Deep Neural Networks with Depthwise Separable Convolution for Music Genre Classification

Yunming Liang, Yi Zhou, Tongtang Wan, Xiaofeng Shu
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications, Chongqing, China
e-mail: lym0302@foxmail.com, zhouy@cqupt.edu.cn

***Abstract*—With the prevalence of internet-based music platforms and the fast increase of music data, considerable attention has been paid to music information retrieval (MIR), like music genre classification. Although the traditional deep neural networks (DNNs) technology for music genre classification has achieved good results, it has a large amount of calculations and parameters due to large data sets and deep networks, resulting in a slow running speed. In an effort to overcome this problem, two types of deep neural networks with depthwise separable convolution are proposed for music genre classification in this paper. The experimental results show the proposed models have better performance on the Extended Ballroom dataset when compared with the traditional models, especially the convolutional neural network with depthwise separable convolution, which achieves an accuracy of 95.10%.**

Keywords-music genre classification; depthwise separable convolution; deep neural network

I. INTRODUCTION

With the widespread use of various internet-based music platforms and an increasing number of music data, it is difficult to manage and organize these music data. Therefore, the automatic music genre classification method is greatly desired, which is also very challenging due to the boundaries between genres still remain fuzzy. Most of the state-of-the-art methods aim to classify the music genres which are top-level labels of music so as to help audiences to categorize and describe various music [1]. Meanwhile, exact classification on the music genre is crucial for music platforms to organize music into different groups. Thus, classification of music genre has attracted wide attentions in the field of music information retrieval (MIR) [2, 3].

Generally, an automatic genre classification method consists of three steps: 1) features such as spectrogram, Mel-frequency cepstral coefficients (MFCCs) and statistical features are extracted from the original audio signal; 2) certain techniques are applied to select the meaningful subsets of the features [4] or to aggregate features to improve the classification accuracy [5, 6]; 3) the appropriate classifier is selected according to the selected features to map feature vectors into different music genres.

The convolutional neural networks (CNNs) have been widely used for various music information retrieval tasks such as music tagging [7, 8], genre classification [6, 9], and user-item latent feature prediction for recommendation [10]. And convolutional recurrent neural networks (CRNNs) also fit the music tagging well on reference [11].

Recently, depthwise separable convolution has been proposed as an efficient alternative to the standard 3-D convolution operation and has been used to achieve compact network architectures in the area of computer vision [12]. And the Xception proposed by Chollet F based on depthwise separable convolution significantly outperforms the Inception V3 on a larger image classification dataset comprising 350 million images and 17,000 classes due to more efficient use of model parameters [13]. Moreover, depthwise separable convolution has also been used in keyword spotting [12] and audio events detection [14], which all gave good results.

Inspired by the above literature, two types of modified deep neural networks are introduced for music genre classification in this paper. One is the CNNs with depthwise separable convolution, DS-CNNs, the other is the CRNNs with depthwise separable convolution, DS-CRNNs. Both the DS-CNNs and the DS-CRNNs take advantage of convolutional layers for local feature extraction and use depthwise separable convolutional layers for the reduction in network parameters and computation. The proposed models are then compared with other traditional CNNs and CRNNs models on Extended Ballroom dataset [15]. The traditional models have nearly the same parameters with the proposed models. The experimental results reveal that the DS-CNNs and DS-CRNNs have better performance on accuracy and calculation over the traditional CNNs and CRNNs.

The rest of this paper is organized as follows. In Section II, depthwise separable convolution (DSC) was briefly introduced. And deep neural networks with DSC were described in detail in Section III. Then experimental setup and results are illustrated in Section IV. Finally, a conclusion is mentioned in Section V.

II. THEORETICAL BASIS

In this section, DSC is introduced in music genre classification since the feature of music can be approximated as images which suggest it may apply well in audio analysis. To eliminate the effect of network capacity, we let the proposed models have roughly the same parameters as the traditional models to be compared.

A DSC can be grouped into two parts, the depthwise convolution and the pointwise convolution. The first part is the filtering stage. It uses a separate 2-D filter kernel to

978-1-7281-5102-1/19/$31.00 ©2019 IEEE

convolve each channel of input respectively as shown in Fig. 1. The input volume **F** has width, height and depth/channel denoted as W_{in}, H_{in}, M. The output **G** from depthwise convolution with width, height and depth being W_{out}, H_{out}, M, where W_{out} and H_{out} are obtained according to the kernel size and strides. Obviously, the number of filter kernels is equal to the number of input channels and the depth of **G** is the same as the depth of **F**.

The second part is pointwise convolution which is the combining stage. Fig. 2 demonstrates the process of pointwise convolution. It uses a 1×1 convolution to implement linear combination of each channel. Assuming the output from DSC is N.

In comparison to standard convolution, the DSC simplifies multiplication operations and parameters.

In more detail, the number of parameters P_{dsc} and multiplication operations Mul_{dsc} using DSC from **F** to **O** are shown in (1) and (2):

$$P_{\text{dsc}} = M(H_k W_k + N) \tag{1}$$

$$Mul_{dsc} = MH_{out}W_{out}(H_k W_k + N) \tag{2}$$

And the parameters P_{c} and multiplication operations Mul_c of standard convolution are given in (3) and (4):

$$P_{\text{c}} = M(H_k W_k N) \tag{3}$$

$$Mul_c = MH_{out}W_{out}(H_k W_k N) \tag{4}$$

Therefore, the DSC has great advantages in calculation and parameter utilization.

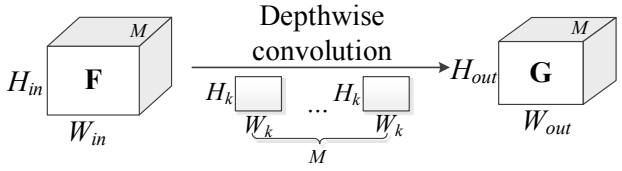

Figure 1. Depthwise convolution.

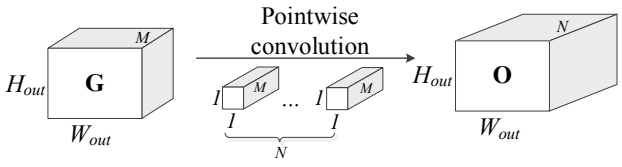

Figure 2. Pointwise convolution.

III. DSC For Genres Classification

In this paper, depthwise separable convolution is applied to music genre classification. By combining and comparing traditional CNN and CRNN, we found the following methods to help and improve this task.

A. Convolutional Neural Networks with Depthwise Separable Convolution (DS-CNNs)

The architecture of DS-CNNs proposed in this paper is shown in Fig. 3. DS-CNNs models consist of two parts: the feature extractor composing of one convolutional layer and multiple depthwise separable convolutional layers, and the classifier with one fully-connected layer.

The functionality of the convolutional layers is not only to extract high-level features but also to convert 2-D features into 3-D features for better use of DSC. In our experiment, the depthwise separable convolutional layers first perform spatial convolution on each input channel using 3×3 filter kernels, which is also called "depthwise convolution", and then performs pointwise convolution, i.e. the 1×1 convolution, to project the channel of the depthwise convolution output to new channel space. Note that all the depthwise convolution and pointwise convolution layers are followed by a non-linear activation function. We also use a technique called batch normalization (BN) [16] to speed up the training process and make the final model more robust.

The input of the net is MFCCs extracted from a music signal, which is a typical representation of audio features. And the output is the genre of each test sample used for accurate calculation.

Prior to the fully-connected layer, global average pooling operation is applied to each output channel after the DSC block, which has the same kernel size and strides as each output channel.

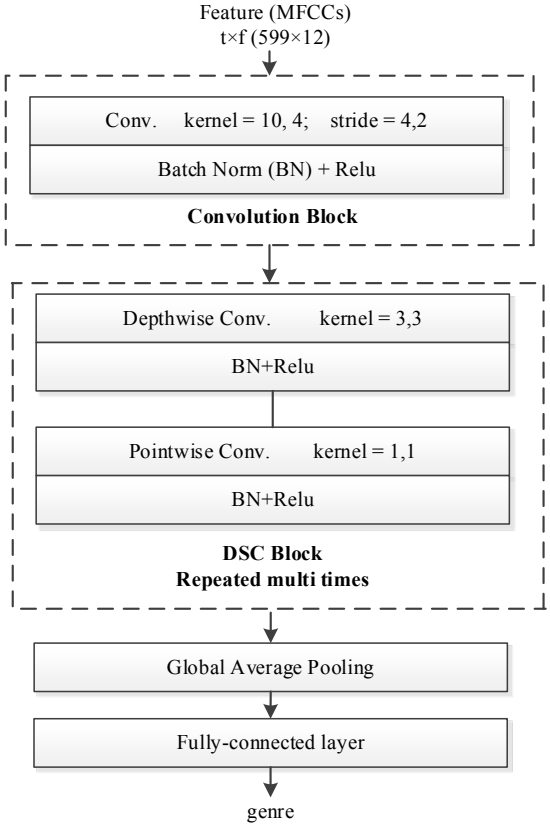

Figure 3. Convolution neural network with depthwise separable convolution (DS-CNN).

B. Convolutional Recurrent Neural Networks with Depthwise Separable Convolution (DS-CRNNs)

Similarly, the DS-CRNNs consist of a standard convolutional layer, depthwise separable convolutional layers, recurrent layer and a fully-connected layer. In comparison with DS-CNNs models, the proposed DS-CRNNs models have an extra recurrent layer between the average pooling layer and the fully-connected layer. Note that the pooling size and stride size of these DS-CRNNs models are both 2×2, which is different from the DS-CNNs. In our experiments, the base cells of recurrent layers are Gated recurrent units (GRU). It uses fewer parameters to achieve better convergence over the long short-term memory (LSTM).

IV. Experiment and Results

The traditional CNNs, CRNNs and the proposed DS-CNNs and DS-CRNNs are compared on Extended Ballroom dataset in the experiments. The sizes of the networks are controlled by varying the numbers of parameters.

A. Dataset Description

The Extended Ballroom dataset is an improved version of the well-known Ballroom dataset and it is useful for genre/rhythm-class recognition systems as well as tempo estimation algorithms [15]. There are 4,180 tracks distributed in thirteen different genres: Chacha (455), Jive (350), Quickstep (497), Rumba (470), Samba (468), Tango (464), Vienne sewaltz (252), Waltz (529), Foxtrot (507), Pasodoble (53), Salsa (47), Slow waltz (65), Wcswing (23). Among them, the amount of tracks in the four genres (PasoDoble, Salsa, SlowWaltz, WcSwing) is relatively low. Each track lasts about 30 seconds and is sampled at 16000Hz and quantized by 16 bits.

B. Experimental Setup

In the experiments, all the tracks are split into 8/1/1 training, validation and test sets. 7188 (599×12) features are extracted from the music track. It contains about 599 frames and each frame has 12 coefficients. In this case, each STFT frame spans 100ms and the sliding window size is 50ms.

The parameters of all the models are divided into three types: S(~360k), M(~500k) and L(~900k), where the parameters of the networks are determined according to the size and classes of the dataset. Moreover, the hyperparameters of all the models including kernel size, strides and feature map are designed as usual and the summary of these hyperparameters are shown in table I. C and DSC represent standard convolutional layer and depthwise separable convolutional layer respectively, where the numbers in parentheses correspond to kernel sizes and strides in time and frequency axes. GRU stands for the recurrent layer with the number of memory elements in parentheses. FM stands for the number of feature map of each convolutional layer. All the models are trained in Google TensorFlow framework using the standard cross-entropy loss and Adam optimizer. With a batch size of 100, all the models are trained in the following way: for the first 2000 iterations, the learning rate is 8×10^{-4}. For the next 1000 iterations, the learning rate is reduced to 10^{-4}, which is further reduced to 2×10^{-5} for the successive last 1000 iterations. The dropout technique with 0.5 dropout rate is used to alleviate the over-fitting problem.

The neural networks are used to extract features from the data and a classifier is trained on the feature extracted from the training and validation set. The training data is augmented by turning a random selected 100ms samples into zeroes to mitigate overfitting and to improve accuracy. The accuracy is calculated on test data. To improve performance reliability, all the models are run repeated 10 times.

TABLE I. Table Type Styles

Paras. Size	***Models***	***Hyperparameters***
S(~360k)	CNN1	C(10,4,4,2) -3C(3,3,1,1) FM: 32-64-128-228
	CRNN1	C(10,4,4,2)-C(3,3,2,2) -2C(3,3,1,1)-GRU(216) FM: 32-64-88-102
	DS-CNN1	C(10,4,4,2) -5DSC(3,3,1,1) FM: 256(5)-262
	DS-CRNN1	C(10,4,4,2)-DSC(3,3,2,2) -2DSC(3,3,1,1)-GRU(232) FM: 32-64-128-218
M(~500K)	CNN2	C(10,4,4,2) -3C(3,3,1,1) FM: 64-128-148-186
	CRNN2	C(10,4,4,2)-C(3,3,2,2) -2C(3,3,1,1)-GRU(218) FM: 32-64-78-216
	DS-CNN2	C(10,4,4,2) -7DSC(3,3,1,1) FM: 256(8)
	DS-CRNN2	C(10,4,4,2)-DSC(3,3,2,2) -5DSC(3,3,1,1)-GRU(236) FM: 32-64-88-128(2)-256(2)
L(~900K)	CNN3	C(10,4,4,2) -5C(3,3,1,1) FM: 32-64-88-128-212-256
	CRNN3	C(10,4,4,2)-C(3,3,2,2) -4C(3,3,1,1)-GRU(228) FM: 32-64-96-128(2)-226
	DS-CNN3	C(10,4,4,2) -9DSC(3,3,1,1) FM: 128(3)-256(3)-416(3)-512
	DS-CRNN3	C(10,4,4,2)-DSC(3,3,2,2) -6DSC(3,3,1,1)-GRU(256) FM: 32-64-128(2)-256(3)-486

C. Results

The music genre classification accuracy of various models is reported in Table II, which includes four types models: the traditional CNNs and CRNNs models, the proposed DS-CNNs and DS-CRNNs models.

As shown in Table I, with the same parameter size, networks with DSC can be designed with more layers and

bigger feature map. The results shown in Table II, demonstrate that with the same dataset and nearly the same number of parameters, the DS-CNNs performs better than other models. Moreover, the DS-CNNs not only perform better than the CNNs when they have approximately equal parameters, but also the multiplication operations of DS-CNNs are much less than CNNs. Similarly, the DS-CRNNs also outperform the CRNNs on accuracy and calculation given approximately equal parameters. The best model in the experiments is DS-CNN3, which achieves an accuracy of 95.10%.

TABLE II. GENRE CLASSIFICATION RESULTS ON EXTENDED BALLROOM DATASET

Models	*Accuracy (%)*	*Parameters*
CNN1	93.14	359,977
CRNN1	91.67	360,223
DS-CNN1	93.42	360,039
DS-CRNN1	92.40	359,697
CNN2	93.65	498,019
CRNN2	93.38	505,029
DS-CNN2	94.28	496,141
DS-CRNN2	93.63	496,597
CNN3	94.36	909,349
CRNN3	93.87	903,629
DS-CNN3	**95.10**	906,221
DS-CRNN3	94.12	905,657

In the case of a larger number of feature channels, the parameters of CNNs and CRNNs will be many times more than the parameters of DS-CNNs and DS-CRNNs. In reality, the number of classes and data, even the time duration of each track are more than that of Extended Ballroom dataset used in this paper, so the automatic classification system will require a deeper network structure. In this situation, the advantage of deep neural networks with DSC is more evident.

V. CONCLUSION

The work presented in this paper is mainly to apply CNNs and CRNNs with DSC to music genre classification. This studied architecture facilitates to use model parameters more efficiently and leads to higher accuracy. The present work relies primarily on the following prior research efforts: The Xception with DSC proved by Chollet F performs well in image classification. In addition, DSC was also used in audio files in [12, 14]. The proposed models proposed are inspired by them due to music is a kind of audio and the feature of music can be seen as images. And the work by Siddharth Sigtia, Simon Dixon proposed three ways to improve feature learning for audio data using neural networks [6], which provides some references on choosing related function in models design.

In this paper, the DS-CNNs and DS-CRNNs are proposed for music genre classification. From experiment results, the DS-CNNs and DS-CRNNs can achieve better performance than traditional the CNNs and CRNNs models on the same size of parameters. Compared to the traditional models, CNNs and CRNNs with DSC can not only reduce the calculation but also improve the accuracy. In the future, DSC can be applied to other music analysis tasks.

REFERENCES

[1] Tzanetakis G, Cook P. Musical genre classification of audio signals[J]. IEEE Transactions on speech and audio processing, 2002, 10(5): 293-302.

[2] Shawe-Taylor J S, Meng A. An investigation of feature models for music genre classification using the support vector classifier[J]. 2005.

[3] West K, Cox S. Finding An Optimal Segmentation for Audio Genre Classification[C]//ISMIR. 2005: 680-685.

[4] Auguin N, Huang S, Fung P. Identification of live or studio versions of a song via supervised learning[C]//2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. IEEE, 2013: 1-4.

[5] Bergstra J, Casagrande N, Erhan D, et al. Aggregate features and a da b oost for music classification[J]. Machine learning, 2006, 65(2-3): 473-484.

[6] Sigtia S, Dixon S. Improved music feature learning with deep neural networks[C]//2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2014: 6959-6963.

[7] Dieleman S, Schrauwen B. End-to-end learning for music audio[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 6964-6968.

[8] Choi K, Fazekas G, Sandler M. Automatic tagging using deep convolutional neural networks[J]. arXiv preprint arXiv:1606.00298, 2016.

[9] Chiliguano P, Fazekas G. Hybrid music recommender using content-based and social information[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 2618-2622.

[10] Van den Oord A, Dieleman S, Schrauwen B. Deep content-based music recommendation[C]//Advances in neural information processing systems. 2013: 2643-2651.

[11] Choi K, Fazekas G, Sandler M, et al. Convolutional recurrent neural networks for music classification[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 2392-2396.

[12] Zhang Y, Suda N, Lai L, et al. Hello edge: Keyword spotting on microcontrollers[J]. arXiv preprint arXiv:1711.07128, 2017.

[13] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.

[14] Gajarsky T, Purwins H. An Xception Residual Recurrent Neural Network for Audio Event Detection and Tagging[C]//Sound and Music Computing Conference. 2018.

[15] Marchand U, Peeters G. The extended ballroom dataset[J]. 2016.

[16] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.