Music Thumbnailing via Neural Attention Modeling of Music Emotion

Yu-Siang Huang¹, Szu-Yu Chou¹,² and Yi-Hsuan Yang¹
¹Music and Audio Computing (MAC) Lab in CITI, Academia Sinica, Taiwan
²National Taiwan University, Taiwan
{yshuang, fearofchou, yang}@citi.sinica.edu.tw

Project: https://remyhuang.github.io/music_thumbnailing/

INTRODUCTION

- The goal of music thumbnailing is to find a short, continuous segment of a song that represents the whole song
- Chorus is usually the most memorable and emotional part
- Without annotations of the chorus sections of any song, we extract a music snippet of a song that happens to correspond to the songs chorus section by learning from emotion labels
- The key is to apply attention mechanism to a convolutional neural network (CNN)
- Not only learn to predict music emotion, but also know where the novel parts are

Figure 1: Example 30-second thumbnails.

METHOD

- Add a so-called attention layer [1] on top of an ordinary CNN
- Assess the importance of different short time audio chunks in predicting the emotion of the song

DATABASE

- Music emotion recognition (for training & testing part I): in-house collection of 31,377 clips with 24 seconds of Pop music as our corpus with 190 possible emotion tags from AllMusic (http://www.allmusic.com/moods/) [2].
- Chorus detection (for testing part II): the popular music subset of the RWC database [3] which contains 100 songs with manually labeled section boundaries

RESULT (PART I)

- Music emotion recognition: the average AUCs of emotion recognition is 0.7663 which outperforms the result from [2].

Figure 2: The proposed attention-based CNN model for music thumbnailing.

RESULT (PART II)

- Chorus detection: compare with the state-of-the-art music segmentation algorithms - MSAF [4]
- Figure 2 shows the percentages of songs (among 100 songs from RWC) that have certain degree of overlaps between the thumbnail and the chorus section

Figure 3: The result of MSAF [4] and the attention-based CNN.

Figure 4: Four example result. The first row is the mel-spectrogram, the second row marks the ground truth chorus sections (yellow regions), the third row marks the 30-second thumbnail and the fourth row shows the attention scores estimated by our model. The peaks fall within chorus sections.

Reference