Deep Learning Approach to Music Genre Classification

Yunfei Luo YUNFEILUO@UMASS.EDU
College of Information and Computer Science
University of Massachusetts Amherst
Amherst, MA, US

Abstract

Music Genre Classification task is essentially a classification problem conducting on time-series with multi-resolution. The trained model has great application potentials in the assistants to professional music works evaluation, recommendation system, and audio information retrieval. Traditional methods such as K-Nearest Neighbors, Decision Tree, and Linear Regression are often used in modeling. Recently, many research works focus on applying the deep learning method such as Multi-Layer Perceptron, Convolutional Neural Network. These works came to the same conclusion that the deep learning method achieves state-of-the-art performance. Other related research works have shown that Recurrent Neural Networks such as LSTM is robust in extracting patterns from time-series data. In this project, I provide empirical analysis indicating that convolutional neural network is a better approach than LSTM, and Mel-spectrogram is a better feature-engineered representation of musical data than Mel-frequency cepstral coefficients (MFCC).

1. Introduction

When adopting machine learning approach to a classification task, it is important to determine the proper methods to use based on the nature of the dataset. The classical methods such as K-Nearest Neighbors and Decision Trees are usually very intuitive and easy-interpreted models for classification. However, music data, as a time series data that contains have multi-resolution nature (there are many elements such channels at different frequencies), requires the models to be able to extract more complex pattern in order to achieve considerable performance on the classification task Tzanetakis and Cook (2002).

Deep learning model have been shown to be robust in many fields. There are many research works done on using neural networks for music genre classification. The trained models are claimed to have great potential benefits for Music Information Retrieval (MIR). However, there are many choices, among which Convolutional and Recurrent Neural Networks (CNN and RNN) are the most common choice for this task. CNN extract local features, while RNN tends to extract more global features though there are potential problems with long-term dependencies. They are the *system* in this problem, and it is hard to say which one is better. Empirical results are needed for analysis.

As music data is essentially a time-series data, it is intuitive to apply recurrent neural networks for extracting features. But many of the popular work are using convolutional rather than recurrent networks, such as Vishnupriya and Meenakshi (2018) and Allamy and Koerich (2021). Wu et al. (2018) indicates that recurrent neural networks could achieve considerable performance.

Convolutional and recurrent neural network have completely different structure, and different behaviour in extracting features from data. In order to get better understanding of these two different architecture, it is worth to addressing the question "is recurrent neural network a better choice?", "If it is, then under what circumstance does the statement hold? And what could be the explanation?", "If it doesn't, what could be the explanation?".

In addition, data is one of the most important section of this project, as it is the *environment* of where the task is performed in. Choi et al. (2017) discussed about several existing preprocessing techniques. A good preprocessed data will let the models to be easier trained, or converge faster, because preprocessing usually involve normalization and pre-feature-extraction based on some existing golden knowledge. Thus, exploring this research question and identify the frontier for it would be helpful for exploration in the main research questions discussed above.

In Allamy and Koerich (2021), the raw amplitude series are directly used for training. Well, most other works use the Mel-spectrogram, or mel frequency cepstral coefficients (MFCC), that includes processing through Short-term Fourier Transformation, and some cosine and log scales.

Finally, it is well known that CNN could benefit from stacking more layers, while LSTM didn't. There are some tradeoffs while selecting models. The number of parameters and time of forward processing need to be considered. This might be a potential research question that could be taken into discussion.

2. Related Work

Tzanetakis and Cook (2002) is a well known research work on musical genre classification. The authors extract three sets of features based on MEL-frequency cepstrum of the audio file, namely timbra texture, rhythmic content, and pitch content. They then feed these feature sets to Gaussian mixture models (GMM) for training, and achieve the best performance among all the methods they experimented on the GTZAN dataset. Tzanetakis and Cook (2002) not only act as a forerunner on the task of music genre classification, but also indicate that MEL-frequency cepstrum and result series from STFT are more valuable representations of music data rather than the raw audio signal.

Vishnupriya and Meenakshi (2018) is one of the recent works that adopt neural network to the task. The authors preprocessed the data by calculating the MEL-frequency cepstrum following Tzanetakis and Cook (2002). They conduct experiments on a large dataset: Million Song Dataset (MSD), where 2D convolutional neural network are applied. They've showed that MEL spectrogram is more friendly to deep learning model comparing to MFCC. Though this paper demonstrates that the deep learning approach is worth to try, they didn't provide comparison between the model and a baseline. Moreover, a potential issue in this work is that MEL-frequency cepstrum is still a time-series data. 2D convolutional neural networks usually works well on image data as they are doing well in extracting local features, thus, the model might miss some information in the pattern across time.

Allamy and Koerich (2021) evaluate the performance of 1D convolutional neural networks (1D-CNN) on GTZAN dataset. Their approach is more reasonable than Vishnupriya and Meenakshi (2018) as the audios are time-series data. The paper provide comparison of the results between several previous works, and showed that 1D-CNN achieve the best per-

formance. Introducing data augmentation could further boost the performance. Although the authors cite the works related to MEL-frequency cepstrum, their experiments are still conducted on the raw audio signal, where the numerical values represent the amplitude.

For indicating the applicability of Recurrent Neural Networks, Wu et al. (2018) conduct experiments on GTZAN with their proposed model. The authors shows that either LSTM and Independent Recurrent Neural Network outperform the forerunner Tzanetakis and Cook (2002). They achieve an averaged accuracy of 96% under 10-fold cross validation setting, where Tzanetakis and Cook (2002) achieve 61% under the same setting 20 years ago.

3. Methodology

3.1. Problem Setup

The machine learning task that will be engaged is the music classification task. More specifically, we have the input data X and the labels Y such that

$$X \in \mathbb{R}^{N \times L \times D}, \quad Y \in \{1, 2, ..., 10\}$$

where N is the batch size, L is the length of audio series, and D is the number of features at each time step. We construct functions (or models) f with parameters θ. We then have the optimization task

$$\min_{\theta} \sum_{n=1}^{N} L(y_n, f_\theta(x_n)) + R(\theta)$$

where L is the loss function (Cross Entropy Loss is adopted), and R is the regularization term (depends on how the optimizer was initialized).

3.2. Data preprocessing

The raw audio data is essentially a series representing the amplitude. As each songs, denoted as x_i, in the GTZAN dataset is 30 seconds long, for 10 genres such as jazz, classical, pop, etc. There are 1000 audio files in total, with 100 for each genre. The raw audio signal has dimension $x_i \in \mathbb{R}^{1 \times 110250}$. In Allamy and Koerich (2021), the authors train models directly on the raw data. However, Choi et al. (2017) have presents several pre-processing methods for audio data, among which Mel-frequency cepstral coefficients (MFCC) have shown to be a better representation of audio comparing to raw signal in the task of speech recognition.

First of all, I transfer all the raw signal to Mel-spectrogram. During the transformation, short-term Fourier Transfer is applied on the series first; then the resulting series is mapped to Mel scale which is a perceptual scale of pitches. Finally, log and discrete cosine transform are applied to generate the final spectrogram. The resulting series have dimension

$$x_{i,mel} \in \mathbb{R}^{128 \times 1024}$$

where the first dimension with value 128 represents the number of Mel channels, and the second dimension is the time axis.

Mel-spectrogram still have a bit large dimension, since there are 128 features required to represent different frequencies at one time step. MFCC is a more latent representation

that maintain the amplitudes information in the Mel-spectrogram. As the MFCC being extracted, the final series have dimension

$$x_{i,mfcc} \in \mathbb{R}^{16\times 1024}$$

where the second dimension still represent the time axis, but only 16 channels needed to represents the information at one time step.

Although MFCC has shown its robustness in the field of speech recognition, it is ambiguous to say which of MEL-spectrogram and MFCC is better on the task of musical genre classification, because musical data is considered as multi-resolution series data. Both of these preprocessed data will be fed into the experiments separately. More details about the experiments design are in the latter section.

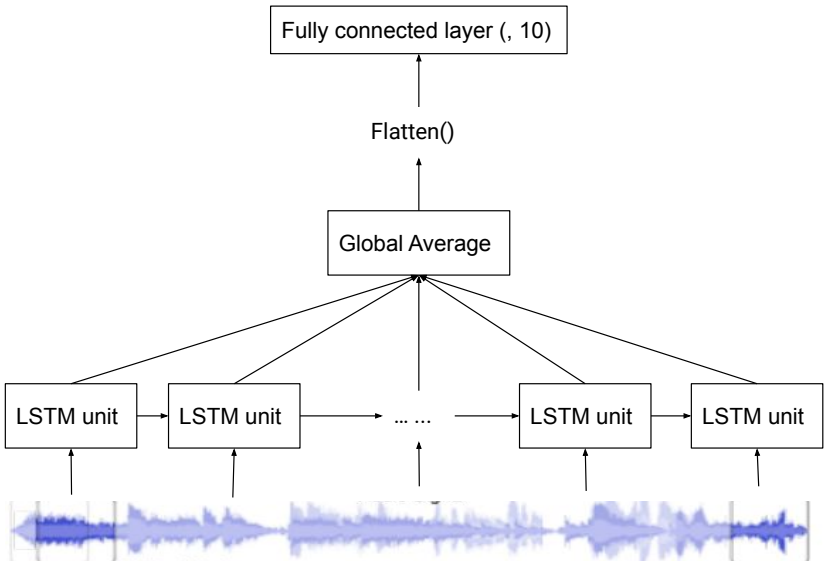
3.3. Models

3.3.1. Shallow 1D CNN

This is the simplest model proposed in this project, and it will be treated as the baseline. The shallow 1D CNN consist of a Convolutional layer, followed by a Max-Pooling layer, followed by a fully connected layer as the output layer. Because the length of 1 second in the preprocessed data (For both Mel-spectrogram and MFCC) is roughly 21, the filter size of 21 is chosen for the first convolutional layer.

This model is expected to capture only the local features of the series data. More specifically, the output that will be fed to the linear classifier is the high frequency pattern lie in the window size of 1 second.

3.3.2. LSTM



A Long-short-term Memory (LSTM) based model is constructed followed Wu et al. (2018)

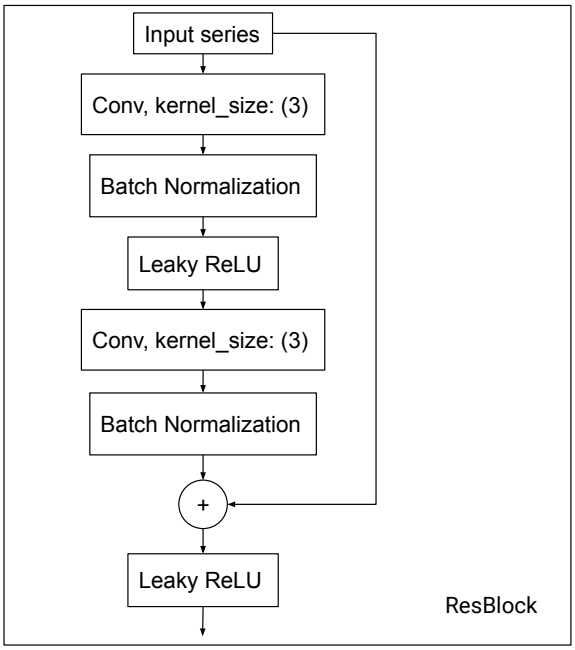
where the average of all the hidden states are fed into the final fully-connected layer, as shown in the figure above.

Each LSTM unit takes one step in the time-series as input. LSTM is chosen here because it is more widely used recurrent neural network than others such as vanilla RNN and GRU. The cell states in addition to the hidden states are designed to mitigate the issue lie in long dependencies. In addition to deal with the long dependencies issue, the global average is used. As a result, the latent feature representation that will be fed to the linear classifier contains information of hidden states from all the time steps, and the model is more likely to capture the patterns lie in low frequencies.

3.3.3. 1D CNN ResNet

This model is constructed followed Allamy and Koerich (2021), where there are 21 layers in total. When using Mel-frequency cepstral coefficients (MFCC) as input to the model, the number of layer is reduced to 13 because the length of MFCC is much shorted than the raw audio signal as used in Allamy and Koerich (2021). In this model, the filters of the several layers at the beginning is expected to extract the high frequency features, same as the shallow CNN. The latter the layer is, the higher the expectation that the filters will extract the low frequency features.

A single block in the ResNet consist of two identical set of layers begin with a convolutional layer followed by batch normalization and leaky-ReLU activation function. The final output is the addition of the input with the output from last batch normalization layer, as shown below



Details of the layers are shown below in table 1 with specification of the number of filters, kernel size, and stride of each layer. The input length is 1024 as described above in the data preprocessing section. The output from the final convolutional layer is expected to deduct the series length into 1 with 512 number of features representing the latent representation

of the original input series. Finally, same as the other two methods above, a linear classifier is used to generate the final prediction scores.

Table 1: Details of the 1D ResNet

Layer	# Filters	Kernel Size	Stride
Input	-	-	-
Conv1D	128	3	3
ResBlock	128	3	1
MaxPool	-	3	3
ResBlock	128	3	1
MaxPool	-	3	3
ResBlock	256	3	1
MaxPool	-	3	3
ResBlock	256	3	1
MaxPool	-	3	3
ResBlock	512	3	1
MaxPool	-	3	3
Conv1D	512	1	1

4. Hypothesis

4.1. LSTM will result in better performance than 1D CNN. (Existential)

LSTM, as a commonly choice when we decide to apply recurrent neural network for modeling, have shown to be robust on extract features from time series data. Music data is considered as time series data with each time step represents amplitude or frequencies.

1D CNN extract local features. From the empirical results from image classification, local features have shown to be more helpful in terms of classification accuracy. However, in music classification task, the underlying pattern among the time steps are more important than the local features. Some people called it as "flow" of the music. As 1D CNN might miss the information lie in "flow".

4.2. 1D CNN can benefit from stacking more layers, while LSTM might not. (Attributional)

According to the practice of ResNet, the authors have shown that it is possible to stack more convolutional layers to be stacked together with the performance being improved at same time. It would also worked for the music classification task with 1D CNN. Because the more layers we have, the more higher level features we could have. And these higher level features also have a relative large Field of View (FoV). As stated in the previous hypotheses that 1D CNN is likely to miss the melody information, increasing the FoV would recover this issue.

However, LSTM is less likely to benefit from stacking more layers. One layer of LSTM already extract features at global level. Stacking more LSTM layers wouldn't make much

difference. Purely raise the number of parameters with no obvious improvement on any entities would only increase the complexity of the model, which will let the model suffer from overfitting.

5. Experiments

Experimental analysis is intended to be conducted. The performance of each of the models described in the "Machine Learning Methods" section will be compared against each other. Each of these model will be trained separately on Mel-spectrogram and MFCC. Therefore, there will be 6 trails in total:

- Shallow 1D CNN trained on Mel-spectrogram
- Shallow 1D CNN trained on MFCC
- LSTM trained on Mel-spectrogram
- LSTM trained on MFCC
- 1D CNN ResNet trained on Mel-spectrogram
- 1D CNN ResNet trained on MFCC

For each trail, the model will be evaluated on the same test set. From the final performance table, I intend to draw conclusion for the hypotheses of whether LSTM is a better approach compared to 1D CNN, and discuss the possible explanation of the behaviors of the models according to the empirical results. Moreover, whether or not MFCC is a better representation of the audio could also be determined from the result table.

5.1. Evaluation

Accuracy ($\frac{\#correct\ predicted}{\#test\ samples}$) will be used as the metric for evaluation. All models will be trained with the same set, and tested with the same set. The train and test set are randomly splitted with portions 80% to 20%. The accuracy on test set is used as the main criterion for the performance of each model.

5.2. Empirical Results

Table 2 present the testing accuracy, the number of epochs till convergence, and the number of trainable parameters for each model. The results for ResNet CNN has been added.

From the table we can see that model trained with Mel spectrogram achieve higher accuracy, and the shallow convolutional neural network achieve the highest performance.

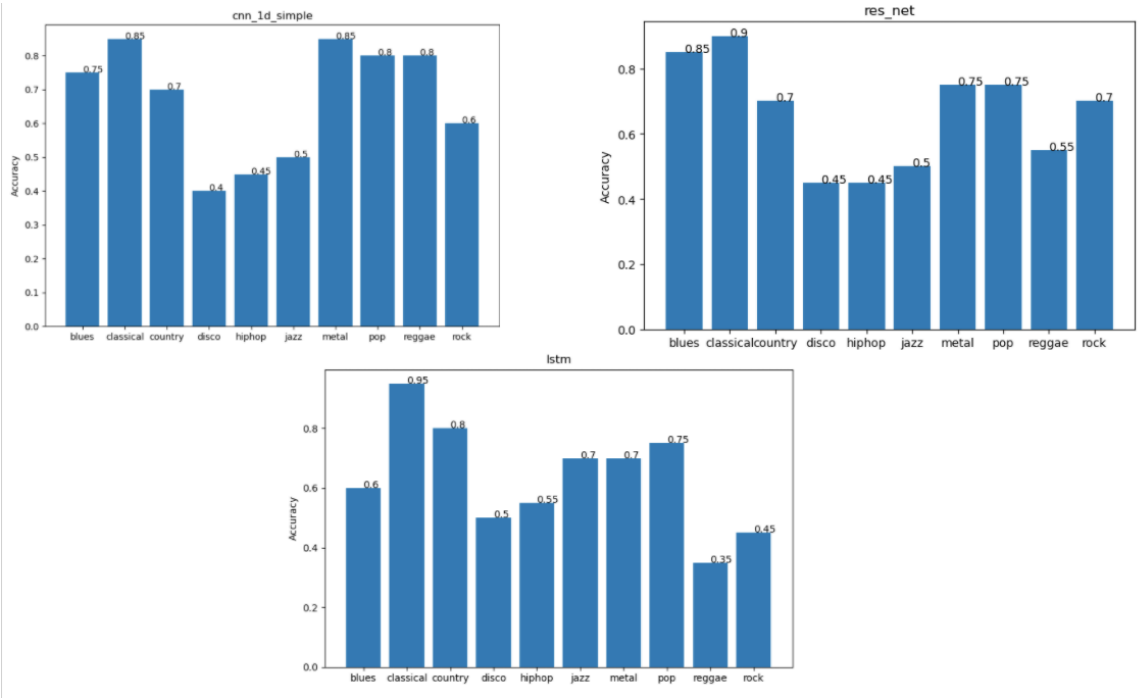
As the rejection of the other two hypotheses is discussed in Empirical Results I, I will only discuss the final hypothesis. Unfortunately, the empirical results reject my final hypothesis that convolutional neural network can benefit from stacking more layers on the task of musical genre classification on the dataset GTZAN. Though stacking more layer could achieve competitive results with shallow version CNN, the high computational costs (large number of parameters) makes the model having low cost performance. Theoretically, stacking more convolutional layers is supposed to raise the field of view (FoV), which is

Table 2: Accuracy of musical genre classification on GTZAN

Model	Accuracy	#Epochs	#Params
3-layers CNN, on MFCC	0.21	20	134282
3-layers CNN, on Mel	**0.705**	100	435338
LSTM, on MFCC	0.25	80	76042
LSTM, on Mel	0.68	**280**	133386
ResNet CNN, on MFCC	0.43	20	2346378
ResNet CNN, on Mel	0.66	80	**2389386**

equivalent of capturing patterns lie in low frequencies of time-series. However, GTZAN is a very small dataset, the huge amount of parameters might lead to overfitting very easily.

5.3. Performance Analysis



I then conduct further visualization analysis on models trained with Mel spectrogram. As shown in the figure above, the testing accuracies of each musical genre are displayed for the three models. From the figure we can see that genres like *classical, country, metal, and pop* are more likely to be correctly classified. Some other genres like *disco, hiphop, and jazz* seems a bit hard to be correctly classified. Moreover, *reggae and rock* seems also very likely to get low accuracies. A general pattern is that musical genre with clear drumbeat is more likely to confuse the models, as the beat will dominate the other features in the spectrogram, and all the beats are very similar to each others in general.

As all the hypotheses have been rejected, several new conclusions have been drawn:

- 1D CNN outperform recurrent neural network, which indicate the significance of capturing patterns lie in high frequencies.
- Mel spectrogram is a better representation of the data than MFCC on the task of musical genre classification, as Mel spectrogram maintain most information within various range of frequencies.

6. Discussion

6.1. MFCC vs. Mel spectrogram

My hypothesis before experiments stated that MFCC is a better representation for the musical data than Mel spectrogram, because it has been shown to be a reliable choice speech recognition task. Moreover, MFCC got a much lower dimensions than Mel spectrogram, and was claimed to be as informative as Mel spectrogram. Thus, I thought using MFCC could speed up the convergence.

However the empirical result shows that it is not the case. The testing accuracy is not promised, and the model is very likely to suffer from overfitting, even many regularization methods have been tried (L2 penalty, dropout, reducing learning rate etc.).

One possible explanation is that Mel spectrogram maintain all the information of magnitude at different frequencies. MFCC is highly likely to missing such information because it only extract the amplitude. Because music is a multi-resolution data, the patterns lie in each level of the frequencies are significance. This is a key aspect that I haven't thought about while came up with the hypothesis.

Another reason that MFCC often lead to overfitting is that the size of the dataset is a bit small (less than 1000 samples). With larger dataset, it might be hard to tell whether models trained with MFCC perform as good as models trained with Mel spectrogram.

6.2. Local vs. Global features

My hypothesis state that local features is should be more helpful for the musical classification task than local features. However, the empirical result shows that local features is as significance as it was in image classification task. Furthermore, train a convolutional neural network is much easier than train a recurrent neural network. As stated in the table 2, although CNN got more parameters, it could benefit from parallel computing which doesn't work for LSTM. The result also indicate that LSTM require much more epochs to converge.

The good news is that LSTM achieve competitive performance than the shallow 1D CNN. The next step is to build deeper 1D CNN to see if the performance could be improved. In addition, analysis of accuracy for each genre in GTZAN dataset will be conducted for further discussion.

7. Conclusion and Future Works

In this project, I explore the behaviors of convolutional and recurrent neural network on the task of musical genre classification. The empirical results indicate that 1D CNN that capture high frequency as local feature outperform LSTM that tends to extract more global patterns. Moreover, Mel-spectrogram is a better feature-engineered representation of musical data

than MFCC, as musical data has the essence of multi-resolution that MFCC will negate much useful information.

The 1D ResNet in this project didn't demonstrate its potential. Exploring data augmentation technique or conducting experiments on larger dataset could be the directions. Furthermore, stacking more convolutional layers is not the only solution to increase the so-called field-of-view (FoV) that allow the model to capture the low frequency features, dilated convolution could be another approach. These further possible explorations are leaved for future works.

References

Safaa Allamy and Alessandro Lameiras Koerich. 1d CNN architectures for music genre classification. *CoRR*, abs/2105.07302, 2021. URL https://arxiv.org/abs/2105.07302.

Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark B. Sandler. A comparison on audio signal preprocessing methods for deep neural networks on music tagging. *CoRR*, abs/1709.01922, 2017. URL http://arxiv.org/abs/1709.01922.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. doi: 10.1109/TSA.2002.800560. URL https://ieeexplore.ieee.org/document/1021072.

S Vishnupriya and K. Meenakshi. Automatic music genre classification using convolution neural network. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–4, 2018. doi: 10.1109/ICCCI.2018.8441340. URL https://ieeexplore.ieee.org/document/8441340.

Wenli Wu, Fang Han, Guangxiao Song, and Zhijie Wang. Music genre classification using independent recurrent neural network. In *2018 Chinese Automation Congress (CAC)*, pages 192–195, 2018. doi: 10.1109/CAC.2018.8623623. URL https://ieeexplore.ieee.org/abstract/document/8623623.