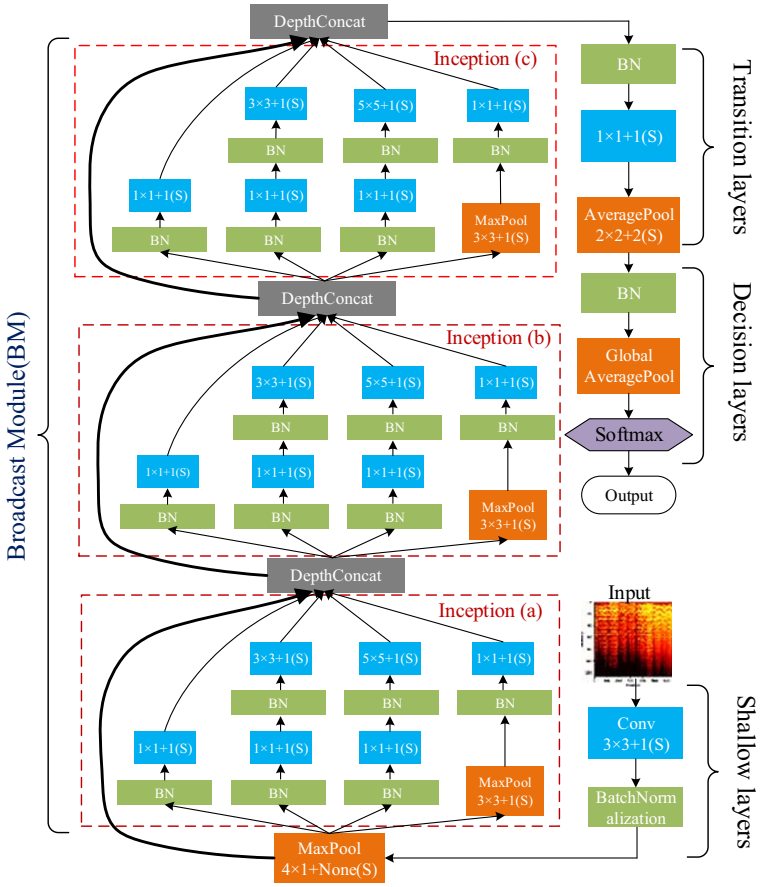


Fig. 2 The corresponding network architecture of BBNN

of the grid. The use of the Inception block is based on [39], although our implementation differs in that, we employ an extra BN layer before each convolution. This makes the network generalization ability significantly enhanced even trained on a small-scale dataset. As shown in the Table 1, the growth rate k of the BM is 128. Each block has $k_0 + k \times (l-1)$ input feature maps, where k_0 is the number of channels in the input $\mathbf{X}_{SL}$. The exact BM configurations used in the experiment are shown in Table 1. For illustration purpose, we divide the structure of Inception block into a top and bottom parts and list them respectively.

3.2 Network structure

As shown in Fig. 2, The BBNN comprises 9 layers when counting only layers with parameters (or 12 if we also count pooling layers). Each of layers implements a non-linear transformation such as Convolution (Conv), Softmax, Batch Normalization (BN) [14], and Pooling operation. Inspired by [14], We execute the BN transform immediately after each convolution operation and then use rectified linear activation (ReLU). A main beneficial aspect of BN is that it regularizes our model and reduces the need for Dropout.

All layers of the proposed network can be summarized in four parts to play different roles as follows: shallow feature extraction layers, BM, transition layers, and decision layers. The whole model aims to learn the all parameters Θ of a composite function $F(\cdot|\Theta)$, which maps input $\mathbf{X}_0$ to the output (genre) p:

$$\begin{aligned} p &= F(\mathbf{X_0}|\Theta) \\ &= f_{DL}(f_{TL}(f_{BM}(f_{SL}(\mathbf{X_0}|\theta_{SL})|\theta_{BM})|\theta_{TL})|\theta_{DL}), \end{aligned} \tag{2}$$

where the index of $f(\cdot)$ represents a composite function of corresponding part of network. Specifically, the shallow layers (the ones close to the input) include a 3×3 convolution, a BN, and a 4×1 max pooling with 1 stride. A relatively small receptive field is used to extract

Table 1 The configuration of BBNN

Type	Layers	Output size	Filter Size/Stride (Number)	Params
SL	Convolution	$647 \times 128 \times 32$	$3 \times 3/1(32)$	320
	Max Pool	$161 \times 128 \times 32$	$4 \times 1/None$	
BM	Inception (a), top	–	$[1 \times 1/1(32)\text{conv}] * 3, [3 \times 3/1\text{max pool}] * 1$	3,168
	Inception (a), bottom	$161 \times 128 \times 160$	$[3 \times 3/1(32)\text{conv}] * 1, [5 \times 5/1(32)\ \text{conv}] * 1$ $[1 \times 1/1(32)\text{conv}] * 1$	35,936
	Inception (b), top	–	$[1 \times 1/1(32)\text{conv}] * 3, [3 \times 3/1\text{max pool}] * 1$	15,456
	Inception (b), bottom	$161 \times 128 \times 288$	$[3 \times 3/1(32)\text{conv}] * 1, [5 \times 5/1(32)\ \text{conv}] * 1$ $[1 \times 1/1(32)\text{conv}] * 1$	40,032
	Inception (c), top	–	$[1 \times 1/1(32)\text{conv}] * 3, [3 \times 3/1\text{max pool}] * 1$	27,744
	Inception (c), bottom	$161 \times 128 \times 416$	$[3 \times 3/1(32)\text{conv}] * 1, [5 \times 5/1(32)\ \text{conv}] * 1$ $[1 \times 1/1(32)\text{conv}] * 1$	44,128
TL	Convolution	$161 \times 128 \times 32$	$1 \times 1/1(32)$	13,344
	Max Pool	$80 \times 64 \times 32$	$2 \times 2/2$	
DL	Global Average Pool	$1 \times 1 \times 32$	–	
	Softmax	$1 \times 1 \times 10$	–	330
Total params				**185,642**

Note that each convolution layer shown in the table corresponds the sequence BN-ReLU-Conv

the local frequency information in a short time span. After activating the local features with BN and ReLU functions, we further add a max-pooling operation. It is considered that human is more concerned about the salient tempo in a short time when recognizing the music genre. The max-pooling layer can filter out the dominant frequency in the short time interval of the mel-spectrogram. Furthermore, it makes the model possess some capacity of translation invariance. According to the above operations, the extracted local information is transmitted into each layer of BM and fused to gather evidence in support of contextual "time-frequency signatures" that are indicative of recognizing different musical genres. The structural details about BM were given in Section 3.1.

The Down-sampling layer is an essential part of convolutional networks. After extracting hierarchical features with BM, we further conduct several down-sampling layers to reduce the size of feature-maps and the number of channels that are significantly increased by the concatenation operations used in BM. These layers between BM and decision layers are referred to as transition layers, which do a BN, ReLU activation, 1×1convolution and 2×2 average-pooling with stride 2 operations.

At the final decision stage of BBNN, instead of adding fully connected layers stacked on the feature maps [15], we utilize global average pooling [24] layer to take the average of each feature map. It is easier to interpret the correspondence relations between feature maps and genres and less prone to overfitting than traditional fully connected layers. Then, the resulting vector of the global average pooling layer is fed into a softmax log-loss function which can produce a distribution over the genre labels (blues, classic, etc.).

The BBNN is designed with full consideration of computational efficiency and practicality. Here, the configurations of BBNN is described in Table 1 for demonstrating the specific architectural parameters, where the size of mel-spectrogram $\mathbf{X}_0$ is 647×128 (30s music duration). In all convolution layers, we pad zeros to each side of the input to keep size fixed. As seen from Tabel 1, the trained model has a tiny size, only 0.18M, which can be applied in individual devices including even those with limited computational resources.

4 Experiments

4.1 Datasets

GTZAN. The dataset has been widely used in many studies with the aim of music genre classification. It was collected and proposed by Tzanetakis in [40]. The labels and numbers of corresponding genres are given in Table 2.

Ballroom. The dataset [2] consists of clear and constant rhythmic patterns, which makes it suitable for recognition tasks. The specific genres and the corresponding number of each genre are listed in Table 2.

Extended Ballroom. The dataset [26] was proposed in 2016 by Marchand, which extended the original Ballroom dataset. Comparing to the original one, the extended version contains six times more tracks of better audio quality. We show in Table 2 the genre class distribution of the dataset. The imbalance of this dataset poses vast challenges for genre classification.

4.2 Preprocessing

In this work, mel-spectrogram is utilized as input to the proposed network, which is accomplished by applying a logarithmic scale to the frequency axis of Short Time Fourier

Table 2 Datasets description

GTZAN		Ballroom		Extended ballroom	
Genre	Track	Genre	Track	Genre	Track
Classic	100	Cha Cha	111	Cha Cha	455
Jazz	100	Jive	60	Jive	350
Blues	100	Quickstep	82	Quickstep	497
Metal	100	Rumba	98	Rumba	470
Pop	100	Samba	86	Samba	468
Rock	100	Tango	86	Tango	464
Country	100	Viennese Waltz	65	Viennese Waltz	252
Disco	100	Slow Waltz	110	Waltz	529
Hiphop	100			Foxtrot	507
Reggae	100			Pasodoble	53
				Salsa	47
				Slow Waltz	65
				Wcswing	23
Total	**1000**	**Total**	**698**	**Total**	**4180**

Transform (STFT). Specifically, we use Librosa tool box [29] to extract mel-spectrograms with 128 Mel-filters (bands) covering the audible frequency range (0-22050 Hz), setting a frame length of 2048 and a hop size of 1024. We can get mel-spectrogram of size 647×128. The block diagram of the whole preprocessing stage is illustrated in the Fig. 3.

Figure 4 shows the mel-spectrograms generated by preprocessing. The raw music signals are randomly selected as samples from GTZAN dataset. We can observe that there is abundant texture information in mel-spectrogram, which has the vital effect of training a robust and accurate BBNN model. Besides, the mel-spectrograms of music signals assigned different genres are distinguishable.

4.3 Training and other details

All files for each dataset are transformed into mel-spectrograms by the preprocessing program presented in Section 4.2. The mel-spectrogram with size 647×128 input to the BBNN. All the models are trained to minimize categorical cross-entropy between the predictions and truthful genre labels utilizing ADAM [19] optimizer. All three datasets use batch size 8 for 100 epochs. We set an initial learning rate is 0.01 and automatically decrease it by a factor of 0.5 when the loss has stopped improving after 3 epochs. In addition, we set up an early stop mechanism, that is, training stops when a monitored quantity has stopped improving even if the epoch does not reach 100. Fig. 5 shows the training and validation loss curves of the BBNN network on GTZAN, Ballroom, and Extended Ballroom datasets. BBNN converges to a low loss whether training set or verification set. Figure 6 shows the accuracies of training, validating and testing sets on each round of 10-folds cross-validation estimations for different datasets, which can reveal the BBNN possessing stiff training stability. We further analyze the BBNN's effect on the test sets of each dataset in more detail below.

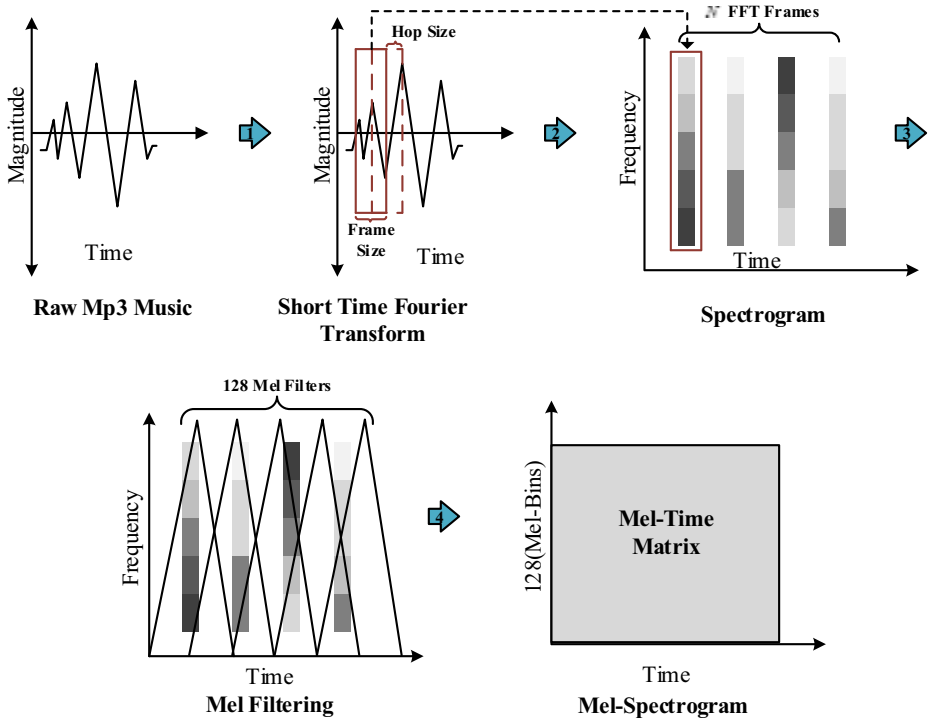


Fig. 3 Preprocessing of music signal to mel-spectrogram

Metric. Following previous works (e.g. [3]), we perform a 10-fold cross-validation to evaluate the classification accuracy across all experiments. The training, testing, and validating sets are randomly partitioned following proportion 8/1/1. The total classification accuracy is calculated as the average of 10-folds cross-validations.

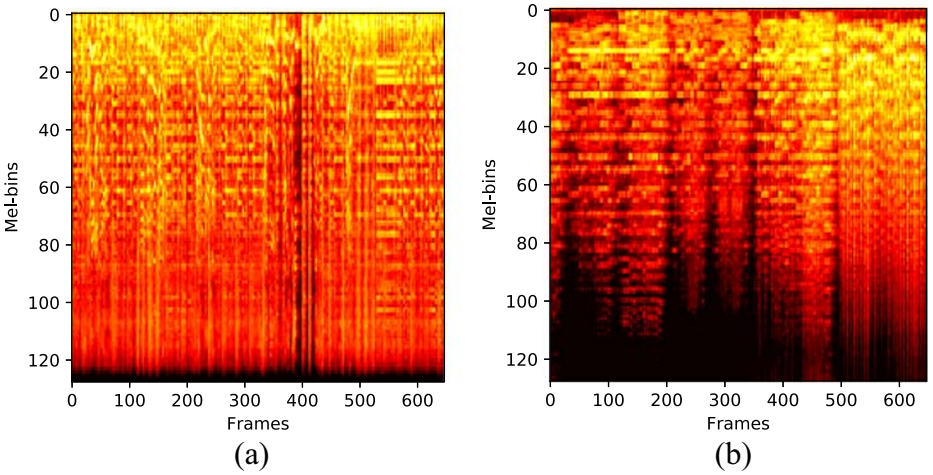


Fig. 4 Mel-spectrograms generated by preprocessing. **a** belongs to a blues genre. **b** belongs to a classic genre

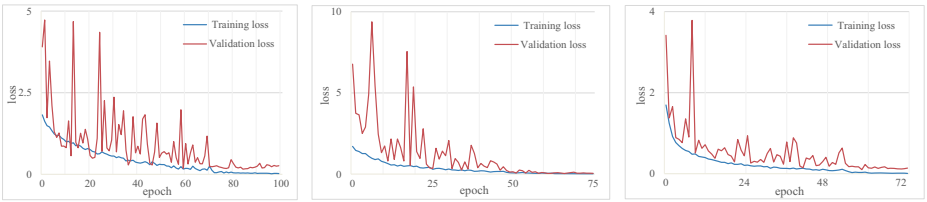


Fig. 5 Loss curves of training and validation on the different datasets: **a** GTZAN, **b** Ballroom, **c** Extended Ballroom

Experiment Platform. Our code is written by Python, based on the Keras [4] and the publicly available toolbox of preprocessing Librosa [29]. All of our experiments are running on NVIDIA TITAN Xp GPU with 12 GB memory.

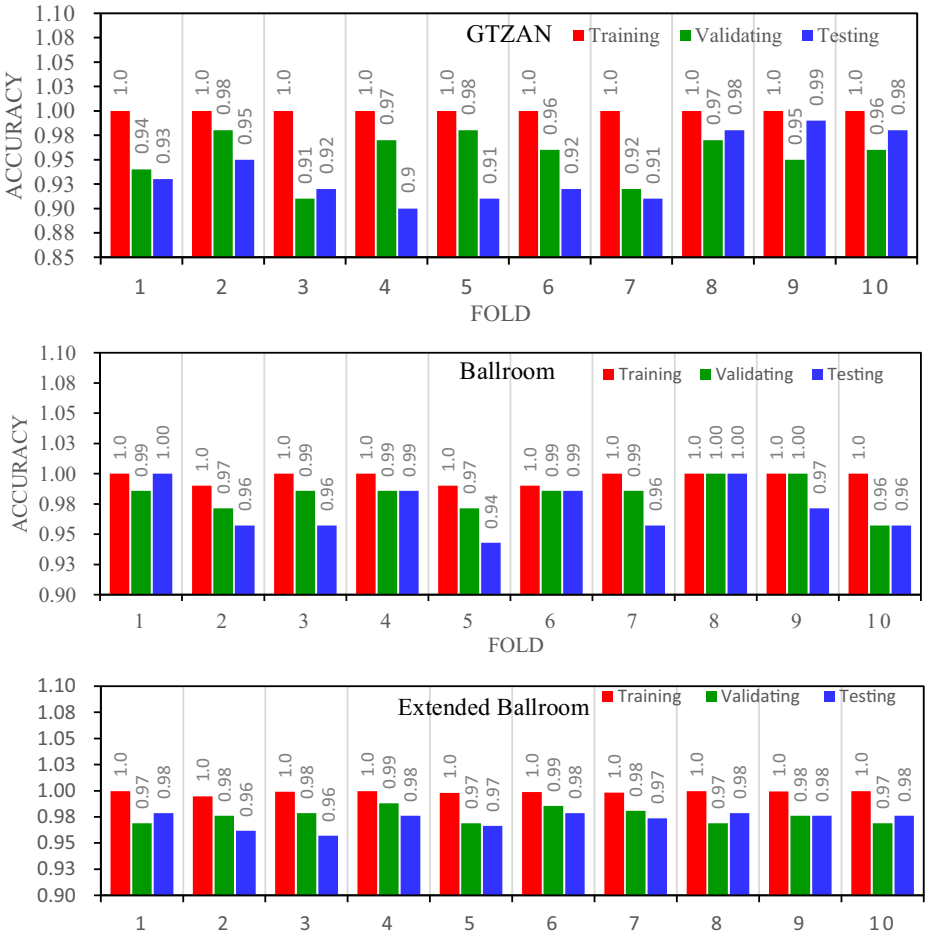


Fig. 6 Confusion matrix of the 10-folds results on different datasets

Table 3 Classification accuracy (%) on GTZAN dataset is compared across recently proposed methods (the best result is marked in bold)

Methods	Preprocessing	Accuracy
AuDeep [7]	mel-spectrogram	85.4
NNet2 [46]	STFT	87.4
Hybrid model [17]	MFCC, SSD, etc.	88.3
Transform learning [3]	MFCC	89.8
BRNN+PCNNA [45]	STFT	90.0
CVAF [31]	mel-spectrogram, SSD, etc.	90.9
MFMCNN [37]	STFT, ZCR, etc.	91.0
Multi-DNN [6]	MFCC	93.4
Ours	mel-spectrogram	**93.9**

The results of all methods have reported in the original papers or related literatures

4.4 Classification results on GTZAN

In Table 3, we compare BBNN with some recent excellent models including 6 different deep learning models and 1 traditional method based on a hand-crafted feature descriptor. Audeep [7] was based on a recurrent sequence to sequence autoencoder, which took full consideration of temporal dynamics of audio data and yielded an accuracy of 85.4%. The transform learning framework [3] obtained an accuracy of 89.8%, which is transplanted to the music genre classification task from the visual domain. Hybrid model, CVAF and MFMCNN relied on different strategies of feature fusion to improve classification accuracy and generated the accuracies of 88.3%, 90.9%, and 91.0% respectively. Multi-DNN generated a slightly lower accuracy than BBNN by using a cascaded DNN network which consumes more resources and strongly depends on an additional database to train the model.

Figure 7(left) shows the confusion matrix of 10-folds results predicted by BBNN on GTZAN. The rows and columns of the matrix represent ground-truths and their predicted labels. The diagonal numbers of matrix respectively represent correct prediction per genre and the off-diagonal entries are confusions between different genres. Confusion matrix can give the individual discrimination relation between the ground-truth and predicted genre label, which provides a better view of the general classification performance of the BBNN model. Table 4 lists the precision, recall rate and F-score of each genre corresponding confusion matrix. From the Fig. 7(left) we can see that the proposed model distinguishes most of

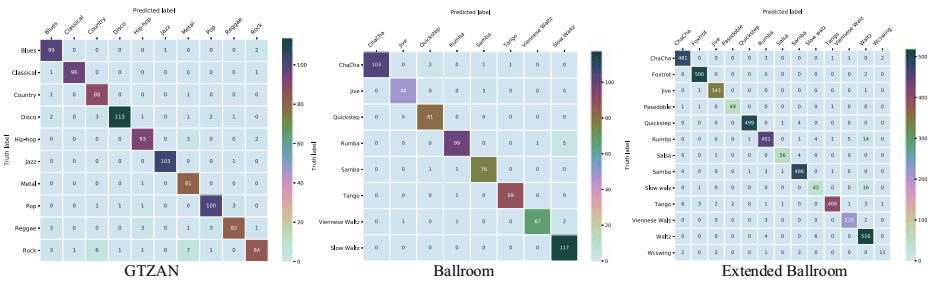


Fig. 7 Confusion matrixes on the different datasets

Table 4 Precision (%), recall rate (%) and F-score(%) of each genre obtained on the GTZAN dataset

Genre	Precision	Recall rate	F-score
Blues	90.8	97.0	93.8
Classical	98.9	97.9	98.4
Country	89.8	97.7	93.6
Disco	98.2	91.8	94.9
Hip-hop	93.0	94.9	93.9
Jazz	99.0	99.0	99.0
Metal	86.1	98.7	92.0
Pop	94.3	94.3	94.3
Reggae	94.2	88.1	91.1
Rock	93.3	80.7	86.6
Average	93.7	94.0	93.7

the genres very well, but it is highly confused to distinguish Rock from Country and Metal. One explanation is that they might share more similar frequency information which makes it more difficult to classify them in nature. Nonetheless, expert advice probably is required to improve the classification accuracy on the Pop and Rock genres. It is further discussed in Section 5. Overall, as shown by the Table 4, most genres have been correctly classified and recall rate and precision of genre Jazz even reach 99%.

4.5 Classification results on ballroom

Table 5 shows the classification results obtained on the Ballroom dataset by BBNN framework and 5 novel methods including 3 deep learning frameworks (MMCNN, MCLNN, Pons et al. [33]) and 2 traditional methods based on different hand-crafted feature representations. The work [33] presented by Marchand et al. is based on the Modulation Scale Spectrum presentation of audio (called MSS) and a modified KNN classifier is used to perform the classification achieving an accuracy of 87.6%. Based on MSS, Marchand et al. then proposed a Modulation Scale Spectrum with Auditory Statistic representation (SOTA) and used an SVM as the classifier, which boosted the recognition accuracy by about 3%. The MMCNN architecture has two layers (CNN + feed-forward) and uses two different filter shapes in the CNN layers (1-by-60 and 32-by-1). It produces more parameters (196,816)

Table 5 Classification accuracy (%) on Ballroom dataset is compared across recently proposed methods (the best result is marked in bold)

Methods	Preprocessing	Accuracy
MMCNN [35]	mel-spectrogram	87.6
MCLNN [30]	mel-spectrogram	90.4
Pons et al. [33]	mel-spectrogram	92.1
Marchand et al. [27]	MSS	93.1
SOTA [28]	MASSS	96.0
Ours	mel- spectrogram	**96.7**

The results of all methods have reported in the original papers or related literatures. Note the exception that SOTA is reported in recall rate

than BBNN to build the classification model and generates relatively lower accuracy. To preserve the inter-frames relation of a temporal signal, Medhat et al. designed a masked conditional neural network (MCLNN) which obtained an accuracy of 90.4%. Pons et al. paid attention to temporal features of audio to use wider kernels of convolution layers that span over the long time duration. They used fewer parameters (92,808) than BBNN for modeling the network but generated a relatively lower accuracy than SOTA and BBNN. For this dataset, our proposed BBNN network (accuracy of 96.7%) outperforms all the compared models.

Figure 7(middle) illustrates more detailed information about the BBNN performance in the form of a confusion matrix. The corresponding precision and recall rate for each genre is further listed in the Table 6. Through the confusion matrix, it is clear exhibited that the strong ability of BBNN to recognize the most genres, for example, Cha Cha and Quickstep. There are relatively easy confusions between the Rumba and Slow Waltz genres. This is due to the reason that the genre boundaries of Rumba and Slow Waltz are not clear cut as Cha Cha or Quickstep. Such confusion leads to their relatively lower precisions and recall rates than other genres as shown in the Table 6. It is noticeable that the BBNN model can achieve good performances in both precision and recall rates for almost genres. The average recall rate is 0.9% higher than SOTA.

4.6 Classification results on extended ballroom

Table 7 reports the comparison of BBNN with new state-of-the-art music genre classification methods in term of accuracy. BBNN achieves an accuracy of 97.2% on this dataset, which surpasses the methods including CNN-based architectures [3, 16, 34] and handcrafted feature approach [28]. Different representations of audio signals including MFCC, mel-spectrogram, and MASSS represent preprocessing programs of audio signals utilized in the corresponding methods. The first work [3] proposed by Choi et al., achieved an accuracy of 86.7% using a transfer learning framework. Specifically, a Vggnet was designed and trained for a source dataset including 244,224 music clips, then, the trained model was adopted to the Extended Ballroom dataset as a feature extractor. DLR [16] also is a transform learning framework as similar as the above mentioned work [3], but the difference is that DLR aims to learn a rhythmic representation for a source task which can be used as an input for the musical genre recognition task. Compared with these transform learning frameworks, BBNN generates a higher accuracy and does not require pre-training on the larger dataset. In RWCNN, randomly weighted CNN architecture was utilized to extract

Table 6 Precision (%), recall rate (%) and F-score(%) of each genre obtained on the Ballroom dataset

Genre	Precision	Recall rate	F-score
Cha Cha	100	96.2	98.0
Jive	97.9	94.1	96.0
Quickstep	96.4	100	98.1
Rumba	97.0	94.2	95.6
Samba	95.0	97.4	96.2
Tango	98.8	98.8	98.8
Viennese Waltz	98.5	94.3	96.4
Slow Waltz	94.3	100	97.0
Average	97.2	96.9	97.0

features for a classifier (e.g. SVM and ELM). Its accuracy is relatively lower than BBNN. The comparative experiment results strongly validate the effectiveness of BBNN model.

Figure 7(right) presents the confusion matrix of the 10-folds results produced by BBNN model corresponding accuracy in Table 7. Rows present the dataset ground-truths; Columns denote labels predicted by the BBNN model. By analyzing the confusion matrix, it can be noticed that the severe occurrence of confusion is from Rumba and Slow Waltz with Waltz. These genres are difficult to distinguish since they contain similar patterns [25].

Table 8 reports the tested precision and recall rate of BBNN model for each genre. Since Slow Waltz is prone to misclassified as Waltz, it has a relatively lower precision and recall rate. Because the training samples of genre Wcswing are very few, only accounting for 0.5% of the total samples, the BBNN can only learn severely limited discriminative information. It makes the model's generalization ability for recognizing genre Wcswing relatively poor resulting in the precision and recall rate of Wcswing are relatively low. The recognition ability of BBNN is still robust on the other unbalanced classes such as Pasodoble, Salsa, and Slow Walz, which have slightly more samples than genre Wcswing. On the whole, the recognition precisions of most genres are more than 90%, of which Quickstep and Tango are as high as 99%. Based on the results, BBNN has better classification performance even in the case of the incredibly unbalanced dataset.

4.7 Analysis

4.7.1 Ablation study

To analyze the contributions of designed components in the proposed architecture, we investigate the importance of different components in this subsection. Specifically, we define the following baseline methods:

- $\text{BBNN}_{w/o\textbf{BM}}$: to investigate the effectiveness of the proposed BM, we evaluate the performance of the architecture where the BM is replaced with a 3×3 convolutional layer.
- $\text{BBNN}_{w/o\textbf{IB}}$: it denotes the variant without using multiple scaled features. We replace the Inception block with the 3×3 convolutional layer, which can validate the two different configurations of multi-scaled features and single-scaled features.
- $\text{BBNN}_{w/o\textbf{Dense Conn}}$: in order to study the effectiveness of dense connections in our proposed network, we design the network with only short connections.

Table 7 Classification accuracy (%) on Extended Ballroom dataset is compared across recently proposed methods (the best result is marked in bold)

Methods	Preprocessing	Accuracy
Transform learning [3]	MFCC	86.7
RWCNN [34]	MFCC	89.8
BRNN+PCNNA [45]	STFT	92.7
DLR [16]	mel-spectrogram	93.7
SOTA [28]	MASSS	94.9
Ours	mel-spectrogram	**97.2**

Note the exception that SOTA is reported in recall rate. the results of all methods have reported in the original papers or related literatures

Table 8 Precision (%), recall rate (%) and F-score of each genre obtained on the Extend Ballroom dataset

Genre	Precision	Recall rate	F-score
Cha Cha	98.1	98.5	98.3
Foxtrot	98.8	99.6	99.2
Jive	98.5	99.4	98.9
Pasodoble	96.0	94.2	95.1
Quickstep	99.6	99.0	99.3
Rumba	96.7	94.5	95.6
Salsa	94.9	91.8	93.3
Samba	97.5	98.7	98.1
Slow Walz	80.7	71.1	75.6
Tango	99.0	95.3	97.1
Viennese Walz	96.8	97.7	97.3
Walz	93.1	98.1	95.5
Wcswing	78.5	57.8	66.6
Average	94.5	92.0	93.1

Table 9 tabulates the comparisons of the precisions (%) with these baselines on different datasets. Compared with BBNN, $\text{BBNN}_{w/o\mathbf{IB}}$ degrades the mAP (mean Average Precision) by 18%, 14%, 7% on three datasets respectively. The reason lies that the convolutional layers with identical scales inadequate to meet the different requirements from various genres in frequency bands and the time intervals. From the results, the baseline $\text{BBNN}_{w/o\textbf{Dense Conn}}$ without the dense connections also largely degrades the performance for all these datasets. Such results confirm that dense connections between multiple Inception Modules are indeed beneficial for music genre recognition. We also find that the baselines including $\text{BBNN}_{w/o\mathbf{BM}}$, $\text{BBNN}_{w/o\mathbf{IB}}$, and $\text{BBNN}_{w/o\textbf{Dense Conn}}$ have very low precisions for fewer genre Slow Waltz and Wcswing, which means they are insufficient to deal with unbalanced data in classification of music. However, The BM used in BBNN can learn and transmit the insufficient features caused by unbalanced data. In summary, our BBNN exploits the effectiveness integrated with both the IB and dense connections.

4.7.2 Time costing for training and testing

To train our BBNN model, it tasks 1.2 hours (for 100 epochs) on GTZAN, 0.7 hours on Ballroom, and 4.4 hours on Extended Ballroom. During testing, our BBNN only takes 0.1 seconds per music. BBNN model can effectively balance the accuracy and the computational complexity, which provides a premise of practical application.

5 Conclusions and future work

In this article, we present a specially designed network for accurately recognizing the music genre. The proposed model aims to take full advantage of low-level information of mel-spectrogram for making the classification decision. We have shown how our model is effective by comparing the state-of-the-art methods, including hand-crafted feature approaches and deep learning models with different architectures trained on different benchmark datasets. Deep learning approaches usually rely on a large amount of data to train a

Table 9 Ablation experiments on different datasets

(a) GTZAN

Method	Blues	Classical	Country	Disco	Hip-hop	Jazz	Metal	Pop	Reggae	Rock	mAP
$BBNN_{w/oBM}$	83.2	96.0	57.6	67.4	83.6	85.8	87.4	76.9	60.0	42.5	74.0
$BBNN_{w/oIB}$	80.4	84.8	66.9	70.4	81.8	84.6	86.8	81.7	68.6	51.7	75.7
$BBNN_{w/oDenseConn}$	83.5	87.6	75.4	69.3	83.7	92.5	90.4	78.4	68.7	67.1	79.6
BBNN	90.8	98.9	89.8	98.2	93.0	99.0	86.1	94.3	94.2	93.3	93.7

(b) Ballroom

Method	ChaCha	Jive	Quickstep	Rumba	Samba	Tango	Viennese Waltz	Slow Waltz	mAP
$BBNN_{w/oBM}$	92.3	79.1	69.6	76.1	90.6	83.4	63.3	78.6	79.1
$BBNN_{w/oIB}$	86.4	90.2	83.1	81.1	87.0	95.4	68.8	78.9	83.8
$BBNN_{w/oDenseConn}$	97.7	88.6	91.5	84.9	91.8	92.9	75.7	86.5	88.7
BBNN	100	97.9	96.3	97.0	95.0	98.8	98.5	94.3	97.2

(c) Extended Ballroom

Method	ChaCha	Foxtrot	Jive	Pasodoble	Quickstep	Rumba	Salsa	Samba	Slow Waltz	Tango	Viennese Waltz	Waltz	Wcswing	mAP
$BBNN_{w/oBM}$	91.3	90.9	92.1	65.4	88.1	87.6	76.5	86.7	10.0	83.3	83.5	76.7	6.6	72.2
$BBNN_{w/oIB}$	86.0	91.8	93.1	73.0	92.7	86.8	83.5	94.1	25.3	92.7	89.4	82.4	13.3	77.2
$BBNN_{w/oDenseConn}$	94.8	94.9	93.0	91.7	96.2	86.5	71.6	92.5	26.7	92.7	98.4	80.6	20.0	80.0
BBNN	98.1	98.8	98.5	96.0	99.6	96.7	94.9	97.5	80.7	99.0	96.8	93.1	78.5	94.5

model. In practice, since the number of annotated music recordings per genre class is often limited [8], therefore, except for accuracy, another major challenge is to train a robust CNN model from a few labeled data. In this work, the three common datasets (GTZAN, Ballroom, and Extended Ballroom) are employed to validate the proposed network structure from different data scale, especially the Ballroom dataset with only 698 tracks. Experiment results demonstrate that BBNN can overcome this challenging and achieve satisfactory accuracies.

In the future works, we will further improve the proposed model through the following ways. Firstly, we will explore the acoustic features (e.g. SSD, RH) adopting fusion strategies [41, 43] as also input into the network. Secondly, in the decision-making stage, we will adopt a new distance metric method [44] to compute the similarity between genres.

References

1. Bertin-Mahieux T, Ellis DP, Whitman B, Lamere P (2011) The million song dataset. In: Ismir, vol 2, pp 10
2. Cano P, Gómez Gutiérrez E, Gouyon F, Herrera Boyer P, Koppenberger M, Ong BS, Serra X, Streich S, Wack N (2006) Ismir 2004 audio description contest
3. Choi K, Fazekas G, Sandler M, Cho K (2017) Transfer learning for music classification and regression tasks. arXiv:1703.09179
4. Chollet F et al (2015) Keras
5. Costa YM, Oliveira LS, Silla CN Jr (2017) An evaluation of convolutional neural networks for music classification using spectrograms. Applied Soft Computing 52:28–38
6. Dai J, Liu W, Ni C, Dong L, Yang H (2015) "multilingual" deep neural network for music genre classification. In: Sixteenth annual conference of the international speech communication association
7. Freitag M, Amiriparian S, Pugachevskiy S, Cummins N (2017) Schuller, B.: audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. J Mach Learn Res 18(1):6340–6344
8. Fu Z, Lu G, Ting KM, Zhang D (2011) A survey of audio-based music classification and annotation. IEEE Trans Multimedia 13(2):303–319
9. Hafemann LG, Oliveira LS, Cavalin P (2014) Forest species recognition using deep convolutional neural networks. In: 2014 22nd international conference on Pattern recognition (ICPR), IEEE, pp 1103–1107
10. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1):10–18
11. Hamel P, Eck D (2010) Learning features from music audio with deep belief networks. In: ISMIR, vol 10, Utrecht, pp 339–344
12. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
13. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: CVPR, vol 1, pp 3
14. Ioffe S, Szegedy C (2015) Batch normalization:, Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167
15. Jakubik J (2017) Evaluation of gated recurrent neural networks in music classification tasks. In: International conference on information systems architecture and technology, Springer, pp 27–37
16. Jeong Y, Choi K, Jeong H (2017) Dlr:, Toward a deep learned rhythmic representation for music content analysis. arXiv:1712.05119
17. Karunakaran N, Arya A (2018) A scalable hybrid classifier for music genre classification using machine learning concepts and spark. In: 2018 International conference on intelligent autonomous systems (ICoIAS), IEEE, pp 128–135
18. Kereliuk C, Sturm BL, Larsen J (2015) Deep learning and music adversaries. IEEE Trans Multimedia 17(11):2059–2071
19. Kingma DP, Ba J (2014) Adam:, A method for stochastic optimization. arXiv:1412.6980
20. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
21. Lee J, Nam J (2017) Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. IEEE Signal Process Lett 24(8):1208–1212

22. Lee H, Pham P, Largman Y, Ng AY (2009) Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in neural information processing systems, pp1096–1104
23. Li TL, Chan AB, Chun A (2010) Automatic musical pattern feature extraction using convolutional neural network. In: Proc Int Conf Data mining and applications
24. Lin M, Chen Q, Yan S (2013) Network in network. arXiv:1312.4400
25. Lykartsis A, Lerch A (2015) Beat histogram features for rhythm-based musical genre classification using multiple novelty functions. In: Proceedings of the 16th ISMIR Conference, pp 434–440
26. Marchand U, Peeters G (2016) The extended ballroom dataset
27. Marchand U, Peeters G (2014) The modulation scale spectrum and its application to rhythm-content analysis. In: DAFX (Digital audio effects)
28. Marchand U, Peeters G (2016) Scale and shift invariant time/frequency representation using auditory statistics: Application to rhythm description. In: 2016 IEEE 26th international workshop on Machine learning for signal processing (MLSP), IEEE, pp 1–6
29. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E (2015) Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th python in science conference, pp 18–25
30. Medhat F, Chesmore D, Robinson J (2017) Automatic classification of music genre using masked conditional neural networks. In: 2017 IEEE international conference on Data mining (ICDM), IEEE, pp 979–984
31. Nanni L, Costa YM, Lucio DR, Silla CN Jr, Brahnam S (2017) Combining visual and acoustic features for audio classification tasks. Pattern Recogn Lett 88:49–56
32. Nguyen QH, Do TT, Chu TB, Trinh LV, Nguyen DH, Phan CV, Phan TA, Doan DV, Pham HN, Nguyen BP et al (2019) Music genre classification using residual attention network. In: 2019 International conference on system science and engineering (ICSSE), IEEE, pp 115–119
33. Pons J, Serra X (2017) Designing efficient architectures for modeling temporal features with convolutional neural networks. In: 2017 IEEE international conference on Acoustics, speech and signal processing (ICASSP), IEEE, pp 2472–2476
34. Pons J, Serra X (2018) Randomly weighted cnns for (music) audio classification. arXiv:1805.00237
35. Pons J, Lidy T, Serra X (2016) Experimenting with musically motivated convolutional neural networks. In: 2016 14th international workshop on Content-based multimedia indexing (CBMI), IEEE, pp 1–6
36. Salamon J, Bello JP (2017) Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters 24(3):279–283
37. Senac C, Pellegrini T, Mouret F, Pinquier J (2017) Music feature maps with convolutional neural networks for music genre classification. In: Proceedings of the 15th international workshop on content-based multimedia indexing, ACM, pp 19
38. Sigtia S, Dixon S (2014) Improved music feature learning with deep neural networks. In: 2014 IEEE international conference on Acoustics, speech and signal processing (ICASSP), IEEE, pp 6959–6963
39. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
40. Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Trans Speech Audio Process 10(5):293–302
41. Wang Y, Lin X, Wu L, Zhang W, Zhang Q, Huang X (2015) Robust subspace clustering for multi-view data by exploiting correlation consensus. IEEE Trans Image Process 24(11):3939–3949
42. Wang Y, Lin X, Wu L, Zhang W (2017) Effective multi-query expansions:, Collaborative deep networks for robust landmark retrieval. arXiv:1701.05003
43. Wang Y, Zhang W, Wu L, Lin X, Zhao X (2017) Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion. IEEE Trans Neural Netw Learn Syst 28(1):57–70
44. Wang Y, Wu L, Lin X, Gao J (2018) Multiview spectral clustering via structured low-rank matrix factorization. IEEE Transactions on Neural Networks and Learning Systems
45. Yu Y, Luo S, Liu S, Qiao H, Liu Y, Feng L (2020) Deep attention based music genre classification. Neurocomputing 372:84–91
46. Zhang W, Lei W, Xu X, Xing X (2016) Improved music genre classification with convolutional neural networks. In: INTERSPEECH, pp 3304–3308