MCRN: A New Content-Based Music Classification and Recommendation Network

Yuxu Mao[1], Guoqiang Zhong[1(✉)], Haizhen Wang[1], and Kaizhu Huang[2,3]

[1] Department of Computer Science and Technology, Ocean University of China, 238 Songling Road, Qingdao 266100, China
gqzhong@ouc.edu.cn

[2] Department of Electrical and Electronical Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China

[3] Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Hangzhou 310058, China

Abstract. Music classification and recommendation have received widespread attention in recent years. However, content-based deep music classification approaches are still very rare. Meanwhile, existing music recommendation systems generally rely on collaborative filtering. Unfortunately, this method has serious cold start problem. In this paper, we propose a simple yet effective convolutional neural network named MCRN (short for music classification and recommendation network), for learning the audio content features of music, and facilitating music classification and recommendation. Concretely, to extract the content features of music, the audio is converted into "spectrograms" by Fourier transform. MCRN can effectively extract music content features from the spectrograms. Experimental results show that MCRN outperforms other compared models on music classification and recommendation tasks, demonstrating its superiority over previous approaches.

Keywords: Music classification and recommendation · Information retrieval · Convolutional neural networks · Music spectrogram dataset

1 Introduction

Deep learning has achieved great successes in many fields, such as object detection [5], natural language processing [12], and information retrieval [15]. Particularly, in recent years, there arises much interest in music classification and recommendation with deep learning models [2,16].

For music classification, its performance heavily relies on how effective the features are extracted from the audios. In light of this, most traditional music classification methods extract audio features manually or using feature engineering methods [1]. However, the discriminability of these features is not high enough, so that the music classification accuracy is relatively low. In recent years,

H. Yang et al. (Eds.): ICONIP 2020, CCIS 1332, pp. 771–779, 2020.
https://doi.org/10.1007/978-3-030-63820-7_88

more and more research tries to use deep learning models for music classification. For example, a deep convolutional neural network (CNN) with small filters for music classification was proposed in [7]. However, the CNN model was only applied on the raw waveform audios, where the potential capability of CNN models on 2D/3D image classification was not fully exploited.

For music recommendation tasks, an important problem is how to recommend music to users in line with their preferences. At present, although there is much research on recommendation systems, music recommendation is still a challenging and complicated problem, due to the diversity of music styles and genres. Among quite a few techniques, collaborative filtering is one of the most successful recommendation algorithms [10]. Herlocker et al. [6] first applied the collaborative filtering to music recommendation tasks. However, the collaborative filtering algorithm has the cold start problem, i. e. it is not effective when recommending new or unpopular items to users.

In this paper, we present a simple yet effective music classification and recommendation network called MCRN. To extract the audio content features, we convert audio signal into a spectrogram via the Fourier transformation, as input to MCRN. Since spectrogram contains the frequency distribution of music and vary in sound amplitude, MCRN can extract valuable content features from them. In addition, we collect a music spectrogram dataset containing nearly 200,000 images. It offers an excellent resource for music classification and recommendation research. The dataset is publicly available to facilitate further in-depth research[1]. With this dataset, we show that MCRN is superior to existing music classification and recommendation models.

2 Related Work

Music Classification. During the past few decades, audio recognition are generally realized based on traditional feature extraction on time series. For instance, in [8], MFCC was used to represent audio features for music classification, while Dario et al. [4] proposed to parameterize the short-time sequence features of music through the multivariate autoregressive coefficient. However, these feature extraction methods are based on feature engineering approaches, which may lead to the loss of valuable information in the audios. To overcome the shortcoming, a convolutional recurrent neural network for music classification was proposed in [3], where CNN and RNN were applied to extract and aggregate features, respectively. Frustratingly, this method only extract features in the time domain, while the frequency domain is not considered. Recently, some music classification methods based on 1D CNN have been explored [2]. These methods suppose that 1D CNN is effective for processing audio sequences in music classification.

Music Recommendation. The collaborative filtering has been widely applied in recommendation systems [9,17]. Particularly, some work used the collaborative filtering techniques for music recommendation [13]. They calculated user

[1] https://github.com/YX-Mao/Music-spectrogram-dataset.

Table 1. Music categories of the collected music spectrogram dataset.

Breakbeat	Deep House	Disco
Downtempo	Drum and Bass	Dubstep Grime
Electro House	Euro Dance	Trance

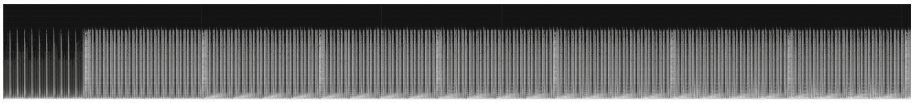

Fig. 1. A spectrogram obtained by Fourier transformation from a Deep_house type audio file.

preferences by constructing a music scoring matrix or recording playing coefficients. However, the collaborative filtering algorithms have a serious cold start problem, which is ineffective when recommending unpopular or new songs. To address it, an interesting music recommendation method was proposed by [14], where an attribute-based (i. e. mellow, unpretentious, sophisticated, intense and contemporary) method was used to characterize music content. Furthermore, with the development of deep learning, content-based music recommendation methods have been built based on CNNs [11,16]. Among others, the closest work to this paper is MusicCNNs [16]. Although the same data preprocessing method is adopted in MusicCNNs and this work, we propose a novel CNN architecture that is more effective than MusicCNNs in feature extraction of music content.

3 The Collected Dataset

To the best of our knowledge, there is currently no publicly available music spectrogram dataset. Thus, to facilitate the research of content-based music classification and recommendation, we have collected a new music spectrogram dataset.

We download 9,000 pieces of music from the JunoDownload website, including 9 types of music Table 1. To extract the audio content features, the audio signal is converted into an "image" by the Fourier transformation. The intensity of color on the spectrogram represents the amplitude of the sound at that frequency.

Concretely, the Sound eXchange (SoX) software and the Fourier transformation are adopt to convert the MP3 audio into a spectrogram. Figure 1 shows a partial spectrogram of a piece of music. To ease the training of CNN models, we crop the entire spectrogram to size 256×256 in this work, such that each piece of music is approximately split to 23 slices with a duration of 5 seconds. The image slices after cutting are shown in Fig. 2. To ensure the size of the data to be consistent, the edge parts that cannot be properly cropped in the spectrogram

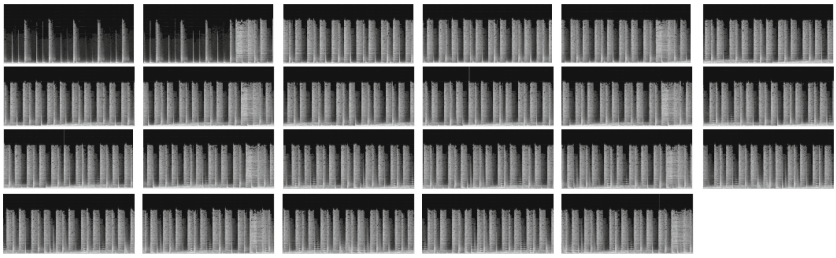

Fig. 2. Spectrogram of a piece of music after split. The size of image is 256×256.

are ignored. Eventually, we obtain a total of 197,463 spectrograms, which can be used to train the models for music classification and recommendation.

4 Music Classification and Recommendation Network

4.1 Architecture of MCRN

MCRN is a simple yet effective convolutional neural network, which architecture is illustrated in Fig. 3. We describe the details of MCRN in the following.

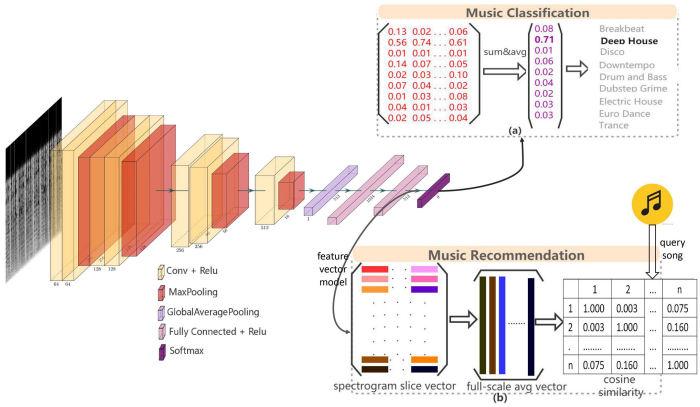

Fig. 3. The overview of MCRN. It can be used for (a) music classification and (b) music recommendation. For more details about the architecture of MCRN and its applications to music classification and recommendation, please refer to the text in Sect. 4.

The spectrogram is passed through a series of convolutional blocks, where seven convolutional layers and four pooling layers are divided into four blocks. Concretely, each of the first two blocks contains two convolutional layers and a max-pooling layer, where the size of the kernels is 3×3, and max-pooling is conducted over a 2×2 pixel window. The structure of the third block is highly

similar to the previous two groups. It also contains two convolutional layers and a max-pooling layer, but rather than using the same size of convolutional kernels, we implement the receptive fields of the two convolutional layers using 5×5 and 3×3 kernels, respectively. In addition, the pixel window of the max-pooling is expanded to 4×4. The final block consists of a convolutional layer with 512 convolutional kernels of size 3×3 and a max-pooling layer. To regularize the entire network to prevent from overfitting, a global average pooling layer is followed. These convolutional layers are followed by two fully connected layers with 1024 and 512 dimensions, respectively, using RELU as the activation function.

4.2 Music Classification

The classification probability of spectrogram in each category can be obtained by MCRN, as shown in Fig. 3(a). Since a full-scale spectrogram was cut into multiple small fragments during the data preprocessing, we calculate the classification probability of the music in each category using spectrograms belonging to the same piece of music:

$$P_i^c = 1/T \sum_{t=1}^{T} a_{i,t}^c, \quad c \in \{Breakbeat, \ldots, Trance\}, \tag{1}$$

where T represents the number of spectrograms obtained by cutting a piece of music, and $a_{i,t}^c$ is the classification probability of the t-th slice of the i-th piece of music in category c. We predict the category C of the music by selecting the class with the highest probability:

$$C = \arg\max P_i^c. \tag{2}$$

The music classification method takes into account all spectrograms in a piece of music, i. e. each segment of music contributes to the final classification. Thus, the music classification results obtained by our proposed method are persuasive.

4.3 Music Recommendation

To represent music content, we integrate the spectrogram features extracted by MCRN together to create a full-scale music feature vector. Specifically, a feature vector is first created for each spectrogram based on MCRN. Please note that since each piece of song corresponds to about 23 spectrograms, 23 feature vectors can be obtained for a piece of music. Then, the feature vector of each piece of music is acquired by averaging the feature vectors of all spectrograms belonging to the music. Finally, our strategy for music recommendation is to calculate the similarity between music based on the cosine distance as follows:

$$\cos(X, Y) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}}, \tag{3}$$

where X and Y denotes two different pieces of music, x_i and y_i represent the feature values of the music, and n is the length of the feature vector. A full description of the music recommendation by MCRN is given in Algorithm 1.

Algorithm 1. Music Recommendation

Require: The music feature extractor M based on MCRN, the number of music N, the spectrogram number of a piece of music T, all music spectrogram slices after data preprocessing $D = \{d_{1,1}, d_{1,2}, \ldots, d_{N,T}\}$, the music for recommendation X

Ensure: The recommended music for X

1: Create feature vectors $F = \{f_{1,1}, f_{1,2}, \ldots, f_{N,T}\}$ from all music spectrogram slices in D according to feature extractor M
2: **for** music index $i = 1$ to N **do**
3: **for** spectrogram index $j = 1$ to T **do**
4: Select the feature vector of T spectrograms corresponding to the i-th music, $f_i = \{f_{i,1}, d_{i,2}, ..., d_{i,T}\}$
5: **end for**
6: Calculate the average feature vector of the i-th music: $F_i = 1/T \sum_{t=1}^{T} f_{i,t}$
7: Add the average feature vector F_i of music to V: $V.append(F_i)$
8: **end for**
9: Calculate the cosine distance between music according to Equation (3)
10: The top 3 pieces of music with the highest similarity to X as its recommendations

5 Experiments

5.1 Implementation Details

We train MCRN 10 epochs with a batch size of 64. Each fully-connected layer follows a dropout with a drop rate of 0.5, and add L2 regularizers with the coefficient set to 0.001. An RMSprop optimizer with a learning rate $\mathrm{lr} = 1\mathrm{e}{-3}$ is adopted for parameter optimization. Since the 2D spectrogram cannot be directly applied to MLP and Softmax regression, we perform dimensionality reduction on the spectrogram to meet their input needs.

5.2 Experiments on Music Classification

To provide more reasonable evaluation, 9,000 pieces of music are divided into training set, validation set and test set according to the ratio of 0.65: 0.25: 0.1. The music of test set only used for the final evaluation of model.

Table 2 shows the classification results obtained by MCRN and some existing music classification approaches, including MLP, Softmax regression and some state-of-the-art deep learning methods. This table shows that MCRN achieves state-of-the-art performance on the music classification, and the total classification accuracy reaches 77.3%. Among 9 types of music, MCRN achieves excellent performance on 7 types of music. Taking the Dubstep Grime as an example, MCRN outperforms the competitors with the superiority result (83.84%). Such results demonstrate the superiority of MCRN on music classification.

5.3 Experiments on Music Recommendation

To verify the effectiveness of MCRN on music recommendation, we conduct recommendation experiments based on each piece of music in test set. The music

Table 2. Results of music classification accuracy by different models on test set.

Model	Breakbeat	Deep House	Disco	Down Tempo	Drum and Bass	Dubstep Grime	Electro House	Euro Dance	Trance	Total acc. (%)
MLP	11.01	41.41	33.00	69.07	61.00	13.03	12.86	53.81	9.09	33.81
Softmax regression	28.62	13.80	28.33	59.79	14.00	13.80	3.81	15.87	44.11	24.69
The model of [11]	75.16	54.25	**78.10**	71.28	89.65	65.06	62.26	58.25	86.53	71.18
The model of [2]	68.70	71.70	76.00	71.20	**92.00**	10.10	68.60	55.20	63.69	64.30
MusicCNNs [16]	67.72	72.86	62.85	58.97	90.42	75.79	**72.91**	57.33	82.87	71.30
MCRN(ours)	**76.77**	**79.80**	71.00	**73.20**	**92.00**	**83.84**	70.50	**60.00**	**89.90**	**77.30**

selected from test set each time is used as the recommendation item, namely query song. As long as the recommended music and query song belong to the same class, we consider that the recommendation is correct. For a fair comparison, similar to MusicCNNs [16], our recommendation experiment also uses cosine distance to measure the similarity between music. The comparison of MCRN with state-of-the-art methods are summarized in Table 3. The table shows that, for music recommendation, MCRN achieves excellent recommendation on top-1 and top-3, with the recommended accuracy is 71.50% and 84.65%, respectively.

Table 3. The comparison of recommendation accuracy of different models on the top-1 and top-3 metrics based on test set. MusicCNNs is an improvement of R-MusicCNNs, where the ReLU activation function in the R-MusicCNNs is replaced with ELU.

Model	top-1 acc.(%)	top-3 acc.(%)
MLP	34.15	60.67
Softmax regression	25.52	46.52
The model of [11]	67.28	82.98
R-MusicCNNs [16]	64.02	77.65
MusicCNNs [16]	64.75	78.01
MCRN (ours)	**71.50**	**84.65**

6 Conclusion

In this paper, a new deep convolutional neural network called MCRN is proposed, which is applied to content-based music classification and recommendation tasks. By learning the audio content features of music, MCRN overcomes the problem of the cold start for music recommendation applications. Importantly, to fully extract audio content features, we convert the audio signal into a form of spectrogram by Fourier transform. We collect a new music spectrogram dataset, which contains nearly 200,000 images. To the best knowledge, this is the first publicly available dataset of music spectrogram. On this dataset, we

conduct extensive music classification and recommendation experiments. Experimental results show that MCRN attains new state-of-the-art results on music classification and recommendation tasks.

Acknowledgments. This work was supported by the Major Project for New Generation of AI under Grant No.2018AAA0100400, the National Natural Science Foundation of China (NSFC) under Grant No.41706010, the Joint Fund of the Equipments Pre-Research and Ministry of Education of China under Grant No.6141A020337, and the Fundamental Research Funds for the Central Universities of China.

References

1. Bergstra, J., Casagrande, N., Erhan, D., Eck, D., Kégl, B.: Aggregate features and AdaBoost for music classification. Mach. Learn. **65**(2–3), 473–484 (2006). https://doi.org/10.1007/s10994-006-9019-7
2. Bian, W., Wang, J., Zhuang, B., Yang, J., Wang, S., Xiao, J.: Audio-based music classification with DenseNet and data augmentation. In: Nayak, A.C., Sharma, A. (eds.) PRICAI 2019. LNCS (LNAI), vol. 11672, pp. 56–65. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29894-4_5
3. Choi, K., Fazekas, G., Sandler, M.B., Cho, K.: Convolutional recurrent neural networks for music classification. In: ICASSP, pp. 2392–2396 (2017)
4. Garciagarcia, D., Arenasgarcia, J., Parradohernandez, E., Diazdemaria, F.: Music genre classification using the temporal structure of songs (2010)
5. He, Y., Zhu, C., Wang, J., Savvides, M., Zhang, X.: Bounding box regression with uncertainty for accurate object detection. In: CVPR, pp. 2888–2897 (2019)
6. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. SIGIR Forum **51**(2), 227–234 (2017)
7. Jongpil, L., Jiyoung, P., Keunhyoung, K., Juhan, N.: SampleCNN: end-to-end deep convolutional neural networks using very small filters for music classification. Appl. Sci. **8**(1), 150 (2018)
8. Kour, G., Mehan, N., Kour, G., Mehan, N.: Music genre classification using MFCC, SVM and BPNN. Int. J. Comput. Appl. **112**(6), 12–14 (2015)
9. Li, D., Lv, Q., Shang, L., Gu, N.: YANA: an efficient privacy-preserving recommender system for online social communities. In: CIKM, pp. 2269–2272 (2011)
10. Liu, X., Qiu, J., Hu, W., Huang, Y., Zhang, S., Liu, H.: Research on personalized recommendation technology based on collaborative filtering. In: ICSC, pp. 41–46 (2019)
11. Murray, M.: Building a music recommender with deep learning. http://mattmurray.net/building-a-music-recommender-with-deep-learning
12. Ren, S., Zhang, Z., Liu, S., Zhou, M., Ma, S.: Unsupervised neural machine translation with SMT as posterior regularization. In: AAAI, pp. 241–248 (2019)
13. Sánchez-Moreno, D., González, A.B.G., Vicente, M.D.M., Batista, V.F.L., García, M.N.M.: A collaborative filtering method for music recommendation using playing coefficients for artists and users. Expert Syst. Appl. **66**, 234–244 (2016)
14. Soleymani, M., Aljanaki, A., Wiering, F., Veltkamp, R.C.: Content-based music recommendation using underlying music preference structure. In: ICME, pp. 1–6 (2015)
15. Yang, X., Wang, N., Song, B., Gao, X.: BoSR: a CNN-based aurora image retrieval method. Neural Netw. **116**, 188–197 (2019)

16. Zhong, G., Wang, H., Jiao, W.: MusicCNNs: a new benchmark on content-based music recommendation. In: Cheng, L., Leung, A.C.S., Ozawa, S. (eds.) ICONIP 2018. LNCS, vol. 11301, pp. 394–405. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04167-0_36
17. Zhuang, F., Zheng, J., Chen, J., Zhang, X., Shi, C., He, Q.: Transfer collaborative filtering from multiple sources via consensus regularization. Neural Netw. **108**, 287–295 (2018)