MUSIC GENRE CLASSIFICATION BY BILSTM

Erica Wei (UNI: cw3137)

Columbia University

ABSTRACT

Musical genres are categorical labels created by humans to characterize music style. A musical genre is characterized by the common similar musical characteristics shared by its members. Those musical characteristics are highly correlated with the rhythmic structure, beats, and harmonic content of the music. Traditionally, musical genre annotation is performed manually by human. However, automatic musical genre classification can help or even replace the human user in this process, not only is cost-effective but also eliminate human errors to generate comprehensive understanding.

In this paper, we proposed an approach to apply efficient technique from speech recognition to extract features and train a classifier to automatically predict music genre based on their musical characteristics. We used librosa to extract timbral texture features, pitch feature and spectrogram features from music to represent the audio characteristics and then apply Bidirectional Long Short-Term Memory (BiLSTM) Neural Network, which implemented in PyTorch to carry out classification. Additionally, a feature analysis is proviede and for experimental training, which gives a sense how the feature combination and dataset inlfuenced the results.

Index Terms— Music Genre Classification, Spectrogram features, Pitch, Texture Features, Long Short-Term Memory Neural Network

1. INTRODUCTION

Large music collections has raised the challenge of how to retrieve, browse, and recommend based on their genre or styles. So Music genre labels are useful to organize songs, albums, and artists into broader groups that share similar musical characteristics. The traditional way is to manually labeled by human, but there are many problems such as different opinion based personal choice, does not have standard rule to classify them, and it also costs huge energy of human, not efficient. However, to automatically labeled by machine will be very useful and more effective, meanwhile to classify by standard rule will be more accurate, in which case eliminate the human errors. Therefore, the automated music genre classification is one effective method and hot research area, which can easily group music by labelling them to appropriate genre. The current state of the art approach are focus on SVM, LSTM and CNN which consider as the time-series problem with the audio nature.

In this paper, we combined three types of features, timbral texture features, pitch feature and spectrogram features, which have been commonly used in speech recognition and music genre classification. Then we proposed a BiLSTM classifier, which is implemented in PyTorch, to realize the genre classification task. Then, we conduct a classification-accuracy analysis to search for the best feature combination and hyper-parameters based on the results of many experiments.

2. RELATED WORK

There are many active research in this topic. Hansi Yang and W. Zhang [1] proposed an approach to improve music genre classification with convolutional neural networks (CNN). They extracted mel-scale spectrogram as the input, then used duplicate convolutional layers whose output will be applied to different pooling layers to provide more statistical information for classification. This approach utilized the fact that music genres are conventional categories. Yingying Zhuang, elt. [2] designed a Transformer classifier to learn dependencies between distant positions in a sequence and in such way to improve accuracy on music genre classification. Despite the modelling, many researchers have found that the improve on feature extractions from music will also improved the classification task a lot in this topic. S. Oramas [3] has shown that the combination of multimodal data representations from music improved a lot. The experiments on single and multi-label genre classification are then carried out, evaluating the effect of the different learned representations and their combinations and results showed that the aggregation of learned representations from different modalities improves the accuracy of the genre classification. Yannis Panagakis [4] also proposed Joint Sparse Low-Rank Representation of Audio Features, smoothing noise in the test samples and identifying the subspaces that the test samples lie onto, to boost the accuracy on music genre classification.

Advisor Prof.Beigi

3. METHODS

3.1. Method Introduction

Our approach is based on related work previously stated. First, we extracted three type of features, timbral texture features, pitch feature and spectrogram features from librosa. Basically, oen value in spectral centroid, 13 values in MFCC, 12 values in chroma features and 7 values in spectral contrast. In addition, because Braz shows that SVM classifier also did a well work on music genre prediction [5], we also extracted features used in SVM, namely roll-off frequency , spectral flux, spectral bandwidth, flatness and zero-crossing rate of an audio time series. There are total 38 features we used to analysis and did experiments in our project. Then we tried different combination of those 38 features to choose best group based on the result feeding into our model.

We used two LSTM model for training, the first one is single directional 2-layer LSTM, this is the baseline for the project as someone has proved to be efficient in music genre classification. Another one is for experimental purpose, as we don't know if it will work before we start this project. It is Bidirectional LSTM model with adding max-pooling, mean-pooling, dense linear, activation function and dropout. Since this model is more complicated and time-consuming for training and tuning hyper-parameter and our time is limited for the semester, we will stop the experiments as the semester ends. Figure 1. shows our work flow for this project.

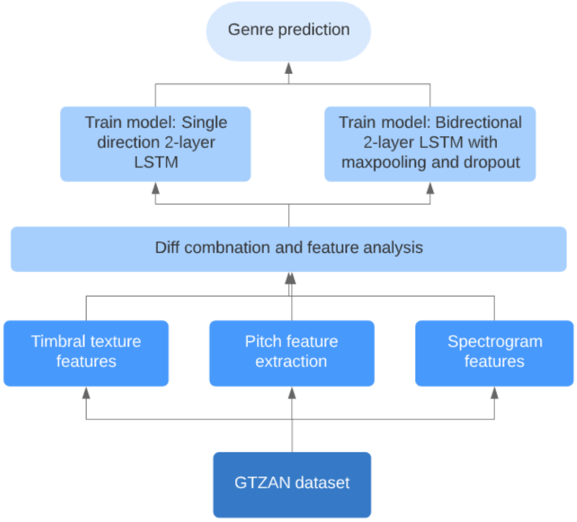
Genre prediction
Train model: Single direction 2-layer LSTM
Train model: Bidirectional 2-layer LSTM with maxpooling and dropout
Diff combnation and feature analysis
Timbral texture features
Pitch feature extraction
Spectrogram features
GTZAN dataset

Figure 1. Model Flowchart

3.2. Feature Analysis and Extraction

A well-generalized representation of features can help the all pattern recognition techniques fully absorb the information of our signal with a great learning processing. Therefore the feature extraction is a key step in this research experiment. In our project, we tried combination of 38 features. Here we choose several important and common feature to introduce.

3.2.1. Spectral Centroid

The spectral centroid is a measure used in digital signal processing to characterise a spectrum. It indicates where the center of mass of the spectrum is located.

$$Centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

where $x(n)$ is the weighted frequency value of bin number n, and $f(n)$ is the center frequency of that bin. Figure 2 shows how the value of spectral centroid based on frequency the energy of a spectrum.

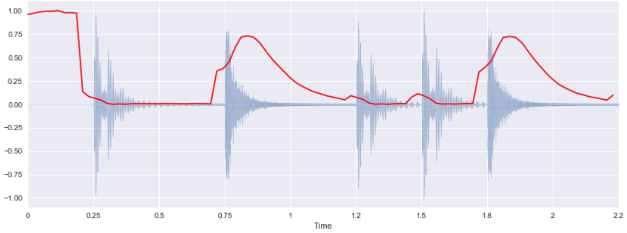

Figure 2.Example of Spectral Centroid from librosa

3.2.2. Mel-Frequency Cepstral Coefficients

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel-scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. MFCC computed by taking the log-amplitude of the magnitude spectrum and then, decorrelate the resulting feature vectors by a discrete cosine transform. Here, we still employ the 13 typical coefficients which are commonly used as speech and music representation.

3.2.3. Spectral Contrast

Spectral contrast considers the spectral peak, the spectral valley, and their difference in each frequency sub-band. It represented the relative spectral distribution instead of average spectral envelope. From the recent research, experiments showed that Octave-based Spectral Contrast feature performed well in music type classification [6]. Figure 3 shows example from librosa computing the spectral contrast for six sub-bands for each time frame.

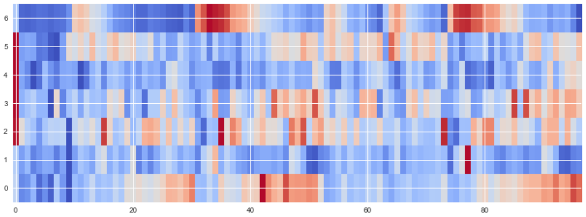

Figure 3. Example of Spectral Contrast from librosa

3.2.4. *STFT for chroma features*

Chroma features are also referred to as ”pitch class profiles”. Chroma features a powerful tool for analyzing music whose pitches can be meaningfully categorized and whose tuning approximates to the equal-tempered scale. Basically it will give 12 value vector to classify. One main property of chroma features is that they capture harmonic and melodic characteristics of music.

The Short-time Fourier transform (STFT), is a Fourier-related transform used to define the sinusoidal frequency and phase content of windows of a signal as it changes over time. In our project, we used continuous-time STFT, simply in the continuous-time case, the function to be transformed is multiplied by a window function which is nonzero for only a short period of time.

$$STFT(x(t))(\tau, w) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-iwt}dt$$

where $w(\tau)$ is the window function and $x(t)$ is the signal to be transformed. The function is a complex function representing the phase and magnitude of the signal over time and frequency.

3.2.5. *CQT for chroma features*

The problem of STFT is that the resolution of the windowing function in the standard STFT is the same for all values of the frequency. Further, the standard STFT has spaced frequencies equally because the exponent increases linearly. The Constant-Q transform improves in this problem. It modifies the STFT such that the frequency bins are logarithmically spaced. And also the windowing function is adjusted with center frequency so that the width of each bin too increases in proportion to the center frequency.

$$X(k) = \sum_{n=0}^{N(k)-1} x(n)W_k(n)e^{-j(\frac{2\pi Qn}{N(k)})}$$

where W_k defines a windowing function that indicates its dependence on k, the bin width increases with the frequency.

In summary, we learned a lot of useful features and for the experiments we will choose the combination with better result.

3.3. Neural Network Architecture

3.3.1. *Long Short-Term Memory*

To learn sequential data, the most appropriate type of neural network is the RNN and long short-term memory (LSTM) neural network has been proved as a special Recurrent Neural Network which solves the vanishing or exploding gradient problem[7]. Recurrent Neural Network, in contrast with traditional dense neural network, will remember previous state and information, with different weights on the all previous information, it has access to some prior knowledge about the data to completely understand it.

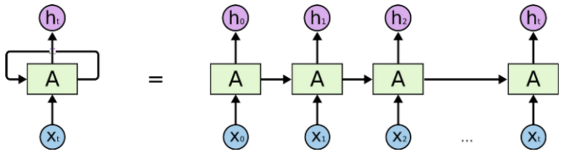

Figure 4. Recurrent Neural Network

Especially, as shown in Figure 5, the three gates: forget gate, input gate and output gate are created inside LSTM cell, which can help the neural network efficiently capture the important information of current input and at the same time, add information from previous time as well as get rid of the redundant information. We believe that LSTM will successfully capture the musical characteristics based on its architecture, which has been widely used in various applications such as natural language translation, image captioning and speech recognition. One model we tried is the simple 2-layer LSTM, which is the baseline of the project.

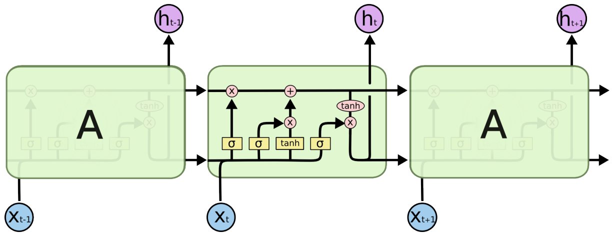

Figure 5. LSTM cell

3.3.2. *BiLSTM*

Another model is for experimental purpose, we used Bidirectional LSTM(BiLSTM) and adding other activation function and drop out, trying to find a better classification model. Bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification problems. The bidirectional, as showin in Figure 6, is just to put two network together with forward and back word ordering, in which case the network will learn the information from both forward and backward directions. This structure allows the networks to have both backward and forward information about the sequence at every time step. Bidirectional network will run inputs in two ways, one from past to future and one from future to past. the advantage of bidirectional LSTM is that it preserves information from the future and using the two hidden states combined, which will optimize result based on both sides information. We believe the music feature will be also different in forward and backward ordering, in which case to learn more information in both directions might give a better understanding to the model.

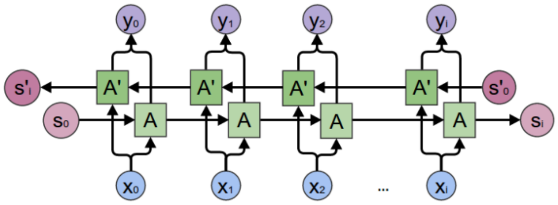

Figure 6. Bidirectional LSTM

4. DATASETS

We used GTZAN dataset, which contains 1000 music files in .wav format for ten genres, with around 100 files belonging to each genre. Every audio file is 30 seconds long within one of the ten genres: Hip-Hop, Rock, Reggae, Classical, Jazz, Blues, Pop, Disco, Country, and Metal. The GTZAN dataset is the most-used public dataset for evaluation in machine listening research for music genre recognition (MGR). The files were collected in 2000-2001 from a variety of sources including personal CDs, radio, microphone recordings, in order to represent a variety of recording conditions. In order to compare with other research results, we think this dataset is good to have a understanding how good a model is by comparing with the benchmark. To get a better sense of musics in different genre, we plot the waveplots(in Figure 7) for 10 example from each genre in GTZAN dataset.

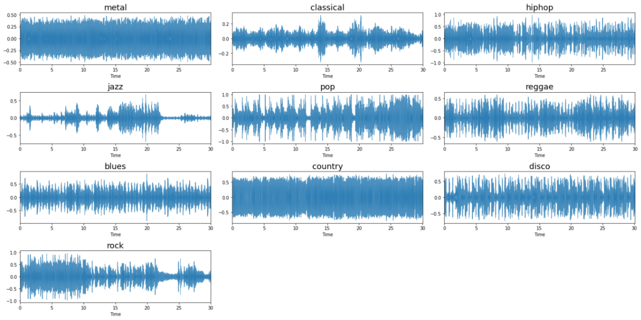

Figure 7. Waveplots for 10 genres

As we can see, some genres are very similar to each other such as disco, blues and pop. In addition, we also plot the spectrogram plots for the same 10 genre examples.

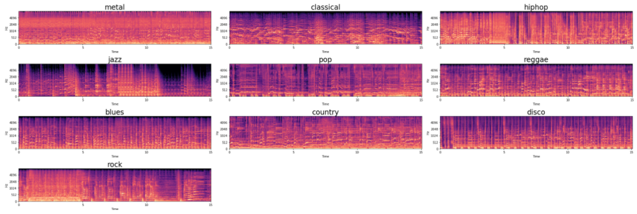

Figure 8. Spectrogram for 10 genres

From the spectrogram of MFCC, it also shows that country, blues and disco are very similar to pop and rock. From recent research[8], Senac has noticed that rock, blues, disco and country are problem to classifier and will be very noisy to the model. By considering that, we did experiments in both 10 genres and 6 more distinguish genres, namely Classical, Hip-Pop, Metal, Jazz, Pop and Reggae.

5. EXPERIMENT DETAIL

5.1. Computation Requirement

The BiLSTM and LSTM classifiers are implemented and evaluted. Both of them are trained and tested on personal computer, with Intel core i7 2.6GHz CPU and 16Gb of memory. Code for development of the LSTM and BiLSTM classifiers are written by using PyTorch module in Python 3. The feature extraction and pre-processing, post-processing are implemented in Python3. The whole training time is around 10 hours to finish all the models for all feature groups.

5.2. Pre-processing

Firstly the preprocessing is made for GTZAN data and split them into training, testing and validation set as ration 80:10:10. There are 100 songs in each genre and 1000 songs in total. Each song is 30 seconds duration. We first split whole datasets and make sure each genre appears in training, testing and validation set equally. Then by choosing six distinguish genres as we mentioned before, Classical, Hip-Pop, Metal, Jazz, Pop and Reggae, we split them into another training, testing and validation sets as same criteria. After split them, we need a fix length of time-series feature. So a process has been made to compute To find a minimum length over all audios so that we can get a fixed length for all of them, the length setting is 128, which is the number of Mel bins in the feature extraction. So for each audio data, we have (number of features, 128) vector to represent.

5.3. Feature Combination

As we mentioned before, we choose 38 features extracted from audio. They are Spectral Centroid, Spectral Roll-off, Zero-Crossing Rate of an aduio, Spectral Flux, n'th-order Spectral Bandwidth, Spectral Flatness. Those are features haven been proved to be useful in Support Vector Machine model in music genre classification. Also, we extracted Spectral Contrast(7 values), MFCC(13 values) and Chroma feature(12 values) by both STFT and CQT. We used following combinations of features:

(1) All 38 features.

(2) 33 features, including MFCC(13 values), Chroma feature(12 values) and Spectral Contrast(7 values).

(2) 32 features, including MFCC(13 values), Chroma feature(12 values) and Spectral Contrast(7 values) and Spectral Centroid.

Above are groups of features with using STFT to generate chroma features. Another three combinations with same structure but instead using CQT to get Chroma features.

5.4. Models and Parameter Tuning

There are two models we used in experiments. The first one is basically simple 2-layers LSTM, shown as Figure 9, we used one dense linear and log soft-max to get scores for predicting genres.

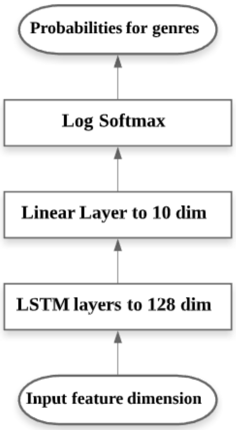

Figure 9. Architecture for LSTM model

The second one is 2 layer bidirectional LSTM, shown as Figure 10, we added max pooling and mean pooling for the output from Bidirectional LSTM, and then concatenate them together. Because the output from bidirectional LSTM is doubled, it was 256 dimensions. And then we used one linear layer one dense linear to conver it to output dimension, then use Relu activation and Dropout with 0.2 value to randomly drop 20% nodes for reducing over-fitting in the neural networks. And then with same log soft-max function to normalize probabilities.

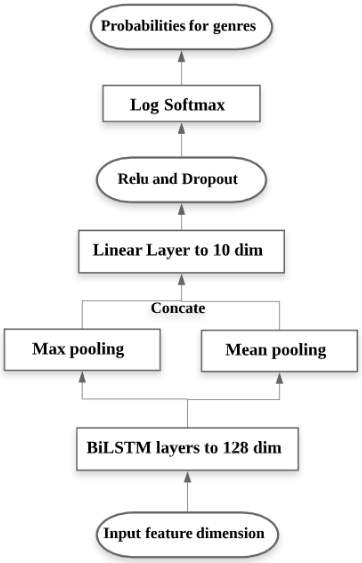

Figure 10. Architecture for BiLSTM model

For both models, Cross Entropy Loss function is used with 0.01 learning rate, and Adam is used as an optimizer to train the model. We track the loss and accuracy for validation set in each training epoch and preserve the best model during the process, in which case we prevent to use last one model that could be over-fitting to the training set. For both models, we run 300 epochs with 30 batches, because BiLSTM is efficient to converge, we ended to run 150 epochs on BiLSTM.

6. PRELIMINARY RESULTS

Since we tried six different groups of features and used two models, we will not present all of results. Here we just show the best result we got from those experiments. The full results will be included in the code submission.

After completed a training process over the training dataset, we achieved the historical training and validation accuracy, which are plot in Figure 11. From plot we can see that the model is very efficient to converge, around 120 epochs it converges. The loss of validation set goes up after 120 epochs because the model is over-fitting to the training set, that's why we won't use the final model, we track the models with lowest loss and highest accuracy on validation set.

We have two criteria to choose best model to save during training, when the loss of validation set is lower than before and when the accuracy of validation set is better than before. Figure 12. shows the loss plot over epochs. By comparing two plots, if we use criteria by the accuracy of validation set,

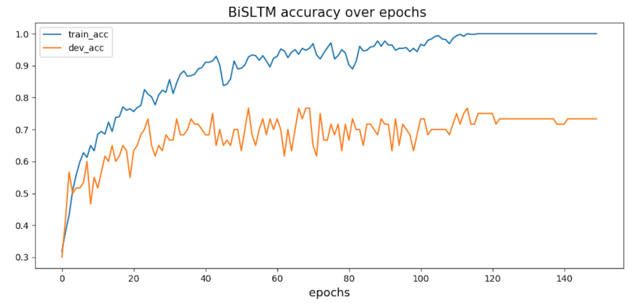

Figure 11. Accuracy over epochs from BiLSTM model

the best model is around 110 epochs. If if we use criteria by the loss of validation set, the best model is saved around 30 epochs. We think this is because the training and validation sets are both small and the information to learn is limit for the model. If we can use larger datasets, it might be better on the loss of validation set. So we keep the model around 120 epochs, by the criteria of accuracy for this specific datasets.

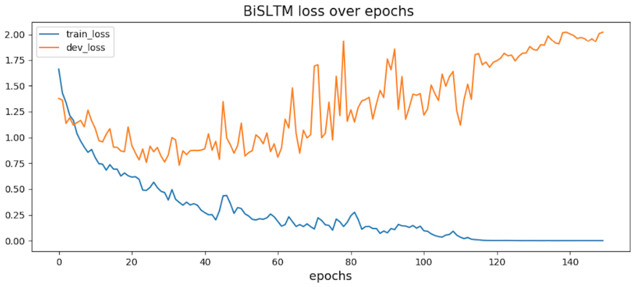

Figure 12. Loss over epochs from BiLSTM model

In addition, we explored the classification accuracy of individual genres by the BiLSTM classifier with the different feature combinations we used before. The overall results are visualized by a histogram plot, as shown in Fig. 13. It can be stated evidently that for our music genre classification study, the 33 feature vector with CQT method is the best one for training BiLSTM model.

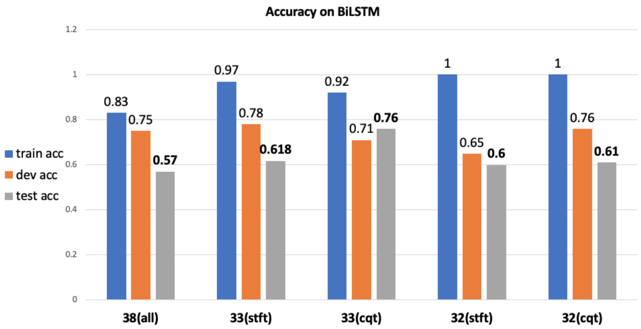

Figure 13. Accuracy on BiLSTM

Comparing to the results on LSTM, as Figure 14, overall BiLSTM performs well on each feature selections.

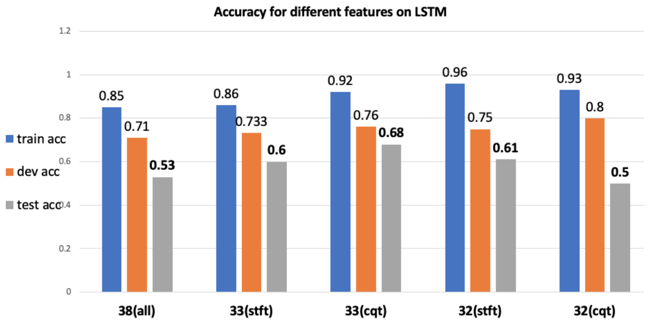

Figure 14. Accuracy on LSTM

Furthermore, the classification improvement happens at the Classical, Pop and Metal genres. It is probably due to the extensive use of musical instruments and beats in the audio, which contains many higher and lower frequency bins. In this situation, CQT can help to reduce the resolution for high frequencies leading to the better pitch representation for the pitch-sensitivity musical genres such as Classical and Pop. And also the Octave-based Spectral Contrast feature considers the strength of spectral peaks and spectral valleys in each sub-band separately, so that it could represent the relative spectral characteristics, and then roughly reflect the distribution of harmonic and non-harmonic components. [6] Therefore, the classifier could perform better when predicting these those genres.

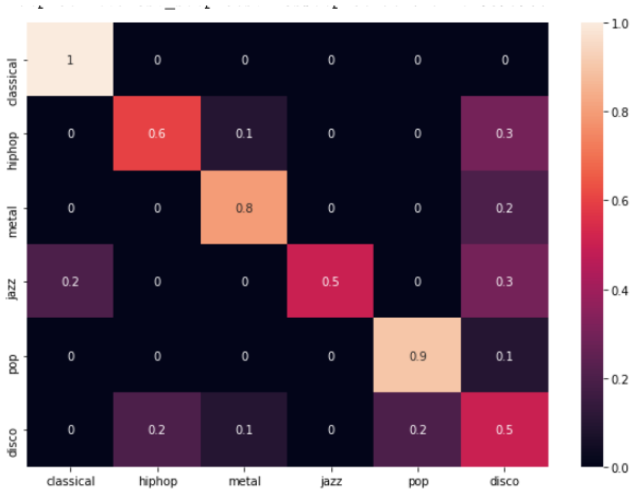

Figure 14. Confusion Matrix from prediction

7. CONCLUSION

To concludes, this project studies and presents a BiLSTM model predicting music genres and comparing with simple LSTM with different feature combinations. By considering timbre and pitch features, as well as the nature of musics, in which case the pitch feature and harmonic components are more significant than voice audio, we showed that spectral contrast is also a important feature in this topic. Also there are some aspects need to be improved in the future work, the dataset is kind of small for the neural network, only 1000 audios in total and only 30 seconds in each audio. Bob Sturm points out that the GTZAN dataset has some bias and drawback[9], so for the future potential work, it could be to use larger and diverse datasets, and try more other networks such as CNN and Autoencoder.

8. REFERENCES

[1] Hansi Yang and W. Zhang, "Music genre classification using duplicated convolutional layers in neural networks," in *INTERSPEECH*, 2019.

[2] Yingying Zhuang, Yuezhang Chen, and Jie Zheng, "Music genre classification with transformer classifier," *Proceedings of the 2020 4th International Conference on Digital Signal Processing*, 2020.

[3] S. Oramas, Francesco Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," *Trans. Int. Soc. Music. Inf. Retr.*, vol. 1, pp. 4–21, 2018.

[4] Yannis Panagakis, Constantine Kotropoulos, and G. Arce, "Music genre classification via joint sparse low-rank representation of audio features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1905–1917, 2014.

[5] Carlos N. Silla Jr., Alessandro L. Koerich, and Celso A. A. Kaestner, "A machine learning approach to automatic music genre classification," *Journal of the Brazilian Computer Society*, vol. 14, pp. 7 – 18, 09 2008.

[6] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai, "Music type classification by spectral contrast feature," in *Proceedings. IEEE International Conference on Multimedia and Expo*, 2002, vol. 1, pp. 113–116 vol.1.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[8] Christine Senac, Thomas Pellegrini, Florian Mouret, and Julien Pinquier, "Music feature maps with convolutional neural networks for music genre classification," 06 2017, pp. 1–5.

[9] Bob L. Sturm, "The state of the art ten years after a state of the art: Future research in music information retrieval," *Journal of New Music Research*, vol. 43, no. 2, pp. 147–172, Apr 2014.