Music Classification with Neural Network

Bai Ke
Department of Physics, Tsinghua University
baik13@mails.tsinghua.edu.cn

Abstract

Music classification is a challenging task due to its subjectivity. The conventional methods are mainly based on feature extraction; the better-extracted feature, the better performance. Because music is a kind of naturally produced process, it has many time-invariant features. And convolutional neural network is especially good at finding this kind of information. So in this report, we introduced convolutional neural network and its variant recurrent convolutional neural network as new tools for music classification. Although this network hasn't been carefully tuned up to now, it has already reached the average level as many conventional methods. We use two different evaluation criterion to evaluate our model.

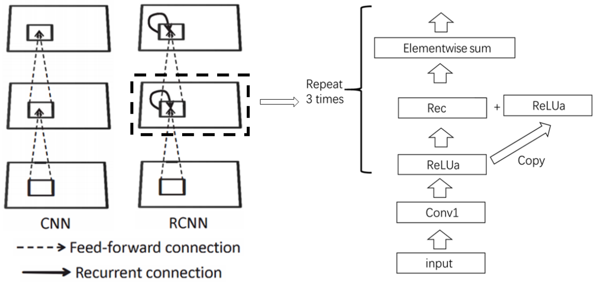

Figure 1. Illustration of the architectures of CNN and RCNN. Left panel: the graphical illustration. Right panel: the detail network architecture.

1. Introduction

The music genre classification is a highly subjective task. One can not always judge the music precisely, since a piece of music may be the mixture of several styles.

Most traditional methods for music classification have three levels: First, low-level features extractor. It is usually a statistical descriptor of music, such as Fourier transform, short time Fourier transform (STFT), spectral centroid. Second, low-level features are combined in certain ways to form mid-level features, like Mel-scale, Bark-scale. These features are designed to mimic the properties of human primary auditory system. Third, a high-level model or discriminator based on mid-level features, which learns conceptual meanings like human beings.

In this report, we used a system only has two levels: a low-level feature and a deep neuron network. We hope the neuron network can extract useful features automatically. The low-level feature in this method is STFT Spectrogram. And we tried two kinds of neural networks, convolutional neural network (CNN) and recurrent CNN (RCNN). Finally, because of over-fitting, the performance of our model only reaches the same level as conventional models.

The following contents are divided into three parts. In the first section, we give a quick review of two models we use. The second part explains the data transference. Then we give a detailed explanation of the two models. At the end of the article, we compare the results of these two models.

2. Backgrounds

2.1. DataSet

The dataset we choose is GTZAN, 23% researchers used this dataset before 2014[3].This dataset has 10 classes. Each class owns 100 samples. Each sample is a piece of music lasting 30 seconds. Up to now, the highest classification rate is 92.24%.[2]

2.2. CNN and RCNN

CNN has a good performance on image classification and audio classification. Comparing to conventional methods, it can automatically extract the hidden information from huge data.

For improving the performance, We also apply the RCNN model to this task. Comparing to the former one, the RCNN introduces recurrent units in a single layer (Fig. 1),which was proved to have better performance than conventional CNN on sequence classification tasks[1].

Table 1. The configuration of CNN for music genre classification.

Layer	Data	Conv1	Pool1	Conv2	Pool2	Conv3	Pool3	FC1	FC2	Classifier
Param	whitening	5x5, Pad 2 St 1, BN, ReLU	Max, 2x2 St 2	5x5, Pad 2 St 1, BN, ReLU	Max, 2x2 St 2	3x3, Pad 1 St 1, BN, ReLU	Max, 2x2 St 2	512 Drop 0.5	128 Drop 0.4	Softmax
Size	33x166x1	33x166x32	16x83x32	16x83x32	8x41x32	8x41x32	4x20x32	1x1x512	1x1x128	1x1x10

St: stride, BN: batch normalization, Drop: dropout ratio, Ave/Max: average/max pooling, FC: fully connect

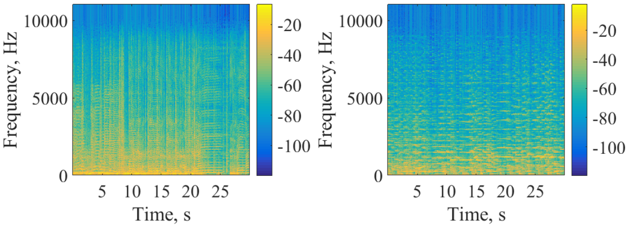

Figure 2. The left picture is the spectrum of a classical music, the right one is that of a disco.

3. Methods

3.1. Data Preprocessing

CNN can uses vector or matrix as its input. In this case, music sample can be treated as vector in time or frequency domain. But it lead to loss of information. Using the amplitude of the sound directly leads to the loss of the important frequent information. And using the Fourier transform on the entire song leads to the loss of the sequential information. A wiser choice is expand the 1D music signal to 2D time-frequency domain to preserve the time and frequency information simultaneously. So we used STFT here. To check that STFT is a valid method, two samples from different categories are shown in Fig. 2. It is obvious that these two pictures have sensible differences.

The window size of STFT is 46 ms. Compared with the around 25 ms window in many traditional vocal analysis, a larger window can extract the feature of melody and chord better. The overlay rate between windows is 75%. The number of filters is 33 for detecting the frequency. The 30-second-long music is cut into 15 pieces.

After the transformation, the input data contains 15000 samples: 1500 samples for each genre. Each sample is a 33×166 matrix.

3.2. Model

Table 1 shows the CNN structure.

The RCNN structure is based on CNN structure. We add the recurrent structure in the convolutional layer, which is shown in Fig 1. We conducted a grid search on the hyperparameter space to determine the network complexity and dropout ratio.

Table 2. The accuracy of three different models

	clip accuracy (%)	piece accuracy (%)
CNN	70.6	82.5
RCNN	72.4	83.0
LPNTF[2]	-	92.2

4. Result and Analysis

The results are shown in Table 2. Even in the best hyperparameter settings, the CNN & RCNN suffered from serious over-fitting and there is a huge gap between them and the state-of-art method.

Because of the limited data, a trade-off between the data size and the complexity of the model should be taken into consideration. If the model is too simple, it is hard to converge at last, leading to poor performance. If the number of parameters is too large, the over-fitting is serious. Enlarging the dropout ratio is a wise chose, but it still can not solve this problem in the experiment. So our model can not reach a high classification right rate.

To make sure that the model converges, the learning rate should also be appropriate to adapt to the model.

5. Discussion

Doing a music genre classification task with CNN largely simplify the tedious feature extraction process.

The result can not compare with the best methods. One the one side, it shows us the importance of data volume to the deep learning. On the other side, we can find that the conventional methods and deep learning have their own strength. it motivates us to combine the conventional methods with deep learning and make a good use of their advantages.

References

[1] M. Liang and X. Hu. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3367–3375, 2015.

[2] Y. Panagakis, C. Kotropoulos, and G. R. Arce. Music genre classification using locality preserving non-negative tensor factorization and sparse representations. In *ISMIR*, pages 249–254, 2009.

[3] B. L. Sturm. A survey of evaluation in music genre recognition. In *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, pages 29–66. Springer, 2014.