AudioBots - Audio Genre Classification Executive Summary

Overview

Historically, songs have been categorized into genres not just for commercial purposes but also to enhance the listening experience and foster cultural exchange through music. With the advent of Music Information Retrieval in the 1990s, researchers began using algorithms to analyze audio files, classifying music based on features like pitch, tempo, and timbre. The abundance of extractable signals from audio files and the rise of deep learning have made genre classification a popular and evolving field among data scientists. In response, we have developed a genre classification system that contributes to these ongoing advancements.

Approach

The burgeoning field of audio classification motivated us to experiment with a range of algorithms and datasets. Our primary goal was to compare the performance of traditional machine learning models with more advanced deep learning models, thereby evaluating the effectiveness of these newer neural network solutions. We began by extracting features from audio files and processing them through XGBoost and other classic machine learning algorithms. For deeper insights, we employed sophisticated deep learning techniques in three ways. First, we used a short-term memory model based on audio features. Second, we generated spectrograms, which are graphical representations of the frequency signals, and analyzed them with Convolutional Neural Networks (CNNs) for image classification. Third, we input raw audio into pretrained transformer models such as DistilHuBERT, Whisper, and Wav2vec, and a version of Google's WaveNet convolutional architecture.

Data Collection and Methods

Finding a dataset to train genre classifiers was challenging due to licensing issues and low-quality audio files. We used two datasets: GTZAN and Free Music Archive (FMA). GTZAN is a well-known dataset comprising 1,000 audio tracks, each 30 seconds long, distributed across 10 genres with 100 tracks each. FMA offers two subsets. FMA-small, a balanced dataset with 8,000 tracks of 30 seconds each across 8 genres, and FMA-medium, a larger, unbalanced dataset featuring 25,000 tracks across 16 genres. We initially developed and tested our algorithms on the GTZAN dataset and then moved onto bigger and more complex FMA datasets, using accuracy and F_1 score as our evaluation metrics.

Results

On GTZAN, XGBoost and Whisper Small emerged as the top-performing algorithms, each achieving an accuracy and F_1 score of 0.92. However, on the FMA-small and FMA-medium datasets, Whisper Medium stood out, delivering an accuracy and F_1 score of 0.63, outperforming all other models. We further tested our algorithms in a practical scenario by addressing a contemporary debate: whether Beyoncé's new album Cowboy Carter is country. Our answer is no - according to our models, only two tracks, "AMEN" and "SMOKE HOUR II" were classified as country.

Future Iterations and Benefits to Stakeholders

Our project encountered significant challenges, primarily concerning dataset quality and the computational costs of training models. Genre classification was particularly difficult due to the multi-genre nature of many songs, highlighting the need for multiclass classification in future efforts. We recommend prioritizing the acquisition or curation of higher-quality datasets to enhance model performance. Future iterations will explore data augmentation techniques for both image and audio inputs, as well as the development of ensemble models to leverage the strengths of individual models.

Our project delivers key insights into genre classification algorithms, which are foundational to recommendation systems used by music streamer services like Spotify, Apple Music, and Amazon Music. Additionally, we provide notebooks for music enthusiasts who wish to explore newcomer albums in greater detail. Finally, we created a HuggingFace Space where users can upload their audio tracks to classify genres.