Music Genre Classification`

Sriya Reddy (190002037)
Sreeja Yadav (190002039)
Kashish Bansal (190002032)

Introduction

- With the growth of online music databases and easy access to music content, people find it increasingly hard to manage the songs that they listen to.

- One way to categorize and organize songs is based on the genre, which is identified by some characteristics of the music such as rhythmic structure, harmonic content and instrumentation

- Being able to automatically classify and provide tags to the music present in a user's library, based on genre, would be beneficial for audio streaming services such as Spotify and iTunes.

Basic Implementation overview

- The first is a deep learning approach wherein a CNN model is trained end-to-end, to predict the genre label of an audio signal, solely using its MEL spectrogram.

- The second approach utilizes hand-crafted features, both from the time domain and frequency domain.These features are then fed to conventional machine learning models which are trained to classify the given audio file.

- The models are evaluated on the Audio Set dataset (Gemmeke et al., 2017). They also compared the proposed models and also study the relative importance of different features.

Dataset

- They made use of Audio Set, which is a large-scale human annotated database of sounds (Gemmeke et al., 2017).

- The dataset was created by extracting 10-second sound clips from a total of 2.1 million YouTube videos. The audio files have been annotated on the basis of an ontology which covers 527 classes of sounds including musical instruments, speech, vehicle sounds,animal sounds and so on.

- This study requires only the audio files that belong to the music category, specifically having one of the seven genre tags.

Dataset

Number of instances in each genre class

	Genre	Count
1	Pop Music	8100
2	Rock Music	7990
3	Hip Hop Music	6958
4	Techno	6885
5	Rhythm Blues	4247
6	Vocal	3363
7	Reggae Music	2997
	Total	**40540**

Dataset

- The raw audio clips of these sounds have not been provided in the Audio Set data release. However, the data provides the YouTubeID of the corresponding videos, along with the start and end times. Hence, the first task performed is to retrieve these audio files.- downloaded in mp4.

- The mp4 files are converted into the desired wav format using an audio converter named ffmpeg.

- Each wav file is about 880 KB in size, which means that the total data used in this study is approximately 34 GB.

Using CNN

- Using deep learning, we can achieve the task of music genre classification without the need for hand-crafted features.

- Convolutional neural networks (CNNs) have been widely used for the task of image classification.The 3-channel (RGB) matrix representation of an image is fed into a CNN which is trained to predict the image class.

- In this study, the sound wave can be represented as a spectrogram, which in turn can be treated as an image.

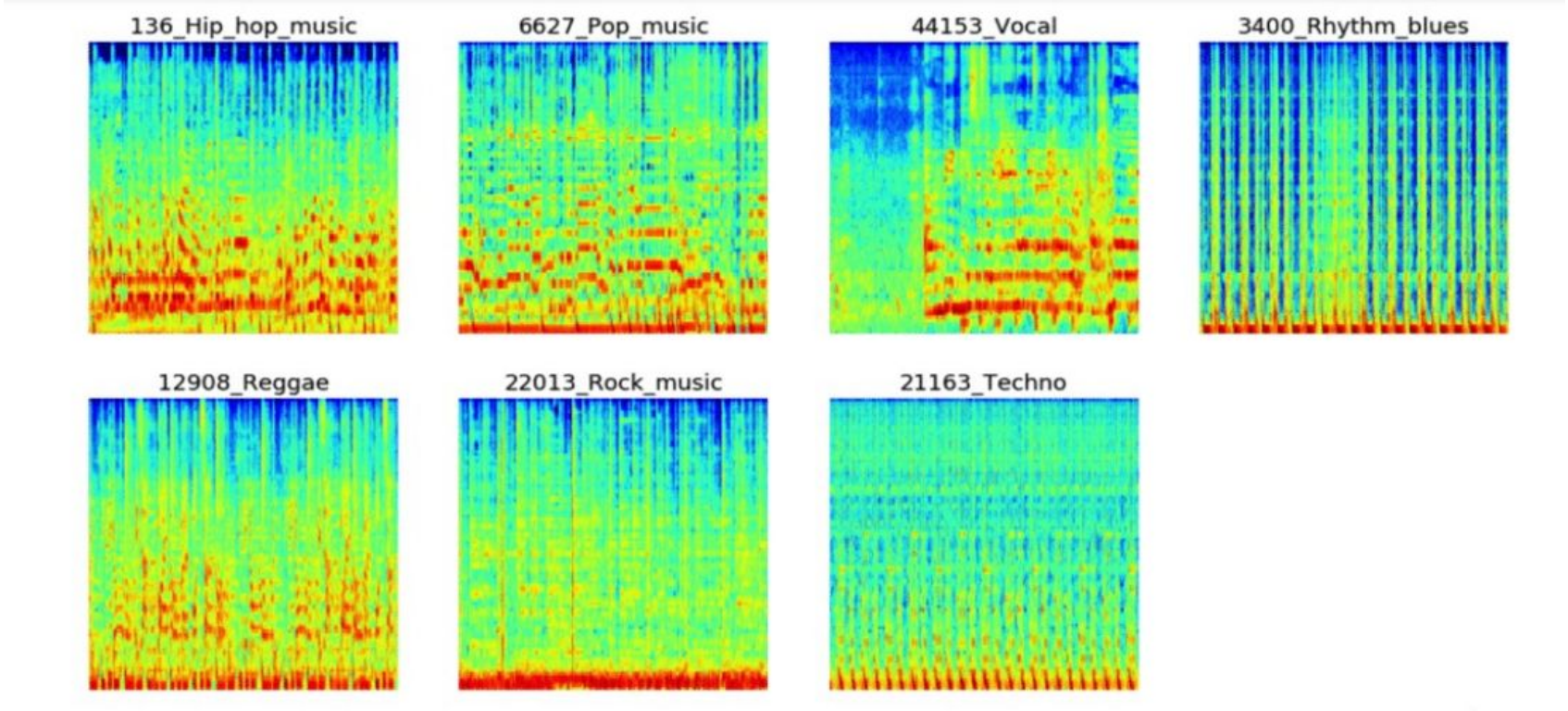
- The 3-channel (RGB) matrix representation of an image is fed into a CNN which is trained to predict the image class.

Spectrogram Generation

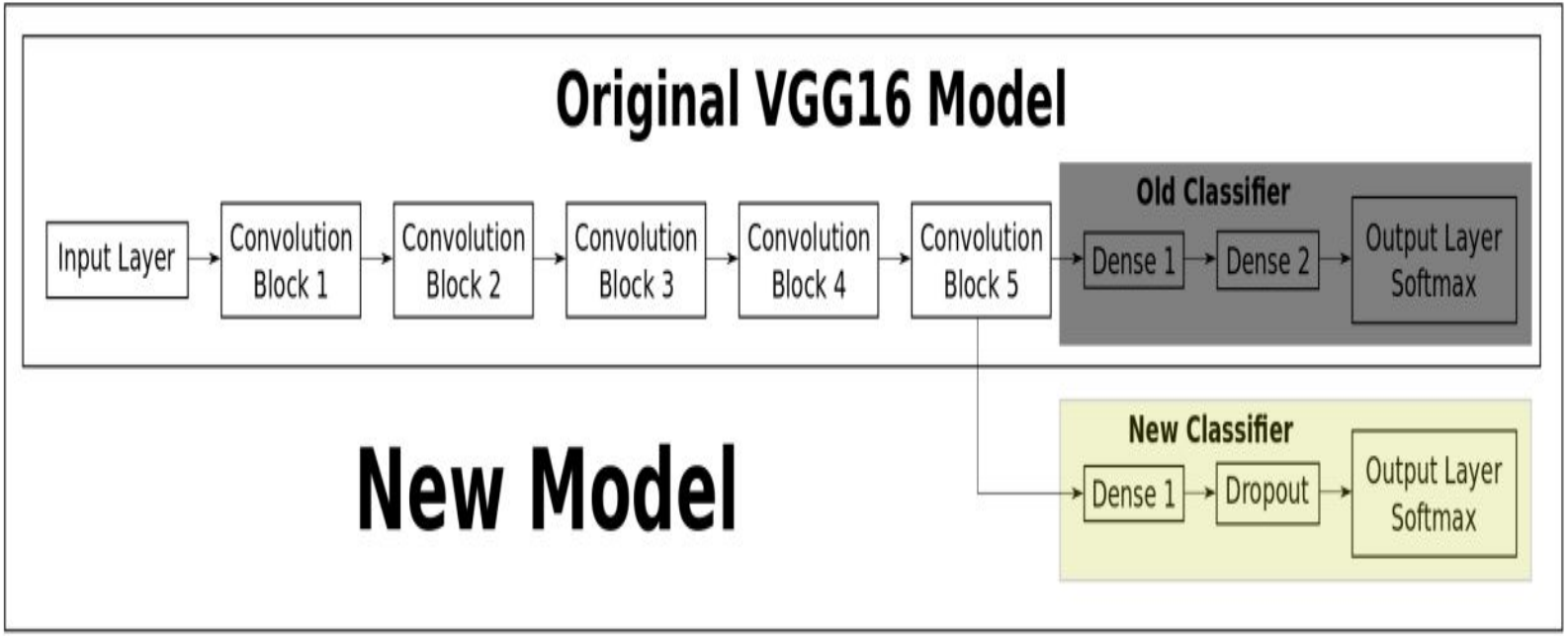
Spectrogram is a way of representing signal with time on x-axis and frequency on y-axis.Here, we are using MEL spectrogram (i.e on y-axis we have MEL frequencies).MEL spectrogram is generated using STFT.As the audio signal frequency content changes overtime hence we apply fast fourier transform on several overlapping windowed segments of signals which is called STFT.

These parameters are used to generate the spectrogram:
Sampling rate (sr) = 22050
Frame/Window size (n fft) = 2048
Time advance between frames (hop size) = 512 (resulting in 75% overlap)
Window Function: Hann Window
Frequency Scale: MEL
Number of MEL bins: 96
Highest Frequency (f max) = sr/2

Spectrogram of different genres



Convolutional Neural networks



Convolutional Neural Networks

CNN is supervised learning neural Network.Early layers of CNN might detect edges then the middle layers will detect parts of objects and the later layers will put the these parts together to produce an output.

Types of layer in a convolutional network: Convolution,Pooling ,Fully connected .

Convolution Layer: The objective of the Convolution Operation is to **extract the high-level features** such as edges,high-lighted patterns from the image, by convolving image with kernel/filter.

Pooling Layer: This method is helpful to extract dominant features.It also reduces size of the feature map obtained from convolutional layer hence reduces computational power.thus maintaining the process of effectively training of the model.

Non-linear Activation: The purpose of the activation function is to introduce non-linearity into the output of a neuron to make neural network more powerful.ex: ReLU activation function.

In CNN after passing through all Convolutional layers and pooling layers output will be passed through dense layer we flatten the 3d matrix to 1d before giving as it as a input to dense layer.

Convolutional Neural Networks

Dense layer: It is called dense layer because each neuron receives input from all the neurons of previous layer.Each Layer in the *Neural Network* contains neurons, which compute the **weighted average of its input** and this weighted average is passed through a non-linear function, called as an "***activation function"***. Result of this activation function is treated as output of that neuron. In similar way, the process is carried out for all neurons of all layers.

Softmax function: It is used when we have two or more classes.It returns the probabilities of each class.If we have total 10 classes, then the number of neurons in the output layer will be 10 . **Each neuron represents one class.**All 10 neurons will return probabilities of the input image for the respective class. *Class with highest probability will be considered as output for that image*.

Convolutional Neural Networks

For the genre classification,a CNN architecture (VGG-16) is used,we download the model architecture with pre trained weights and extract conv base .This model consists of 5 convolutional blocks,followed by a set of densely connected layers,which outputs the probability that whether the image belongs to above classes.
There are two possible settings while implementing the pre-trained model:
Transfer learning: The weights in the conv base are kept fixed but the weights in the feed-forward network are allowed to be tuned to predict the correct genre label.
Fine tuning: In this setting, we start with the pre-trained weights of VGG-16, but allow all the model weights to be tuned during training process
The cross-entropy loss is computed ,this loss is used backpropagate the error and compute gradients and update neural network until the loss becomes minimum.

Manually Extracted Features

- Hand-crafted features to be fed into a machine learning classifier.
- Features can be broadly classified as time domain and frequency domain features. The feature extraction was done using librosa , a Python library.

  Time Domain Features:
- **1. Central moments:** This consists of the mean, standard deviation, skewness and kurtosis of the amplitude of the signal.
- **2.Zero Crossing Rate (ZCR):**A zero crossing point refers to one where the signal changes sign from positive to negative (Gouyon et al., 2000). The entire 10 second signal is divided into smaller frames, and the number of zero-crossings present in each frame are determined.Finally, the average and standard deviation of the ZCR across all frames are chosen as representative features.
- **Root Mean Square Energy (RMSE):** RMSE is calculated frame by frame and then we take the average and standard deviation across all frames.

Manually extracted features

Tempo: In general terms, tempo refers to the how fast or slow a piece of music is; it is expressed in terms of Beats Per Minute (BPM).Since the tempo of the audio piece can vary with time, we aggregate it by computing the mean across several frames.

Frequency Domain Features:
The audio signal can be transformed into the frequency domain by using the Fourier Transform. We then extract the following features.

- **Mel-Frequency Cepstral Coefficients (MFCC):**To obtain mfcc coefficients, we would do following:
- (sound signal frame in time domain)->FFT->mel freq. scale filter->log->DCT

Manually extracted features

- **Chroma Features:**This is a vector which corresponds to the total energy of the signal in each of the 12 pitch classes. (C, C#, D, D#, E ,F, F#, G, G#, A, A#, B).The chroma vectors are then aggregated across the frames to obtain a representative
- **Spectral Contrast:** Each frame is divided into a pre-specified number of frequency bands. And, within each frequency band, the spectral contrast is calculated as the difference between the maximum and minimum magnitude
- **Spectral Centroid:** For each frame, this corresponds to the frequency around which most of the energy is centered . It is a magnitude weighted frequency calculated .

Manually extracted features

- Spectral Roll-off: This feature corresponds to the value of frequency below which 85% (this threshold can be defined by the user) of the total energy in the spectrum lies
- For each of the spectral features described above, the mean and standard deviation of the values taken across frames is considered as the representative final feature that is fed to the model.
- The features that contribute the most in achieving a good classification performance were identified and reported.

Classification and evaluation

- Logistic Regression
- SVM
- Random Forest
- Extreme Gradient Boosting (XGB)

Results:

- **Accuracy:** Refers to the percentage of correctly classified test samples.

	Accuracy
Logistic Regression (LR)	0.53
Random Forest (RF)	0.54
Support Vector Machines (SVM)	0.57
Extreme Gradient Boosting (XGB)	0.59
VGG-16 CNN	0.64

Most Important Features

- To carry out this experiment, they chose the XGB model
- To do this, they ranked the top 20 most useful features based on a scoring metric
- The metric is calculated as the number of times a given feature is used as a decision node among the individual decision trees that form the gradient boosting predictor
- MelFrequency Cepstral Coefficients (MFCC) appear the most among the important features.
- Their experiments show that MFCCs contribute significantly to this task of music genre classification.
- The mean and standard deviation of the spectral contrasts at different frequency bands are also important features.

Most important features

- We study how much of performance in terms of accuracy, can be obtained by just using the top N while training the model
- Tt was seen that with only the top 10 features, the model performance is surprisingly good. In comparison to the full model which has 97 features, the model with the top 30 features has only a marginally lower performance.
- Comparing XGB performance keeping only top N features
- N - Accuracy pairs
- 10 -0.47, 20 - 0.52 ,30 ,0.55 97-0.59
- Experiment where 2 models are trained one with only time domain features and the other with only frequency domain features. This experiment further confirms the fact that frequency domain features are definitely better than time domain features when it comes to modelling audio for machine learning tasks

Conclusion

- Our goal was to classify music signals into their respective genres.
- Two methods are explored here
- The first involves generating a spectrogram of the audio signal and treating it as an image. An CNN based image classifier, namely VGG-16 is trained on these images to predict the music genre solely based on this spectrogram
- The second approach consists of extracting time domain and frequency domain features from the audio signals, followed by training traditional machine learning classifiers based on these features.
- Also the most important features were reported.