Music Genre Recognition from Audio Tracks and Metadata

Berkay Erkan ***Can Kocagil*** ***Efe Eren Ceyani*** ***Mehmet Çalışkan*** ***Hammad Khan Musakhel***

{berkay.erkan,can.kocagil,mehmet.caliskan,eren.ceyani,hammad.musakhel}@ug.bilkent.edu.tr

Keywords— Audio Processing, Music Information Retrieval, Music Genre Recognition, Multi-class Time Series Classification

I. Introduction

Automatic recognition of audio tracks from music waves has been a challenging task in the context of Music Information Retrieval (MIR). The mathematical representation of natural music, composed of rhythm, tones, intervals, patterns, and harmonies, are complex subjects by human specialists, as there is geometric interpretation in the humming of the strings and inherent subjective nature. The discrimination of genres from purely audio tracks is stochastic by nature and deeply controversial as the genres share common statistical knowledge. For instance, the musical composition emerges from the intersection of multi-cultural genres, which are known as fusion genres, e.g., country rock is a fusion of country music and rock music. Contextually, the difficulty level of autonomous discrimination of genres are changing over time, as the music trends are changing, which makes the music genre classification task temporally dynamic, and introduces a higher degree of conceptual complexity.

II. Dataset Description

We are going to design a computational system of end-to-end learning mechanism, for automatic recognition of music genres from Creative Commons Licensed audio files, with hand-crafted track-level features. As the dynamics of common genres are temporally varying, representable and scalable datasets are required to perform cognitive tasks. Free Music Archive (FMA) [1] is a dataset for common MIR analysis tasks and is supervised for browsing, searching, and organizing large music collections. FMA dataset is an MIR benchmark dataset that consists of 106,574 tracks from 16,341 artists and 14,854 albums, arranged in a hierarchical taxonomy of 161 genres [1]. By its large composition, it is structured by 30 seconds long and high-quality audio, pre-computed features, together with track- and user-level metadata, tags, and free-form text such as biographies, for scaling other music domain processing tasks [1]. FMA is a metadata-rich, future proof, and permissively licensed dataset that offers quality audio, pre-computed music-level features, reproducibility and long-term sustainability. Considering the computational comfort zone, the FMA dataset provides four versions, divided according to the sizes: small (30 seconds long 8,000 tracks, eight balanced genres), medium (30 seconds long 25,000 tracks, 16 unbalanced genres), large (30 seconds long 106,574 tracks, 161 unbalanced genres) and full (106,574 untrimmed tracks, 161 unbalanced genres). Realizing our computational resources, restricted to open-source computing, we'll utilize the small or medium version of the datasets, but in the case of any computing device availability, we'll try to process a full FMA dataset that is 879 GiB of raw information.

III. Problem Description and Milestones

We're going to implement trend-aware classifier-based learning algorithms, specialized for capturing the complex interplay of cultures to recognize genres in an autonomous fashion. Our classes will be music genres, supervised and annotated by the artists themselves, from contemporary to jazz. By acknowledging the socio-cultural context, we'll design cognitive machine learning algorithms, to capture temporal dynamics of audio, from conventional ML algorithms like SVM, to feature sensitive and time-aware deep learning algorithms such as Convolutional Neural Networks (CNNs). Various ML-based techniques will be utilized to accomplish the scope of the project, as of start, simple and naive algorithms such as Decision Trees and SVM will be introduced to create our baseline model to benchmark, then signal-aware and temporally dynamic classification algorithms such as CNNs and LSTMs will be proposed to compare our multi-class results. Initially, we will be computing a small-sized dataset that is class-balanced, as such, our main goal will be to predict the single top genre on the balanced small subset. Then, in the case of large or full-size datasets, our main objective will be maximizing the accuracy of multiple top genres. From the computational perspective, our task is a multi-class classification of signals and metadata that require supervised performance evaluation metrics. Single-top class accuracy and ROC-AUC will be the main metrics to compare in small or medium dataset sizes. Hence, our primary milestone will be preparing the data pipeline and creating conventional ML-based classifiers to construct a baseline. Then, our second generic milestone will be creating a winning approach that accumulates by a significant margin in accuracy.

IV. Challenges and Conclusion

Musical genre recognition classifies the musical collections based on the audio-level, user-level and track-level features. There are expected challenges to overcome such as poor labelling, and non-triviality of extraction of features. In the former one, musical genres are loosely defined, as people often argue over the genre of a song. In our case, the genres are supervised by the artists themselves, hence it'll be smoother to perform data-centric computational tasks. In the latter one, automatic extraction of semantic features are not differentiable and conceptualized, hence there is no common way to extract features in the context of music information retrieval. The Mel Spectrogram [2] based representation algorithms are proposed to overcome these challenges, by converting audio signals to visual, and represent how the spectrum of frequencies vary over time. The latter challenge will be tried to overcome with spectrum-based representations before feeding the ML model. From a more musical perspective, the notations of composers and audios created by its artists consist of musical representations: counting, rhythm, scales, intervals, patterns, symbols, harmonies, time signatures, overtones, tone, and pitch [3], that is highly unstructured and hard to interpret mathematically. Hence, ad-hope statistical and structural pattern recognition frameworks will be fit to capture conceptual and contextual sides of the music for the extraction of genres.

References

[1] Defferrard, Michaël, et al. "Fma: A dataset for music analysis." arXiv preprint arXiv:1612.01840 (2016).

[2] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.

[3] "Mathematics and Music." American Mathematical Society. Web. 30 Oct. 2021.