Music Genre Identification

Cathy Guang, Katrina Li

Introduction

Music classification has many potential applications. Not only can music be managed more effectively, but it also enables the user to search the genre of music they are interested in. It is a milestone for potential future studies such as classification of mood, instrumentation or styles. Music genre recognition is a non-trivial issue because the boundaries between genres are fuzzy. In this project, we trained a deep learning model to recognize the genre of a piece of music, in one of the ten genres listed below: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae and rock. We achieved reasonable results by feeding the music input in the form of mel frequency cepstral coefficients (MFCC) into a 2D convolutional neural network (CNN).

Data

We downloaded the GTZAN dataset from kaggle:

<https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification>

The GTZAN dataset is the most-used public dataset for evaluation in research for music genre recognition. The raw data contains a thousand sound files, evenly split into ten genres, all having a length of 30 seconds. The files were collected from a variety of sources including personal CDs, microphone recordings, radio, to better represent different recording conditions. We used the librosa library to analyze the characteristics of those sound files. Mel Frequency Cepstral Coefficients (MFCCs) were used to represent features of each audio file.

Methods

1. Small Scale Neural Network

As an initial attempt in simple neural network.ipynb, we extracted 40 mel frequency cepstral coefficients (MFCCs) for each audio file, took them as input X, and fed them into a fully-connected neural network. The resulting accuracy was low, which was only around 0.2, a bit more than random guessing.

In order to achieve an improved result, we extracted features from each audio file again by dividing it into segments, so that the input is of shapes: (sample, timesteps, mfcc features). Instead of fitting inputs into the small model, we proceed to use Convolutional Neural Network.

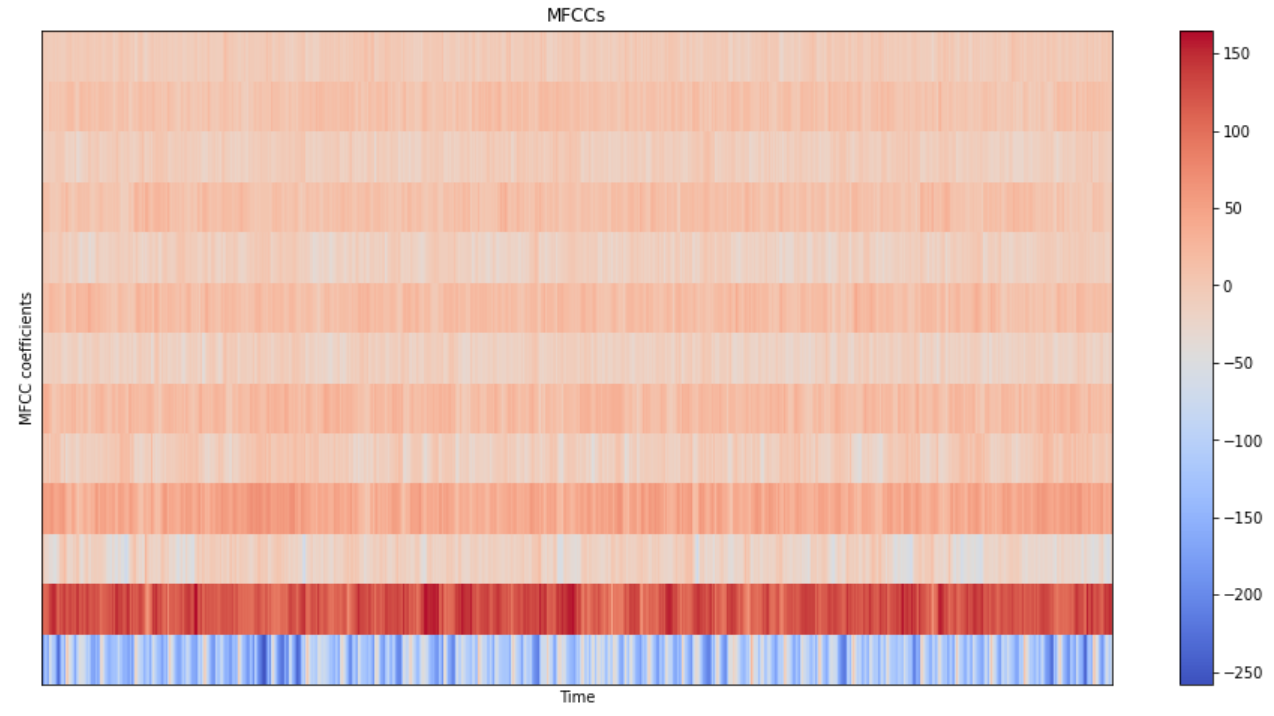


Fig.1 MFCCs for an example file. 13 stripes in this graph represent 13 MFCCs

1. 2D Convolutional Neural Network

In extract\_features.ipynb, we splitted each audio file into 10 segments and extracted 13 MFCCs (mel frequency cepstral coefficients) for each of those segments. For each segment, we distributed a label that represents the genre of the music. We created a three-dimensional array to store the label and MFCCs and dumped the array to a json file to be used for training.

In CNN.ipynb, we loaded the data from our json file, set our X (input) to be the MFCCs and Y (output) to be the label. We set X to be of shapes: (sample, timesteps, mfcc features), Timesteps is the number of time windows, calculated by the number of samples divided by hop length, each of which indicating the MFCC value in the particular time frame. For each segment we assigned a Y value that represents the genre/label of the segment. Then we did a train test split for our data, we put 60% of the data into training, 15% into validation and 25% into testing

We added several layers in the convolutional neural network. We set the hidden layers to be Rectified Linear Unit activation functions (ReLU) because it is simple to implement and effective, and specifically, it is less susceptible to vanishing gradients that prevent deep models from being trained. For the output layer we use the softmax activation function because it is a multiple classification problem. Then we use he\_normal kernel initialization because it was developed specifically for nodes and layers that use ReLU activation as well as a standard approach for it.

We set the learning rate to be 0.01, loss to be ‘sparse\_categorical\_crossentropy’ and metrics to be ‘accuracy’. We also set up an early stopping for the training. The accuracy in the last epoch was 0.88, which was a much better result than our previous fully connected neural network.

Evaluation and Results

The accuracy for the test data was 0.66. We drew the confusion matrix shown in Fig.2. The numbers 0-9 in our matrix represent the genres 0-‘disco’, 1-‘jazz’, 2-‘country’, 3-‘metal’, 4-‘blues’, 5-‘reggae’, 6-‘rock’, 7-‘hiphop’, 8-‘pop’, 9-‘classical’ respectively. We believed it was reasonable that our model confused more between ‘blues’, ‘reggae’, and ‘rock’, but recognized ‘pop’ and ‘classical’ quite well. Humans can also quickly distinguish between classical, pop, with other music genres, but tend to confuse blues, rock, and reggae.

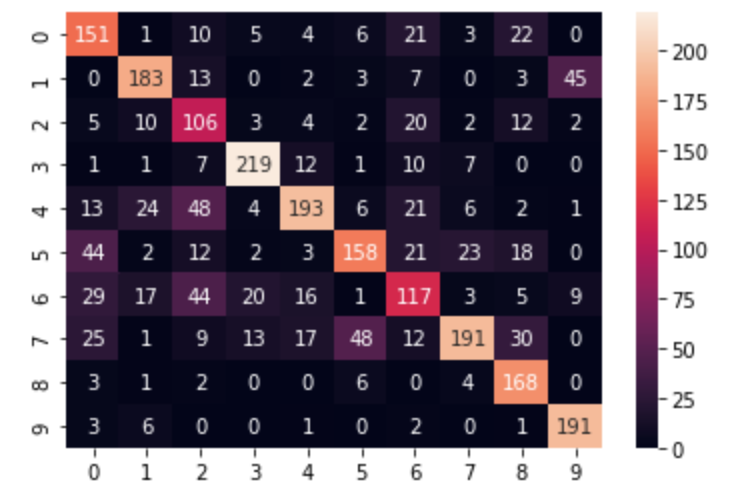


Fig.2 Confusion matrix of the model prediction.

Discussion

Audio can itself be represented as an image (Spectrogram/MFCC). For example in spectrograms, time and frequency are represented as x,y and amplitude as pixel value. We chose CNN model because it is of high accuracy for image classification and recognition. It can recognize textures and timbre in our music files efficiently. For future studies, it will be helpful to compare the performance of other models such as LSTM.

Previous studies had proposed different models and feature extraction methods for music classification. These include training based on Short-time Fourier Transform (STFT), Wavelet Transform (WT), and MFCCs, and some additional parameters were used to obtain feature vectors. The best overall accuracy (74.2%) was achieved when the test was carried out using Wavelet Coefficient Histograms (DWCHs) based on One-Against-All (OAA) approach (Li et al.), trained with middle parts of songs. Instead of using timbral textural and rhythmic content features, it is also worth researching into using different combinations and of different segments of the music.

Reference:

Youtube channel-Valerio Velardo - The Sound of AI

<https://www.youtube.com/channel/UCZPFjMe1uRSirmSpznqvJfQ>

Why we choose ReLU for hidden layers

[How to Choose an Activation Function for Deep Learning (machinelearningmastery.com)](https://machinelearningmastery.com/choose-an-activation-function-for-deep-learning/)

Why use he\_normal for ReLU

[Weight Initialization for Deep Learning Neural Networks (machinelearningmastery.com)](https://machinelearningmastery.com/weight-initialization-for-deep-learning-neural-networks/)

Li T, Ogihara M, Li Q. A Comparative study on content-based music genre classification. In: Proceedings of the 26th annual international ACM SI-GIR conference on research and development in information retrieval. Toronto: ACM Press; 2003. p. 282–9.