MUSIC GENRE/MOOD/COMPOSER CLASSIFICATION: MIREX 2014 SUBMISSIONS

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

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

ABSTRACT

In this submission, we applied feature detector on spectrogram to capture higher level features. We used feature maps to capture percussion and harmonic components, respectively. Finally the down sampled features are connected to a multi-layer network classifier.

1. INTRODUCTION

This work is inspired by convolutional neural network(CNN). We propose using CNN on spectrogram. 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 catenated and connected to a multi-layer perceptron (MLP).

2. FEATURE EXTRACTION

2.1 Receptive Fields & Feature detector

We first applied STFT on 8k sampled monaural audio and get spectrogram. Window size is 512, overlap is 256. Then we applied CNN inspired by image processing on spectrogram. First, we introduce the feature detector. They are some small blocks of size 7*7, 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 such as percussion and harmony in spectrogram.

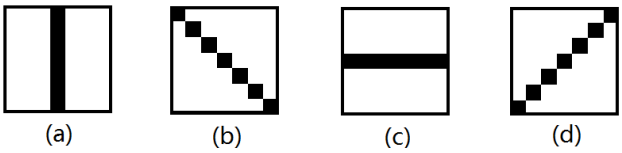

Figure 1. Feature detectors.

Fig. 1(a): capture percussive 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 and get four feature maps, as shown in Fig 2.

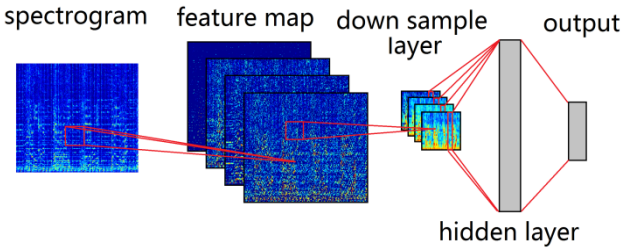

Figure 2. Structure of convolution neural network.

2.2 Sub-sample Layer

We use a 9*9 maximum sub-sample matrix on each of feature map to reduce dimension. So the number of weights decrease to $\frac{1}{81}$ of original. We choose maximum operator because it is the simplest to implement.

3. CLASSIFICATION

Finally the output of sub-sample layer is connected to multi-layer perceptron (we use some code from UFLDL http://deeplearning.stanford.edu/wiki/index.php/UFLDL_Tutorial). We applied aggregation on time scale to get better results. The input nodes' number is 28(per feature-map) * 4(number of maps) * 5(aggregation on time scale) = 560. The hidden nodes' number is 300. The output nodes' number is 10(classes). All the code is written in Matlab.

4. . CLASSIFICATION

[1] 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.

[2] 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.

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

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