Deep Learning for Content-based Music Genre Classification

Zesen Wang
Shanghai Jiaotong University
University of Michigan - Shanghai Jiao Tong University Joint Institute
Email: UncleSam@sjtu.edu.cn

Zhi Lin
Shanghai Jiaotong University
University of Michigan - Shanghai Jiao Tong University Joint Institute
Email: linzhilynn@sjtu.edu.cn

Zhengyang Feng
Shanghai Jiaotong University
University of Michigan - Shanghai Jiao Tong University Joint Institute
Email: crukedshfeng@sjtu.edu.cn

Wendi Wang
Shanghai Jiaotong University
University of Michigan - Shanghai Jiao Tong University Joint Institute
Email: avage.Wwendy@sjtu.edu.cn

***Abstract*—Currently music are likely to be classified and recommended based on user behaviour. But a content-based music genre classification is still necessary because for some new songs with few user behaviour, the correct classification of it decides whether the new songs can be properly recommended to target users. Therefore, in this paper, we propose a new deep learning model to classify music purely based on their content. The main characteristics of the model is that it uses different sizes of kernel in different convolution layer, and the last 3 convolution layers are concatenated as the input of fully-connected layer. This makes the model able to deal with the problem of different scales caused by different rhythms.**

I. Introduction

Lots of facts make automatic music genre classification (AMGC) intelligent systems vital nowadays. The ease of downloading and storing music files on computers, the huge availability of albums on the Internet, with free or paid download, peer-to-peer servers and the fact that nowadays artists deliberately distribute their songs on their websites, make music database management a must. Another recent tendency is to consume music via streaming, raising the popularity of on-line radio stations that play similar songs based on a genre preference. In addition, browsing and searching by genre on the web and smart playlists generation choosing specific tunes among gigabytes of songs on personal portable audio players are important tasks that facilitate music mining.

Also, music genre classification is an ambiguous task that few methods can obtain high accuracy. Lots of artificial methods like Gaussian mixture model are invented to deal with music classification, but they have limits. The artificial neural network is known to have good performance on classification on image classification. Even though music is one dimension data, some popular audio preprocessing method like MFCC provides neural network with enough data for it to learn.

In our paper, we propose a network model to solve this problem.

II. Data & Preprocessing

A. Dataset

We have searched for several datasets with their available metadata, and we select GTZAN Genre Collection, of which contains 1000 audio tracks each 30 seconds long. The dataset is used for the well known paper in genre classification "Musical genre classification of audio signals" [1]. It contains 10 genres, each represented by 100 tracks. The tracks are all 22050Hz Mono 16-bit audio files in .wav format. The 10 genres are classical, jazz, metal, pop, country, blues, disco, metal, rock, reggae and hiphop.

B. Preprocessing

The raw audio file is 22050Hz Mono 16-bit audio file which contains too much information for an input of a neural network. After doing some researches on audio processing, we find that Mel-scale feature and MFCC are widely used in audio processing. We finally choose Mel-scale feature because it is more used to process music and songs while MFCC is more common in the preprocessing of speech recognition.

Firstly, we visualize the Mel-scale feature to see whether some features are potentially able to be extracted, but it seems that the differences along time series are not significant (Figure 1).

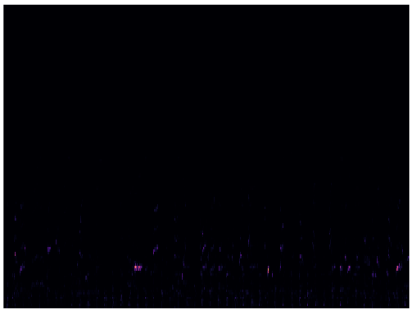

Fig. 1. Mel-power Spectrum

Then a logarithm is taken on the raw mel-power data. As the image shows, both the differences along time and the differences along frequencies are significantly more obvious, which means this preprocessing may have the potential to provide learnable data for the network.

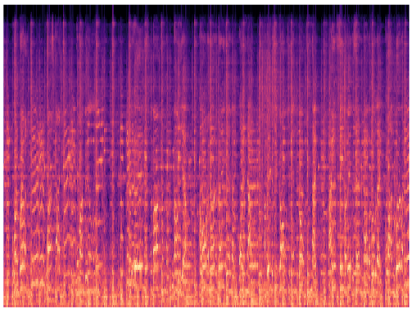

Fig. 2. Log Mel-power Spectrum

III. Model

A. Framework

In Figure 3, it shows the structure of our model. Our method is based on a feed-forward network. The data is processed by a number of convolution layers, and the last three layers are concatenated as the input of "FC1". Then it uses two fully-connected layer to generate the 10-dim output which is corresponding to the confidences of 10 genres.

B. Features

1) Multi-scale Convolution Kernel: Unlike some popular networks used in image recognition (VGG-16 [2], for example), in the convolution layers, convolution kernels with varying sizes are used. We introduce this feature aiming at solving unstable feature extraction due to different rhythms of different songs.

2) Multi-scale Feature Maps: In the last part of our network, we concatenate the last three convolution layers as the input of "FC1". We have this idea because of the inspiration from well-known model SSD [3]. The convolution layers' sizes decrease as the number of layers increase, which allows features with different scales to be collected at the fully-connected layer.

C. Deep Learning Techniques

We use the most popular activation function ReLU in our deep model which is the activation function that makes training the whole model at one time possible. We introduce two fully-connected layers to solve potential non-linear separation problems which work significantly better than one fully-connected layer in practice.

After each convolution layer, one batch normalization layer follows. Batch normalization layer normalizes the output value for each layer, which solves the problem of explode/vanish error gradients to large extent.

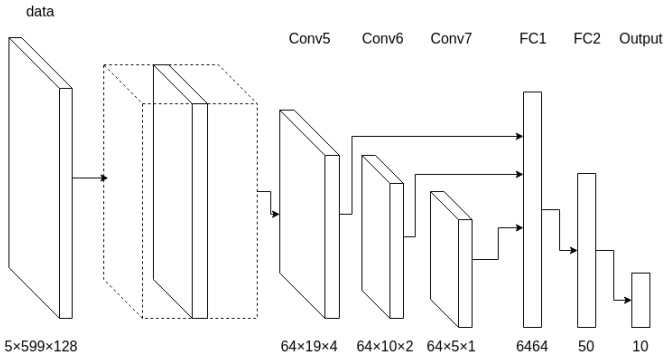

Fig. 3. Classification Model

D. Loss Function

We use the softmax loss function.

$$S_j = \frac{e^{o_j}}{\sum_{k=1}^{C} e^{o_k}}$$

$$Loss = -\sum_{j=1}^{C} y_j \log S_j$$

where o_j is the j^{th} output of the output layer, $y_j = 1$ means that this sample belongs to genre j, and C is the total number of genres (10 in this case).

IV. Experiment & Result

A. Setup

In the experiment, we use 60% of data as training set, and we use other 40% as test set. The training data and test data are balanced on genres.

We use stochastic gradient descent as the main method to train the model. The training ends until the accuracy on test set has no significant increase.

B. Experiment

1) Accuracy: The first experiment is to train the model on the training set, and test the model on the test set and training set to measure the accuracy and top-3 accuracy.

2) k-NN Classifer: The second experiment is to use the output of "FC2" as the deep feature of the input. The result is evaluated as a projection of input data on 50-dimension space by the generalization error and k-NN classifier error.

This experiment aims to test whether the deep feature can be used as index for classification and cluster, which can be treated as a judgment for similar recommendation.

C. Visualization

In the third experiment, principle component analysis is applied to the deep feature, and the feature is projected to 3-dimension space and 2-dimension space. We visualize the result to see whether the classes are seperable.

	Our Method	NN [4]	DBN [4]	GMM [1]
Train Accuracy (%)	**99.8**	94.69	75.3	\
Test Accuracy (%)	**72.5**	60.46	61.15	61±4
Train Top-3 Acc. (%)	100.0	\	\	\
Test Top-3 Acc. (%)	92.2	\	\	\

TABLE I
ACCURACY ON TRAINING SET AND TEST SET

D. Result

We compare our result with the result from Tzanetakis, George, and Perry Cook [1] and the result from Feng, Tao [4] which used the same data set as in this paper.

The result of the first experiment is shown in Table I. The result shows that our method outperforms other method like auto-encoder, deep belief network and Gaussian mixture model which are used in other papers.

It has to be mentioned that the result of Feng, Tao [4] is the accuracy on 4-genre classification, which means for 10-genre classification, the accuracy is expected to be lower.

The result of the second experiment is shown in Table II.

The result of the third experiment is shown in Figure 5 and Figure 4.

	Generalization Error	5-NN Classifier Error
Training Set (%)	1.0	0.3
Test Set (%)	35.3	33.5

TABLE II
GENERALIZATION ERROR AND K-NN CLASSIFIER ERROR

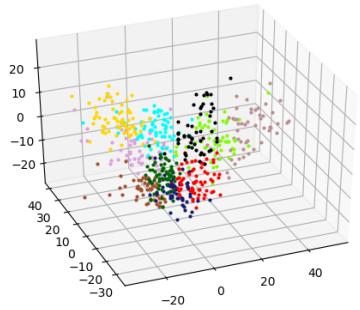

Fig. 4. PCA Result where n-component=3

V. CONCLUSION

This paper introduces a deep learning method to solve music genre classification. A key feature of our work is that the network has features with different scales concatenated, which makes full of information from different levels. Traditional feed-forward network may lead to information loss during the processing of layers.

The results show that the neural network we trained has the potential to solve the problem of music classification. And the deep feature it extracts has correlation with its genres.

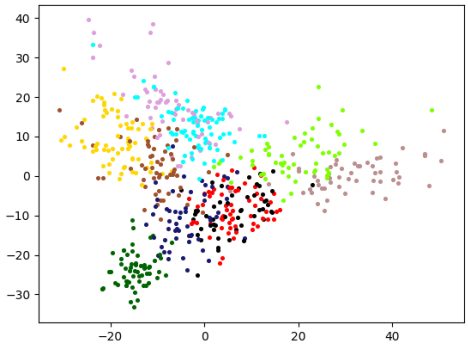

Fig. 5. PCA Result where n-component=2

Apart from the task done by our network, the deep feature can be used as the index for similar music recommendation by making use of the extracted features. Also, the base network we trained can be used to solve multiple tasks like recognizing the author of any songs. The promising structure for deep network has much potential in other tasks.

REFERENCES

[1] Tzanetakis, George, and Perry Cook. "Musical genre classification of audio signals." *IEEE Transactions on speech and audio processing* 10.5 (2002): 293-302.

[2] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv*:1409.1556 (2014).

[3] Liu, Wei, et al. "Ssd: Single shot multibox detector." *European conference on computer vision*. Springer, Cham, 2016.

[4] Feng, Tao. "Deep learning for music genre classification." *private document* (2014).