Music Genre Classification Using CNNs

1st Ruchira Ray
Computer Science Engineering
Next Tech Lab
SRM Institute of Science and Technology
ruchiraray99@gmail.com

2nd Mrunal Sonawane
Electrical and Electronics Engineering
Next Tech Lab
SRM Institute of Science and Technology
mrunal.s15@gmail.com

3rd Sriram Vijendran
Electronics and Communications Engineering
Next Tech Lab
SRM Institute of Science and Technology
vijendran.sriram@gmail.com

4th Sree Harsha Nelaturu
Electronics and Communications Engineering
Next Tech Lab
SRM Institute of Science and Technology
nelaturu@mit.edu

***Abstract*—Music Genre Classification is one of the most popular open problems in Music Information Retrieval (MIR). One of the key challenges of dealing with audio data in the time-domain is preserving the phase information of the samples and another key weakness is the large amount of information present in one audio sample. The task of genre classification often requires complex algorithms which need a significant amount of time to run and hence are not scalable. In order to address this, we use the spectrogram based features of audio, from a curated dataset, to train a deep convolutional neural network on a GPU to classify them according to their genres. This frequency domain based representation not only allows for higher accuracy of classification, but also enables near real-time classification of data not represented in the dataset, even without hardware acceleration.**

***Index Terms*—CNN, genre classification, short-time Fourier transform (STFT), sampling, spectrogram, Nyquist Shannon, backpropagation, deep learning.**

I. Introduction

Music is an art form experienced and savored by everyone on this planet. Music can be divided into different genres in many different ways. Some popular genres include pop, rock, country, jazz, R&B, hip-hop, etc. The genre of a particular song as defined in the Cambridge Dictionary is a style, especially in the arts, that involves a particular set of characteristics which describes it [1].

Categorizing music into these conventional genres having indistinct boundaries has always been a challenging task. Despite the lack of a detailed criteria for defining genres, the classification of music based on genres is one of the most commonly used. Genre classification gives us the ability to be as descriptive as possible about someone's music. Being able to give some musical points of reference in the form of a genre description is critical. Understanding genre is essential for understanding audiences. Fans of a particular genre tend to have similar interests and tend to flock to the same kinds of venues , same kinds of shops, listen to the similar radio stations, watch the same movies and use the same websites. This kind of information is vitally important when it comes to promoting and marketing music and putting together live shows. Identifying genre also help musicians and their representatives choose the right labels to approach with their music [2].

The borders between genres have blurred with time. New genres such as indie rock, steampunk, EDM etc. and hybrid genres like fusion jazz, electro swing, etc. have come up. A genre is a collection of patterns thus songs that share similar patterns can be grouped together. These patterns can be found in any music dynamic. The recent success with deep neural network architectures on large-scale datasets has inspired numerous studies in the machine learning community for various pattern recognition and classification tasks such as audio classification, automatic speech recognition, natural language processing and computer vision.

II. Approach

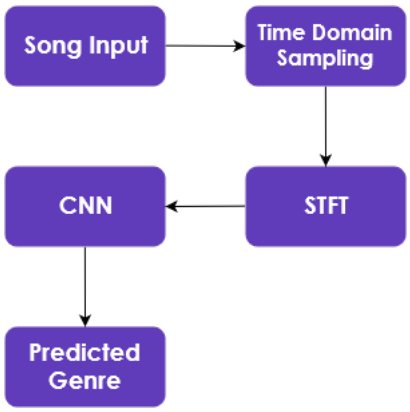

Fig. 1. Pipeline for Approach

A. Data Pre-processing

The GTZAN dataset [3] has 1000 music files divided in 10 genres. The sampling rate is 22050 Hz thus a 30-second song contains 661500 samples. To reduce this large vector of data

to a more reasonable amount, we discard the stereo channel to reduce redundancy. With the help of LibROSA, a python package for music and audio analysis, we convert the audio to a mel-spectrogram (LibROSA provides the basics necessary to create music information retrieval systems). By the Nyquist Shannon sampling criterion, the 22050 Hz sampling rate allows us to reconstruct frequencies up to 11025 Hz. The parameters used to create the spectrogram are n_fft of 2048, hop length of 512 and are max normalized. The spectrogram is then converted to chunks which are of the shape 1 x 128 x 128. These chunks are then shuffled randomly and converted into pickles. The Train-Test-Validation splits in the ratio 70:20:10.

1) Dataset: The dataset used is the GTZAN dataset which was published in 2002 in [4] for the purpose of automated music genre classification using audio signal processing techniques. This dataset contains 1000 tracks, divided into 10 genres, each having 100 tracks and each lasting 30 seconds. These tracks are all 22050 Hz Mono 16-bit audio files in .wav format.

This dataset has been used in over 100 publications, as it is one of the very few large-scale datasets that are publicly available. Even though there are several limitations to this dataset,as discussed by Bob L. Strum in [5], it is still a popular choice for audio processing as it provides raw data and is publicly available unlike other datasets like 1 million song [6]. This dataset is compact, resulting in efficient training, testing and evaluation.

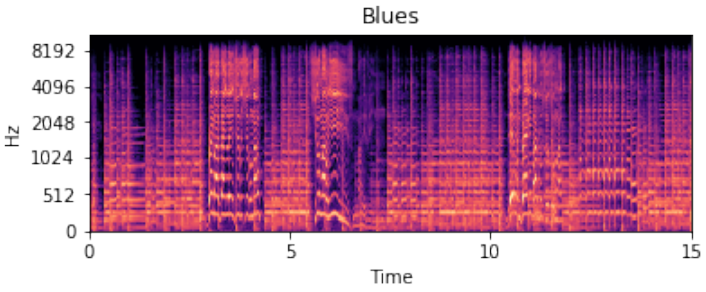

Fig. 2. Spectrogram of a blues track

2) Spectrogram: A spectrogram is a visual representation of any signal and its frequency spectrum. It allows one to make inferences about the energy of a signal as a function of time and the variation of energy with time [7].

$$S(t, w) := |STFT(t, w)|^2 \tag{1}$$

The vertical axis of a spectrogram represents frequency and the horizontal axis represents time. A spectrogram serves as the input for a convolutional neural network.

B. Model

As depicted in Fig. 1, an audio file of length 30 seconds is given as an input to the model. A time domain sampling operation is performed on the given audio producing a sequence of samples. STFT is then applied to the sequence of samples to calculate the magnitude of frequency spectrum for each sample. The outputs are then concatenated to form an image called the Spectrogram. The spectrogram generated is then analyzed by the model by first forward propagating through the image and then backpropagating the error between the expected and predicted label (genre) of the sound. Iterating the process for the whole dataset results in a model which can recognize patterns in the spectrogram and hence predict the genre of the audio.

1) Convolutional Neural Networks: A convolutional neural network (CNN) [8] is a specific type of artificial neural network that uses perceptrons, a machine learning unit algorithm, for supervised learning, to analyze data. CNNs apply to image processing, natural language processing and other kinds of cognitive tasks. Like other kinds of artificial neural networks, a convolutional neural network has an input layer, an output layer and various hidden layers. Some of these layers are convolutional, using a mathematical model to pass on results to successive layers. This simulates some of the actions in the human visual cortex.

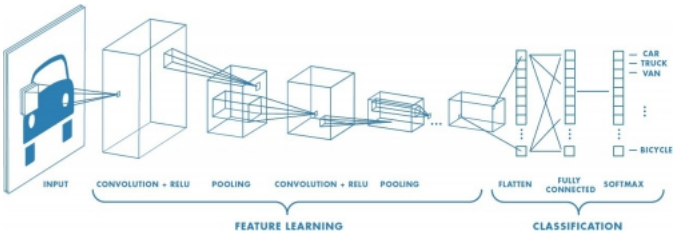

Fig. 3. Convolutional Neural Network

a) Neural Networks: An artificial neural network (ANN) is based on the neural network of the human brain [9]. It follows the same idea of multiple neurons being connected to each other and passing on information. They are intended to replicate the way humans learn. ANNs are extremely good in finding out patterns in data which humans cannot.

b) Deep Learning: Deep learning is a sub field of machine learning [10]. Deep learning works on the concept of what comes naturally to humans: learn by example. The main goal accomplished by deep learning is that it learns patterns in the data. Most popular applications of deep learning are self-driving cars, virtual assistants, facial recognition, medical research, etc.

c) Backpropagation: Backpropagation is a method used in deep learning to calculate a gradient that is needed in the calculation of the weights to be used in the network [11]. It is the partial derivative of the loss function with respect to a weight in the network ($\frac{\partial L}{\partial w_{ij}}$).

III. Testing and Evaluation

A binary technique is used to infer the results of testing, one includes using the dataset itself and the other includes using songs from other sources. A similar pre-processing is applied to the song before attempting to classify them. The inference time for a full-length song averaged at 17s. This demonstrates the capability to run inference in near real-time

using pre-trained deep learning models. The main indicators of its confusion were perceivably similar genres or hybrid genres. For example, in the case of modern pop songs, their features tend to correlate with rap at the time of the creation of the dataset.

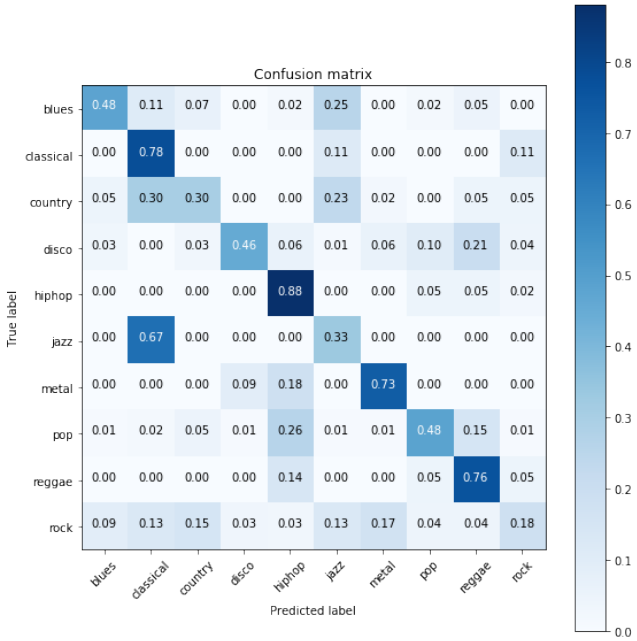

Fig. 4. Confusion Matrix

IV. Result and Conclusion

In this work, we used a CNN to classify music by generating a spectrogram of the particular audio. The reported training accuracy of the model is 96% and test accuracy is 82%. We observe that the network performed reasonably well during the testing.

We used a CNN based image classifier to classify music and we found that it is much faster and more accurate than the statistical pattern recognition.

One improvement we can make to this work is to use a CRNN model [12]. A CRNN is a hybrid of a CNN and an RNN. A CNN acts like a feature extractor and the RNN acts like a temporal summarizer. This model would allow for a better performance with respect to the number of parameter and training time, indicating the effectiveness of its hybrid structure.

Acknowledgment

We would like thank SRM Institute of Science and Technology, the Chancellor and Director E&T for guiding us and providing resources to help us with the paper and also for facilitating this opportunity.

References

[1] *Definition of genre*
https://dictionary.cambridge.org/dictionary/english/genre

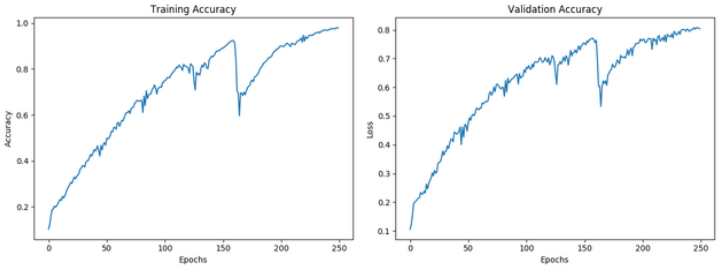

Fig. 5. Accuracy of the network trained on GTZAN dataset

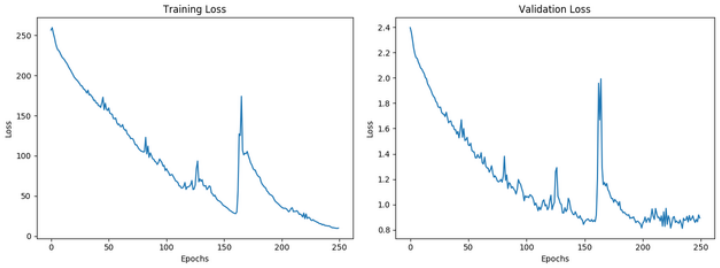

Fig. 6. Loss of the network trained on GTZAN dataset

[2] *Importance of genre classification*
https://www.thebalancecareers.com/music-genre-what-is-it-and-why-does-it-matter-2460500
[3] *GTZAN Genre Collection Dataset*
http://marsyas.info/downloads/datasets.html
[4] George Tzanetakis, Student Member, IEEE and Perry Cook, Member, IEEE *Musical Genre Classification of Audio Signals.* 2002
[5] Bob L. Sturm *The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use.* 2013
[6] *Million Song Dataset*
https://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset#subset
[7] Mller, Meinard
Fundamentals of Music Processing
[8] *Convolutional Neural Networks*
https://sushscience.wordpress.com/2016/12/04/understanding-alexnet/
https://www.techopedia.com/definition/32731/convolutional-neural-network-cnn
[9] *Artificial Neural Network*
https://www.techopedia.com/definition/5967/artificial-neural-network-ann
[10] *Deep Learning*
https://www.mathworks.com/discovery/deep-learning.html
[11] *Backpropagation*
https://en.wikipedia.org/wiki/Backpropagation
[12] Keunwoo Choi, Gyorgy Fazekas, Mark Sandler and Kyunghyun Cho *Convolutional Recurrent Neural Networks for music classification*