Multi-scale Context Based Attention for Dynamic Music Emotion Prediction

Ye Ma
Dept. of Computer Sci. & Tech.
Tsinghua National Laboratory for
Information Science and Technology
Tsinghua University
Haidian District, Beijing, China
ma-y17@mails.tsinghua.edu.com

Xinxing Li
Dept. of Computer Sci. & Tech.
Tsinghua National Laboratory for
Information Science and Technology
Tsinghua University
Haidian District, Beijing, China
lixinxin14@mails.tsinghua.edu.com

Mingxing Xu*
Dept. of Computer Sci. & Tech.
Tsinghua National Laboratory for
Information Science and Technology
Tsinghua University
Haidian District, Beijing, China
xumx@tsinghua.edu.cn

Jia Jia
Dept. of Computer Sci. & Tech.
Tsinghua National Laboratory for
Information Science and Technology
Tsinghua University
Haidian District, Beijing, China
jjia@tsinghua.edu.cn

Lianhong Cai
Dept. of Computer Sci. & Tech.
Tsinghua National Laboratory for
Information Science and Technology
Tsinghua University
Haidian District, Beijing, China
clh-dcs@tsinghua.edu.cn

ABSTRACT

Dynamic music emotion prediction is to recognize the continuous emotion information in music, which is necessary for music retrieval and recommendation. In this paper, we adopt the dimensional valence-arousal (V-A) emotion model to represent the dynamic emotion in music. In our opinion, music and V-A emotion label do not have the one-to-one correspondence in the time domain, while the expression of music emotion at one moment is the accumulation of previous music content for a period of time, so we propose Long Short-Term Memory (LSTM) based sequence-to-one mapping for dynamic music emotion prediction. Based on this sequence-to-one music emotion mapping, it is proved that different time scales' preceding content has an influence on the LSTM model's performance, so we further propose the Multi-scale Context based Attention (MCA) for dynamic music emotion prediction. We evaluate our proposed method on the database of *Emotion in Music* task at *MediaEval 2015*, and the results show that our proposed method outperforms most of the models using the same features and achieves a competitive performance with the state-of-the-art methods.

CCS CONCEPTS

• **Applied computing → Sound and music computing**; • **Information systems →** *Music retrieval*; • **Computing methodologies →** *Machine learning*;

KEYWORDS

Music emotion prediction, Multi-scale context, Long Short-Term Memory, Attention mechanism

1 INTRODUCTION

As an important art form, music plays an important role in our daily life. Emotion, as the core of music, is what musicians or singers want to express through music. Emotion prediction is an important component of music information retrieval, which can provide a new trait for music retrieval and offer suitable personalized service according to user's mental state in music recommendation and therapy.

Due to the complexity and temporal variation of emotions in music, it may be ambiguous and inaccurate to mark a piece of music with one single annotation. Therefore, to depict the flow of emotions expressed in music, dynamic emotion prediction need to be done along the music. Besides, to be more precise, instead of taking it as classification, we adopt the dimensional Valence-Arousal (V-A) emotion model proposed by Russel to map emotion into 2D space [20], which is a regression problem.

In the past few years, the research of dynamic music emotion prediction has made great progress. Considering the temporal continuity of music, the mainstream emotion prediction model has changed from traditional machine learning technology to time sequential model, such as Recurrent Neural Network (RNN) [19, 23], Long Short-Term Memory (LSTM) [10, 12]. Based on these time sequential models, some researchers further improve the performance by attempts to capture the hierarchical structure information in music, a deep bidirectional LSTM (DBLSTM) model based multi-scale fusion method is proposed, in which DBLSTM models trained with feature sequences of different scales are considered to contain music structure information [15, 16].

However, in the existing methods, both the typical machine learning methods and time sequential models take this problem as a one-to-one mapping from acoustic features to emotion labels, even

[*] Corresponding author: Mingxing Xu (xumx@tsinghua.edu.cn)

MM '17, October 23–27, 2017, Mountain View, CA, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123408

though sequential model considers contextual information in the process of mapping. In our opinion, there is no one-to-one or direct relationship between music and V-A emotion label in the time domain. Limited by the psychological and physiological capability of annotators, a specific time point's annotation can be highly influenced by the music before that time point, i.e. the emotion in music at a specific time is the accumulation of a short piece of music before that point.

The recent development in attention mechanism, which attracts wide notice, adds another functionality to the RNN architecture specifically to address the problem that the sequences of input and output are not synchronized. Attention mechanism was first proposed in the field of computer vision [18], but it really draws people's attention by making progress in neural machine translation [4, 17], which accomplishes translating and aligning at the same time and solves the problem of long sentence's translation well.

Attention mechanism was also applied to other deep learning application fields. In speech recognition, some researchers adopted attention mechanism to build an end-to-end Hidden-Markov-Model-free (HMM free) recognition system [5], which helps streamline the training procedures, while some others utilize attention mechanism to capture future context for unidirectional LSTM, which gets the similar performance to bidirectional LSTM without the time delay [21]. In speech emotion recognition, attention represents the uneven distribution among frames containing emotional information [14]. In [13], the author proposed a new use of attention to fuse information across different modalities for video description.

In this paper, we propose a Multi-scale Context based Attention (MCA) model for dynamic music emotion prediction, of which the scale is a new use of attention mechanism: to give different time scales' preceding context respective attention weights. As we mentioned above, the music emotion at a specific time is the accumulation of a piece of music content before that time point. As we cannot verify how much previous content is suitable for the emotion prediction, we pay different attention to the previous context of different time scales, where the weights of different scales are dynamically computed by the model. We believe that multi-scale models fused with attention can learn the deep representation of music structure dynamically, which will utilize characteristics of different time scales of music, leading to a better performance.

It is worth mentioning that the DBLSTM-based multi-scale fusion method [15] also proposed to utilize information of different time scales, the difference between it and our proposed method is that, firstly, our method is a sequence-to-one mapping while the DBLSTM-based multi-scale fusion method is a sequence-to-sequence mapping in essence, and that, secondly, DBLSTM-based fusion method trains model in two steps, training DBLSTM and fusion model separately, while our MCA method accomplishes training in one process. What's more, in our opinion, different kinds of emotion have different means of expression, some emotion's expression may last for a period of time, while some other's may be quite shorter, so Multi-scale Context based Attention method can model the expression of music emotion better.

To be more specific, in the overall architecture of dynamic music emotion prediction, our proposed attention mechanism lies in a higher level than ordinary models such as RNN or LSTM, which is independent of underlying models, and could be considered as an attention method for blending inputs of different time scales.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 describes the Multi-scale Context based Attention (MCA) model we proposed. Section 4 provides experiment settings and process. Section 5 gives the experimental results and analysis. Finally, we give the conclusions drawn from the experiments and some possible future work in Section 6.

2 RELATED WORK

Dynamic music emotion prediction has attracted much attention, many efforts have been done in this field recently. The traditional machine learning methods, such as Multiple Linear Regression (MLR) [7], Support Vector Regression (SVR) [8], have been applied in solving this problem. Yang and Cai consider the labeling process as a continuous conditional random field (CCRF) process, where the V-A values not only depend on the specific music segment's acoustic content, but also their preceding segments, so the CCRF model with SVR as the base classifier is adopted to model continuous emotions in dimensional space [6]. Considering the context information, LSTM model makes a breakthrough at *Emotion in Music* task at *MediaEval 2014* [10].

In [15, 16], the authors propose a DBLSTM-based multi-scale fusion method for dynamic music emotion prediction, in the first step of which, different scales' DBLSTM models are trained to predict Valence (or Arousal), and in the second step, the first step's results of Valence (or Arousal) are combined as feature to predict the final Valence (or Arousal) value. The authors interpret that music has contextual continuity and hierarchical structure, which can influence the flow of emotion in music. DBLSTM models can capture contextual information while different time scales contain various emotion information, which can be associated with the structure of music, so this method can utilize both contextual information and hierarchical structure of music.

In [24], the authors analyse the music emotion prediction problem from the perspective of emotion space. The authors separately calculate the standard deviation of V-A values within a song and among a number of songs, and found that the latter is 10 times larger than the former one. Based on the analysis result, emotion is decoupled to two scales: global scale and local scale, global-scale emotion dynamics can be seen as the base platform of music emotion, and local scale can be seen as small changes on the platform. Then a double-scale SVR method is proposed to predict global-scale and local-scale emotion dynamics. The authors point out that their error in the experiment mainly comes from the prediction of global-scale emotion dynamics.

Attention mechanism first makes a breakthrough in Neural Machine Translation (NMT) [4, 17]. The NMT problem generally adopts the encoder-decoder architecture to accomplish the sequence to sequence mapping, where the input and output sequence may have different lengths and then encoder and decoder are constructed as two separate RNN networks. Before the proposal of attention mechanism, the encoder RNN network encodes the source text sequence to a fixed dimensional vector, which is called context vector, then the decoder RNN network decodes the whole target text sequence from the context vector. This processing method requires the context vector contains all the information in the source text sequence, which is problematic, especially for the translation of long sentences. As the target word in the output sequence is only relevant to some specific words in the input sequence, attention mechanism aims to select the relevant encoded hidden vectors through an informative sequence of weights, called the attention weights, and the context vector is the weighted sum of the encoded hidden vector sequence.

In [14], the authors take speech emotion recognition as a sequence-to-one learning problem, the motivation of applying attention mechanism is that the emotion information is not evenly distributed in the utterance, where some frames contain significant emotion information, while others may not. Attention mechanism tries to make the emotion-significant hidden vectors contribute a majority portion to the construction of context vector, while the effect of the irrelevant ones is minimized through the attention weights.

In [13], the authors adopt an attention-based multimodal fusion method for video description, which is similar to our idea. They propose to expand the attention model to selectively attend to specific modalities such as image features, motion features, and audio features. One big difference between our model and theirs is that ours can utilize different scales of input feature and exploit the deeper structure of data sequences. Besides, their models and ours can both be adopted in one framework as they are in different stage of the process.

3 PROPOSED METHOD

For each music clip, we have a feature sequence $\mathbf{X} =< x_1, x_2, ..., x_T >$, the annotation of Valence $\mathbf{V} =< v_1, v_2, ..., v_T >$, and that of Arousal $\mathbf{A} =< a_1, a_2, ..., a_T >$,where T is the length of music clip scaling in the time granularity of labeling. Denote $\mathbf{X}_t^l =< x_{t-l+1}, x_{t-l+2}, ..., x_t >$, and define the sequence-to-one mapping of LSTM model,

$$y_t^l = \text{LSTM}(\mathbf{X}_t^l) \tag{1}$$

where l is the length of feature sequence input into LSTM model, representing the scale of LSTM model, and y_t^l is the t-th prediction value in the prediction sequence given by LSTM considering the preceding l feature vectors before t. LSTM model can also be replaced by attention-based LSTM (A-LSTM) model.

Based on the result of LSTM or A-LSTM model, the Multi-scale Context based Attention (MCA) model can be represented as:

$$\begin{aligned} y_t &= \text{MCA}\left(y_t^{l_1}, y_t^{l_2}, ..., y_t^{l_N}\right) \\ &= \text{MCA}\left(\text{LSTM}(\mathbf{X}_t^{l_1}), \text{LSTM}(\mathbf{X}_t^{l_2}), ..., \text{LSTM}(\mathbf{X}_t^{l_N})\right) \end{aligned} \tag{2}$$

where N is the number of time scales and y_t is the result of MCA model which utilizes the information from different scales' preceding context.

In the following of this Section, Section 3.1 shows single scale LSTM model, which contains the LSTM model and A-LSTM model. In Section 3.2, the proposed MCA model using LSTM or A-LSTM is introduced.

3.1 Single scale Long Short-Term Memory

Long Short-Term Memory. Long Short-Term Memory (LSTM) is a functionally powerful sequential model, which is a redesign of Recurrent Neural Network (RNN). By adding input, forget, output gates to the memory block, LSTM is better at exploiting and storing information for longer periods of time compared to RNN. We utilize LSTM to perform a sequence-to-one mapping, the decoding process is proceeded based on the last hidden vector. Figure 1 shows the process of LSTM containing 5 time steps. For each time step $t^{'}$ in the process,

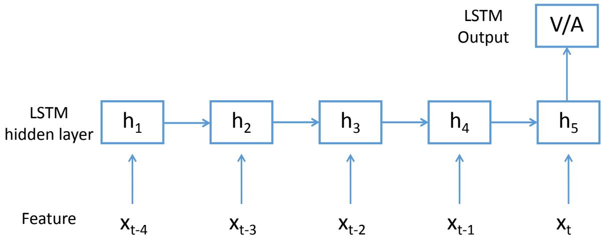

Figure 1: Framework of LSTM

$$h_{t'} = \text{LSTM}(h_{t'-1}, x_{t'}), \tag{3}$$

where the LSTM function is computed as

$$\text{LSTM}(h_{t'-1}, x_{t'}) = o_{t'} \tanh(c_{t'}), \tag{4}$$

where

$$o_{t'} = \sigma(W_{xo}x_{t'} + W_{ho}h_{t'-1} + b_o) \tag{5}$$

$$c_{t'} = f_{t'}c_{t'-1} + i_{t'} \tanh(W_{xc}x_{t'} + W_{hc}h_{t'-1} + b_c) \tag{6}$$

$$f_{t'} = \sigma(W_{xf}x_{t'} + W_{hf}h_{t'-1} + b_f) \tag{7}$$

$$i_{t'} = \sigma(W_{xi}x_{t'} + W_{hi}h_{t'-1} + b_i), \tag{8}$$

where $\sigma()$ is the element-wise sigmoid function, and $i_{t'}$, $f_{t'}$, $o_{t'}$ and $c_{t'}$ are respectively the input, forget, output gate and the cell activation vectors for the $t^{'}$-th input vector, the weight matrices W_{zz} and the bias vectors b_z are identified by the subscript $z \in \{x, h, i, f, o, c\}$.

The prediction of Valence or Arousal is based on the last time step's hidden vector h_t

$$y_t = \tanh(W_{hv}h_t + b_v). \tag{9}$$

Attention-based LSTM. The basic idea of attention mechanism based LSTM (A-LSTM) model is to select relevant encoded hidden vectors through an informative sequence of weights, which are called attention weights, in the decoding phase. This mechanism coincides with the characteristic of music emotion's expression that emotion is not evenly distributed in music and only some moments in music arouse people's emotion response strongly. The average of encoded hidden vectors' sequence can be seen as a special case of attention mechanism, i.e. uniform attention or non-attention. In speech emotion recognition, it has been proved that uniform attention is a better processing method than only taking the one at the last time step [14], so the attention mechanism which gives different encoded hidden vector different weight is more reasonable.

Figure 2 is the schematic of A-LSTM modified from the basic LSTM model in Figure 1. The difference between A-LSTM and LSTM

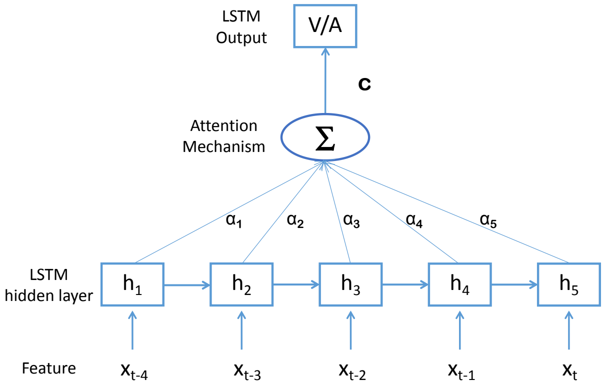

Figure 2: Framework of attention-based LSTM

is that in the phase of predicting Valence-Arousal, attention-based LSTM replaces h_t with c_t in Eq.9

$$y_t = \tanh(W_{hv} c_t + b_v), \tag{10}$$

where c_t is the weighted sum of encoded hidden vectors, which can be computed as

$$c_t = \sum_{t'=1}^{l} \alpha_{t'} h_{t'}, \tag{11}$$

where

$$\alpha_{t'} = \frac{\exp\left(e_{t'}\right)}{\sum_{t'=1}^{l} \exp\left(e_{t'}\right)} \tag{12}$$

$$e_{t'} = \tanh(W_A h_{t'} + b_A), \tag{13}$$

where $\alpha_{t'}$ is the computed attention weights, $\sum_{t'=1}^{l} \alpha_{t'} = 1$. W_A, b_A are the parameter matrix and bias in the attention functionality. l is the length of input sequence, i.e. the length of preceding context sequence.

3.2 Multi-scale Context based Attention model

We extend attention mechanism to multi-scale context fusion. As we proposed that the expression of music emotion at a moment is the accumulation of the previous context before that moment in music, the scale of the previous context can influence the prediction of music emotion. There is no conclusion about how much previous context is most beneficial to the prediction of emotion, so we propose the Multi-scale Context based Attention (MCA), let the model choose by itself, giving larger weights to the relevant context vectors while minimizing that of the irrelevant context vectors.

Figure 3 shows the architecture of MCA model. Attention mechanism is applied to the context vectors of LSTM models with different scales. The LSTM model can be replaced by A-LSTM model described in Section 3.1. If so, the context vector is the weighted sum of the hidden vectors. If not, the context vector is the last hidden vector of LSTM model. Compared to some simple fusion methods, where the context vectors from the same sub-network share the same weights independent of the context vectors, in MCA model the attention weights change according to the context vector (Eq.16 and Eq.17), which can better model the expression of emotion in music. With the MCA model, the prediction of Valence-Arousal is based on the weighted sum of multi-scale context vectors, instead of the context vector from a single time scale:

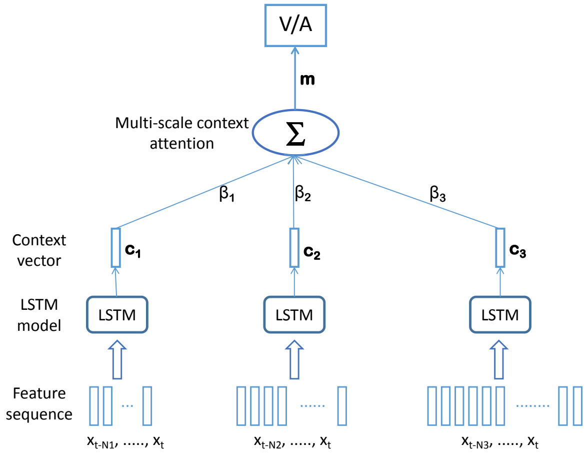

Figure 3: Framework of Multi-scale Context based Attention model

$$y_t = \tanh\left(W_{hv} m_t + b_v\right), \tag{14}$$

where m_t is the fusion result, the weighted sum of multi-scales' context vectors, which can be computed as

$$m_t = \sum_{i=1}^{N} \beta_t^i c_t^i, \tag{15}$$

where

$$\beta_t^i = \frac{\exp\left(e_t^i\right)}{\sum_{i=1}^{N} \exp\left(e_t^i\right)} \tag{16}$$

$$e_t^i = \tanh(W_{FA} c_t^i + b_{FA}), \tag{17}$$

where W_{FA}, b_{FA} are the parameter matrix and bias, and N is the number of time scales, which is specified in the experiment. c_t^i is the context vector of the i-th time scale, and β_t^i is the corresponding attention weight.

If $\beta_t^1 = \beta_t^2 ... = \beta_t^N = 1/N$, the method degenerates to a multi-scale average fusion, which becomes a special case of attention mechanism, i.e. uniform MCA.

4 EXPERIMENT SETTINGS

4.1 Data collection and annotation

The data we use comes from the *Emotion in Music* at *MediaEval 2015* [3], and the training and test sets of our experiments are the same as other participants' in the task. The training set contains 431 music clips of 45 seconds from different songs, extracted from random (uniformly distributed) start point of a song. The test set

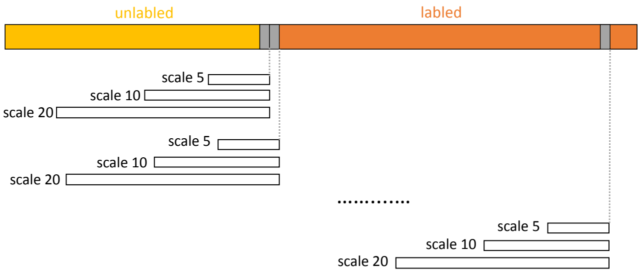

Figure 4: Data preparation for MCA training

consists of 58 complete songs with an average duration of 234 ± 105.7 seconds.

Arousal and Valence were annotated separately for each song in the training and test set by 5-7 annotators, who listened to the entire song before annotation in order to get familiar with the music and reduce the reaction time lag. The temporal resolution of annotation was 500 ms. Since the dynamic annotations of the first 15 seconds of each song or clip were not stable, they were not provided by the organizers of the task. The Cronbach's α of training set is 0.76 ± 0.12 for Arousal, and 0.73 ± 0.12 for Valence, while that of test set is 0.65 ± 0.28 for Arousal, and 0.29 ± 0.94 for Valence [3].

4.2 Feature extraction

Our experiments use the baseline universal feature set from the organizers of MediaEval 2015 [3]. The features are extracted with the openSMILE toolbox [11], consisting of means and standard deviations of 65 low-level acoustic descriptors (LLDs) and their first-order derivatives in non-overlapping segments of 500 ms. The LLDs contain energy-related LLDs (e.g. RMS energy, zero-crossing rate, and etc.), spectral LLDs (e.g. Mel Frequency Cepstral Coefficients, Spectral flux, centroid, entropy, slope, and etc.) and voicing related LLDs (e.g. pitch, Prob. of voice, Log. HNR, and etc.), which are extracted using the openSMILE toolbox with a frame size of 60 ms and a frame shift of 10 ms.

All features have been normalized to zero mean and unit variance in advance.

4.3 Model training

Data preprocessing. In order to obtain input features of different scales, we apply multiple sliding windows of different size but of the same stride on the input features extracted in Section 4.2. Taking features of two scales for instance, two sliding windows contain 5 frames and 10 frames respectively, but both step forward once by 1 frame. Thus, these two segments of input features share the same annotation at each specific point.

Considering the acoustic property of music and its emotional influence upon people, we decide to adopt three different window sizes, which are 5, 10, and 20 frames long (i.e. 2.5, 5, and 10 seconds respectively), and all step forward once by one frame, as illustrated in Figure 4.

Specifically, as mentioned in Section 4.1, the input music data provided by MediaEval 2015 contains 15-second-long unlabeled music data before the labeled ones, which cannot be used directly in training process. However, features extracted from these segments still hold the potential information of music emotion, which are bases of subsequent labeled data and naturally can be used in generating different scales of data. Therefore, edge cases needn't be handled specially.

LSTM Training. In our experiments, LSTM models are trained for Valence and Arousal separately. In order to find the best model hyper-parameters (hidden units, layer numbers, etc.), we randomly select 50 music clips from training set for validation.

The weights in LSTM layers are initialized randomly with zero mean and different standard deviations, which are also hyper-parameters to be determined. Attention length of LSTM equals to the length of input scales (5, 10, 20 frames respectively). Models are trained with RMSProp algorithm [22], whose initial learning rate is 0.005. The training is stopped after 50 epochs, each of which means one full iteration of input data.

To avoid over-fitting, half of LSTM cells might be dropped out randomly. L2 regularization term of weight matrix is also added to training loss. In addition, training data are randomly shuffled in each epoch.

In order to alleviate the fluctuation of model output, a 25-frame-long triangle filter is applied to the output sequence of each music, as a post processing step to smooth out the random noise.

All models are implemented and trained with TensorFlow [1]. The best model hyper-parameters of each set of experiments are selected based on validation set's loss and then evaluated on the test set.

5 RESULTS AND DISCUSSION

To demonstrate the effectiveness of Multi-scale Context based Attention model, we've done three sets of experiments, which are single scale's LSTM models, uniform MCA and MCA respectively. The prediction accuracy is evaluated with RMSE (root mean square error) on the test set.

In all of the following tables, A-LSTM represents attention-based LSTM model. h is the hidden size of each layer. n represents the number of hidden layers and l represents the length of feature sequence which is the input of LSTM models. Each row of tables represents one set of experiments, and the model size shown is the best model selected according to the validation set's loss. Thus, the model size of each row of tables may vary.

5.1 Performance of single scale LSTM

Table 1 and Table 2 show the RMSE results of Valence and Arousal of single scale LSTM models respectively.

First, we compare the performance of LSTM model and attention-based LSTM (A-LSTM) model from Table 1 and Table 2. In Table 1, we can see that for Valence, if the sequence length is 10 or 20, adding attention functionality to LSTM model can greatly improve the model's performance, while the attention-based LSTM model performs poorer than LSTM model if the sequence length is 5. Table 2 shows that for Arousal, adding attention functionality achieves a little worse result for any sequence lengths.

Second, we observe the performance variation along with the change of the sequence length of attention-based LSTM model and LSTM model separately. For Valence, the difference of sequence

Table 1: Valence RMSE of single scale model

l	Model	Model Size	RMSE
5	LSTM	h=200, n=3	0.358
	A-LSTM	h=200, n=2	0.366↑
10	LSTM	h=200, n=4	0.352
	A-LSTM	h=200, n=2	0.308↓
20	LSTM	h=100, n=4	0.355
	A-LSTM	h=100, n=2	0.294↓

Table 2: Arousal RMSE of single scale model

l	Model	Model Size	RMSE
5	LSTM	h=200, n=4	0.244
	A-LSTM	h=100, n=2	0.255↑
10	LSTM	h=100, n=4	0.261
	A-LSTM	h=100, n=2	0.268↑
20	LSTM	h=200, n=3	0.268
	A-LSTM	h=100, n=2	0.272↑

Table 3: Valence RMSE of multi-scale model

Method	Model	Model Size	RMSE
uniform MCA	LSTM	h=100, n=3	0.382
	A-LSTM	h=100, n=3	0.292
MCA	LSTM	h=100, n=2	0.357
	A-LSTM	h=150, n=3	**0.291**

length does not influence the performance of LSTM models, and for Arousal, LSTM model performs worse along with the increase of sequence length. As for attention-based LSTM model, with the increase of sequence length, the RMSE error of Valence is reduced while that of Arousal increases. Our experiments indicate the different trends of variation of Arousal and Valence while changing sequence length, which coincides with the results in the DBLSTM-based multi-scale fusion method [15]. In our view, this experimental result is related to the nature of Valence and Arousal, as Valence represents positive or negative emotion and Arousal describes the energy of emotion.

5.2 Performance of MCA models

We compare our proposed model, Multi-scale Context based Attention (MCA) model, with the uniform MCA defined in Section 3.2. The uniform MCA is a special case of attention mechanism, which takes music segments of different time scales and average them uniformly and constantly, so we call it uniform MCA.

Table 4: Arousal RMSE of multi-scale model

Method	Model	Model Size	RMSE
uniform MCA	LSTM	h=100, n=2	0.257
	A-LSTM	h=100, n=2	0.270
MCA	LSTM	h=100, n=2	0.260
	A-LSTM	h=150, n=2	**0.241**

Table 5: RMSE of different regression models

Model	Valence	Arousal
MLR [3]	0.366	0.270
SVR [8]	0.366	0.255
LSTM [9]	0.373	0.242
RNN+smooth [19]	0.365	0.247
SVR+CCRF [6]	0.343	0.241
GPR [2]	0.295	0.285
DBLSTM+ELM [16]	0.318	0.239
DBLSTM based multi-scale fusion [15]	**0.285**	**0.225**

From Table 3 and Table 4 we can find that, our proposed MCA model using A-LSTM outperforms uniform MCA model in Valence and Arousal, especially in Arousal. This result supports our assumption that multi-scale models fused with attention can utilize different time scales' characteristics of music and learn the deep representation of music structure dynamically, which enhances the model performance to some extent.

In addition, MCA model using A-LSTM (Table 3, Table 4) also has better performance compared to single scale ones (Table 1, Table 2) in both Valence and Arousal, indicating that different segment lengths of music may contain different potential information for emotion analysis. Therefore, fully utilizing the information of different scales leads to a better emotion prediction result.

5.3 Experimental results of related work

The results in Table 5 come from the *Emotion in Music* task at *MediaEval 2015* using the same dataset and baseline features mentioned above, so the models below are comparable to ours.

The MLR (Multiple Linear Regression) method is provided by the organizers of *MediaEval 2015* [3]. The SVR (Support Vector Regression) method uses Radial Basis Function (RBF) kernel function [8]. The LSTM method uses a deep LSTM-RNN with 3 hidden layers [9]. The RNN+smooth method performs the regression with a RNN of 10 hidden units and smoothing with a moving average filter [19]. The SVR+CCRF method adopts the Continuous Conditional Random Field model with SVR as the base classifier to model continuous emotions [6]. The GPR method uses Gaussian Processing regression to predict the emotion per segment [2]. The

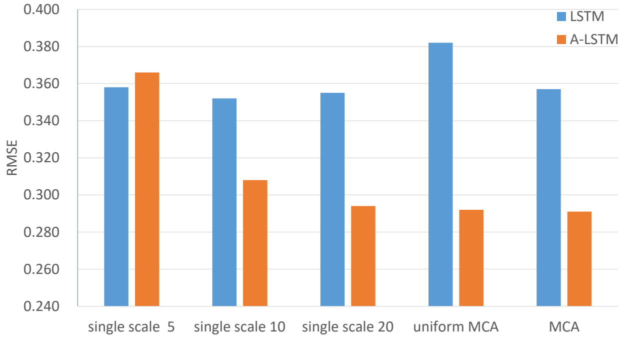

Figure 5: Valence RMSE of all models

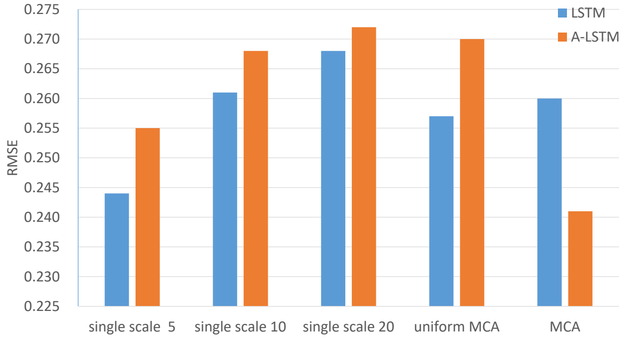

Figure 6: Arousal RMSE of all models

DBLSTM+ELM model uses Extreme Learning Machine (ELM) to fuse the results of DBLSTM models with different scales [16]. The DBLSTM based multi-scale fusion method proposes a multi-scale fusion model based on deep bidirectional LSTM, which achieves the best RMSE on Valence and Arousal [15].

5.4 Discussion

We've merged the results of all tables to figures (Figure 5 for Valence, and Figure 6 for Arousal) to be more intuitive. If we just examine the first three orange bars (A-LSTM of single scale models) of the figures, we can find that the increasing of sequence length can enhance the performance of Valence, though weaken that of Arousal. This could be on account of the different characteristic of Valence and Arousal. Valence relies more on long-term influence of input while Arousal relies more on short-term influence. If examining all of the bars of Arousal, one obvious finding is the rightmost bars of figures, i.e. our proposed MCA model using A-LSTM, outperforms single scale models and uniform MCA models.

As we can see from figures and tables, performance gain in Valence is not so much as that in Arousal. In our opinion, this could be account for the following reasons: First, Valence is actually harder to predict than Arousal, which can be discovered through the fact that all Valence results are worse than Arousal results. Second, the test set's annotation of Valence has a poorer inter-annotation agreement than that of Arousal [3], which could influence our results to some extent.

When compared to state-of-the-art models, our model surpasses the performance of all single models, except the DBLSTM based multi-scale fusion one, which exploits the information contained in multiple models and blend them to gain a better performance. Thus, it comes as no surprise that our proposed model performs a little bit worse in comparison.

Besides, the DBLSTM based multi-scale fusion model has some advantages for ours to incorporate: First, we've only utilized the preceding content of the sequence, without using the subsequent content, which can still hold potential information of music emotion. Second, our MCA model might lack the correlation in the emotion label sequence, which means that the inner states of LSTM model in different sequences of same music may not be consistent, as our model is a sequence-to-one mapping, instead of sequence-to-sequence. These two differences could explain the little gap between the performance of DBLSTM-based multi-scale fusion method and ours. We would like to address them in future.

6 CONCLUSION

In this paper, we propose a Multi-scale Context based Attention (MCA) model for dynamic music emotion prediction and compare the performance of our model with single scale models and uniform MCA models. The results show that our MCA using attention-based LSTM outperforms most of the models using same features and achieves a competitive performance with the state-of-the-art methods. Our model obtains the mapping from a sequence of acoustic features to one emotion label, based on the assumption that music emotion at a specific time is the accumulation of a piece of music. Because there is yet no definitive conclusion about the most effective time scale for emotion prediction, and Valence and Arousal represent different characteristics along with the change of time scale, the proposed MCA model can help make up for these deficiencies by giving the more suitable time scale larger attention weights dynamically adjusted by the data.

As for future work, we plan to further evaluate our finding using more data and on some other databases of music emotion prediction. Furthermore, our proposed method can also be effective in other field of studies, such as sentiment analysis of natural language, emotion recognition of speech and affective impact prediction of movie, all of which contain sequence of input and sequences of variant scales have different influences on results. Actually, problems containing this characteristic could all benefit from our proposed method and remain to be further investigated.

ACKNOWLEDGMENTS

This work was supported by the National High Technology Research and Development Program of China (863 program) (2015AA016305) and the National Natural Science Foundation of China (61433018, 61171116).

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al.

2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Anna Aljanaki, Frans Wiering, and Remco C. Veltkamp. 2015. MediaEval 2015: A Segmentation-based Approach to Continuous Emotion Tracking. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, 2015.* http://ceur-ws.org/Vol-1436/Paper82.pdf

[3] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. 2015. Emotion in music task at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *Computer Science* (2014).

[5] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2015. End-to-End Attention-based Large Vocabulary Speech Recognition. *Computer Science* (2015).

[6] Kang Cai, Wanyi Yang, Yao Cheng, Deshun Yang, and Xiaoou Chen. 2015. PKU-AIPL' Solution for MediaEval 2015 Emotion in Music Task. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, 2015.* http://ceur-ws.org/Vol-1436/Paper57.pdf

[7] Yu An Chen, Ju Chiang Wang, Yi Hsuan Yang, and Homer Chen. 2014. Linear regression-based adaptation of music emotion recognition models for personalization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2149–2153.

[8] Michal Chmulik, Igor Guoth, Miroslav Malik, and Roman Jarina. 2015. UNIZA System for the "Emotion in Music" task at MediaEval 2015. In *MediaEval 2015 Multimedia Benchmark Workshop*.

[9] Eduardo Coutinho, George Trigeorgis, Stefanos Zafeiriou, and Björn W Schuller. 2015. Automatically Estimating Emotion in Music with Deep Long-Short Term Memory Recurrent Neural Networks. In *Mediaeval 2015 Multimedia Benchmark Workshop, Satellite of INTERSPEECH*. 1–3.

[10] Eduardo Coutinho, Felix Weninger, Björn Schuller, and Klaus R Scherer. 2014. The Munich LSTM-RNN Approach to the MediaEval 2014 Emotion in Music Task. In *MediaEval Workshop*. 5–6.

[11] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 835–838.

[12] Alan Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 273–278.

[13] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Kazuhiro Sumi, John R Hershey, and Tim K Marks. 2017. Attention-Based Multimodal Fusion for Video Description. *arXiv preprint arXiv:1701.03126* (2017).

[14] Che-Wei Huang and Shrikanth S. Narayanan. 2016. Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016.* 1387–1391. https://doi.org/10.21437/Interspeech.2016-448

[15] Xinxing Li, Jiashen Tian, Mingxing Xu, Yishuang Ning, and Lianhong Cai. 2016. DBLSTM-based multi-scale fusion for dynamic emotion prediction in music. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME.2016.7552956

[16] Xinxing Li, Haishu Xianyu, Jiashen Tian, and Wenxiao Chen. 2016. A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 544–548.

[17] Minh Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *Computer Science* (2015).

[18] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent Models of Visual Attention. 3 (2014), 2204–2212.

[19] Thomas Pellegrini and Valentin Barriere. 2015. Time-continuous Estimation of Emotion in Music with Recurrent Neural Networks. In *Working Notes Proceedings of the MediaEval 2015 Workshop*. 60.

[20] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[21] Jian Tang, Shiliang Zhang, Si Wei, and Li Rong Dai. 2016. Future Context Attention for Unidirectional LSTM Based Acoustic Model. In *INTERSPEECH*. 3394–3398.

[22] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4, 2 (2012).

[23] Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation* 1, 2 (1989), 270–280.

[24] Haishu Xianyu, Xinxing Li, Wenxiao Chen, Fanhang Meng, Jiashen Tian, Mingxing Xu, and Lianhong Cai. 2016. SVR based double-scale regression for dynamic emotion prediction in music. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 549–553.