Musical Lyric Classification

Deep Learning Project Proposal

Dorian Desblancs
École Normale Supérieure Paris-Saclay
4 Avenue des Sciences, 91190 Gif-sur-Yvette, France
dorian.desblancs@mail.mcgill.ca

Kodjo Mawuena Amekoe
École Normale Supérieure Paris-Saclay
4 Avenue des Sciences, 91190 Gif-sur-Yvette, France
kodjo.amekoe@ens-paris-saclay.fr

1. Motivation and Problem Definition

Musical genre recognition is a vast problem that researchers have been trying to tackle for years. It also has numerous applications these days, as more and more companies aim to curate musical playlists and recommendations for their users. This is notably the case for Spotify, Deezer, and Apple Music. These companies track users' tastes so as to recommend new songs to their customers. Hence, the ability to automatically decipher song genre similarity is incredibly valuable, and could potentially make these companies' algorithms far more accurate.

The first attempts at solving the problem relied on extracting signal features such as the spectral centroid, spectral flux, and time-domain zero-crossings. Perceptually motivated features extractors such Mel-Frequency Cepstral Coefficients were also developed in order to garner information on signals for later recognition. The above methods notably influenced Tzanetakis and Cook's [9] seminal paper *Musical Genre Classification of Audio Signals*. The authors developed beat histogram features using a series of signal transformations (notably filtering and downsampling). These histograms were extracted from a variety of songs in a 10-class dataset. From there, histograms were clustered using KNN. Each cluster was then evaluated for its accuracy. Tzanetakis and Cook's approach was the first approach to extract genres well. It achieved approximately 61%.

Tzanetakis and Cook's [9] paper also heavily influenced the work of future researchers in the field of musical classification. The dataset they created, now commonly known as GTZAN [9], is the most heavily used dataset in the field. It contains 1000 30-second song excerpts that span a total of 10 genres. Most genre recognition systems are evaluated on this dataset. It however contains many flaws and limitations. Most notably, some excerpts are mislabeled and the dataset is quite limited in its number of genres. A deeper exploration into the dataset's flaws can be found here [7].

In the past few years, however, Convolutional Neural Networks have taken over the musical genre recognition problem. In 2014, Gwardys and Grzegorz achieved 78% on the GTZAN dataset. They used a CNN to extract spectrogram features. These were then classified using an SVM [4]. More recently, Ghosal and Kolekar achieved a mind-boggling 94.2% [3] accuracy score on the dataset using an ensemble of Convolutional Neural Networks applied to a variety of signal features.

The methods outlined above are incredibly impressive, but are unfortunately hard to interpret. The GTZAN dataset is too small, so any classifier performance must be taken with a grain of salt. More importantly, signal processing-based methods haven't yet shown an ability to generalize their knowledge to other problems such as musical sentiment analysis.

This is why our project aims to classify songs using their lyrics only. We believe that the lyrical content of a song is already incredibly representative of its genre. We will start by classifying songs by genre in order to see how well a variety of models and algorithms perform on this task. This will allow us to test our hypothesis that lyrical content can be an indicator of a song's genre. We will then explore unsupervised learning in order see how lyrics are able to capture the fluid nature of musical genre and resemblance. As the reader probably knows, rock and roll can be divided into a wide variety of sub-genres, such as punk rock, alternative rock, and metal. Unsupervised learning approaches to the musical genre recognition problem could potentially detect these subtle differences.

In general, research linked to lyric-based music classification isn't very frequent. Two papers released in 2017 did however explore the fields in an interesting way. The first can be found here [2], and introduced the MoodyLyrics dataset. This dataset annotates 2500 songs by sentiment, and greatly influenced us for our project. They notably extracted their lyrics from the same website we plan to data mine. The authors achieved 75% on binary lyrical sentiment analysis using SVM.

Around the same time, Alexandros Tsaptsinos [8] published the state-of-the-art paper for lyrics-based genre classification. He achieved 49.77% using an LSTM model on a 20-genre, 449 458 song dataset. Unfortunately, his dataset is not public. We hope to achieve similar supervised learning results on our smaller, evenly-distributed dataset.

2. Methodology

The first step to our problem is to gather a dataset. In order to do so, we will be mining Spotify playlists. Popular playlists are usually genre-based, and contain tracks from a wide range of artists. We will be extracting artist and song names, playlist genre, and the Spotify genre tags associated with each song we extract to construct a dataset. Note that the Spotify Developer API contains all the useful information for scraping such information. From there, we will use the website Lyrics.com [1] to extract lyrics from every song we extracted. Note that we already have a Lyrics.com scraper, so the dataset construction shouldn't take too long.

From there, we will explore a variety of algorithms for genre classification. A good starting point seems to be CNN's, LSTM's, and RNN's. These algorithms are already well-established in the natural language processing community, and have proven to be extremely high-performing. Moreover, these models commonly use pre-trained word embeddings for classification. We plan on using the state-of-the-art GloVe word-embeddings [6] to achieve maximum accuracy on our genre classification problem.

In parallel, we will explore unsupervised methods for clustering our lyrics. One popular approach seems to be to use doc2vec [5] and then using a clustering algorithm for classification (k-means, GMMs, or hierarchal clustering). Pre-trained word embeddings for each song will of course also be explored.

3. Evaluation

Our supervised learning problem should be fairly easy to evaluate. Our goal is to develop and test a series of deep learning models that will maximize accuracy on our dataset. During this process, we will be able to compare multiple architectures. We will also see how our models perform on the GTZAN dataset by mining its tracks' lyrics. Our models' performance on the popular dataset should be an indicator of how good our approach to genre recognition is. More importantly, our models' performance will allow us to compare our approach to signal processing methods for genre recognition. This will help answer the following question: are lyrics expressive of a song's genre?

Our approach to data collection will allow us to evaluate our unsupervised learning models soundly. Since each song we will be studying comes from a playlist, we will be evaluating our approach by comparing how closely our clusters resemble the songs' playlist distribution. For example, if a cluster is full of hip-hop songs coming from the Rap Caviar playlist on Spotify, we know that hip-hop tracks are being split appropriately.

References

[1] Welcome to lyrics.com. 2

[2] Erion Çano and Maurizio Morisio. Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pages 118–124, 2017. 2

[3] Deepanway Ghosal and Maheshkumar H Kolekar. Music genre recognition using deep neural networks and transfer learning. In *Interspeech*, pages 2087–2091, 2018. 1

[4] Grzegorz Gwardys and Daniel Michał Grzywczak. Deep image features in music information retrieval. *International Journal of Electronics and Telecommunications*, 60(4):321–326, 2014. 1

[5] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014. 2

[6] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[7] Bob L Sturm. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013. 1

[8] Alexandros Tsaptsinos. Lyrics-based music genre classification using a hierarchical attention network. *arXiv preprint arXiv:1707.04678*, 2017. 2

[9] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002. 1