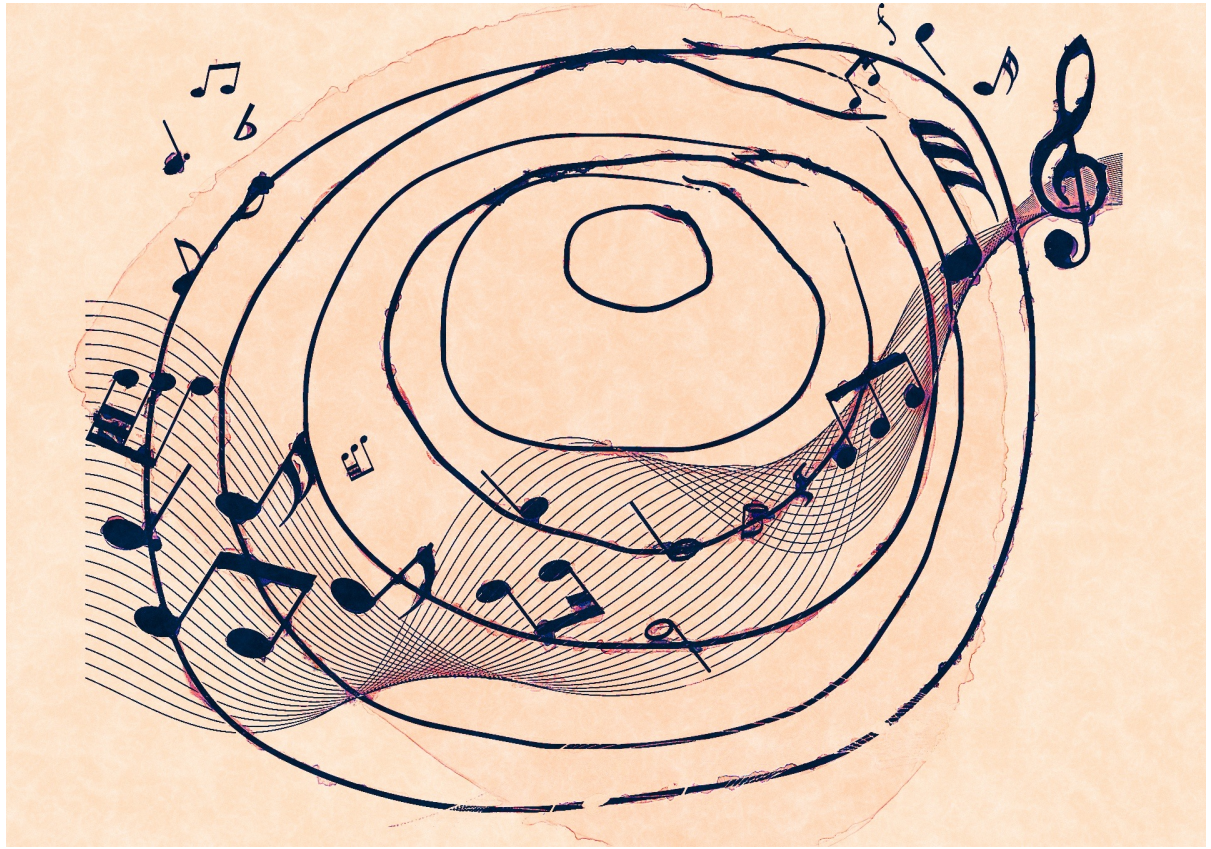
Music Genre Classification

Shruti Misra and Manuja Sharma



Goal

Compare the performance of Gaussian Mixture Models, Support Vector Machines, Convolutional Neural Networks (CNNs), Convolutional Recurrent Neural Networks (CRNNs) to classify audio tracks into 10 different musical genres

Dataset

- The Million Song Dataset (MSD) consists of audio features and metadata for 1,000,000 contemporary music tracks from 44,745 unique artists. Data for each track includes textual features such as artist and album names, numerical descriptors such as duration and audio features.

- The Million Song Dataset does not include any genre labels. However, external groups have proposed genre labels for some of the tracks by accessing external music tagging databases such as the CD2C tagtraum genre annotations. (Schreiber (2015)).

- We use the CD2C variant with non-ambiguous annotations, that is, tracks with multiple genre labels are not included. We classify tracks into 10 different genres; Rap, Rock, RnB, Electronic, Metal, Blues, Pop, Jazz, Country, Reggae.

- We used a subset of the Million Song Dataset, that consists of 40,000 training samples and 10,000 test samples The training and test sets are balanced. Additionally, the dataset we used only consists of the audio features for each track and does not contain textual or numerical descriptors.

Features

- The features for each audio track are partitioned into bins of 120 segments.
- A segment corresponds to an automatically identified, roughly continuous sections of audio with similar perceptual quality.
- The number of segments per track is fixed to 120. For each segment, the following audio features are available: **12 timbre features, 12 chroma features and loudness** at start of segment concatenated in that order.
- Consequently, each segment is characterized by a 25 dimensional vector. Furthermore, each audio track is represented by a total of 3,000 features (120x25 = 3000). Every feature has been normalized by subtracting the mean and dividing by the standard deviation.

Baseline

Random Classification

- Randomly assigned class to dataset

Gaussian Mixture Model (GMM)

- Each segment as a data point
- Input size: 40,000 x 25
- Training: 8 minutes
- Validation: 5 fold CV

Support Vector Machine (SVM)

- Input Size: Data x 3000
- Training: 1 hr on GPU (K80)

Neural Network

Convolutional Neural Network (CNN)

- Used for image classification
- Layers: 7
- Activation: ReLu, Softmax
- Loss: Cross Entropy
- Training: 15 mins on GPU (K80)
- Validation: 10% of training set

Convolutional Recurrent Neural Network (CRNN)

- Used for images with texts classification
  - Recently for audio classification
- Layers: 9
  - CNN: 4 layers
  - LSTM: 2 layers
- Input Size: 120x25x1
- Activation: ReLu, Softmax
- Training: 25 mins on GPU (K80)
- Validation: 10% of training set
- Loss: Cross Entropy
- Optimizer: Adam

Results

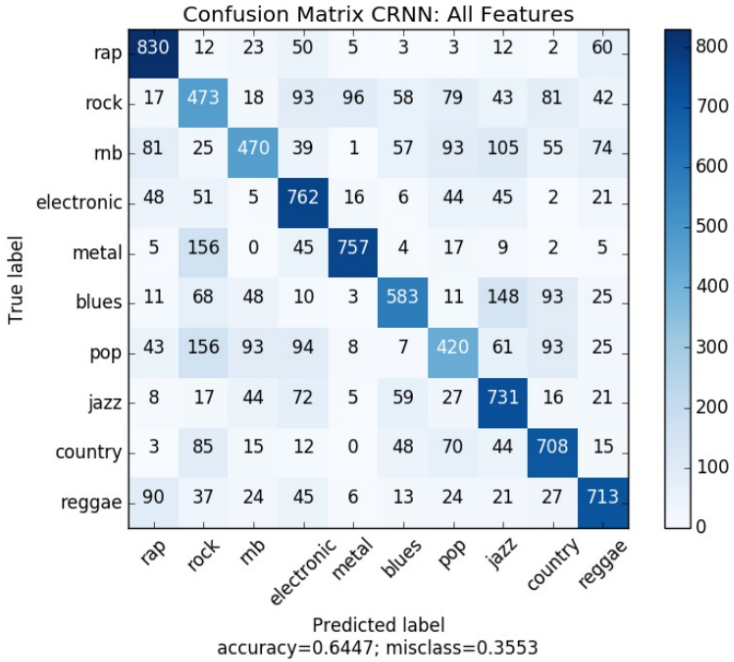
Model	Accuracy	F1 Score (Mean)
CRNN	64.47 %	33.75
CNN	56.1 %	29.4
GMM	51.05	26.76
SVM	26 %	15.7
Random	10 %	5.11

Related work with the dataset has 66% accuracy

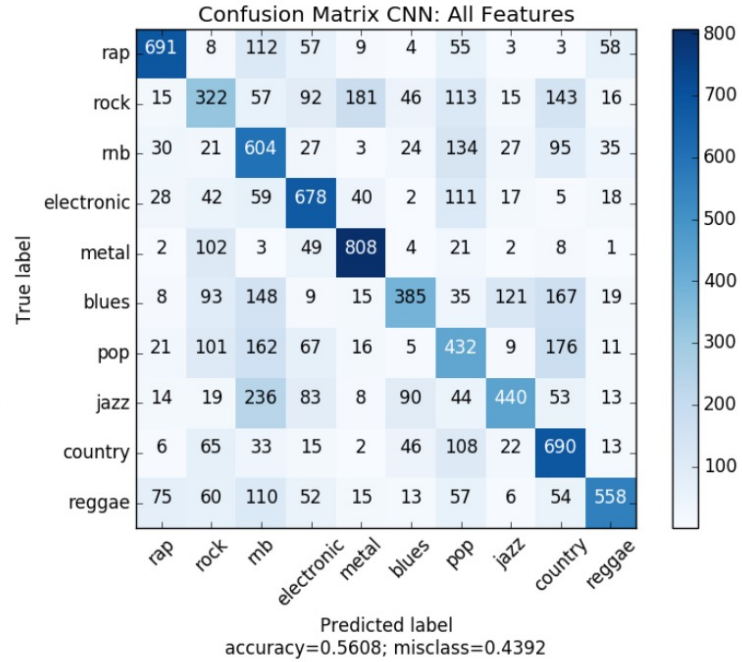
CRNN and GMM Timbre features alone performed better than Chroma features

Evaluation Metrics for different models

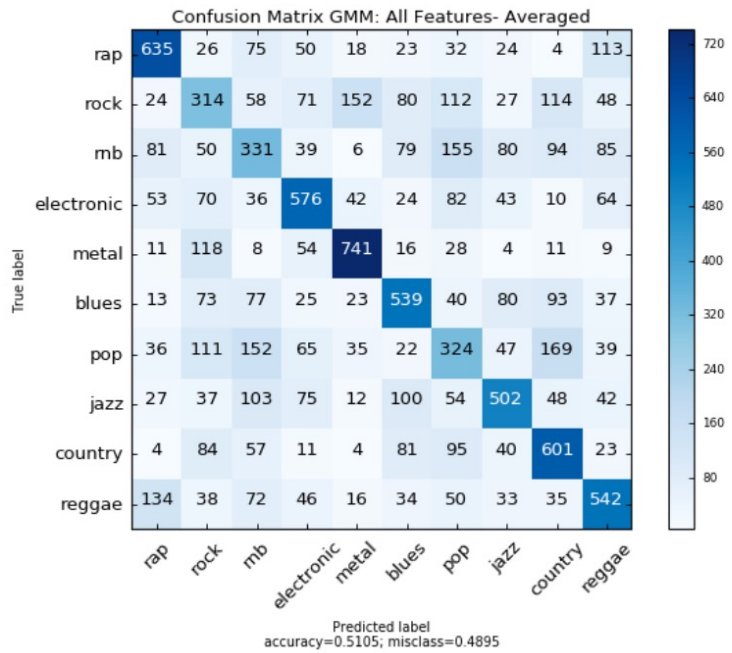
Model	Accuracy	F1 Average	Training Time(mins)
CRNN training	63.85%		25
CRNN testing	64.47%	33.75	
CNN training	58.6%		15
CNN testing	56.1%	29.4	
GMM Training	51.20%		8
GMM Testing	51.05%	26.76	
SVM Testing	30%	15.7	70
Random	10%	5.11	1

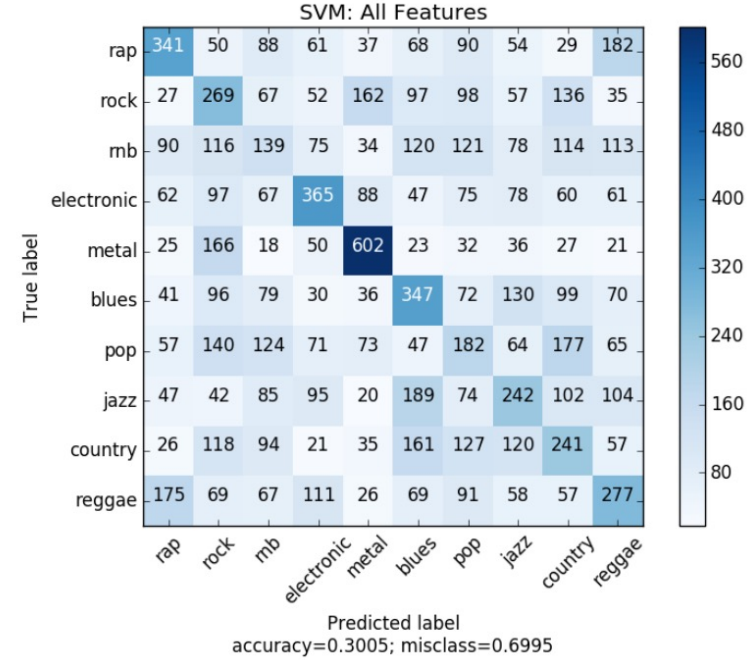


(a) Confusion Matrix CRNN



(b) Confusion Matrix CNN





Conclusion

- The random classification as expected performed the worst with around 10% accuracy.
- Our most prominent baseline was the Gaussian Mixture Model, where each genre corresponded to a GMM. Our best GMM classifier was trained over all features averaged over all segments and garnered an accuracy of 51.05%.
- We also trained a Support Vector classifier as a baseline. However, this model did not fare too well, providing an accuracy of only 26.64%.
- We then implemented CNN and CRNN neural network models and both outperformed all the baselines.
- CRNN performed the best with an accuracy of 64.47%.
- We also concluded that timbral features provided more information about music genre than chroma or pitch features. As part of future work, we can look at obtaining raw audio data and extracting our own features to characterize different genres more accurately.
- Furthermore, incorporating textual data such as year of release, artist and lyrics might also help improve performance.