DJNET: A DREAM FOR MAKING AN AUTOMATIC DJ

Yu-Siang Huang[1], **Szu-Yu Chou**[1,2], **and Yi-Hsuan Yang**[1]
[1]MAC Lab, CITI, Academia Sinica, Taiwan
[2]National Taiwan University, Taiwan
{yshuang,fearofchou,yang}@citi.sinica.edu.tw

ABSTRACT

We present a research project called DJnet, whose goal is to make a fully-automatic DJ to create music medley, mashup, remix and even electronic dance music (EDM). In this demo paper, we demonstrate the research results of our two recent works in the DJnet project. The first one is about *music thumbnailing*, which aims to find a short and continuous snippet (or, highlight) of a song to represent the whole song. The second one is on *music medley generation*, whose goal is to find an optimal ordering of a given set of music clips from different songs. We hope to let more people appreciate and dig deeper in DJ music through this project. Example results also can be found online at https://remyhuang.github.io/DJnet/.

1. INTRODUCTION

Disc Jockeys (DJs) are professional audio engineers whose role is to generate music like electronic dance music (EDM) and to manipulate musical elements to create music medley, mashup and remix. Toward the goal of making a fully-automatic DJ, we initiate the 'DJnet project' and focus on the following two sub-topics to begin with.

The first sub-topic is on generating a *music thumbnail* (highlight) of a song, as Figure 1 exemplifies. Among different sections of a song, the chorus sections are usually considered as the most memorable and emotional part [1]. In light of this finding, we are motivated to extract a music snippet of a song that happens to correspond to the song's chorus section by learning from emotion labels, without annotations of the chorus sections of any song [2].

In the second sub-topic, we focus on the ability to *combine* a given set of snippets into a longer piece of music, in a way that people can listen to with comfort and pleasure. We call this *music medley* generation. For computers to understand and deal with such a sequential problem, we propose a unsupervised learning method to let computers find the optimal ordering of sets of snippets [3].

This paper presents the main ideas of the proposed methods. Readers are referred to [2] and [3] for details.

Coldplay - Hymn For The Weekend
(Screenshot from Internet)

*The most representative 30 seconds is **03:30~04:00***
(extracted by our proposed method)

Figure 1. An example for extracting the most representative 30-second snippet of a song. More examples can be found in our project website.

2. PROPOSED METHODS

2.1 Music Thumbnailing

Methodological-wise, we propose a neural network model, illustrated in Figure 2, to perform music emotion recognition and music thumbnailing at the same time [2]. Specifically, we investigate adding a so-called attention layer on top of an ordinary convolutional neural network (CNN) to help the CNN assess the importance of different short time audio chunks in predicting the emotion label of the song. We expect that the *attention scores* estimated by the attention layer can not only help emotion prediction, but also indicate the location of at least one chorus section.

To generate the thumbnail of a user-specified length for any arbitrary song, we consider a simple method that simply uses a running window over the attention scores to pick a consecutive collection of chunks with the highest aggregated (i.e. summed) attention scores. Although the method is simple, it provides a direct way to assess what is captured by the attention layer.

In our implementation, we use the MER31K dataset [4] for training the emotion classifier. It contains 31,422 30-second songs, with song-level emotion labels (in total 190 possible categories) crawled from AllMusic (http://www.allmusic.com/moods). For the input to our

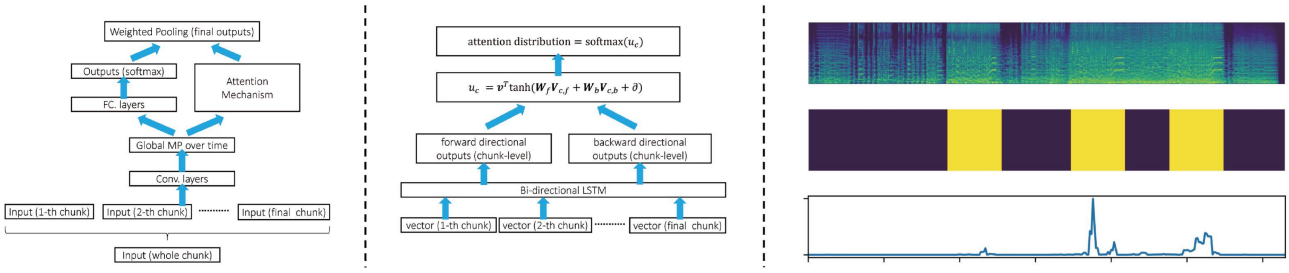

Figure 2. The proposed attention-based CNN model for music thumbnailing [2] and an example result (rightmost). In the example result, the first row is the mel-spectrogram, the second row marks the ground truth chorus sections (yellow regions), and the third row shows the attention scores estimated by our model. The peak falls within a chorus section.

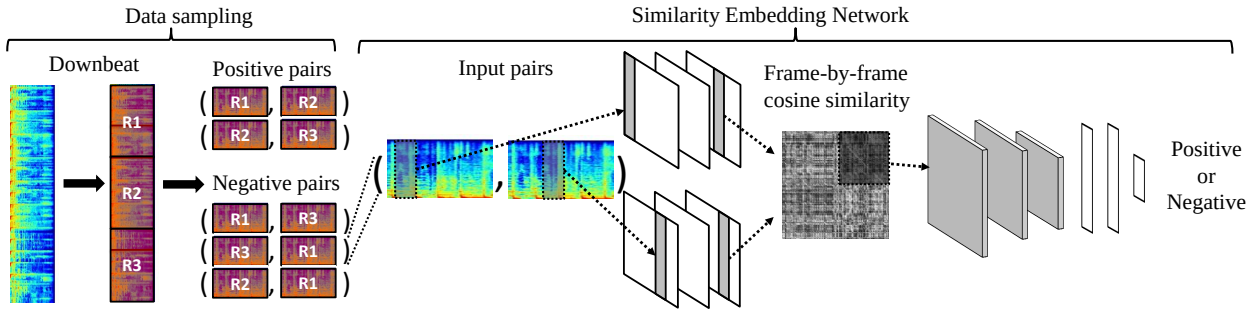

Figure 3. The proposed similarity embedding network for solving music puzzle games and generating music medleys [3].

model, we sample the songs at 22,050 Hz and use a Hamming window of 2,048 samples and hop size 512 samples to compute the magnitude spectrograms, which are then transformed to 128-dim log mel-spectrograms. We use a dataset of full-length songs with chorus labels for evaluation (but not for training) and find promising result [2].

2.2 Music Medley Generation

Toward this goal, we firstly formulate the task of assembling multiple multi-second, non-overlapping music fragments in proper order as the *music puzzle games*, drawing the analogy that a fragment is like a puzzle piece. We then propose the *similarity embedding network* (SEN), depicted in Figure 3, that learns to solve these music puzzle games by learning patterns from the similarity matrix of a pair of music fragments (using Siamese CNNs) and predicting whether the pair is in correct order or not [3].

Any music collection can be used in our puzzle games, since we do not need any human annotations. In our implementation, we use again the MER31k dataset [4], with the same input data preprocessing method mentioned above (but not using the emotion labels this time). In the training stage, we segment consecutively each song into three segments without overlaps. The pairs that are consecutive and in correct order are treated as positive pairs, and the others negative. For the global ordering, we evaluate the "fitness" of any ordering of the fragments by summing the model output of the composing consecutive pairs. We then pick the ordering with the highest fitness score as our solution for the game for that song. The results of music puzzle games and music medleys can be seen in our project site `https://remyhuang.github.io/DJnet/`.

3. CONCLUSION

We have presented the result of two sub-topics in the DJnet project. Specifically, we train an attention-based model to predict the emotion labels of songs and at the same time detect the chorus sections. Then, we propose an unsupervised learning method to learn to generate music medleys by playing puzzle games. In future works, we will investigate other interesting sub-topics such as music mashup.

4. REFERENCES

[1] Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Trans. Audio, Speech, and Language Processing*, 14(5):1783–1794, 2006.

[2] Yu-Siang Huang, Szu-Yu Chou, and Yi-Hsuan Yang. Music thumbnailing via neural attention modeling of music emotion. *Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2017.

[3] Yu-Siang Huang, Szu-Yu Chou, and Yi-Hsuan Yang. Similarity embedding network for unsupervised sequential pattern learning by playing music puzzle games. *arXiv preprint arXiv:1709.04384*, 2017.

[4] Yi-Hsuan Yang and Jen-Yu Liu. Quantitative study of music listening behavior in a social and affective context. *IEEE Trans. Multimedia*, 15(6):1304–1315, 2013.