Music Genre recognition

Sagatbekov Dinmukhammed

Introduction

The aim of this project is to build and compare different models trained on well-known GTZAN dataset. 3 models were tested on 2 dataset variations and different split techniques.

Dataset description

The famous GTZAN dataset is considered to be the MNIST for music. It has 1000 samples for 10 music genres with 30 seconds each. It is popular since the concept of music genres and single-label classification is easy and simple. However, there are some problems with this dataset: the audio quality varies by samples, there is heavy artist repetition, which are very often ignored during dataset split and the labels not 100% correct.

Feature selection

There are many features that could be extracted from audio files.

1) Zero Crossing Rate - information about the rate at which the signal changes its sign, or crosses zero, over a given period.
2) Perceptual features that are characteristics of the sound that are related to human perception. These features can include loudness, pitch, duration, and timbre.
3) Chroma features - derived from the chromagram, which is a representation of the energy distribution of musical pitch classes in an audio signal.
4) Harmonics
5) Tempo BPM (beats per minute)
6) Spectral Centroid - indicates where the ”center of mass” for a sound is located.
7) Spectral Rolloff - a measure that provides information about the shape or steepness of the spectral content of a signal.
8) Mel-Frequency Cepstral Coefficients (MFCCs) - coefficients that collectively represent the short-term power spectrum of a sound signal. They are widely used in audio signal processing, speech processing, and music analysis. The audio signal is divided into short frames, often on the order of 20-40 milliseconds each to ensure continuity in the analysis and prevent loss of information at frame boundaries; frames are typically overlapped.
9) MFCC Mean - The MFCC mean is the average value of the MFCC coefficients over a certain duration or set of frames
10) MFCC Variance- represents the spread or dispersion of the MFCC values. It measures how much individual MFCC values deviate from the mean.

In total there are 56 features, 40 of them are MFCC vars and means.

Dataset split

During the model training on the 30-second dataset, a stratified random split was employed. However, when working with the 3-second dataset, a challenge emerged. The 3-second dataset consists of 10,000 instances, with each set of 10 instances belonging to the same song as in the previous dataset. Many research papers typically use stratified random splits. However, this method poses the risk of parts from the same songs ending up in both the test and train sets. To address this, an alternative split tactic was implemented. In this approach, all parts belonging to one song were grouped together and assigned to either the test or train set. In conclusion, models were trained on all three of these datasets, each employing a different strategy for splitting the data.

Model selection

For this project I used two Fully-Connected neural networks and XGBoost.
Model 1: Fully Connected Neural Network (7 layers). Architecture: 7 hidden layers with ReLU activation, dropout rate of 0.5, and softmax output layer. Results: 30s dataset: 76% 3s dataset (random stratified split): 92% 3s dataset (exclusive split): 50%.
Model 2: Fully Connected Neural Network (4 layers) Architecture: 4 hidden layers with ReLU activation, dropout rate of 0.5, and softmax output layer. Results: 30s dataset: 78% 3s dataset (random stratified split): 92% 3s dataset (exclusive split): 47%
Model 3: XGBoost Classifier Results: 30s dataset: 72% 3s dataset (random stratified split): 72% 3s dataset (exclusive split): 90.5%

The fully connected neural networks with 7 and 4 layers achieved competitive performance, with the 3-second dataset showing superior accuracy compared to the 30-second dataset. The random stratified split proved to be more effective than the exclusive split for both neural network architectures.

Model 3, based on the XGBoost classifier, performed well on the 3-second dataset with an exclusive split, demonstrating the ability of gradient boosting algorithms to capture complex patterns in short audio segments.

Conclusion

The fully connected neural networks demonstrated stable but not impressive performance, particularly on the 3-second dataset with a random stratified split. However, it is still unclear how justified this split is. The XGBoost classifier also showcased promising results, especially in scenarios where exclusive split was applied. Exclusive split should be considered in further research. Additionally, a bigger dataset is required for better understanding of how to build music genre classifiers.