DEEP LEARNING FOR MUSIC GENRE RECOGNITION

Mikhail Belov

Columbia University

ABSTRACT

Automated Music Genre Recognition has been a challenging and interesting problem in the field of Music Information Retrieval that is yet to be resolved. Having read a variety of papers that present their approaches to the task at hand, it was apparent that there is a potential for an improvement of the current algorithms deployed for this classification task. Moreover, most of classification results presented in these papers are neither impressive nor compelling which served as a motivation for contributing to this area of MIR. With Deep Learning emerging as one of the most popular and promising Machine Learning techniques, it was decided to use Deep Neural Networks for the implementation of Music Genre Recognition algorithm.

In this paper, a new Convolutional Neural Network Architecture is proposed, and its performance is evaluated on three different music features: tempogram, chroma frequencies, and spectrogram. The goals of this research paper are to determine which of the three features yields best classification performance on the proposed CNN, and to provide a discussion on the reasons why certain music attributes are more meaningful for discerning music genres than others.

Index Terms — MIR, Deep Learning, DNN, CNN

1. INTRODUCTION

Music Genre Classification is a problem of recognizing a song genre based on the audio content of the song. It is a not a trivial task to tackle as there is a lot of overlap between musical genres. It is hard to draw a fine line between each music genre even for an experienced musician, not to mention a computer algorithm. Consequently, designing an autonomous system that can capture the musical differences in a given song comes with some challenges.

One of these challenges is identifying a set of features that would represent the distinctions between music genres in the most compact and efficient way. After going through numerous research papers on Music Genre Recognition, I figured out that the trend is to use MFCCs [1], mel-spectrogram [2], zero crossing rate [1], spectral roll-off [1] and chroma frequencies [3] as music genre descriptors. Even though MFCCs and mel-spectrograms have been most commonly used in this area of research, I am skeptic about these features being efficient for music genre recognition since these features are more suitable for speech recognition. Significant audio signal attributes lie in a wider range of frequency spectrum, and, therefore, their extraction cannot be handled using the same approaches.

Consequently, I decided to use tempogram, chroma frequencies and spectrogram as music features after further research on Music Information Retrieval. All three features are intuitively relevant music genre descriptors from both musical and signal processing perspective.

Recent major progress in the field of Machine Learning and AI made it possible to find complex patterns in large amounts of data. Deep Neural Networks have already produced remarkable results in a variety of classification tasks. One of deep learning techniques – Convolutional Neural Networks – proved to be effective in the field of computer vision [4]. Since tempogram, spectrogram, and chroma frequencies are all visual representations of an audio signal, I decided to implement my own Convolutional Neural Network architecture that would perform music genre classification on these music features.

2. METHODS

2.1. Features

Tempogram, which is a time-pulse representation of a music signal, is constructed by applying Fourier Transform to the onset detection function that characterizes musical events in an audio track [5]. This way local tempo of a music piece is estimated across its entire length. Because of the correlation between a song's genre and its local tempo, local autocorrelation of onset strength envelope provides us with patterns that are distinctive and relevant to music genre recognition which is exactly what we are looking for in a good feature.

Chroma frequencies feature is a representation for music audio in which the entire spectrum is projected onto 12 bins representing 12 distinct semitones (or chroma) of the musical octave [3]. Since there are definitive rhythmic patterns and musical note intervals of certain music genres, chroma frequencies is a useful feature for genre classification.

Spectrogram is a time-frequency representation of frequency spectrum of a signal [6]. It is one of the most fundamental tools in the area of digital signal processing, and it proved to be efficient for music genre recognition as well [3]. The process of chroma frequencies and spectrogram generation is similar since both features are constructed by

projecting an audio signal onto N number of bins. However, as it is discussed later in the paper, spectrogram is a much more powerful feature for music genre recognition than chroma frequencies because of higher number of bins that the audio signal is projected onto.

5.1. Classification via Supervised Deep Learning

Convolutional Neural Network is deployed as both a mechanism for extracting low-level features from the visual representations of the signal, and a classification algorithm. The proposed CNN architecture has three hidden convolutional layers. Max pooling is performed between the convolutional hidden layers to resample the features. Then, last convolutional layer is followed by the fully connected layer and N-unit output layer. Number of neurons in the output layer corresponds to the number of genres that the songs are classified into. After applying Softmax function to the output layer, genre prediction probabilities are calculated for a given music track.

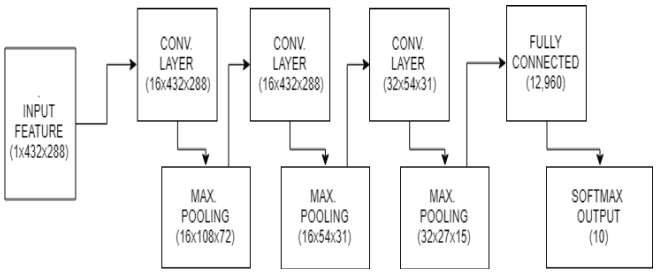

Fig. 1. Proposed CNN Architecture.

Two datasets were used for feature extraction and training, validation and testing of the CNN: GTZAN [7] and FMA (Free Music Archive) [8]. GTZAN consists of 1000 30-second song segments of 10 genres: blues, classical, country, metal, rock, reggae, pop, hip-hop, disco and jazz. FMA consists of 8000 30-second segments of 8 genres: pop, rock, folk, electronic, instrumental, international, hip-hop and experimental. Both datasets are balanced and commonly used in other MIR projects and papers.

As a way of expanding these datasets, five features per track have been extracted from the music segments in GTZAN and FMA. Having more data to train a DNN allows for better classification accuracy and prevents the model from overfitting. To ensure no overlap between the extracted features from a single audio sample, the features were extracted at different 5-second time intervals.

While experimenting with parameters of feature extraction, it was determined that hop length of 512 and 22,050 Hz sampling rate produce optimal tempogram for 5-second audio in terms of detail preservation and feature matrix size. Same hop length and audio sampling rate was used when extracting chroma frequencies.

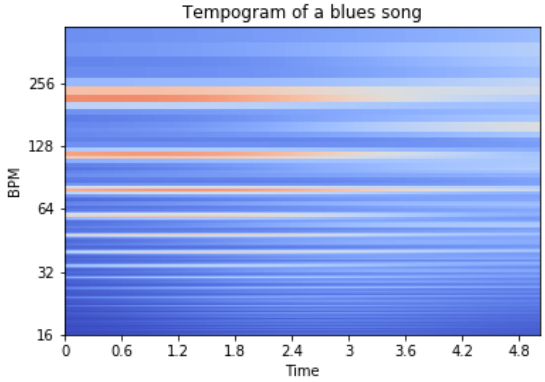

Fig. 2. Blues Song Tempogram.

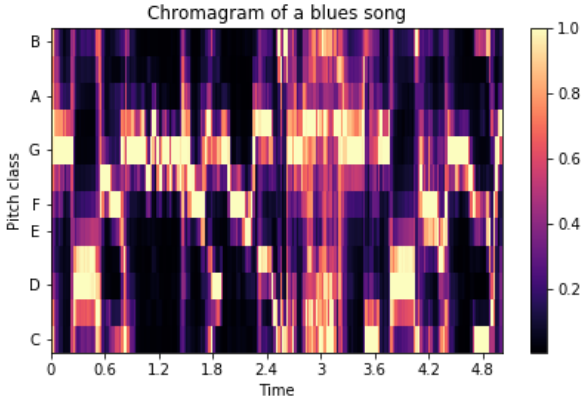

Fig. 3. Blues Song Chromagram.

2048 FFT length, 128-point overlap and Hann window were used as input parameters when calculating spectrograms of an audio signal.

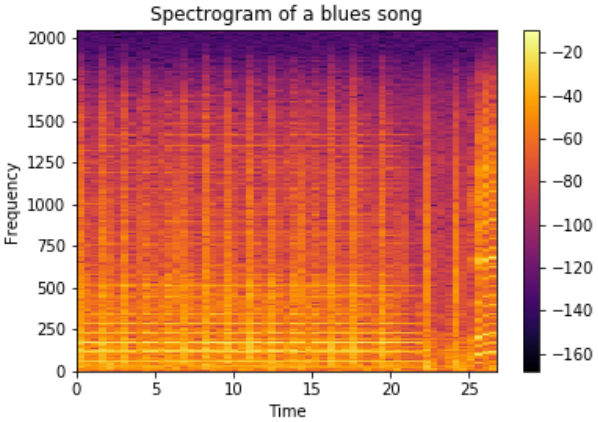

Fig. 4. Blues Song Spectrogram.

3. RESULTS

Proposed CNN model was trained on expanded GTZAN and FMA databases. It produced suboptimal genre classification accuracy on FMA (maximum of 40% accuracy using spectrograms) despite the type of input features that were used for model training. Intuitive reason for this outcome is a significant amount of overlap between some genres presented in the dataset. There are a lot of similarities in musical attributes between experimental and electronic music genres such as common use of computer-generated sounds which is reflected in the features' visual content. Same argument can be made regarding international and folk genres, where instrumental music can be often mistaken for folk music, and the other way around. Therefore, CNN model was not able to learn enough meaningful patterns that would allow it to discern certain musical genres with higher accuracy.

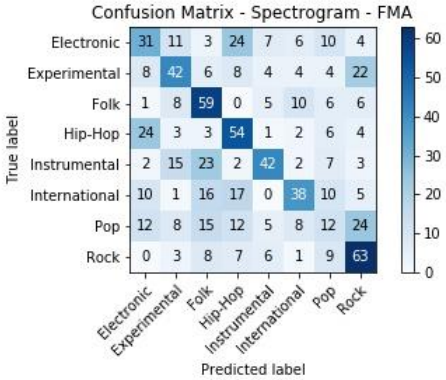

Fig. 5. Confusion matrix of classification results using Spectrogram on FMA (classification accuracy in percentages).

Major improvement in classification accuracy was achieved on GTZAN dataset. GTZAN has been widely used in the field of MIR by other researchers which proves its overall quality and, consequently, can be considered a benchmark. Spectrogram turned out to be the best feature out of the three considered for this classification task. CNN model that was trained on songs' spectrograms reached 70% accuracy classifying 10 genres, while tempogram and chroma frequencies provided only 49% and 42% accuracy, respectively.

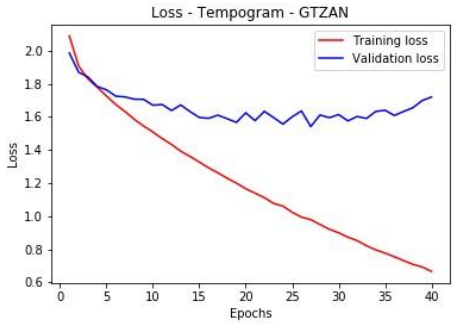

Fig. 6. CNN Cross Entropy Loss on GTZAN using song's tempogram.

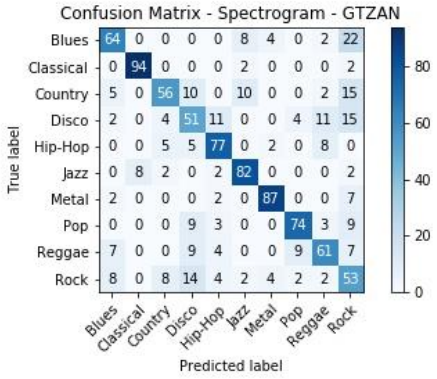

Fig. 7. Confusion matrix of final classification results (classification accuracy in percentages).

Dataset	Audio Feature			
	Tempogram	Spectrogram	Chroma Frequencies	Accuracy
GTZAN	49%	70%	42%	
FMA	38%	43%	33%	

Fig. 8. Genre Classification Accuracy Table.

4. CONCLUSION

Music Genre Recognition continues to be a difficult problem in the field of MIR. Machine Learning and Deep Learning in particular, make it feasible to reach high genre classification accuracy, however, quality and size of the dataset used for training Deep Learning models is crucial for achieving optimal test accuracy.

Also, spectrogram proved to be the most effective feature for MGR. This outcome makes sense in retrospect. Essentially, CNN model learned from spectrogram images how loud on average a music track of each genre is, and it would identify song's genre based on this parameter. This argument can be supported by high classification accuracy of extremely loud music genres such as metal, and relatively quiet genres such as jazz and classical music.

Taking this project a step further, it would be interesting to research other audio signal features that are more compact than spectrogram, and provide more or just as much information about the loudness of a music piece, and use it as an input feature for Convolutional Neural Network.

In addition, merging multiple datasets together and experimenting with other DNN models is something to consider while moving forward with the research in this area of MIR. CNN+LSTM models that are highly effective in computer vision tasks of action recognition and image description might be worth trying and experimenting with on a large music database [9].

12. REFERENCES

[1] Bahuleyan, Hareesh. “Music Genre Classification Using Machine Learning Techniques.” *ArXiv.org*, 3 Apr. 2018, arxiv.org/abs/1804.01149v1.

[2] Lachmish, Matan. *Music Genre Recognition Using Deep Learning*. 1 May 2016, github.com/mlachmish/MusicGenreClassification.

[3] Pandey, Parul. “Music Genre Classification with Python.” *Towards Data Science*, Towards Data Science, 13 Dec. 2018, towardsdatascience.com/music-genre-classification-with-python-c714d032f0d8.

[4] Lee, Joseph. “Intuitive Deep Learning Part 2: CNNs for Computer Vision.” *Towards Data Science*, 24 Feb. 2019, towardsdatascience.com/intuitive-deep-learning-part-2-cnns-for-computer-vision-472bbb2c8060.

[5] Tian, Mi, and Mark Sandler. “On the Use of Tempogram...” *On the Use of the Tempogram to Describe Audio Content and Its Application to Music Structural Segmentation - IEEE Conference Publication*, 2015, ieeexplore.ieee.org/document/7178003.

[6] “Spectrogram.” *Wikipedia*, Wikimedia Foundation, 23 Apr. 2019, en.wikipedia.org/wiki/Spectrogram.

[7] Tzanetakis, George. “GTZAN Dataset.” *Marsyas.info*, marsyas.info/downloads/datasets.html.

[8] Defferrard, Michaël. “Mdeff/Fma.” *GitHub*, 9 May 2017, github.com/mdeff/fma.

[9] Brownlee, Jason. “CNN Long Short-Term Memory Networks.” *Machine Learning Mastery*, 19 July 2017, machinelearningmastery.com/cnn-long-short-term-memory-networks/.