MIR for Emotion Classification: Proposing a new DL Architecture

Why We Love Music?

Understanding why we love music revolves around temporal expectancies, their associated predictions, and reward value generated by these predictions. First, the auditory cortex represents patterns of sounds as opposed to individual sounds. The interactions between auditory areas and frontal cortices are critical in allowing working memory to knit separate sounds into abstract representations. These generate expectancies that are rooted in templates in an individual’s history of listening, which are stored in auditory cortices.

The reward system promotes certain behaviors, where the dopaminergic circuits establish reward value of relevant stimuli. An important part of this system is devoted to rewarding prediction, where fulfillment of prediction leads to dopamine release. These interactions suggest greater informational cross-talk between the systems responsible for pattern analysis and systems responsible for assigning reward value. It is known that both dorsal and ventral striatum respond to moments of peak pleasure (“chills”) induced by music (i.e. dopamine release occurred in the striatum during these moments), and the more listeners liked a musical piece, the greater the cross-talk between the striatum and auditory system. In fact, there is a “sweet spot” somewhere on the spectrum from completely unpredictable to completely expected that maximizes the reward response. Musicians define this most rewarding music as having the ability to surprise the listener with novelty within a predictable framework. In [2], the author concludes that there is causal evidence that musical pleasure is directly linked to reward system activity.

With this rewarding nature of music in mind, how do we represent music in an emotional space?

Mathematically Defining Emotion: To predict arousal and valence in Russell’s emotional space

Quick note: Valence refers to how positive or negative the event is. Meanwhile, arousal refers to whether an event is exciting/agitating or calming/soothing.

The model shown below has been adapted from Russell’s “A Circumplex Model of Affect” in order to quantize the variables associated with emotion. For simplicity, we divide our “emotional space” based on the four cartesian quadrants: 1. Happy, 2. Tense, 3. Dark or Bored, 4. Pleasant.

We pose this problem as a classification one with an underlying regression problem. That is, we will first regress on extracted musical features to predict arousal and valence values. Then, using these values, we can classify the song into the four emotional categories stated above.
Russell’s 2-D Emotional space

Music Datasets

- Spotify’s Sequential Skip Dataset
- 1000 Songs Dataset

Initially, clustering algorithms were implemented on the sequential skip dataset for determining if there are any clusters with respect to emotion. Unfortunately, the data is primarily labeled for user preferences and the song clips were not provided. Hence, emotion classification proved to be difficult with this dataset.

However, the 1000 Songs Dataset is publicly available and contains 744 45 second clips of songs with arousal and valence annotations (both static and dynamic available). For each of these clips, we have the mean valence and arousal. These values were determined with the help of crowdsourcing (participants were chosen for their qualifications and correctness of their answers).
Since emotions can change in a single song, the authors suggest a dynamic method for computing the valence and arousal values (which both range from -1 to 1). Therefore, we could determine the emotion of the entire song or even a small section of it.

The authors approach this as a regression problem. Since convolutional layers can help identify relevant local patterns and the recurrent layers summarize information over time, they explore a way to combine these two ideas to extract both pieces of information (as emotions are dependent on the current music as well as its interaction over time, which would be difficult with just CNNs).

This architecture was replicated in Python Keras. The loss stabilized around the fourth epoch and the root mean squared error on the validation set for valence (between 0 and 10 in this case) was approximately 2.678. This clearly is not at all accurate, and this approach will probably not satisfy the predictions we are looking for. To fix this, we can try to extract our own meaningful musical features and perform regression.

**Extracting The Right Features**

The preprocessing was performed on the 1000 songs dataset, which provides the original music files and annotations of arousal and valence (both dynamic and static). The new processed dataset contains the following features:
- **Tempo**
  - the speed at which a passage of music is or should be played
  - Shape: (1,)

- **Spectral Centroid**
  - Indicates where the "centre of mass" for a sound is located and is calculated as the weighted mean of the frequencies present in the sound

- **Spectral Rolloff**
  - measure of the shape of the signal. It represents the frequency below which a specified percentage of the total spectral energy.
  - Shape: (744, 200)

- **Mel-Frequency Cepstral Coefficients**
  - Concisely describe the overall shape of a spectral envelope. It models the characteristics of the human voice.
  - Shape: (744, 20, 200, 1)

- **Chroma Frequencies**
  - the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.
  - Shape: (744, 12, 200, 1)

- **Timbre**
  - Contains the following:
    1. **Hardness** (related to the attack times and the associated bandwidth - difference between low and high frequencies)
    2. **Depth** (a weighted sum of fundamental frequency, approximate duration/decay-time, weighted mean of lower frequency ratios)
    3. **Brightness** (prevalence of upper mid and high frequency content)
    4. **Roughness** (buzzing, harsh, raspy sound quality of narrow harmonic intervals)
    5. **Warmth** (the tilt towards the bass frequencies)
    6. **Sharpness** (related to the spectral balance of frequencies)
    7. **Boominess** (prevalence of lower frequencies)
    8. **Reverberation** (persistence of sound after the sound is produced)
    9. **Zero Crossing Rate** (scalar): Simply the number of times the signal crosses the x-axis (i.e. changing from negative to positive or vice versa). Computed by \(\frac{\text{np.where(np.diff(np.sign(data)))}[0].size}{\text{duration}}\)
    10. **Loudness** (scalar): The Python pyloudnorm library provides a couple of concise methods to compute this.
  - Shape: (744, 10)

- **Static Standardized Valence**
  - Intended to be used as ground truth for training
  - Shape: (744,)
The proposed deep learning architecture for extracting valence and arousal from music is shown above. Given the musically meaningful features (as described above), we simply need to use dense and convolutional layers, pooling, and dropout for the network to learn these values. That is, we also approach this as a regression problem. Similarly, we have two branches of the same network: one predicts valence and the other predicts arousal. The convolutional layers help identify local patterns in the chroma frequencies and Mel-Frequency Cepstral Coefficients. Meanwhile, the dense layers combine and summarize the extracted information from each of the features to predict the final values for valence and arousal. The neural network was tested in Python Keras: the code and hyperparameters used can be found on the GitHub repository.

Currently, during initial testing, this architecture has provided promising results for predicting valence (reducing the regression errors presented by Malik et. al). A follow-up with the results of more rigorous testing will be analyzed and released.
References