Music Genre Classification

Introduction

Music genre classification is a challenging task in the area of music information retrieval (MIR). In theory, there could be more than 50 music genre nowadays and some are derived from the traditional stream. In this project, we are going to start with classifying the more popular and well-known genre. We will try modelling with various classifiers and compare their usage along with the pros and cons.

By classifying different genre of music, we could collaborate with entertainment companies and music distributors to pick the most suitable songs to their customers. The potential development could also be cooperating with social scientist to see how different kind of music reacts with people in different emotional situation.

Dataset

In our project, we have used a subset of the dataset from Free Music Archive (FMA). A total of 41409 songs with 6 different genres are selected and the 6 genres are: Electronic, Experimental, Folk, Hip-hop, International and Rock. The distribution of our dataset can be referred to the table below. Full track was used for feature generation which we would further discuss in the next section.

Genres	No. of Tracks
Electronic	9372
Experimental	10119
Folk	2803
Hip-Hop	3552
International	1389
Rock	14174

Table 1. Number of tracks in genres chosen for classification in FMA

Features

In order to deal with audio records, we pre-computed some features listed in the table below. These features can be extracted by using the Python library called **Librosa**. Each feature set (except zero-crossing rate) is calculated on a FFT window with size 2048 and hoop length 512, with seven statistics measures were computed over all windows. These measures are mean, standard deviation, skew, kurtosis, median, minimum and maximum. Next, we are going to discuss each of the features in more details.

Features	Description	Dimensions
Chroma_cens	Chroma Energy Normalized	84
Chroma_cqt	Constant - Q Chromagram	84
Chroma_stft	Short Time Fourier Transform Chromagram	84
MFCC	Mel-Frequency Cepstral Coefficient	140
Spectral_bandwith	Spectral Bandwidth	7
Spectral_centroid	Spectral Centroid	7
Spectral_constrast	Spectral Contrast	49
Spectral_rolloff	Spectral roll-off	7
Tonnetz	Tone Network	42
zcr	Zero Cross Rate	7
rmse	RMS Energy	7
	Total Dimension :	518

Table 2. Features on genre classification and their dimensionality

Chroma Feature Analysis

Chroma features of a music audio is the feature in which the entire spectrum is projected onto 12 notes in western music (**C, C$^{\#}$, D, D$^{\#}$, E, F, F$^{\#}$, G, G$^{\#}$, A, A$^{\#}$, B**). The chroma representation shows us the intensity of each musical tone at each time frame. We can turn this back into an audio signal by transforming the 12 chroma values to 12 modulated sinusoids, which then turned to cover one octave.

1. Short Time Fourier Transform (Convert Sound into Image)

The Short-Time Fourier Transform (STFT) is a general tool for audio processing. It is used for determining the sinusoidal frequency and phase information of local sections of time. In the discrete time series, the audio data is broken up into chunks of frames.

$$X_m(\omega) = \sum_{n=-\infty}^{\infty} x(n)w(n-mR)e^{-j\omega n} = \mathrm{DTFT}_{\omega}(\boldsymbol{x} \cdot \mathrm{SHIFT}_{\mathrm{mR}}(\boldsymbol{w}))$$

$x(n) = \text{input signal at time n}$
$w(n) = \text{length M window function}$
$X_m(\omega) = \text{DTFT of windowed data centered about time mR}$
$R = \text{hop size between successive DTFTs}$

$$\text{, where} \sum_{n=-\infty}^{\infty} w(n-mR) = 1, \forall n \in \mathbb{Z}$$

The graphs below indicate the STFT of one of the audio track [190.mp3] - **Castle of Stars**. The larger sized window allows the frequencies to be precisely seen.

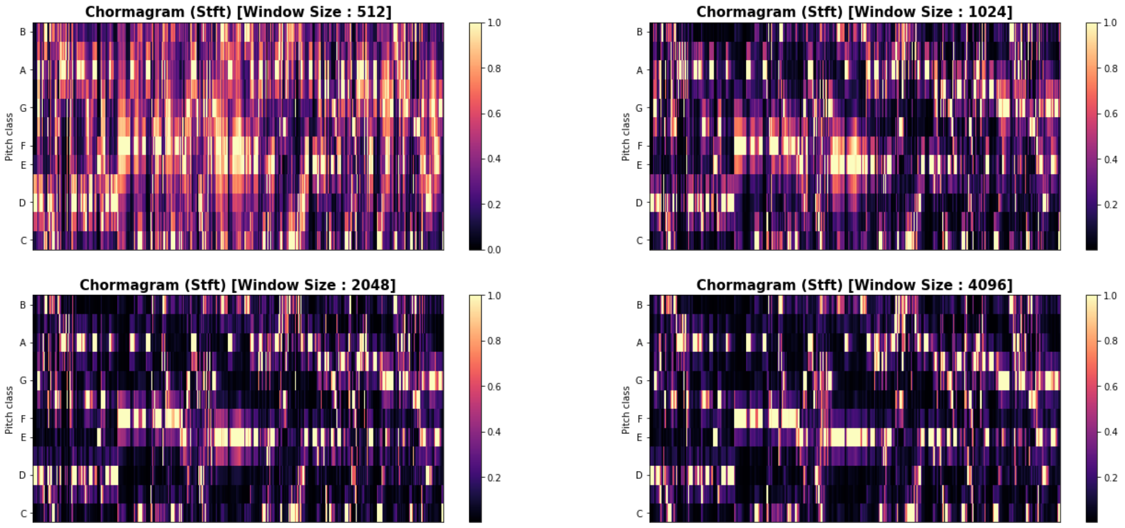


Figure 1. Chormagram of different window sizes

2. Constant-Q Transform

The Constant-Q Transform is closely related to the Fourier transform. In contrast to the Fourier Transform, the transform is a series of logarithmically spaced filters.In general, the transform is well suited to musical data. It has been seen in its advantages compared with Fast Fourier Transform. The output is amplitude/phase against log frequency, fewer frequency bins are required to cover a given range effectively. As the range of human hearing covers approximately ten octaves from 20 Hz to 20 kHz, the reduction of output data is significant.

Mel Frequency Cepstral coefficient (MFCC)

MFCC is commonly used as a feature extraction technique in audio recognition and genre classification in nowadays. Here is the process of computing MFCC:

1) Pass the audio signal and increase the energy of high frequency
   There is more energy in low frequency compare to high frequency. By doing this, we would get more information in high frequency. Then we split the audio signal into overlapping frames and multiple them by a hamming window.
   The frame size is usually between 10 and 25 ms and frame shift every 5-10 ms.

2) Perform discrete fourier transformation
   In order to get the frequency components, we would perform discrete fourier transform in each time frame. With the frequency we got, we would map it to Mel scale and take log of each Mel frequencies.
   Mel scale is approximately linear below 1 kHz and logarithmic above 1 kHz which is similar to human's hearing. The reason is because not all the frequency is equally sensitive to human and human tends to be less sensitive to frequency >1kHz.

3) Take discrete cosine transformation
   This final step gave us the MFCCs.

The process flow can be represented as following:

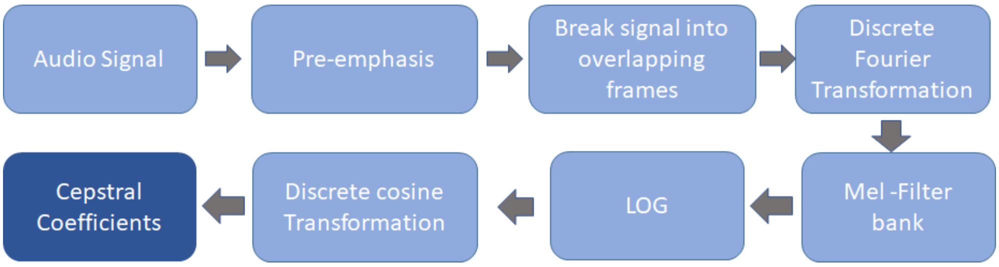

Audio Signal
Pre-emphasis
Break signal into overlapping frames
Discrete Fourier Transformation
Mel -Filter bank
LOG
Discrete cosine Transformation
Cepstral Coefficients

Figure 2. Flow diagram of MFCC

Spectral Bandwidth

It could represent the by the following formula for computing the order ***p*** spectral bandwidth.

$$\left(\sum_{k} S(k)(f(k)-f_c)^p\right)^{\frac{1}{p}} \text{ ,where } \left\{ \begin{array}{r} S(k): spectral\ magnitude\ at\ frequency\ bin\ \boldsymbol{k} \\ f(k): frequency\ at\ bin\ \boldsymbol{k} \\ f_c: spectral\ centroid \end{array} \right.$$

Spectral centroid

Spectral centroid is typically used as a measurement of the "brightness" of sound. It is like the weighted average, where F[k] is the amplitude of bin k in discrete Fourier transformation spectrum.

$$Spectral\ Centroid = \frac{\sum_{k=1}^{N} k \cdot F[k]}{\sum_{k=1}^{N} F[k]}$$

Spectral Contrast

Octave-based spectral contrasts considers the spectral peak, spectral valley and their differences in each sub-band. In contrast to MFCC, where the sums fourier transform amplitudes is considered, which loses the relative spectral information. Spectral peaks and valleys are estimated using:

$$Peak_k = \log(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k,i})$$

$$Valley_k = \log(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k,N-i+1})$$

And their difference is given by : $SC_k = Peak_k - Valley_k$

α- neighborhood factor
k - th sub-band
N - size of fourier transform vector

Spectral Rolloff

The spectral roll-off measures where the frequency below a specified percentage of the total spectral energy concentrated, by default the percentage is 85%.

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n]$$

Tonnetz

Tonnetz is a lattice diagram representing tonal space which is first described by Leonhard Euler in 1739. The visual representation of the Tonnetz can be used to show harmonic relationships in a music. Chords become geometric structures on the plane. By introducing Enharmonic and Octave Equivalence, one can then reduces the set of all notes to 12 pitch classes and wraps the plane into a hypertorus which has been described in neo-Riemannian theory.

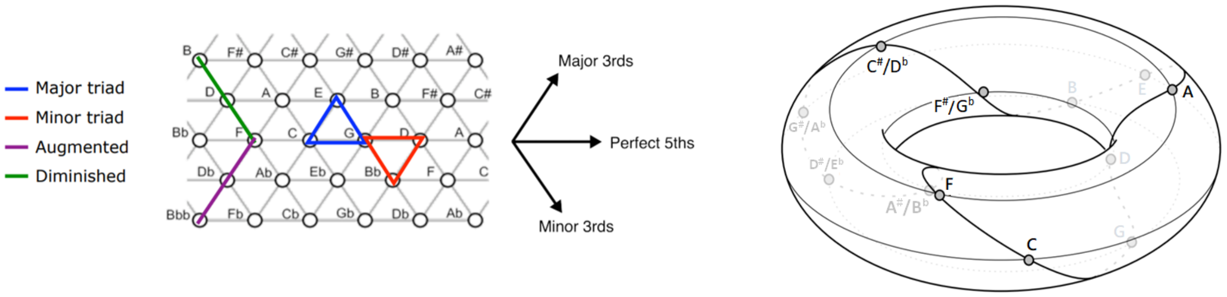


Figure 3. Tonnetz lattice diagram

The 6D interior space of the hypertorus can be represented as three 2D circles of Perfect 5ths, Major 3rds and Minor 3rds. Chords can be described by their 6D centroids in this space (Harte and Gasser, 2006).
Harte and Gasser's tonal centroid of a chroma vector can be computed as:

$$\zeta_n(d) = \frac{1}{\|\boldsymbol{c_n}\|_1}\sum_{l=0}^{11} \boldsymbol{\Phi}(d,l)\boldsymbol{c_n}(l) \quad ,\text{where } 0 \le d \le 5$$

where $\boldsymbol{\Phi} = [\boldsymbol{\phi_0}, \boldsymbol{\phi_1}, \ldots, \boldsymbol{\phi_{11}}]$

$$\boldsymbol{\phi_l} = \begin{bmatrix} r_1 sin(l\frac{7\pi}{6}) \\ r_1 cos(l\frac{7\pi}{6}) \\ r_2 sin(l\frac{7\pi}{6}) \\ r_2 cos(l\frac{7\pi}{6}) \\ r_3 sin(l\frac{7\pi}{6}) \\ r_3 cos(l\frac{7\pi}{6}) \end{bmatrix} ,0 \le l \le 11$$

Zero Crossing Rate (ZCR)

Zero crossing rate measures where the signal has changed from positive to negative or negative to positive.

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{\mathbb{R}_{<0}}(s_t s_{t-1})$$

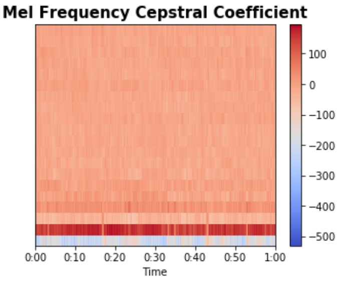
where S is the signal of length T and $1R_{<0}$ is the indicator function

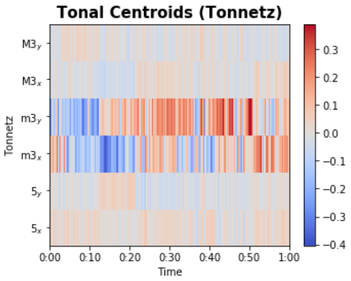
RMS Energy (RMSE)

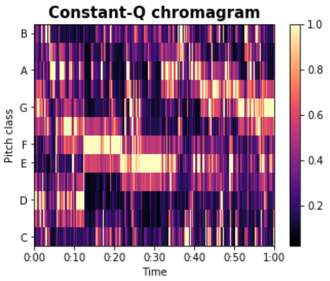
The energy of a signal corresponds to the total magnitude to the signal, and for audio, it is roughly equivalent to how loud the signal is. It is computed for each frame and is defined as $\sum_n \quad |x(n)|^2$and the Root-Mean-Square Energy (RMSE) is defined as $\sqrt{\frac{1}{N}\sum_n \quad |x(n)|^2}$.

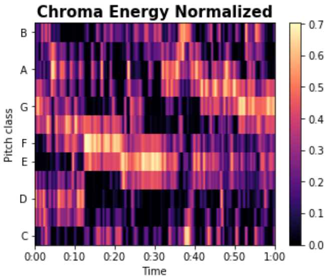
Example Illustration

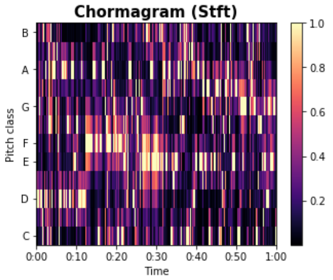
For each raw audio file in the dataset, the following 10 spectral features (except zero-crossing rate) can be extracted from the **librosa** library. The figures below visualize them in a spectrogram.

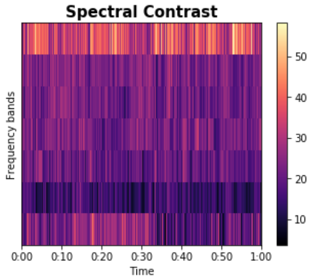

Mel Frequency Cepstral Coefficient

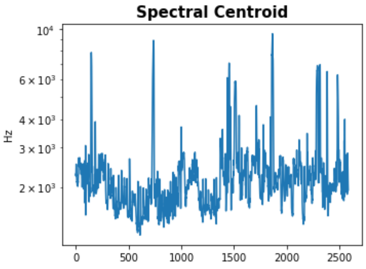

Tonal Centroids (Tonnetz)

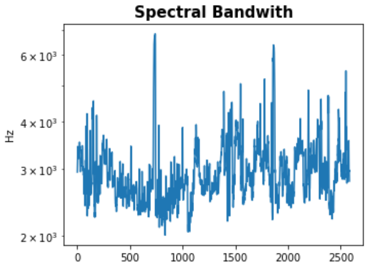

Constant-Q chromagram

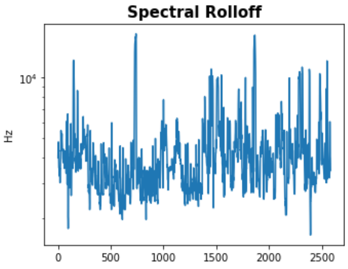

Chroma Energy Normalized


Chormagram (Stft)


Spectral Contrast


Spectral Centroid


Spectral Bandwith


Spectral Rolloff

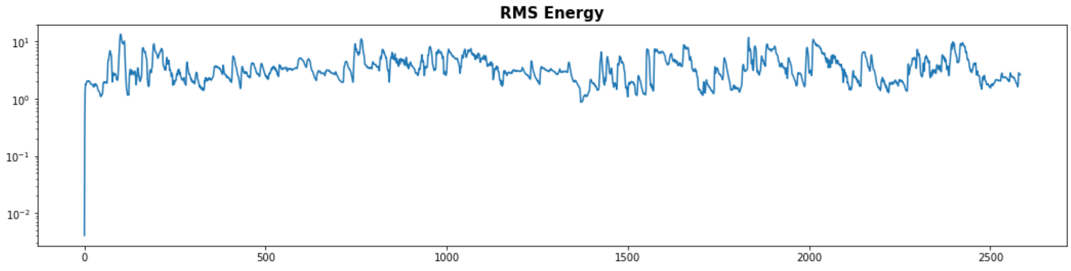

RMS Energy

Figure 4. Librosa library diagram output of FMA dataset

Dimensionality Reduction

1. PCA

There are over 300 features in our training and many of them are co-related. We have applied PCA to reduce it's dimensionality so to reduce a large set of feature variables into a smaller set but still contain most of the information. It is a mathematical transformation that extract the orthogonal component from a large feature variables so the resulting features would be as unrelated as possible. It could be used to speed up the training.

2. Automatic choice for dimensionality

Choosing dimensionality to apply PCA on is a frequent issue that occur when we choose to use PCA. Minka[7] has proposed solution for a probabilistic view to automatically choose the dimensionality of PCA using maximum likelihood estimation, assuming the noise and the principal component are spherical Gaussian.

Classifiers

1. Naive Bayes Classifier (GaussianNB)

Dealing with continuous data, a common assumption is that the continuous values associated with each class are distributed to a Gaussian distribution. This classifier can be employed to our problem without the assumption as the spectral features of the audio are all continuous. For the satisfaction of the conditional independence assumption, we choose those features highly unrelated to each other. Then, the probability distribution can be computed in this way:

$$p(\boldsymbol{x}|y = c\,,\boldsymbol{\theta}) = \prod_{j=1}^{D} \mathcal{N}(x_j|\mu_{jc}\,,\sigma_{jc}^2)$$

2. K-Nearest Neighbors Classifier (KNN)

KNN is a non-parametric method in this classification problem. The prediction is very simple that the training phase is fast. The algorithm is based on the measure of feature similarity (usually euclidean metric). An object is classified by a majority vote of its neighbors, with the object being assigned to the most common class among its k nearest neighbors (k is a positive integer, k must not be a multiple of class number).

3. Random Forest Classifier

Random Forest is a ensemble learning method that construct multiple trees in training. The algorithm consists of two parts. In the first part, each decision tree is trained with a random sample of the dataset which is known as bootstrap aggregation. For each decision tree, a random subset of features is using for making prediction. The final prediction of the classifier is based on the majority vote from the each decision tree classifiers.

4. Gaussian Discriminant Analysis (or Quadratic Discriminant Analysis)

Gaussian Discriminant Analysis is a generative learning algorithm used for classification. The fundamental assumption in this algorithm is that each class conditional densities are Gaussian such that

$$p(\boldsymbol{x}|y = c\,,\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_c}\,,\boldsymbol{\Sigma_c}) \quad \text{, where } \boldsymbol{\theta} = \{\boldsymbol{\mu_c}\,,\boldsymbol{\Sigma_c} \mid c = 1, \dots\dots, C\}$$

As the final decision boundary is generally parabolic, the GDA is also called as **quadratic discriminant analysis.**

5. AdaBoost Classifier

Adaboost, similar to Random Forest Classifier is used to boost the performance with a set of weka learners. The output the weak learners is combined into a weighted sum as the final output of the boosted classifier. The boosting method can achieve similar classification results with much less tweaking of parameters or setting. One only needs to choose which weak classifier to solve the given classification problem and the number of boosting iteration being used during the training phase. The disadvantage is that it can be sensitive to noise and outliers which make it to be less susceptible to the overfitting problem.

6. Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

7. Multi-Class Neural Networks

A neural network model is a major role in deep learning machine which is inspired by the structure and function in the brain cell. Based on our work in the feature extraction, we input the resulting feature vectors to a feedforward network and output with a softmax function. The network is trained under the categorical cross entropy.

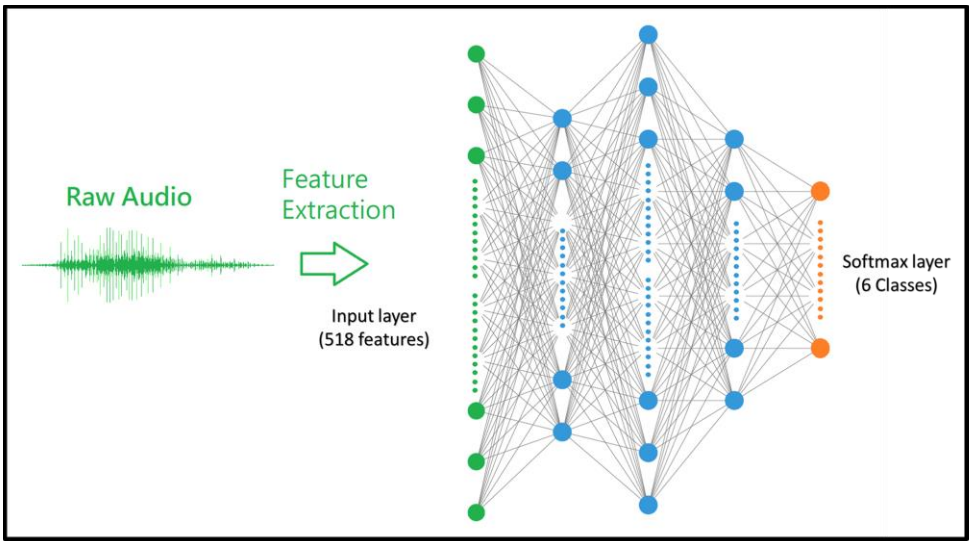


Figure 2. Feed Forward Neural Network setup with 4 layers

Features Selection

By making use of the random forest classifier, we can analyse which features are the most important by using the function *featuer_importance_*, below are the top 10 features which contribute 80% of the accuracy already.

Top 10 features
spectral_centroid_median_1
spectral_centroid_mean_1
spectral_rolloff_median_1
mfcc_median_1
mfcc_max_4
spectral_contrast_mean_4
mfcc_median_3
spectral_rolloff_mean_1
rmse_std_1
mfcc_mean_3

Table 3. Top ten features used in the classification model

Given there are around 500 features, if we make use of 10% of the most important features, we could already achieve 95% of the accuracy obtained. We are satisfied with using fewer dimensions for efficiency purpose.

In light of increasing the accuracy as much as possible, we are going to apply all features to the model training. We will leave the features selection part to our future work if we collect more data points.

Result

We have benchmarked 9 classifiers and they give the following result. Neural network has achieved the best accuracy at 72% and the second one being Random Forest with 128 estimators at 67%. Some models such as Naive Nayses give a relatively low result at 42% since the conditional independence of feature is not a good assumption in our case.

Classifier	Accuracy
Gaussian Discriminant Analysis	29%
Naive Bayes Classifier	42%
KNeighbors Classifier	47%
Decision Tree Classifier	51%
AdaBoost Classifier	56%
Gradient Boosting	58%
Logistic Regression	61%
Random Forest Classifier	67%
Multi-Class Neural Networks	72%

Table 4. Benchmark result of accuracy of different models in music genre classification

Future Work

Convolutional neural networks (CNNs) have been actively used for music classification. CNNs take entry of a spectrogram which is considered as an image instead of a vector of audio features. CNNs assume that the feature are in different levels that can be extracted by convolutional kernels.

And nowadays, another approach is commonly used for sequential data is by combining CNNs with recurrent neural networks (RNNs). This model is called convolutional recurrent neural network (CRNN), the idea is to alternate CNN by replacing the last convolutional layers with a RNN. The CNNs used for local feature extraction and recurrent neural networks for temporal summary on the features.

Reference

1. Davis, S. Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366
2. X. Huang, A. Acero, and H. Hon. Spoken Language Processing: A guide to theory, algorithm, and system development. Prentice Hall, 2001.
3. B McFee et al. librosa 0.5.0, 2017.
4. Ieeexplore.ieee.org. (2018). Music type classification by spectral contrast feature - IEEE Conference Publication. [online] Available at: http://ieeexplore.ieee.org/document/1035731/ [Accessed 21 Nov. 2018].
5. Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. Ann. Statist. 29 1189-1232.
6. Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A Dataset for Music Analysis. In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR’2017), 2017.
7. T. Minka, Automatic Choice of Dimensionality for PCA, 2000.
8. F. Pachet, D. Cazaly, "A classification of musical genre", Proc. RIAO Content-Based Multimedia Information Access Conf., 2000-Mar.
9. B. Logan. Mel frequency cepstral coefficients for music modeling. In Proc. Int. Symposium on Music Information RetrievalISMIR, 2000.]]

Contributions

Group Member	Tasks
AU YEUNG Sai Man	Data preprocessing, Video editing, Project documentation, Presentation Material
CHENG Carmen	Features extraction, Video editing, Demonstration preparation, Presentation Material
CHEUNG Chi Kan	Model implementation, Data analysis, Dimensionality Reduction
KWOK Ka Chun	Model implementation, Project documentation, Dimensionality Reduction
YEUNG Gillian Chi Ling	Data preprocessing, Demonstration preparation, Features extraction