Music Thumbnailing via Neural Attention Modeling of Music Emotion

Yu-Siang Huang, Szu-Yu Chou and Yi-Hsuan Yang

Research Center of Information Technology Innovation, Academia Sinica, Taiwan
E-mail: {yshuang, fearofchou, yang}@citi.sinica.edu.tw Tel: +886-2-2787-2300

***Abstract*— The goal of music thumbnailing is to find a short, continuous segment of a song that represents the whole song. In light of the observation that a main function of music is to communicate emotions, in this short paper we investigate whether a representative part selected by some automatic mechanism for emotion recognition corresponds to the chorus section of music. To address this research question, we introduce a so-called attention layer with long-short term memory cells to a deep convolutional neural network for music emotion classification. The attention layer estimates the importance of each 3-second chunk of a song in predicting the song-level emotion labels of the song. To this end, a collection of 31K songs with emotion labels are used. We then generate a thumbnail for each song based on the importance scores and examine whether the thumbnail corresponds to any chorus section of the song, using another dataset with annotations of chorus sections. Our experiment shows that for 35% of the songs our thumbnails contain a whole chorus section, and that for 80% of the songs the thumbnails overlap 50% in time with a chorus section.** [1]

I. INTRODUCTION

For many music information retrieval (MIR) systems, music structure analysis is an important component [1]. Among different sections of a song, the chorus sections are usually considered as the most memorable part [1, 2]. From a chorus section, it is possible to get the main idea of the song. Therefore, automatic detection of chorus sections from audio may lead to good summary (thumbnail) of a song, which can be useful in many music browsing and retrieval problems.

On the other hand, automatic music emotion recognition has been studied for years to enable emotion-based organization and retrieval of music [3-5]. The majority of existing work focuses on the prediction of the song-level emotion labels of songs. However, as emotions may change over time within a song, models that consider emotion of a song as time-varying have also been proposed [6].

A few attempts have been made to investigate whether we can improve the accuracy of emotion recognition by exploiting the outcome of music structure analysis [7]. One of the most interesting findings is that, if we are given information regarding which parts of a song correspond to the chorus sections, we can improve song-level recognition of arousal and dominance related emotions by focusing on only the chorus sections [7]. In light of this finding, we are motivated to address the following conceptually "reverse" problem: whether a music emotion recognition model can reveal anything about the structure of a song? Or, more specifically: whether a music emotion recognition model can be used to generate a music thumbnail of a song that happens to correspond to the song's chorus section, without annotations of the chorus sections of any song? To the best of our knowledge, such a research question has not been addressed in the literature.

Specifically, two audio datasets are employed in this study. The first dataset contains song-level emotion labels, whereas the second dataset contains annotations of the temporal occurrence of the chorus sections. We use the first dataset to train an audio-based emotion recognizer, and then the second dataset to evaluate how well the emotion recognizer can generate music thumbnails that correspond to the chorus sections.

Methodological-wise, we propose a neural network model to perform music emotion recognition and music thumbnailing at the same time. In this paper, we investigate adding a so-called attention layer [8-10] 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 of the song. We expect that the importance scores estimated by the attention layer can not only boost the accuracy of emotion recognition, but also indicate the occurrence of chorus sections. We call the resulting model a model for neural attention modeling of music emotion.

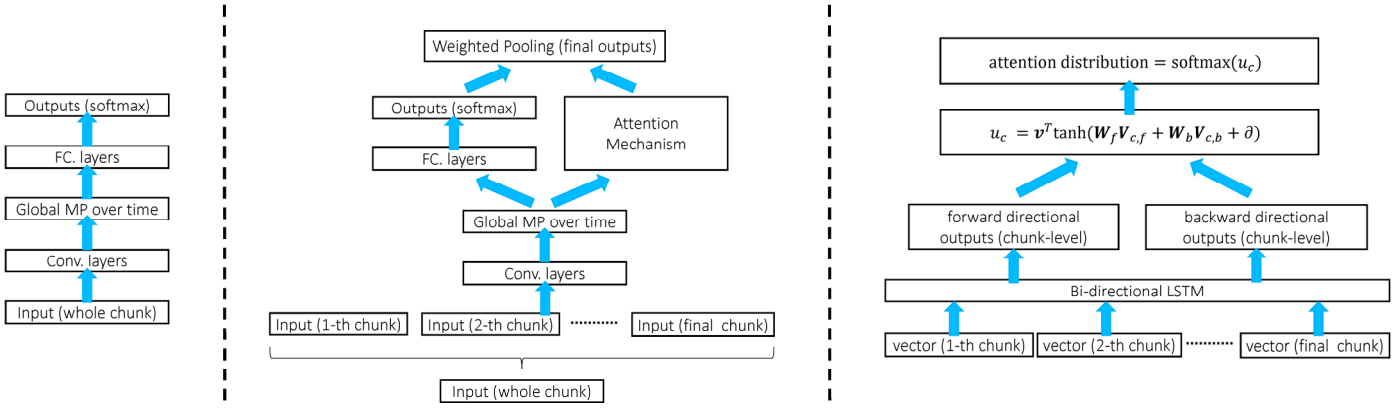
We compare the performance of our model in recalling the chorus sections against that of a state-of-the-art music structure analysis algorithm [11]. It turns out that our model performs better than this prior art, although our model is trained specifically for emotion recognition, not structure analysis.

II. METHOD

A. General Convolutional Neural Network

Figure 1(a) shows the diagram of the general CNN model adopted in this paper. Given the mel-spectrogram of a song of arbitrary length, the model firstly uses a stack of a few convolutional ('Conv') layers for feature extraction, followed by a global max pooling ('MP') over time for temporal aggregation, and then a few fully-connected ('FC') layers to learn the mapping between the extracted features and emotion labels. Our model uses rectified linear unit (ReLU) as the activation function in all the layers, and the cross-entropy as the cost function.

[1] https://remyhuang.github.io/DJnet/



(a) (b) (c)

Fig. 1 The diagram of (a) the general CNN model, (b) the attention-based CNN model, and (c) the attention layer used in attention-based CNN.

B. *Attention-based Convolutional Neural Network*

Figure 1(b) shows the diagram of the proposed attention-based CNN model, which is extended from the general model. There are mainly two differences. First, the attention-based model split a song into several short-time chunks of fixed length. Second, the model takes the output of the global MP layer as input to an additional attention layer to weigh the contribution of different audio chunks. As shown in Fig. 1(c), the first layer of the attention layer is a recurrent layer with bidirectional long-short term memory (LSTM) cells. Given a song with n chunks, the attention score u_c of the c-th chunk is calculated as follows:

$$u_c = \boldsymbol{v}^T \tanh\left(\boldsymbol{W}_f \boldsymbol{V}_{c,f} + \boldsymbol{W}_b \boldsymbol{V}_{c,b} + \partial\right), \quad (1)$$

where $\boldsymbol{V}_{c,f}$ and $\boldsymbol{V}_{c,b}$ denote the output of the forward and backward LSTM for the chunk, respectively, and $\boldsymbol{v}$, $\boldsymbol{W}_f$, $\boldsymbol{W}_b$ and ∂ are learnable parameters (that, once learned, are applied to any chunk of any song). The attention distribution $\boldsymbol{a}$, a n-dimensional vector, is then calculuated by passing the attention scores through a softmax function:

$$\boldsymbol{a} = \text{softmax}(u_c). \quad (2)$$

The final prediction of the attention-based CNN model fuses the output of the FC layers (on the left hand side of Fig. 1(b)) and the attention layers (right hand side) by:

$$\hat{\boldsymbol{y}} = \sum_{c=1}^{n} a_c \boldsymbol{o}_c, \quad (3)$$

where $\boldsymbol{o}_c$ denotes the output of the fully-connected layers (after softmax) for a chunk, a_c the attention score (also after softmax) of that chunk, and $\hat{\boldsymbol{y}}$ the final emotion estimate of the song. We can see that the attention mechanism allows the model to weigh different chunks differently, based on the output of the memory cells in LSTM. Therefore, the final estimate of the model can likely boost the contribution of the relevant chunks and suppress the unimportant ones.

C. *Music Thumbnailing via the Attention-based CNN*

To generate a music thumbnail of a predefined legnth for a song, we consider a simple method that simply uses a running window over the attention distribution 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.

D. *Implementation Details*

In our implementation, we use four layers of convolutional layers, all of which are one-dimensional in the time dimension. The number of convolutional filters in each layer is 64, 128, 128, 256, the filter length is 3, 4, 4, 4, and the stride size is 2, 2, 1, 1 respectively. After convolutions and global MP, we obtain hidden features with 256 dimensions. For the attention-based CNN, we segment a song into successive chunks of 3 seconds, without overlaps. The number of hidden features in bidirectional LSTM is set to 512. All the models are trained using stochastic gradient descent with momentum 0.9, and the batch size is set to 30.

We sample the songs at 22,050 Hz and use a Hamming window of 2048 samples and hop size 512 samples to compute the magnitude spectrograms, which are then transformed to 128-dimensional log mel-spectrograms (as done in many previous work) to be used as input to our CNN models. All features are standardized by z-score normalization, with normalization parameters calculated from the training set of emotion recognition. According to this setting, each audio chunk has approximately 130 time frames.

III. Experiment Setup

A. *Music Emotion Recognition*

We use the MER31K dataset [12] for training the emotion recognizer. It contains in total 31,422 MP3 files, with song-level emotion labels harvested from the music guide service website AllMusic (http://www.allmusic.com/moods). Among the 190 possible emotion categories, the most popular emotion is associated with 248 songs, the least popular emotion 28 songs, and the average number of songs per emotion cate-

gories is about 165. We discard songs which shorter than 25 seconds and use the remaining 31,377 songs in our experiment. To report the accuracy of our emotion recognizer, we follow the settings in [12] and randomly hold out 6,000 songs as the validation set (for parameter tuning), 6,000 songs for testing, and the remaining 19,377 songs for training. We consider the problem as a multi-label classification problem and use the average area under the receiver operating characteristic curve (AUC) [12] as the performance metric. The average AUC across all the 190 emotion categories is reported.

B. *Music Structure Analysis Dataset*

For evaluating the correspondence between music thumbnails and chorus sections, the popular music subset of the RWC database [13] is used. It contains 100 songs with manually labeled section boundaries [14].

In our experiment, we seek to generate a music thumbnail to represent the corresponding song. This thumbnail may be longer or shorter than each individual chorus section of the song. Therefore, to evaluate the correspondence between the generated thumbnails and the chorus sections, we pick the chorus section in a song that has the largest overlap in time with the thumbnail. Specifically, we report the percentage of that chorus section that is covered by the thumbnail. If there is no overlap at all, the correspondence is 0.

IV. Results

A. *Music Emotion Recognition*

As the baseline method, we consider the emotion recognition model proposed by Yang and Liu [12]. They investigated a variety of mid-level audio features characterizing the loudness, rhythm, timbre and tonal aspects of music, in total 243 features, and used support vector machine (SVM) with radial basis kernel for classifier training. This model can be considered as a good instance of non-deep learning based model.

The average AUCs of this baseline model and our CNN models are shown in Table I. It can be seen that the general CNN model can attain similar average AUC as the best model reported by Yang and Liu [12]. This result demonstrates the effectiveness of deep learning in bypassing feature design and in learning directly from the primitive mel-spectrogram for audio classification problems. Moreover, we see that the proposed attention-based model further improves the accuracy and slightly outperforms the non-deep learning based model. Statistical test shows no significant difference among the two models, but we consider this as a promising result.

B. *Music Thumbnailing and Chorus Detection*

We compare our model against state-of-the-art music segmentation algorithms for music thumbnailing. Specifically, we use Music Structure Analyze Framework (MSAF) [11], a Python toolkit that provides open source implementations of various structural segmentation algorithms proposed in MIR. According to the performance study reported in [11], we consider the Structural Features for music segmentation and the 2D-Fourier magnitude coefficients-based clustering method for grouping the resulting musical segments. Following Goto [1], the segments in the largest cluster (i.e. the one with the largest number of segments) are assumed to be the chorus sections. Among the segments in this largest cluster, we consider the one that is closest to the middle of the song as the music thumbnail.

TABLE I
THE RESULT OF MUSIC EMOTION RECOGNITION

Method	Average AUC
Rhythm features only + SVM [12]	0.6229
Tonal features only + SVM [12]	0.7060
Hybrid features + SVM [12]	0.7614
General CNN	0.7562
Attention CNN	0.7663

The music thumbnails selected by the aforementioned have different lengths. For fair comparison, we set the length of the thumbnail generated by model to be the same as that from MSAF.

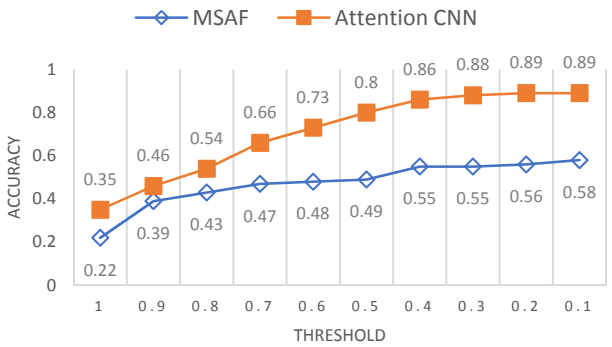


Fig. 2 The results of MSAF and the attention-based CNN.

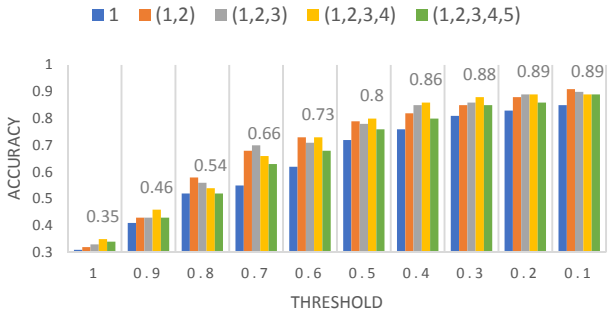


Fig. 3 The result when using different scales of chunks. The values on top of each set of bars are the results of the case '(1, 2, 3, 4).'

Figure 2 shows the percentages of songs (among the 100 songs from RWC) that have certain degree of overlaps (from 100% on the left to 10% on the right) between the thumbnails and the chorus sections. We can see that the thumbnails generated by the proposed method have a greater overlaps with the chorus sections of songs, comparing to those generated by MSAF. For 35 songs, our thumbnails contain a whole chorus section (i.e. 100% correspondence), and for 80 songs the thumbnails overlap 50% in time with a chorus section. In comparison, from the result of MSAF, only 22 songs have

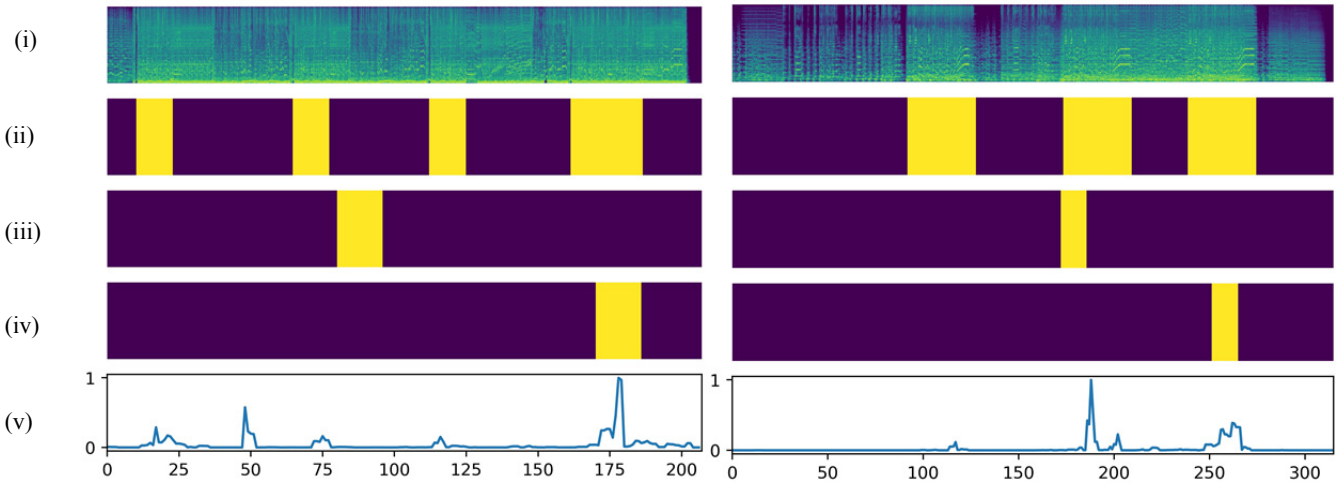


Fig.4 Two examples of illustration, left illustration is 'Disc1/001.mp3', and the right is 'Disc7/009.mp3'in Popular Music Database. The first subplot, (i), is Mel-spectrogram, (ii) is the ground truth of chorus segments where the yellow regions, (iii) is the extracted segment by MSAF, (iv) is by the attention-based CNN model, and (v) is the significant probability distribution by the attention-based CNN.

100% correspondence, and only 49 songs have 50% correspondence. The performance difference is quite remarkable. It provides empirical evidence that the proposed neural attention model for emotion recognition can generate thumbnails that correspond to the chorus sections of songs.

Figure 3 shows the result when we use chunks of different lengths in our attention-based CNN models. For example, '(1, 2, 3)' denotes that we still train our attention-based CNN models with 3 seconds, but for a testing song we split it into 1-second chunks, 2-second chunks and 3-second chunks as input to the network and then take the average of the result for thumbnail generation. It can be found that using multiple scales of chunks slightly improve the correspondence between thumbnails and chorus sections.

Finally, Figure 4 shows two real examples. From top to bottom, the figure shows the mel-spectrogram, ground-truth occurrence of the chorus sections (the yellow regions), the thumbnails selected by the MSAF baseline and by our model respectively, and the attention scores calculated by our model.

V. Conclusions

In this paper, we have proposed a new neural network model by using an attention mechanism for music emotion recognition. We found that the proposed method not only leads to accurate emotion recognizer, but also provides a new way to detect the chorus sections of a song without any priori information about song structure. The result demonstrates the effectiveness of the neural attention model, and the strong link between chorus sections and music emotions.

References

[1] M. Goto, "A chorus section detection method for musical audio signals and its application to a music listening station," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, 1783-1794, 2006.

[2] A. Eronen and F. Tampere, "Chorus detection with combined use of MFCC and chroma features and image processing filters," in *Proc. International Conference on Digital Audio Effects*, pp. 229-236, 2007.

[3] Y.-H. Yang et al., "A regression approach to music emotion recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 2, 448-457, 2008.

[4] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," ACM Transactions on Intelligent Systems and Technology, vol. 3, no. 3, article 40, 2012.

[5] J. Y. Lee et al., "Music emotion classification based on music highlight detection," in Proc. IEEE International Conference on Information Science and Applications, pp. 1-2, 2014.

[6] M. Caetano and F. Wiering, "The role of time in music emotion recognition," in *Proc. International Symposium on Computer Music Modeling and Retrieval*, 2012.

[7] X. Wang et al., "Enhance popular music emotion regression by importing structure information," in *Signal and Information Processing Association Annual Summit and Conference*, pp. 1-4, 2013.

[8] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in arXiv, 2014.

[9] E. Mansimov et al., "Generating images from captions with attention." in arXiv, 2015.

[10] Y. Xu et al., "Attention and Localization based on a Deep Convolutional Recurrent Model for Weakly Supervised Audio Tagging," in arXiv, 2017.

[11] O. Nieto and J. P. Bello, "Systematic exploration of computational music structure research," in *Proc. International Society of Music Information Retrieval Conference*, 2016.

[12] Y.-H. Yang and J.-Y. Liu, "Quantitative study of music listening behavior in a social and affective context," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1304-1315, 2013.

[13] M. Goto et al., "RWC Music Database: Popular, Classical and Jazz Music Databases," in *Proc. International Society of Music Information Retrieval Conference*, pp. 287-288, 2002.

[14] M. Goto, "AIST Annotation for the RWC Music Database," in *Proc. International Society of Music Information Retrieval Conference*, pp. 359-360, 2006.