International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 12 | Dec 2019 www.irjet.net p-ISSN: 2395-0072

Music Genre Recognition Using Convolution Neural Network

Krishna Mohana A J[1], Pramod Kumar P M[1], Harivinod N[1], Nagaraj K[1]

[1]Assistant Professor, Dept. of Computer Science & Engineering,
Vivekananda College of Engineering and Technology Puttur, Karnataka, India

---

Abstract – *Music is an ever changing or evolving field. We see a lot of songs, instrumentals, symphonies and other music forms releasing every day. All this music is of different genre moods and themes. Music streaming platforms need to classify these songs to improve User Experience (UX). We can use various features (like vocals, instruments) of a song to classify them into different forms. However, identifying these features is a challenging task. In order to identify the features of a song that can appropriately classify them into their respective genre with reduced human interaction, we go for deep learning techniques. The proposed project is based on these techniques. Our project mainly focuses on the basis of implementing music genre recognition using Convolution Neural Network (CNN/ConvNet) model which trained using labeled Mel Spectrogram is obtained from audio dataset.*

Key Words: Genre recognition, ConvNet, Labeled Mel Spectrogram, Deep learning.

1. INTRODUCTION

Music is one of the most fascinating and important invention of the human era. Music, in a generic fashion, can be defined as a rhythmic combination of vocals and/or sound generated by instruments. Right from the beginning, music has been an important part of people's life. It is a doorway to the hearts of people. Theoretically, music is a form of data which has evolved to comprise a sophisticated pattern. The type of music we listen is dependent on the mood we're in i.e., the type of music and the mood of the listener are closely related. For emotions, we call it mood, however for music, we call it genre. We have different forms of music across the globe with drastically varying pattern, composed using variety of musical instruments that may be percussions or string instruments and unique singing patterns. There are about 250 recognizable genres in the world of music. Not every genre is totally different from one another. There are genres that belong to a same group, however, within that group, it can be differentiated based of several features. For example, Rock genre is a broad group which can be further differentiated into hard rock, progressive rock, soft rock, punk rock, pop rock, folk rock and so on. Along with that, there are thousands of songs which are created by incorporating the musical melody of different genres into one combined form. Classifying the songs into its respective genres can be done in various ways. One method is to have experts of each genre to identify. But this is a non-feasible method. However, the technological advancement particularly in the field of Computer Science have enabled us to train a machine to work like human (basic ideology of Machine Learning/Deep Learning) to do different tasks like high scale computation, classification etc. In the light of that, in our project, we propose a model that identifies the genre of the music.

It is a challenging task given that the data we need to classify is audio, which is a form of analog data. The direct digital format of the audio file cannot be used to classify the audio. Therefore, we need to represent the audio clips using more suitable format. In our project, we use Mel Frequency Spectrogram representation of the audio input. Mel Spectrogram mathematically is a short-term power spectrum of sound, based on a linear cosine transform of a log power spectrum on a non-linear mel scale of frequency. Mel Spectrogram has been standardized by the European Telecommunication Standards Institute (ETSI).

Each genre of music has a recognizable pattern which in mel spectrogram representation can be accurately identified by the model. The intensity level of the pixels in mel spectrogram representation is the deciding factor in identifying the genre of the input audio file.

2. LITERATURE

Michael Haggblade, Yang Hong and Kenny Kao have discussed the non-trivial problem of music genre classification [1]. They compared the performance of 5 different classification techniques based on accuracy, handling unstructured data and other parameters. It was observed that among the 5 techniques namely Kullback-Leiber (KL) divergence, k-Nearest Neighbor (k-NN), k-Means, Directed Acyclic Graph Support Vector Machine (DAG SVM) and Artificial Neural Network (ANN), ANN was the best technique. Even though k-NN technique had an accuracy of 96%, the accuracy dropped when the number of genres to be classified increased and there were difficulties in handling the unstructured data format of the music.

Audio clip is an analog data form. In order to perform classification of audio files, the audio file must be represented in a digital format. Haggblade used Mel Frequency Cepstral Coefficients (MFCC) to represent audio file.

Sharaj Panwar, Arun Das, Mehdi Roopaei, Paul Rad discussed the deep learning approach for mapping music genre [2]. Deep feature learning methods have been aggressively applied in the field of music tagging retrieval. Genre categorization, mood classification, and chord detection are the most common tags from local spectral to temporal structure. Convolutional Neural networks (CNNs) using kernels extract the local features that are in different levels of

hierarchy. CRNN architectures as a powerful music tagging. The AUC-ROC index for the proposed architecture is 0.893 which shows its superiority rather than traditional structures on the same database.

Narek Abroyan suggested Convolution Neural Network is viable option for real time classification problems [3]. During the last several years, artificial intelligence researchers and specialists achieved notable results in visual, voice and natural language processing tasks by using new methods and approaches of deep learning, such as convolutional neural networks. However, not much research is going on of usage such networks for elaboration of real-time data. Abroyan suggested that ConvNets can be trained and used for real-time data classification, opening doors to a wide range of areas where it can find applications.

Convolution Neural Network has been suggested as a viable technique whenever the audio input comes into picture. Che Wei Huang et al. used the ConvNet along with recurrent neural network on speech audio clips to identify the emotion of the speaker [4]. The method of discriminative feature learning employed by convolution layers uses the mel frequency spectrogram representation of audio clips and identifies the emotion of the speaker.

Wootaek Lim et al. also proposed that ConvNet is the best technique for speech emotion recognition [5]. He suggested that even though at simpler form, traditional low-level descriptors have high accuracy, these descriptors to be trained with complex unstructured data and extracting good features is a difficult task to accomplish.

3. METHODOLOGY

In order to identify the genre of a song, first we need to build a ConvNet and train the same with labeled audio dataset. We employ supervised learning method for the training of the ConvNet. In our project, we considered 10 well known music genres. For each genre, we consider about 100 audio files belonging to that genre for the training. Fig 1 shows the first phase of the project: building dataset. There are many audio datasets available in the Internet like GTZAN, FMA etc., however we chose to build our own dataset.

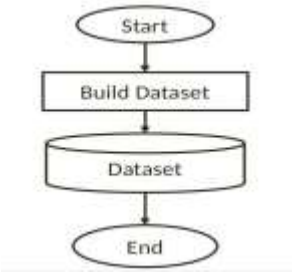

Fig -1: Build Dataset

The built dataset is then used to train the ConvNet model required for future genre recognition which is shown in Fig 2. We developed our project using Python machine learning libraries. Python provides the necessary tools for Mel spectrogram generation, building of ConvNet.

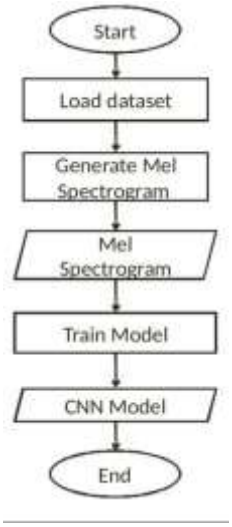

Fig -2: Train ConvNet Model

Fig 3 represents the trained model deployed for genre recognition. The trained model uses its training experience on the Mel spectrogram of test audio file and analyses it. Convolution layers of the ConvNet map the test audio file to the corresponding genre of the music.

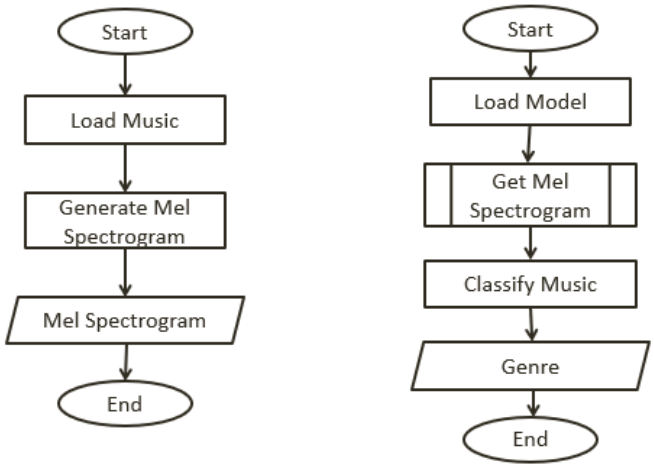

Fig -3: Classify Music

4. ALGORITHM

In our project, we have used Convolution Neural Network (CNN/ConvNet) model for the genre classification. Convolution Neural Network is a deep learning algorithm which is mainly used to differentiate images based on their spatial features. The preprocessing required in a ConvNet is much lower as compared to other classification algorithms. ConvNet is able to successfully capture the spatial and temporal characteristics in an image. The basic architecture of a ConvNet consists of Input layer, Convolution and pooling

layers, Fully Connected Activation layer. Convolution layer comprises of various filters or kernels that aid in extracting the major features of an image. The main objective is to extract the high-level features such as edges and intensity, from an input image. There can be numerous convolution layers where the first filter captures low level features like edges, color, gradient, orientation etc., and subsequent layers capture higher level features.

Pooling layer is responsible for reducing the spatial size of the convolved feature in order to decrease the amount of computational power required to process the data through dimensionality reduction. Furthermore, it is use in extracting dominant features of the input. There are two types of pooling namely Max Pooling and Average pooling. Conventionally Max pooling is preferred over Average pooling because Max pooling essentially performs Noise Suppression, which is a positive feature.

Fully Connected Activation layer is usually a cheap way of learning non-linear combinations of the high-level features as represented by the output of the convolution layers. Over a series of epochs, the model is able to distinguish between dominating and certain low-level features in the image and classify them using the Softmax Classification technique. Softmax function or Normalized exponential function is a function that takes as input a vector of K real numbers, and normalizes it into a probability distribution consisting of K probabilities.

Fig -4: Convolution Neural Network Architecture

5. EXPERIMENTAL RESULT

Mel spectrogram is generated at both training phase as well as the testing phase. In training phase, the Mel spectrogram gives an insight to the model about the level of intensity corresponding to that genre of the music. In testing phase, Mel spectrogram of the test input enables the model to classify it into corresponding genre. Note that each genre has a specific level of intensity output in the spectrogram. This feature is exploited to classify the music.

Fig 5 display the output of the model as observed in the web interface built. The input given was the well-known pop song by Adele, "Someone Like You". The corresponding output shows Pop (green) as the major genre along with hip hop and reggae as other genres.

Fig 6 shows the output of the model for a famous reggae song by Bob Marley, "The Redemption Song". The major output shown is reggae with a minor mix of pop, hip hop and blues.

Fig 7 provides the Mel spectrogram output for the pop song previously mentioned

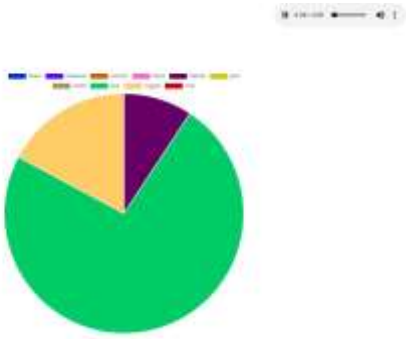

Fig -5: Screenshot of the output obtained for a pop song

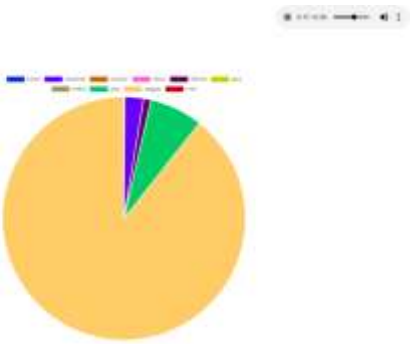

Fig -6: Screenshot of the output obtained for a reggae song

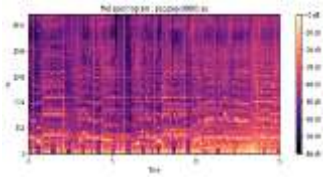

Fig -7: Screenshot of the Mel spectrogram for pop song

6. CONCLUSIONS

Genre Recognition is interesting feature in multimedia applications like music player, streaming applications etc. The project developed is completely software oriented and can be run in any PC with optimum hardware. We have implemented a Convolution Neural Network which was trained with spectrogram of audio files and the trained model was deployed in a web interface. Despite there being numerous classification algorithms and techniques, the Convolution Neural Network has a better performance over other mechanism.

REFERENCES

[1] R. Islam, Mingxing Xu and Yuchao Fan, "Chinese Traditional Opera database for Music Genre Recognition", 2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2015, pp. 38-41, doi: 10.1109/ICSDA.2015.7357861.

[2] Sharaj Panwar, Arun Das, Mehdi Roopaei, Paul Rad, "A deep learning approach for mapping music genres", 12th System of Systems Engineering Conference (SoSE), 2017, pp. 1-5, doi: 10.1109/SYSOSE.2017.7994970.

[3] Narek Abroyan, "Convolutional and recurrent neural networks for real-time data classification", 2017 Seventh International Conference on Innovative Computing Technology (INTECH), 2017, pp. 42-45, doi: 10.1109/INTECH.2017.8102422.

[4] Che-Wei Huang, Shrikanth Shri Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition", july 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 538-588, doi: 10.1109/ICME.2017.8019296.

[5] Wootaek Lim, Daeyoung Jang and Taejin Lee, "Speech Emotion Recognition using Convolutional and Recurrent Neural Networks", In 2016 Annual Conference of the International Speech Communication Association, 2016, pp. 1-4.