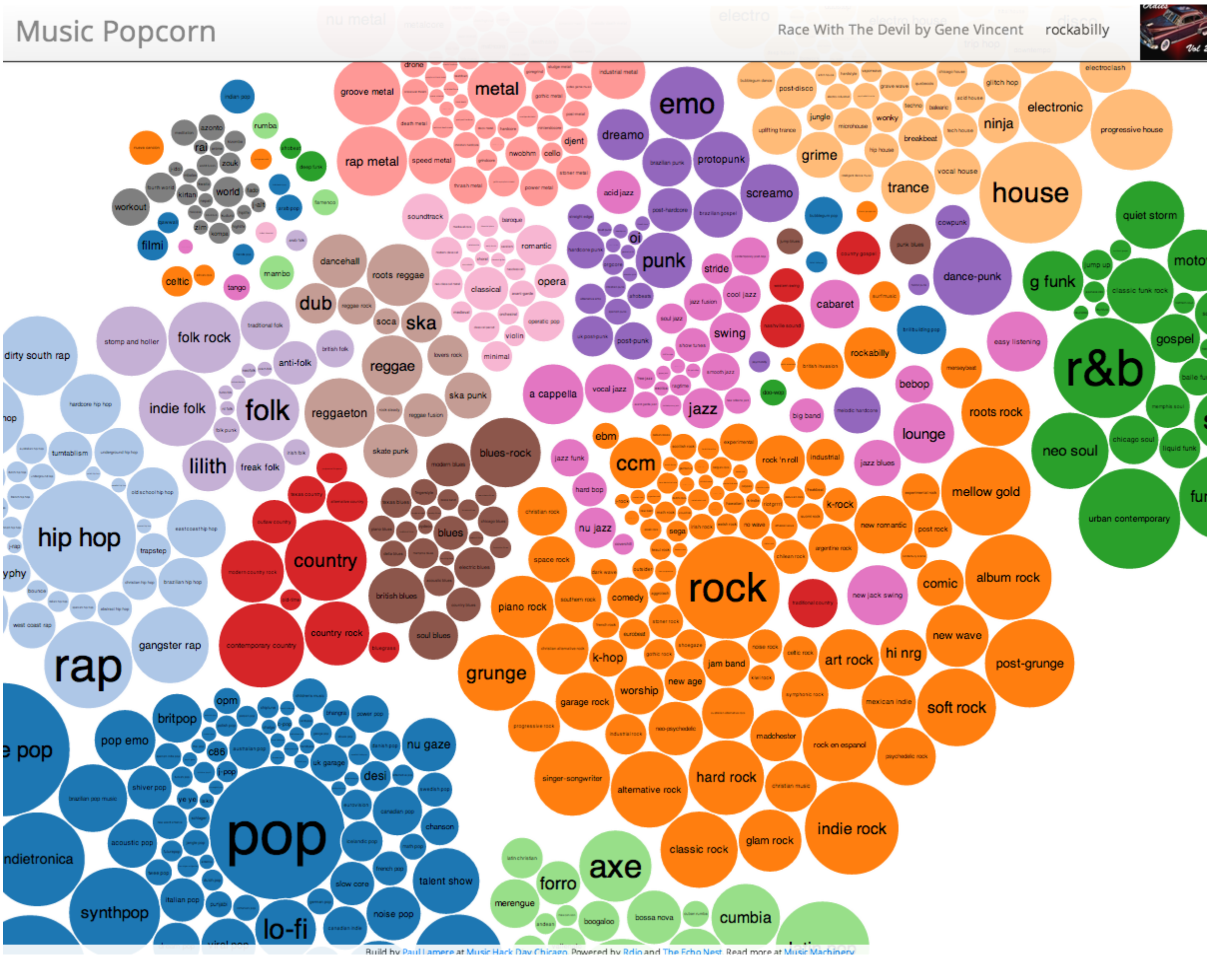
Music Genre Classification using Machine Learning Techniques



Team Members :

Vasukumar Kotadiya (2019171)

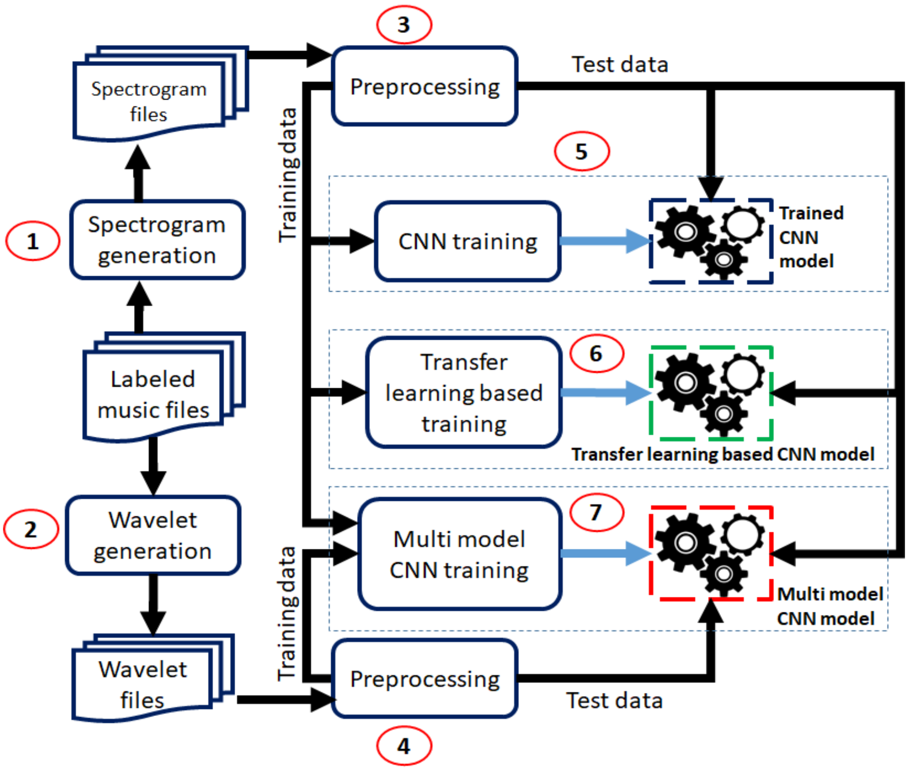
Yash Raj Kesarwani (2019178)

Instructor: Dr. Kusum Kumari Bharati

Introduction:

With the growth of online music databases and easy access to music content, people find it increasing 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 (Tzanetakis and Cook, 2002). 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. This study explores the application of machine learning (ML) algorithms to identify and classify the genre of a given audio file.

This work aims to provide a comparative study between 1) the deep learning-based models which only require the spectrogram as input and, 2) the traditional machine learning classifiers that need to be trained with hand-crafted features. We also investigate the relative importance of different features.



Motivation:

Categorizing music files according to their genre is a challenging task in the area of music information retrieval (MIR). In this study, we compare the performance of two classes of models. 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 spectrogram. The second approach utilizes hand-crafted features, both from the time domain and frequency domain. We train four traditional machine learning classifiers with these features and compare their performance. The features that contribute the most towards this classification task are identified. The experiments are conducted on the Audio set data set and we report an AUC value of 0.894 for an ensemble classifier which combines the two proposed approaches.

Improvement from Previous Version:

In previous version, we had used K-nearest Neighbour (KNN) and Artificial Neural Networks (ANN). And here we have used Convolution Neural Networks (CNN).

Firstly, we had solved the problem statement to classify music according to their genre by KNN in which we had achieved the accuracy of 67.8%. When we used CNN which is much more advanced data classification model, we had achieved a slight increase in accuracy hence reaching the accuracy of 73%.

ANN is ideal for solving problems regarding data. Forward-facing algorithms can easily be used to process image data, text data, and tabular data. CNN requires many more data inputs to achieve its novel high accuracy rate.

Literature Survey :

Music genre classification has been a widely studied area of research since the early days of the Internet. Tzanetakis and Cook (2002) addressed this problem with supervised machine learning approaches such as Gaussian Mixture model and knearest neighbour classifiers. They introduced 3 sets of features for this task categorized as timbral structure, rhythmic content and pitch content. Hidden Markov Models (HMMs), which have been extensively used for speech recognition tasks, have also been explored for music genre classification (Scaringella and Zoia, 2005; Soltau et al., 1998). Support vector machines (SVMs) with different distance metrics are studied and compared in Mandel and Ellis (2005) for classifying genre.

In Lidy and Rauber (2005), the authors discuss the contribution of psycho-acoustic features for recognizing music genre, especially the importance of STFT taken on the Bark Scale (Zwicker and Fastl, 1999). Mel-frequency cepstral coefficients (MFCCs), spectral contrast and spectral roll-off were some of the features used by (Tzanetakis and Cook, 2002). A combination of visual and acoustic features are used to train SVM and AdaBoost classifiers in Nanni et al. (2016).

With the recent success of deep neural networks, a number of studies apply these techniques to speech and other forms of audio data (AbdelHamid et al., 2014; Gemmeke et al., 2017). Representing audio in the time domain for input to neural networks is not very straight-forward because of the high sampling rate of audio signals. However, it has been addressed in Van Den Oord et al. (2016) for audio generation tasks. A common alternative representation is the spectrogram of a signal which captures both time and frequency information. Spectrograms can be considered as images and used to train convolutional neural networks (CNNs) (Wyse, 2017). A CNN was developed to predict the music genre using the raw MFCC matrix as input in Li et al. (2010). In Lidy and Schindler (2016), a constant Q-transform (CQT) spectrogram was provided as input to the CNN to achieve the same task.

References

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing 22(10):1533–1545.

[2] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 .

[3] Nicolas Scaringella and Giorgio Zoia. 2005. On the modeling of time information for automatic genre recognition systems in audio signals. In ISMIR. pages 666–671.

[4] Loris Nanni, Yandre MG Costa, Alessandra Lumini, Moo Young Kim, and Seung Ryul Baek. 2016. Combining visual and acoustic features for music genre classification. Expert Systems with Applications 45:108–117.

[5] Thomas Lidy and Andreas Rauber. 2005. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In ISMIR. pages 34–41.

[6] George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. IEEE Transactions on speech and audio processing 10(5):293– 302.