Analysis of Music information retrieval techniques

Anupama Dhekane, MM33458
University of Maryland Baltimore County.

- ## Abstract:

Music information retrieval is a sub-discipline of information retrieval that extracts information from audio signals. It includes audio analysis tasks such as genre classification, song identification, chord recognition, sound event detection, mood detection. In this paper we concentrate on techniques of Music information classification. The techniques discussed here are the improvisation of well known techniques like Mel frequency cepstral coefficients (MFCC) using convolutional recurrent neural networks. Here onwards we discuss feature extraction, Neural networks, architectures in use and improvisations.

- ## Introduction:

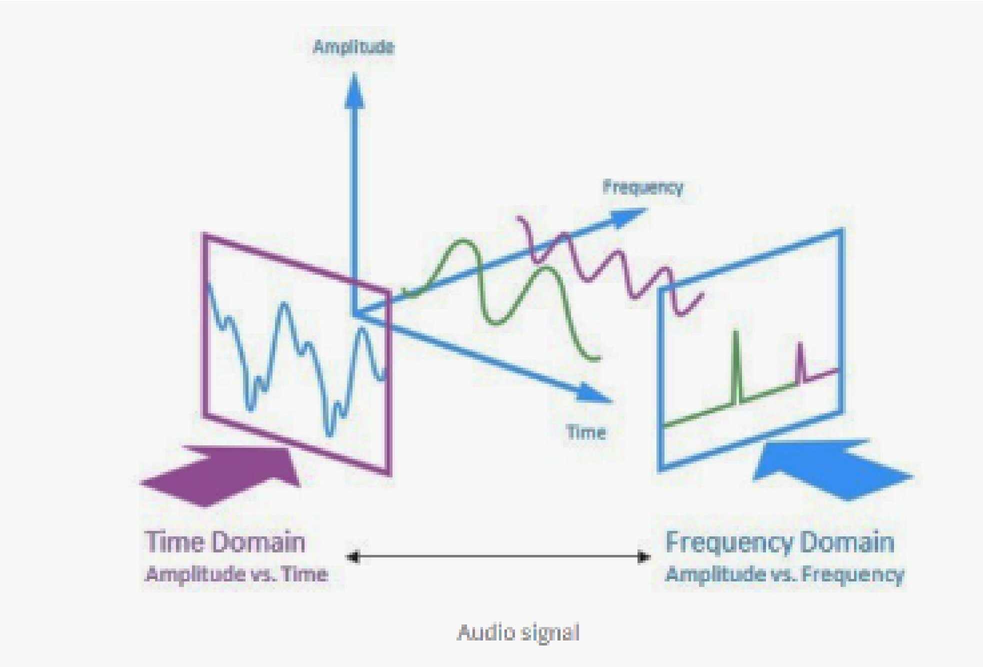
Motivation: To study various deep learning models for audio analysis, overcoming the inability of traditional machine learning models to extract high dimensional data. Traditional machine learning algorithms use vector summaries of frequency content in an audio window, specifically MFCCs as they summarize both pitch and quality of the sound(necessary features for classification) . Thus in turn losing temporal structure of data. CRNN infuses temporal data with frequency to get better results.

- ## Survey of the relevant work:

In here we explain the ideas behind the paper we analysed like what is audio signal, various forms interpretation of an audio signal, CNN , RNN and how both of these are used to construct a CRNN model,

What is an audio signal?

A three dimensional signal in which three axes represent time, amplitude and frequency. Librosa, a well known python library, is used to display the audio files in different formats such as waveplot, spectrogram or colormap.



Features:

1. Mel-frequency cepstral coefficient(MFCC)
2. Key
3. Chords
4. Melodies

5. Main pitch per beat
6. Beats per minute or rhythm

We will discuss MFCC architecture in brief here:

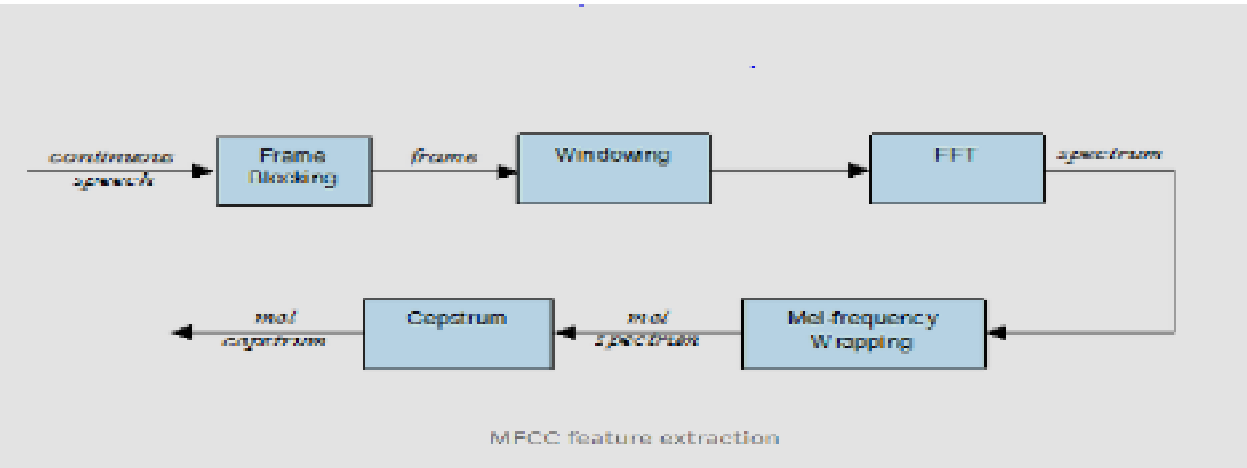


Fig 1.1

1. Frame Blocking: The input signal is segmented into frames of 20-30 ms with optional overlap of ⅓-½ of the frame size. Usually the frame size is equal to two in order to facilitate the use of forward fourier transform(1.1)[]
2. Windowing: Ideally, it uses a hamming window as a processing step to reduce spectral leakages, basically putting some constraints on frequency spread for better suited feature analysis.
3. FFT: STFT(1.2) converts signals such that we can know the amplitude of the given frequency at any given time. Using STFT we can determine the amplitude of various frequencies.
4. Melfrequency warping:
5. Cepstrum: cepstrum is the information of rate of change in spectral bands.

Convolutional Neural Network:

Convolutional neural networks, a variant of neural networks, is used heavily in audio signal processing. Some of the important components of CNN,

Activation functions: There are various activation functions like tanh, sigmoid, relu etc. It serves a purpose similar to neurons in the human body, a trigger that sets up the action.

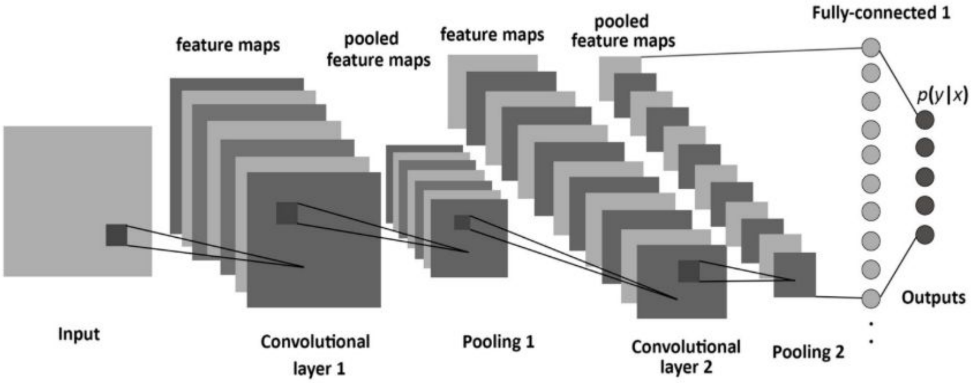
Node: Analogous to biological neurons, there are several layers made up of nodes in CNN,

Convolutional Layer: responsible for creating feature maps to predict the class probabilities for each feature. It takes input signal , applies a filter basically multiplying the input signal with the kernel to get the modified signal. It reduces the input matrix to a more understandable precise matrix.

Pooling layer: Scales down and maintains most essential information.

Fully connected layer: Applies weights to input generated by feature analysis to predict an accurate label

Fully connected output layer: generates final probabilities to determine the result.



Recurrent Neural Networks:

Named so as it performs the same function for every input of data while output of current input depends on the past one computation. RNNs use their internal state (memory) to process sequences of inputs. In RNNs all inputs are related to each other.

CRNN:

Convolutional network followed by RNN builds up the convolutional recurrent neural network. Three layers of convolutional layer followed by followed by permute and reshape layer which is very necessary for CRNN as the shape of feature vector differs from CNN to RNN.CNN are built over 3-D vectors while RNN are built using 2-D vectors.

Representation of audio signal for feature extraction:

Following are the feature extraction methods from sound waves:

STFT: provides time-frequency representation with linearly-spaced centre frequencies. The linear centre frequencies are not always desired in music analysis hence STFT is not a popular choice amongst researchers.

Melspectrogram: A 2D represntation that is optimised for human auditory perception, compresses STFT in frequency axis thus achieving efficiency and preserving perceptually important data. It's not invertible to audio signals. Melfrequenices are composed using below formula:

M = 2595log10(1+f/700)

Constant-Q transformm(CQT): Provides 2D representation with logarithmic scale centre frequencies.

Chromagram: A pitch class profile which provides energy distribution on a set of pitch classes, often with western's music's 12 pitches. Chromogram is considered as a CQT representation folding in the frequency axis.

Dataset used for analysis:

Artist 20 dataset from labrosa, million song dataset, Librispeech, Voxceleb for artist and genre classification. The technique which uses CRNN, outperforms the baseline results of ML algorithms used on the same database.

Splitting Database:

In the work discussed we split the dataset by artists and by albums(90/10 split). The train set is then split using the sma 90/10 split to create a validation subset. Stratification is used to ensure that equal no of songs of each artist are present on both sides. For album split, two albums of each artist are randomly removed from initial dataset, one is used for testing while the other one for validation.

Audio-processing:

A short-fourier transform is applied to a raw audio signal to get a spectrogram for every song. Once created spectrogram frequencies are converted to decibels by previously specified formula. Later on. Spectrograms are split into train, test data. Such spectrograms are created for the entirety of the song in contrast to previous practices. There are various techniques that could be used for the split using which we get variable performance of the model.

Results have been discussed in the next point as the field of MIR is still evolving, there are new developments regarding complexity introductions.

- **Comparison of the relevant work:**

Previous similar works:

Training a SVM model to categorize the artists and songs:

The previous works involve working on MFCC's which is really famous and important feature in audio processing but since it losses the temporal(frequency changes over the period) data, the results predicted by such models are temporary and not adaptive meaning, if the previously identified singer is presented in new form to the model it won't be able to identify the singer.

Using F1 score instead of accuracy:

They have used F1 score instead of the accuracy score as all audio slices within each song are used for training and evaluation. We experience class imbalance in this, but it can be mitigated by weighting the F1 score by number of supporting samples in each class.

In terms of architecture:

CRNN architecture explained earlier is preferable for any classification problem in MIR domain. It summarizes the pattern in frequency and also includes temporal data which gives improved results. Hence this model gives improvised results when the artist changes his/her style unlike the rigid ML models.

Comparisons based on results:

1. Traditional models (like SVM, Gaussian models) perform better for shorter audio length but as we add the temporal data the performance of deep learning models improve over the time.
2. Splitting data based on albums gives better results than splitting on basis of songs, observably because the songs include variations in pitch, frequency and style performed by the artist so huge datasets are required to get better performance.
3. Pooling layer used in CRNN architecture is actually detrimental to data where spatial(changes in frequency at a given point of time) location has importance.
4. The bottleneck layer, every audio sample is fully converted to a vector space which are used for classification, by reducing the data furthermore we can see the visual segregation of audio samples.

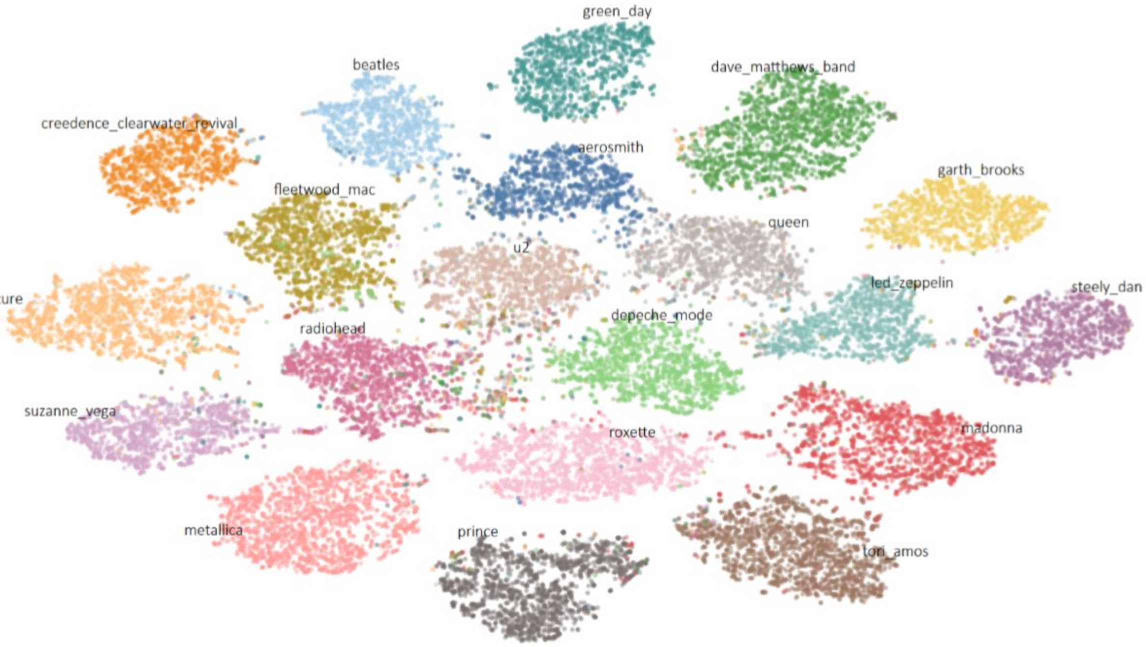


Fig. 3. t-SNE of learned audio representations using audio length of ten seconds (frame level; song split)

[Zain Nasrullah ,14 Mar 2019]

5. CRNN outperforms the traditional models by a minimum of 10% in classifying the data. Though there are few improvements regarding the pooling layer it's still a flexible model than traditional rigid models.
6. With such a good classification F1 score, this model can be used for copyright issues in the music industry.
7. The model classifies artists well and is pretty flexible but to make it even more flexible there are many improvements that can be done.

<u>Some improvements that can be introduced to the model:</u>

1. Interchanging the purpose, instead of training the model at the beginning it can be trained for genre classification which will help in building up the lower layers of the network and we can get faster and better results.
2. Testing if the pooling layer can be replaced by a combination of activation functions or layers as much of the important temporal data that we capture from spectrogram slicing is lost in this layer.
3. Not limiting the work to one area of the industry and instead introducing step by step complexities in terms of layers and classifications.

- **Conclusions:**

We can conclude following things:

- Deep learning models like in our case CRNN well outperform the traditional ML models but at the same time they are complex to implement and train.
- We can certainly agree that including temporal(time-based) data instead of just frequency related data definitely improves the performance.
- To get even better results we need much more complex layers in CRNN and utilize the bottleneck layer even more for better audio visualizations.

Refrences:

Nasrullah, Zain, and Yue Zhao. "Music Artist Classification with Convolutional Recurrent Neural Networks." *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019.

Choi, Keunwoo, et al. "A tutorial on deep learning for music information retrieval." *arXiv preprint arXiv:1709.04396* (2017).

Kim, Jaehun, et al. "One deep music representation to rule them all? A comparative analysis of different representation learning strategies." *Neural Computing and Applications* 32.4 (2020): 1067-1093.