Music Genre Classification Using Convolutional Neural Network

Qiuqiang Kong
South China University of Technology
qiuqiangkong@gmail.com

Xiaohui Feng
South China University of Technology
85004320@qq.com

Yanxiong Li
South China University of Technology
150305211@qq.com

ABSTRACT

Feature extraction is a crucial part of many MIR tasks. Many manual-selected features such as MFCC have been applied to music processing but they are not effective for music genre classification. In this work, we present an algorithm based on spectrogram and convolutional neural network (CNN). Compared with MFCC, the spectrogram contains more details of music components such as pitch, flux, etc. We use feature detector as filter to convolve spectrogram to get four feature maps, which can catch trends of spectrogram in both time and frequency scale. Then sub-sample layer is applied to reduce dimension and enhance resistance to translation in pitch and tempo. Finally the extracted high-level features are connected to a multi-layer perceptron (MLP) classifier. A classification accuracy of 72.4% is obtained on Tzanetakis dataset by uing the proposed features, which performs better than MFCC.

1. INTRODUCTION

Music genre classification task has a wide scope of applications. Since 2002, many manual-selected low-level acoustic features have been proposed. Tzanetakis [1] and Fu [2] reviewed the current low-level features and mid-level features and summarized their performances on genre classification. Since the single manual-selected feature can not reach high classification accuracy, Bergstra [3] used aggregate features and adaboost classifier to realize genre classification. Fu [4] discussed several feature-level and decision-level combination methods. Their experiments showed combinational features performed better than single feature.

In recent days, deep learning methods have been used in extracting features. Hamel [5] proposed a feature extraction system using Deep Belief Network (DBN) on Discrete Fourier Transforms (DFT) of audio and use nolinear SVM as classifier. Andrew Y. Ng [6] used shift-invariant sparse coding (SISC) to learn a succinct high-level representation of the inputs. Andrew Y. Ng also used convolutional deep belief networks (CDBN) to classify audios.

In this paper we propose to use convolutional neural network on spectrogram. Compared with traditional features like MFCC, spectrogram contains all the details of music. First we retain only the amplitude of spectrogram and discard the phase of spectrogram. Then use feature detectors (filters) to convolve the spectrogram and get feature maps. Then a sub-sample layer is applied to reduce the dimension. Finally, the extracted features are concatenated and connected to a multi-layer perceptron (MLP).

2. CONVOLUTIONAL NEURAL NETWORK

The Convolutional neural network (CNN) was first used in digit recognition, which is a variation of MLP inspired by biology. CNN combine three architectural ideas to ensure some degree of shift, scale, and distortion invariance: local receptive fields, shared weights, and sub-sampling. These concepts can be modified and used in music classification based on convolution of spectrogram.

2.1 Receptive Fields & Feature detector

The concept of receptive fields was first discovered by Hubel and Wiesel in cat's visual system. We use local feature detector (filter) to imitate receptive fields. In image processing, the local feature detector can be applied to the entire image. The output is the convolution of input image and the feature detector. We may have several feature detectors to seize different kind of edges. Each output of a feature detector is called a feature map.

In audio processing, we can get a spectrogram by applying Short-time Fast Fourier Transform (SFFT) on a piece of music. The horizontal axis and vertical axis represents time-scale and frequency-scale respectively. In the spectrogram, the harmonic component has fixed pitch so is continuous in time scale. The percussion component is instantaneous so spectrogram is continuous in frequency scale. Tradition features like MFCC is obtained from a single frame so it lacks ability in dynamic analysis. We introduce the variation of CNN inspired by image processing to spectrogram to solve this problem. First, we introduce the feature detector. They are some small blocks of size r*r, shown in Figure 1. The black point represents 1 and white point represents 0. Each of the feature detector can capture different kinds of features in spectrogram, as follows:

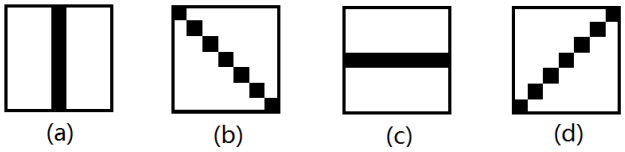

Figure 1. Feature detectors.

Fig.1(a) capture percussion component. Fig.1(b) capture down slide component. Fig.1(c) capture harmonic component. Fig.1(d) capture up slide component.

Then we apply convolution operation on spectrogram using these filters, then we get four feature maps, as shown in Fig 2. Because different genre will have different of these components, so it is reasonable to use these filters to obtain high-level feature.

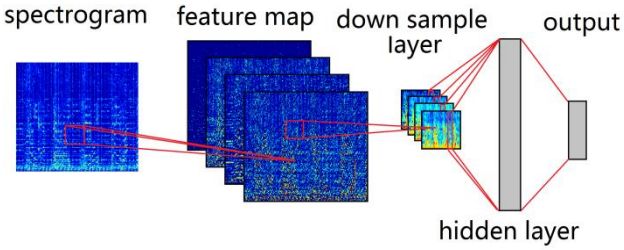

Figure 2. Structure of convolution neural network.

2.2 Sub-sample Layer

Once the feature map is obtained, we apply a sub-sample layer on each of the feature map. There are two reasons for explaining the sub-sample layer. One is that Applying a sub-sample layer can reduce dimension. If we use the original feature map directly, then the number of weights will be very big. Another reason is the output of a sub-sample layer can be invariant to translation of pitch offset and tempo shrinkage and extension.

We use a s*s maximum sub-sample matrix on each of feature map. So the number of weights will decrease to $\frac{1}{s^2}$ of original. There are many sub-sample strategies besides maximum operator. We choose maximum operator because it is the simplest to implement. The sub-sample layer is shown in Figure 2.

Finally the output of sub-sample layer is connected to multi-layer perceptron (MLP).

3. EXPERIMENTS EVALUATION

We use Tzanetakis1 dataset to evaluate our algorithm. This dataset consists 10 genres and each genre contains 100 30-second audio clips.

Our experiment tests MFCC, FFT, CQT and COV_FFT on Tzanetakis' dataset using 2-fold cross-validation. COV_FFT is the feature extraction method proposed in our paper. Because there is no parameter to configure in softmax model, so it is suitable to judge whether a feature is good or not. We will provide both softmax regression and multi layer perceptron on these features. The results are shown in Table 1.

Features	Accuracy
MFCC+SM	34.4%
CQT+SM	51.4%
FFT+SM	55.6%
COV_FFT +SM	66.4%
MFCC+MLP	46.8%
CQT+MLP	57.4%
FFT+MLP	64.2%
COV_FFT+MLP	72.4%

Table 1. Accuracy of different features using softmax (SM) and multi layer perceptron (MLP).

Table 1 shows the performance of COV_FFT > FFT > CQT > MFCC. The accuracy of our method using CNN is 72.4% which is best in our experiment.

4. CONCLUSION

In future work, we will continue investigating convolution neural network using learned feature detectors. The feature detector in our paper is still manual-selected. We are interested in how to learn the feature detectors. This may result in a better result than our current method. Finally, we will investigate the performance using more layers in CNN. It may be possible to get more abstract and higher-level feature using more layers.

5. ACKNOWLEDGEMENTS

This work was supported by the Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (Item No.2012LYM_0012),the Fundamental Research Funds for the Central Universities, South China University of Technology, China (Item No.2013ZM0090, and 2013ZZ0053), the National Natur al Science Foundation of China (No. 61401161)

6. REFERENCE

[1] Tzanetakis G, Cook P. "Musical genre classification of audio signals". *Speech and Audio Processing, IEEE transactions on*, 2002, 10(5): 293-302.

[2] Fu Z, Lu G, Ting K M, et al. "A survey of audio-based music classification and annotation". *Multimedia, IEEE Transactions on*, 2011, 13(2): 303-319.

[3] Bergstra J, Casagrande N, Erhan D, et al. "Aggregate features and AdaBoost for music classification". *Machine learning*, 2006, 65(2-3): 473-484.

[4] Fu Z, Lu G, Ting K M, et al. "On feature combination for music classification". *Structural, Syntactic, and Statistical Pattern Recognition*. Springer Berlin Heidelberg, 2010: 453-462.

[5] Hamel P, Eck D. "Learning Features from Music Audio with Deep Belief Networks". *ISMIR*. 2010: 339-344.

[6] Grosse R, Raina R, Kwong H, et al. "Shift-invariance sparse coding for audio classification". *arXiv preprint arXiv*:1206.5241, 2012.