CONVOLUTIONAL NEURAL NETWORKS FOR MUSIC MOOD CLASSIFICATION TASKS

Wenhao Bian
Beijing University of Posts and Telecommunications
winniebian@163.com

Second author
Affiliation2
Author2@music-ir.edu

Third author
Affiliation3
author3@music-ir.edu

ABSTRACT

In this submission, we used a convolutional neural network(CNN) as the feature extractor and training model and the melspectral transform is used to preprocess the original audio music. For training, we utilized a five-layer convolutional neural network trained with a dataset similar to MIREX mood organized in five clusters [1]. What's more, the CNN parameters such as filter kernel sizes and learning rate were carefully chosen after a range of experiments.

1. INTRODUCTION

When faced with the multitude of music, users usually search for their interested songs by external characteristics, such as singer, band and year. Although it is simple, this classification ignores the direct user's feeling of music so that users still find it hard to get music that suits their interests from a tremendous amount of music. So current music information retrieval(MIR) is catching on. Not only genre or other timbre description labels, but mood also becomes the popular classification method. However it is difficult to annotate massive music with these labels. To take account of this problem, we propose CNN as our training model. More recently, deep learning-based approach, such as convolutional neural network(CNN), has been used in many fields and has achieved a great success. And CNN is so good at dealing with high-dimensional data and extracting features that classification problem can be handled well by it.

In addition to training, the preprocessing of audio data is important. Based on the previous MIREX tasks, Melspectrogram is proved successfully as an input to a training model. Due to this fact, we trained a CNN with MIREX-like mood dataset from audio melspectrogram to classify different clusters of moods.

2. PROPOSED METHOD

The proposed method is composed of two main steps, audio data preprocessing and network model training. In this section, we discuss the two steps in detail.

2.1 Audio Preprocessing

Before being input to the network model, original audio is preprocessed by melspectral transform. First of all, audio signal is segmented into frames. Then, we transform the frames with Short Time Fourier Transform(STFT). Following the STFT, a Mel filterbank is applied to the spectrogram. Subsequently, we perform a Lg transform of all value. At last, melspectrogram is increased to four dimensions as an input to a CNN. the function from Librosa library can implement the above steps. We used melspectrogram with 96 mel-bands as input. So the 512 samples are used as the FFT windows and 256 samples are used as hop length representing audio of frames between STFT columns.

2.2 Network Model

The proposed network is a five-layer CNN-based model. Based on the mood dataset, 903 songs in five cluster are divided into 90% and 10% for training and validation. And we performed the input data normalization by the same method as batch normalization. The similar structure is used in the five layers that every convolution layer is followed by Batch Normalization (BN), Rectified Linear Units (ReLU), and Max Pooling. At the same time, in order to prevent over fitting, we used Dropout [3] of 0.5 as the last step of the five layers. On the prediction layer, categorical cross entropy loss with Softmax activation is used. To optimize the loss, we used RMSProp [2], adaptive learning rate method by setting initial rate of 0.001.

The kernel's sizes in the five convolution layers are 3×3 respectively, and the corresponding amounts of feature maps are 32, 128, 128, 192, 256, respectively. The window sizes of each max pooling layers are 2×4, 2×4, 2×4, 3×5, 4×4, and stride sizes are 2×4, 2×4, 2×4, 3×5, 4×4. The output from the Softmax layer can be thought of as a probability distribution. So the prediction becomes a 5-D probability vector, in which a label corresponding to the index of maximum value is the predicted label.

In order to get higher accuracy, the batch size and epoch were also set carefully.

2.3 Implementation

The proposed system is implemented in Python using Librosa for transforming audio to melspectrogram and tensorflow for training CNN.

3. RESULTS AND CONCLUSIONS

Figure 1 shows…
Table 1 shows…

String value	Numeric value
Hello MIREX	2010

Table 1. Table captions should be placed below the table.

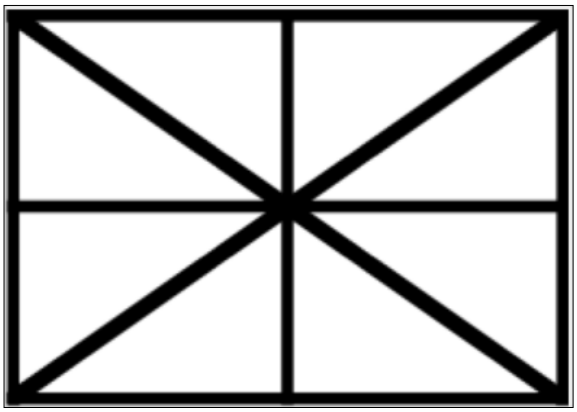

Figure 1. Figure captions should be placed below the figure.

4. REFERENCES

[1] Panda R., Malheiro R., Rocha B., Oliveira A. & Paiva R. P. (2013): Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis. *10th International Symposium on Computer Music Multidisciplinary Research CMMR*,2013.

[2] Sebastian Ruder: *An overview of gradient descent optimization algorithms*, arXiv 1609.04747v2, pp. 6-15, 2017.

[3] HintonGE, SrivastavaN, KrizhevskyA, etal: *Improving neural networks by preventing co-adaptation of feature detectors[J]*. Computer Science, 2012, 3(4): pgs. 212-223.