Music Emotion Recognition

In the next 10 minutes, I’ll go over the task of Music Emotion Recognition in the field of MIR, which is also the choice of our final project: to be able to predict emotional values of a given audio file.

First of all, why do we want to do this. One of the reasons why we listen to music is because it conveys emotions directly. We feel happy, energetic, or sad, or even intellectual when we listen to different kinds of music. While we don’t necessarily remember the name of a piece of music we like all the time, we almost definitely remember how it made us feel. In the current day where enormous amount of musical content can be found on the internet, the ability to automatically organize music based on its emotional quality is important for any application to help us find the music we want, whether we are commuting to work, making music as a producer, or finding background music as a video editor and so on...

However, there are few common challenges associated with the task. The first being the subjective nature of emotions. How I feel about a piece of music may be deeply personal and life circumstantial, which is probably not how you feel about the same music. Also, how to translate emotional qualities to numerical datatypes that is easy for computers to understand. According to (Xinyu Yang, Yizhuo Dong, Juan Li, 2017)’s summary on MER methods, one popular way is to use Russel’s dimensional model to represent emotions on a valence / arousal space. Which, looks like this:

Chart, radar chart

Description automatically generated

This model divides the emotions we experience to two dimensions, valence and arousal. Valence means how attractive a certain thing is, or in other words, how much do we like it. For me, spicy noodle soups have very high valence, because it is a very good thing, and rat poison will have negative valence~~, because, I don’t like to be rat poisoned~~. Arousal on the other hand can be understood as excitement, energy, or lack thereof. If we combine the two dimensions, we can describe a whole lot of emotions. For example, high valence and high arousal would describe ecstatically winning a lottery ticket, and low arousal and low valence would describe tired, sluggish feelings.

So, for our final project, we used a deep-learning based approach that incorporated Russel’s dimensional representation of emotion as labels. The dataset we used is from Deezer’s music mood recognition repository on Github, which contains valence arousal labels for 18,644 tracks. Since Deezer did not provide any audio files, we had to go to their API to manually download 30 second mp3 previews for each track. In the end, we were able to get an annotated dataset of 14,317 tracks in mp3 format after discarding ones with broken links.

For the time being, we are re-implementing an audio-only deep learning algorithm described in Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, Manuel Moussallam’s paper, which uses a CNN model to train and predict input and outputs. The architecture is as follows. To save time, I’ve pre-written the layers and dimensions, and I’ll quickly go over it before showing some preliminary code that we are working on right now.