Music Genre Recognition using Deep Learning Methods

Anonymous submission

Paper ID

Abstract

Today, music represents an important, and probably inseparable part of a person's life. There exist many genres of music, each of them having different aspects. This difference between genres raises an important problem, the one of classifying songs based on these genres. One approach to this problem is based on the acoustic characteristics of music, combined with Deep Learning principles such as the Convolutional Neural Network (CNN).

1. Introduction

1.1. Context

In recent years, we have seen the rise of music streaming platforms, such as Spotify, or Apple Music. These platforms not only provide high quality music, they also offer their users a large variety of songs, the giants of this industry providing up to 100 million songs. Such large music collections create the challenge of how to retrieve, browse and recommend songs to a user. One way to ease the access of large music collections is to classify the songs based on their corresponding genre.

Music genres are useful to organize and classify songs, albums, or artists into groups that share similar musical characteristics. Music genres have been used way before the rise in popularity of the streaming services, thus the problem of music classification has existed for a long time. The rapid increase in the size of data collection has made the topic of automatic music genre classification ever more relevant.

1.2. Motivation

Music Genre Recognition represents an important tool when it comes to large music data sets, which are very common these days, due to the surge in popularity of music streaming platforms. These platforms are mostly interested in the classification of music genres, both because it represents an effective method of grouping songs based on their features, and because it can represent an improvement of the recommendation system that most of the platforms offer to their users.

1.3. Objectives

The objective of this project is to create an application that is accurate enough so that it can take a segment extracted from any given song and be able to classify that song into one of the genres provided by the data set.

2. Literature Review

As the automatization of music genre recognition increased in popularity, more solutions to this problem have appeared that primarily involve Deep Learning principles. The most common method used, is that of using Convolutional Neural Networks (CNN).

The first significant work on this topic was done by Tzanetakis and Cook[1]. They applied time-frequency analysis methods such as Mell Frequency Cepstral Coefficient (MFCC), Spectral Centroid and wavelet transformations in the feature extraction process on the GTZAN Dataset.

Over the last decade, there has been a surge in Convolutional Neural Network (CNN) architectures, which achieved satisfying performances in fields like image recognition and language processing. Similar to images, music also consists of hierarchical structures, which justify a possible use of CNNs when dealing with the problem of music classification.

3. Data

3.1. Dataset

The used dataset is the GTZAN dataset. This dataset has been widely used in many studies with the aim of music genre classification. It contains one thousand music segments, with a sampling frequency of 22,500 Hz, 16 bits resolution and a duration of 30 seconds. The genres present in the GTZAN dataset are blues, classical, country, disco, hiphop, jazz, metal, pop, reggae and rock. Each genre contains

100 different examples.

In order to achieve a higher accuracy on our model, a larger collection of data is needed. Therefore, each music file will be split into 10 pieces, each having roughly 3 seconds. This leads to an increase of the dataset from 1000 sound files to 10000.

Considering the time domain representation, some music genres can be distinguished quite simply, as it can bee seen in Figure 1. The jazz genre clearly stands out when being compared to blues or reggae. However, some types of music can have very similar characteristics that add to the challenge of music classification. Therefore, more informative features have to be utilised in the classification process, such as the spectrograms or MFCCs, which take into consideration both time and frequency.

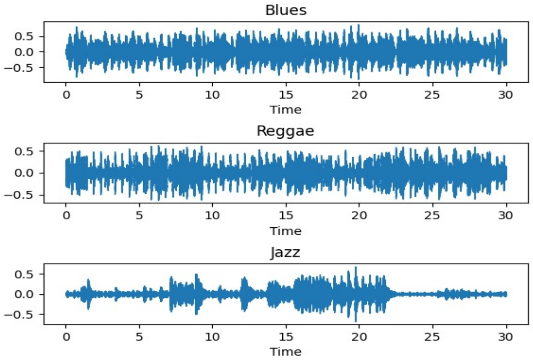

Figure 1. Differences and similarities between the waveforms of music genres

3.2. Data pre-processing

For each audio file in the dataset, a label has been attributed, corresponding to the genre it belongs to. Then, for each of the ten segments of an audio file, a MFCC is computed, which represents the main feature that the application uses.

Mel Frequency Cepstral Coefficients, or MFCCs, are a frequency domain feature that capture timbral and textural aspects of the sound. An advantage that the use of MFCCs provides is that they approximate the human auditory system. Originally introduced for speech recognition, they have recently been also introduced in music analysis.

4. Proposed solution

For the problem of Music Genre Recognition, a specific class of Artificial Neural Networks is used: the Convolutional Neural Network (CNN). The CNN is widely used in image and video recognition, image classification, medical image analysis, natural language processing and many others. CNNs use relatively little pre-processing, when compared to other image classification algorithms, thus making this approach a simple, yet effective one. This type of Neural Network is often compared to the way the brain achieves vision processing in living organisms.

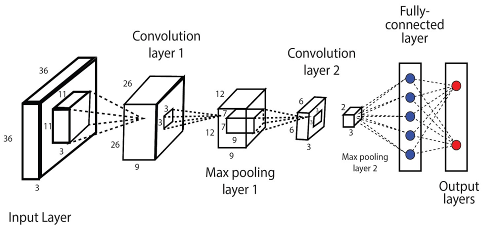

Figure 2. Architecture of a CNN

A convolutional neural network consists of an input layer, several hidden layers, and an output layer. For this experiment, the following layout has been implemented:

- The initial model is a Sequential model.
- 3 two-dimensional Convolutional Layers are added. Each of these layers uses the ReLU activation function.
- After each Convolutional Layer, a Max Pooling Layer and a Dropout Layer are added.
- A Dropout Layer is also added, but only for the 2nd and 3rd layers of the CNN.
- Each Dropout Layer has a probability of 0.3.
- The output of these convolutional layers is then flattened and fed into a Dense Layer. Flattening is used here in order to obtain a one-dimensional array.
- The final layer is a Fully Connected Layer, using the ReLU activation function as well.
- Finally, the Output Layer is a Dense Layer, having 10 units, to match the number of genres in the dataset. The activation function used for this Layer is SoftMax.

5. Experimental Results

Since the GTZAN data set consists of ten different genres, accuracy was used as the main performance metric.

Upon experimenting with various data in the structure of the CNN, the following results have been obtained:

Acc.	Modifications
68%	Use Dropout only on one layer
69%	0.00005 Adam learning rate
70%	Test-Validation split to (0.2, 0.4)
75%	Use Batch Normalization instead of Dropout
80%	Final setup

Table 1. Validation accuracy comparison, 300 epochs run.

The second table also shows results of the proposed model compared to other state-of-the-art systems. Grzegorz and Grzywczak [2] have obtained an accuracy of 78% by extracting features from spectrograms using a deep convolutional network trained for image classification. Afterwards, they used a SVM for genre prediction. Baniya et al. [3] reported an accuracy of 87.9% using rich statistics and low-level music features. In [4], Panagakis et al. used rich, psycho-psychologically inspired properties of music along with a sparse representation based classifier to achieve an accuracy of 91%. The same authors achieved an accuracy of 93.7% in [5] by using topology preserving non-negative tensor factorization.

System	Obtained accuracy
Proposed System	80%
Grzegorz and Grzywczak [2]	78.0%
Baniya et al. [3]	87.9%
Panagakis et al. [4]	91.0%
Panagakis et al. [5]	93.7%

Table 2. Comparison with state-of-the art systems

In the end, it is observed that the accuracy of this model can vary, depending on small changes in the structure of the CNN. Further testing would probably increase the accuracy of the model, but, for the moment, the obtained accuracy represents a satisfactory result. Nevertheless, this result comes close to state-of-the-art models that implement methods that are more complex, such as the Support Vector Machine (SVM).

With rigorous examples and case studies, it is demonstrated that the perfect system in the GTZAN dataset would not be able to surpass the accuracy score of 94.5% due to the inherent noise in the some of the repetitions, mis-labelings and distortions of the songs.

5.1. CNN Drawbacks

Related to the problem of Genre Recognition, the CNN performs better than other Neural Networks, such as the LSTM (Long Short-Term Memory) [6]. Despite this, as a general Neural Network, the CNN has a few drawbacks.

- Large amounts of training data are required for a CNN to be effective.
- In case of multiple layers, the training process can be lengthy.
- CNNs tend to be slower in general.

6. Conclusion

In conclusion, the model proposed in this paper represents a simple, yet effective method of solving the issue of Music Genre Recognition by using Convolutional Neural Networks.

As it can be observed, variations in the structure of the CNN, such as the number and positioning of the Dropout Layers, changing the number of epochs or batch size, can improve the accuracy provided by this model, thus coming close to the existing state-of-the-art methods.

7. References

[1] Tzanetakis, G., and Cook, P.: ‘Musical genre classification of audio signal’, IEEE Trans. Speech Audio Process., 2002, 10, (3), pp. 293–302

[2] Y. Panagakis and C. Kotropoulos, “Music genre classification via topology preserving non-negative tensor factorization and sparse representations,” in Acoustics speech and signal processing (ICASSP), 2010 IEEE international conference on. IEEE, 2010, pp. 249–252

[3] Y. Panagakis, C. Kotropoulos, and G. R. Arce, “Music genre classification via sparse representations of auditory temporal modulations,” in Signal Processing Conference, 2009 17th European. IEEE, 2009, pp. 1–5.

[4] B. K. Baniya, J. Lee, and Z.-N. Li, “Audio feature reduction and analysis for automatic music genre classification,” in Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on. IEEE, 2014, pp. 457–462.

[5] G. Gwardys and D. Grzywczak, “Deep image features in music information retrieval,” International Journal of Electronics and Telecommunications, vol. 60, no. 4, pp. 321–326.

[6] G. Gessle and S. Åkesson, ”A comparative analysis of CNN and LSTM for music genre classification”, 2019