Music Genre Classification using Convolutional Neural Networks (CNN)

Saurabh Palve, Sourabh Dubey, Manas Dhanait, Nitish Purswani, Kushal P. Birla

Student, Assistant Professor
Department of Computer Engineering
K. K. Wagh Institute of Engineering Education and Research, Nashik, India

Abstract: People’s life depends a lot on music. Like-minded people come together via music, and it serves as the community’s glue. The types of songs that are written or heard within a community might be used to identify that community. The proposed system can identify different music genres in audio recordings. These audio recordings will be categorized using low frequency attributes and sound wave attributes. We plan to create a Convolutional Neural Network (CNN) model for this project and train it using the GTZAN database. 1000 different types of audio tracks of 30 seconds each are available in the database. It has ten different music genres, each of which has 100 tracks. The tracks are all in .wav format.

Index Terms – Music, Genre, GTZAN, Frequency, Sound.

I. INTRODUCTION

The main idea is to design and develop a system which will recognize musical genres from audio data. These audio recordings will be categorized based on their low-level frequency and time domain characteristics into one of the 10 classes which are the most popular genres in music. Music of various genres is constantly created by numerous creators worldwide. For music aspirants and creators, it is very important to have the knowledge of genres in music. Be it for the one they are listening to, or the one they are creating. This tool will help music enthusiasts worldwide to recognize genres of any music or audio file they wish to know. This mere fact was a major motivation for the development of this idea.

II. LITERATURE REVIEW

“Music Genre Classification” has been published on a project for Music Genre Classification using the GTZAN dataset, where Jitesh Kumar Bhatia and his co-authors use the K-Nearest Neighbor (KNN) algorithm in order to classify music successfully into 10 genres. The KNN algorithm is a supervised learning algorithm that helps in the modeling of a classifier based on the proximity of a central data point. [1]

“Music Genre Classification using Machine Learning” is a study by Anirudh Ghidiyal and his co-authors in which the proposed models are also trained and compared using the GTZAN dataset. Most of the models are learned using audio file trains, while a small number of models are trained using spectrograms. The study represents the usage of a Convolutional Neural Network, optimized by the Adam Optimization algorithm in order to boost the accuracy and efficiency of the overall results. To update network weights more effectively, Adam optimization, an extension of stochastic gradient descent, can be employed in place of the more traditional stochastic gradient descent. [2]

“Music Genre Classification of audio signals” by P. Cook and G. Tzanetakis is a study in which three main feature sets namely timbral texture, rhythmic content and pitch content are proposed and are used to classify the music in one type of genre and by using the feature sets, classification of 61% for ten musical genres is achieved. This result is comparable to results reported for human musical genre classification. A beat histogram is used to identify the difference in genres. For classification purposes, a number of standard statistical pattern recognition (SPR) classifiers are used and the results are compared with human music genre classification. Using the proposed feature sets classification of 61% (nonreal time) and 44% (real time), was achieved in a dataset consisting of ten musical genres. [3]

“Music Genre Classification using Machine Learning Algorithms: A comparison" is a paper which demonstrates a trade-off between classification accuracy and understandability of the music features. The authors have focused on examining and describing music genres in a quantitive way using various ML models. By leveraging the MFCC features from GTZAN and human-understandable features from Spotify, this paper demonstrates a trade-off between classification accuracy and understandability of the music features. [4]
“Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification” by Mingwen Dong from the Rutgers University in the United States has stated that Human level accuracy in Music Genre Classification is 70% and a normal classification model achieves max of 60% accuracy to overcome this problem. CNN is used on 1000 music tracks which are converted into Mel-spectrogram and are evenly split into training, validation, and testing set give a result 70% accuracy which is equal to the Human level accuracy. [5]

In the study of Tarannum Shaikh and his co-authors named “Music Genre Classification using Neural Network”, a CNN model is used with 2 Datasets namely GTZAN dataset and Musicnet dataset to compare the accuracy and loss among the both datasets based on the spectrogram images generated from the songs time-slices extracted from the respective datasets. Accuracy and Loss attained on GTZAN dataset was 92.65% and 57.37% respectively, meanwhile Accuracy and Loss attained on Musicnet dataset was 91.70% and 25.46% respectively. [6]

“librosa: Audio and Music Signal Analysis in Python” was reviewed for a deeper study of the Librosa python library. Research field of music information retrieval is introduced with Python and to ease the transition of music information retrieval researchers into Python; Librosa: a Python package for audio and music signal processing is made. The design considerations and functionality of librosa are mentioned. The project is under active development to make efficiency improvements and enhanced functionality of audio coding. Use of the library Librosa is also promoted for music and audio analysis for the music information retrieval researchers. [7]

III. PROBLEM STATEMENT AND METHODOLOGY

3.1 Problem Statement
To design and develop a system using Convolutional Neural Networks (CNN) which will recognize musical genres from audio data based on their Mel Frequency Cepstrum Coefficients (MFCCs) numerical values.

3.2 Methodology

3.2.1 Dataset Preparation
The GTZAN dataset was downloaded from Kaggle and it was made sure it has all audios and corresponding labelled audio genres in it including rock, pop, jazz, disco etc. An audio file in jazz named “jazz.00054” was removed as it was corrupted. This was a part of dataset cleaning. The dataset was split into training and testing sets with 70%-30% division respectively.

3.2.2 Pre-processing
The audio samples were converted into suitable CNN input which includes extracting Mel Frequency Cepstrum Coefficients (MFCCs): a widely used feature extraction technique that captures spectral characteristics of the audio, and normalizing them.

3.2.3 Model Architecture Design
A suitable CNN architecture was chosen which consists of convolutional layers, pooling layers, and fully connected layers. Experiments with different configurations were done, such as varying the number of layers, filter sizes, and activation functions, to find the optimal architecture. Techniques like dropout and batch normalization were used to prevent overfitting.

3.2.4 Model Training and Evaluation
The CNN model was initialized with appropriate layers and neurons and random weights. The pre-processed audio samples were fed into the CNN model by passing their features as input in the form of array. The model was optimized by using the “sparse categorical crossentropy” loss function and Adam Optimizer of Keras library in Python with a learning rate of 0.0001. The model was trained on the training set for 50 epochs and batch size of 32. The model’s performance on the validation set during training was monitored to detect overfitting or underfitting. The trained model was evaluated on the validation set using appropriate metrics such as accuracy, precision, and recall.

3.2.5 Model Deployment
The model was integrated into a user-friendly interface that accepts audio inputs and predicts the corresponding genre. Live deployment is a part of future.
3.3 Architectural Design

3.3.1 Block Diagram

The block diagram is as follows,

![Block Diagram](image)

The model trained by the CNN acts as a trained classifier to make predictions of the genre.

3.3.2 Architectural Design

The Architectural Design for future live deployment is as follows,

![Architectural Design](image)

3.4 Algorithm Details

Convolutional Neural Networks (CNNs): Convolutional Neural Networks (CNNs) are a type of deep learning architecture well-suited for audio data analysis. This CNN model consists of six layers, including convolutional, pooling, and fully connected layers, which enable the network to learn hierarchical representations from input data. The number of neurons range from 1024 to 10 in a decreasing order with every layer.

Feature Extraction: MFCCs (Mel-frequency cepstral coefficients) features are extracted from the pre-processed audio files. These features capture the acoustic characteristics of the audio signals and serve as input to the CNN model in the form of an array.

Training and Optimization: The CNN model is trained using the training subset of the GTZAN dataset. The training process involves the following steps:
Loss Function: Categorical cross-entropy loss function is selected to measure the discrepancy between the predicted genre labels and the ground truth labels.

Optimization Algorithm: Adam optimization algorithm is employed to iteratively update the model’s weights and biases during training, minimizing the loss function.

Hyperparameter Tuning: Hyperparameters, such as learning rate, batch size, and regularization techniques (e.g., dropout), are tuned to optimize the model’s performance. A dropout of 0.2 is applied to every layer except the output layer of the CNN to avoid overfitting and increase accuracy.

IV. RESULTS AND DISCUSSION

4.1 Result Analysis

The experiments were done in phases with the properties of the CNN model such as number of layers, number of neurons in each layer, number of dropouts in each layer and no of epochs while training to achieve the perfect accuracy. The most important trials of all are described below.

First Phase: Five Layers were used in the CNN model with the neurons in each layer being 256, 128, 64, 64 and 10. The model was trained for 50 epochs and ReLU activation function was used. An accuracy of 79.05% was achieved but the model experienced overfitting.

Second Phase: Five Layers were used in the CNN model with the neurons in each layer being 256, 128, 64, 64 and 10 just like the first phase. The model was trained for 50 epochs and ReLU activation function was used. An accuracy of 79.17% was achieved because overfitting was managed by using dropout of 0.3 for each layer.
Third Phase: Six Layers were used in the CNN model with the neurons in each layer being 1024, 512, 256, 256, 64 and 10 respectively. The model was trained for 30 epochs and ReLU activation function was used. An accuracy of 79.53% was achieved but the model experienced overfitting.

Fourth Phase: Finally, the best accuracy and prediction was achieved by using six layers in the CNN model with the neurons in each layer being 1024, 512, 256, 128, 64 and 10 respectively. The model was trained for 50 epochs and ReLU activation function was used. Overfitting was managed by using dropout of 0.2 for all the layers except the output layer. With such combinations the model achieved an accuracy of 85.74%. Although this accuracy and all the accuracy values above also vary by a percent or two depending upon the weights and biases in the neurons in the layer every time the CNN Model is run.
4.2 Performance Parameters

4.2.1 Confusion Matrix

The confusion matrix for this study is as follows,

![Confusion Matrix]

The middle diagonal line represents all the true positives of the 10 genres in the dataset. i.e. the actual and the predicted genre is the same for the values which are in dark blue diagonal line.

4.2.2 Accuracy

It is the ratio of correctly classified points (prediction) to the total number of predictions. Its value ranges between 0 and 1. An accuracy of 85.74% was achieved but this may vary slightly according to the weights and biases in the neurons at runtime.

4.2.3 Precision

Out of all the positive predicted, what percentage is truly positive. A Precision Score of 86.07% was achieved but this may vary slightly according to the weights and biases in the neurons at runtime.

4.2.4 Recall

Out of the total positive, what percentage are predicted positive. A Recall Score of 85.74% was achieved but this may vary slightly according to the weights and biases in the neurons at runtime.

4.3 Conclusion

In conclusion, the music genre classification project using convolutional neural networks (CNN) on the GTZAN dataset from Kaggle has successfully demonstrated the feasibility and effectiveness of automated genre classification in the field of music analysis. Throughout the project, various components and techniques were employed to achieve accurate genre predictions.

The CNN model architecture played a vital role in achieving accurate genre predictions. Through the training phase, the model learned to identify intricate patterns and representations within the audio data, enabling it to differentiate between different music genres effectively. The optimization of hyperparameters and the selection of appropriate loss functions and optimization algorithms further enhanced the model’s performance.

The evaluation of the model using various metrics such as accuracy, precision, and recall provided insights into its effectiveness. The high accuracy achieved in genre classification demonstrated the potential of CNNs in accurately categorizing music genres based on audio content.

Overall, the music genre classification system developed in this project provides a valuable tool for automatically categorizing music into different genres. It has the potential for various applications, such as music recommendation systems, playlist generation, and content organization in music streaming platforms.

In conclusion, the project demonstrates the power of convolutional neural networks in music genre classification and highlights the importance of efficient data pre-processing, model training, and evaluation. By leveraging deep learning techniques, this project contributes to the field of music analysis and opens possibilities for further advancements in the automated categorization and understanding of music genres.
REFERENCES


