Music Genre Classification using Machine Learning Techniques

CS 698 - Computational Audio

Hareesh Bahuleyan

UNIVERSITY OF WATERLOO

Problem Statement

- Music genres are a way to classify music based on rhythmic structure, harmonic content and instrumentation

- Automatically recognition
  - Organize digital libraries
  - Provide recommendations

Data

Google *Audio Set*

- 2.1 Million audio samples (of 10 seconds)
- 527 classes of sounds
- Selected 7 labels

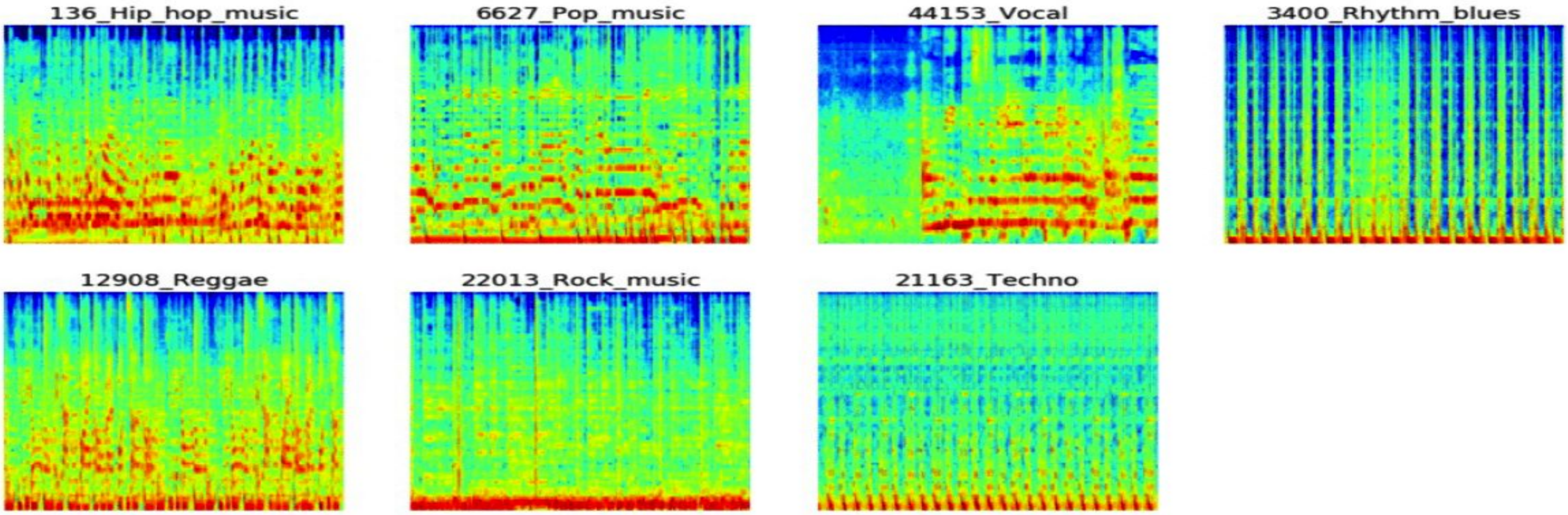
- Not the actual audio, just the YouTubeIDs, start and end times
- 880 KB per wav file,
- Approximately 34 GB data

	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**

Convolutional Neural Networks

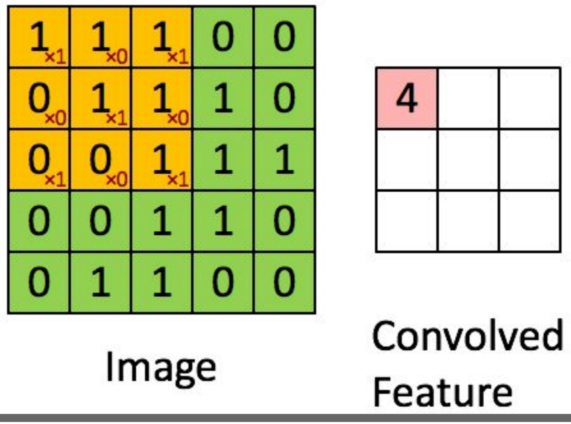
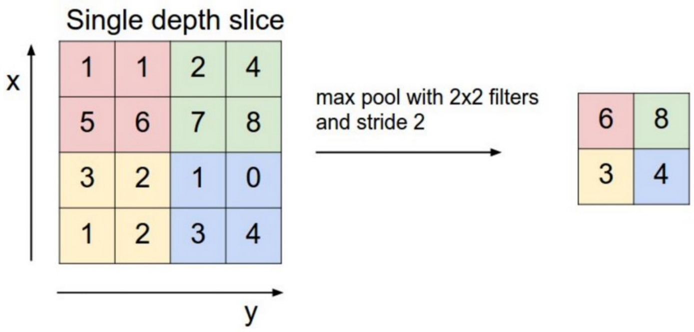
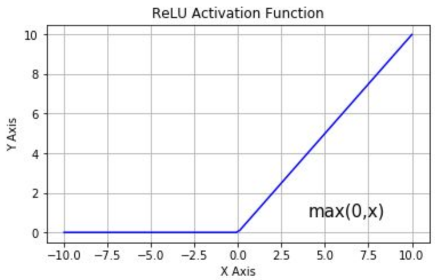
MEL Spectrograms

- 2D colormap representation of the signal
- **STFT**: Window size = 2048, Hop size = 512, Hann window function, Number of MEL bins = 96



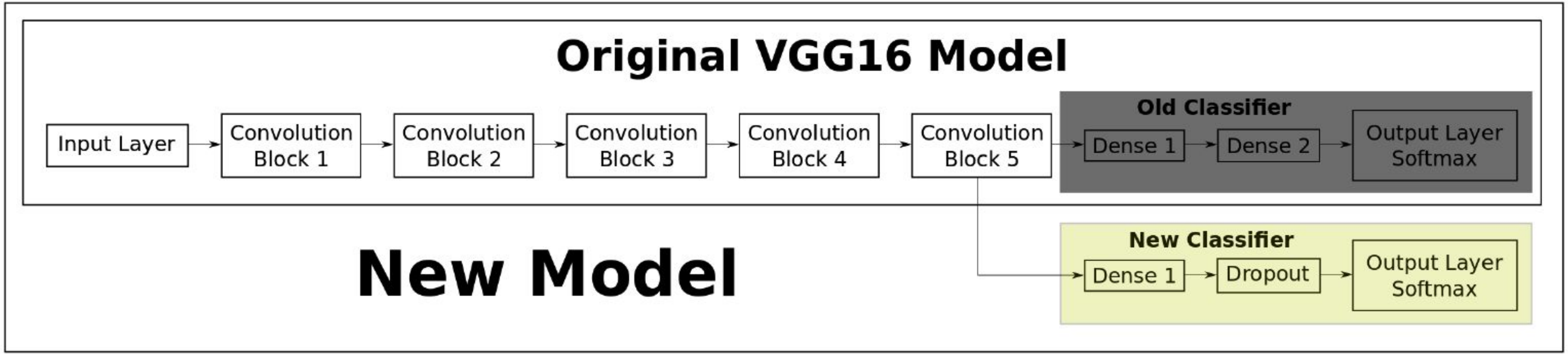
CNN - Image Classification

- Consider spectrogram as an image and train a CNN classifier
- Matrix of pixel values - 3 channel RGB input

Convolution Block		
Convolution	**Pooling**	**Non-Linear Activation**
		

Source: http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/

VGG-16



- **Transfer Learning**
  - Weights of conv base are fixed
- **Fine Tuning**
  - Both conv base and feed-forward network are trainable

Feature Engineering Approaches

Feature Extraction

Time Domain

1. Mean
2. Variance
3. Skewness
4. Kurtosis
5. Zero Crossing Rate
6. Root Mean Square Energy
7. Tempo

Frequency Domain

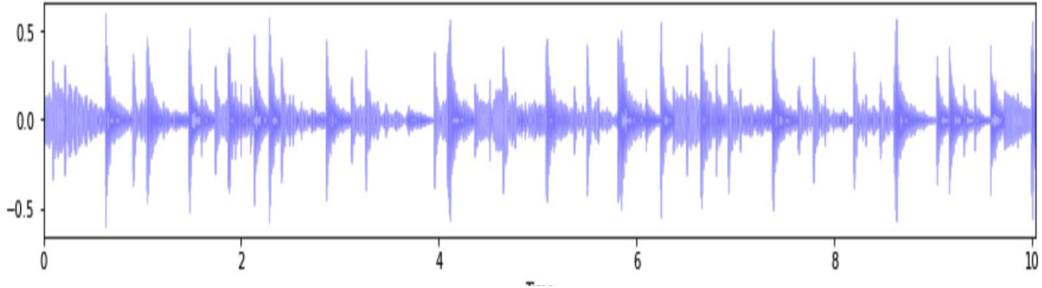
1. MEL Frequency Cepstral Coefficients (MFCCs)
2. Chroma Features
3. Spectral Centroid
4. Spectral Band-widths
5. Spectral Contrast
6. Spectral Roll-offs

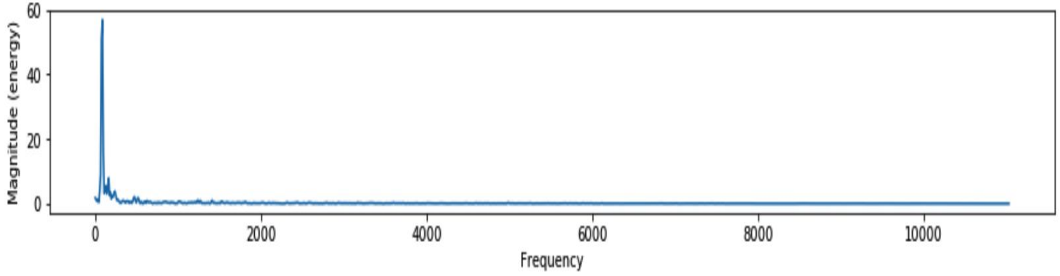
Classifiers

1. Logistic Regression
2. Random Forest
3. Support Vector Machines
4. Extreme Gradient Boosting

- Total Number of Features = 97

Spectral Features





- **Spectral Centroid**

$$f_c = \frac{\Sigma_k S(k) f(k)}{\Sigma_k S(k)}$$

- **Spectral Band-width**

$$(\Sigma_k S(k) f(k) - f_c)^{\frac{1}{p}}$$

- **Spectral Contrast**
  - Divide spectrum into frequency bands
  - Maximum magnitude - Minimum magnitude in each band
- **Spectral Roll-off**
  - Frequency below which 85% of the total energy in the spectrum lies
- **Chroma Features**
  - 12-element feature vector
  - Indicates how much energy of each pitch class, {C, C#, D, D#, E, ..., B}

Results

Comparison of Models

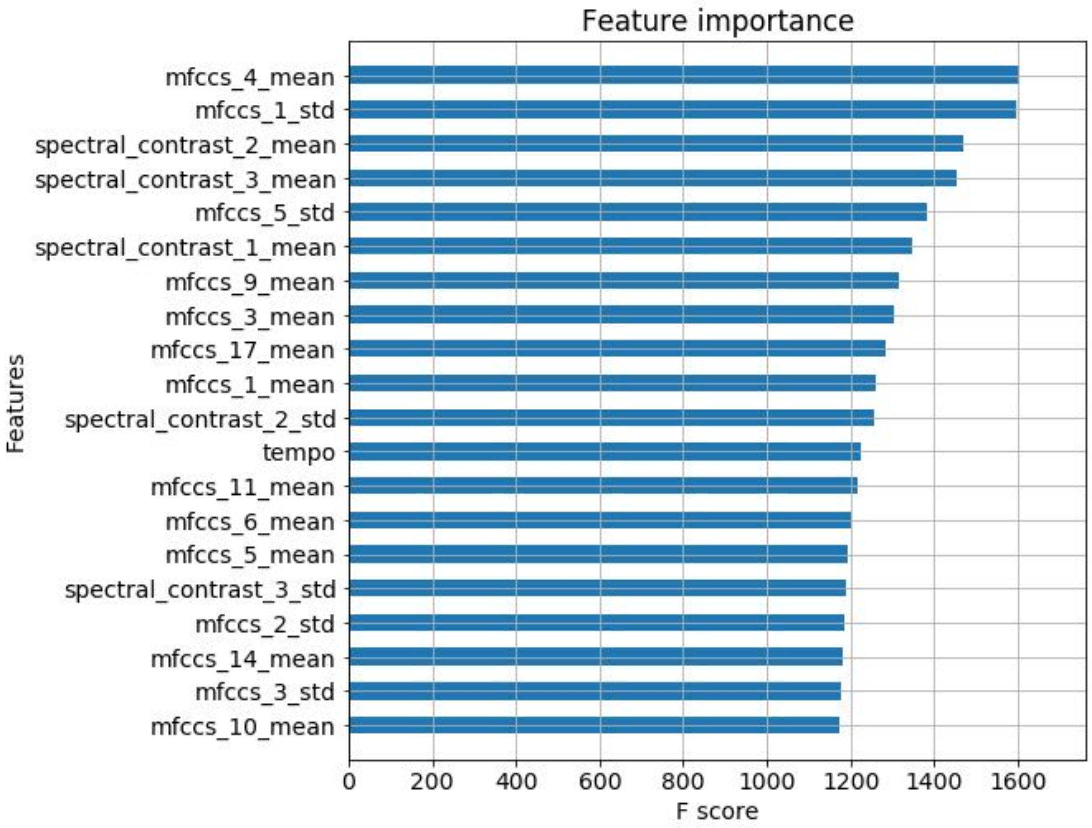
- **Metrics:** Accuracy | F-score | AUC

	Accuracy	F-score	AUC
Spectrogram-based models			
VGG-16 CNN Transfer Learning	0.63	0.61	**0.891**
VGG-16 CNN Fine Tuning	**0.64**	**0.61**	0.889
Feed-forward NN baseline	0.43	0.33	0.759
Feature Engineering based models			
Logistic Regression (LR)	0.53	0.47	0.822
Random Forest (RF)	0.54	0.48	0.840
Support Vector Machines (SVM)	0.57	0.52	0.856
Extreme Gradient Boosting (XGB)	**0.59**	**0.55**	**0.865**
Ensemble Classifiers			
VGG-16 CNN + XGB	**0.65**	**0.62**	**0.894**

Baseline uses flatten vector of pixels

Ensembling classifiers is beneficial

Feature Importance Study



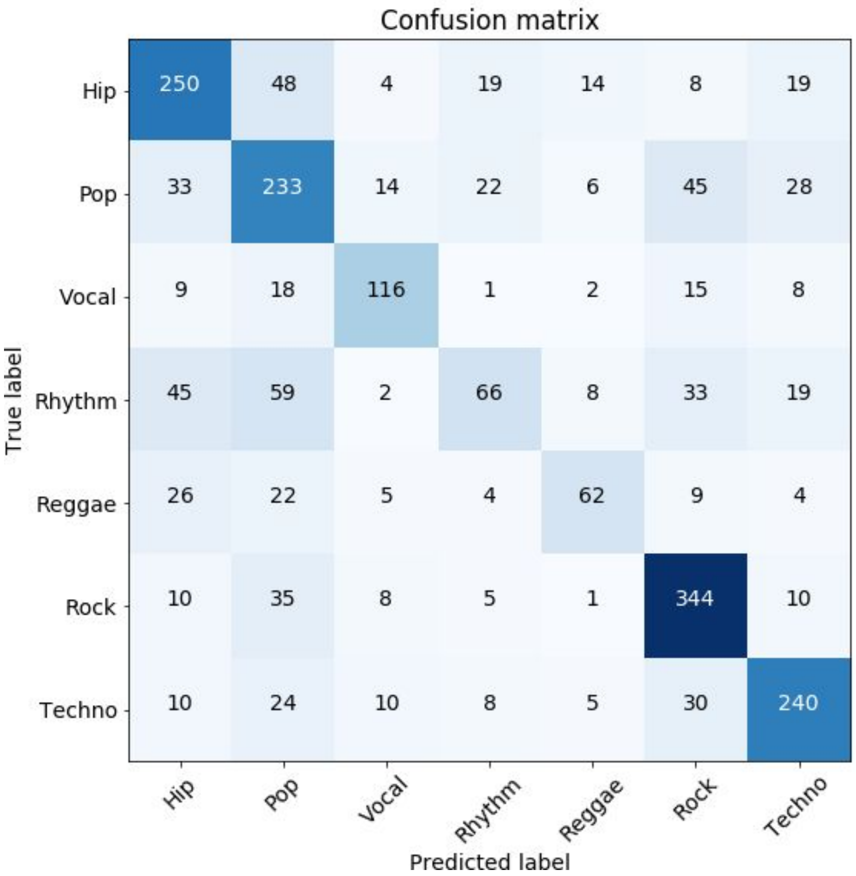
Keep only most important top *N* features

N	AUC
10	0.803
20	0.837
30	0.845
97	0.865

Time domain vs. Frequency domain

Model	AUC
Time Domain only	0.731
Frequency Domain only	0.857
Both	0.865

Confusion Matrix



Good at predicting some classes. Eg: Rock

Many mis-classifications for Rhythm blues, Pop genre

Classes are also unbalanced

Thank You