A DEEP BIDIRECTIONAL LONG SHORT-TERM MEMORY BASED MULTI-SCALE APPROACH FOR MUSIC DYNAMIC EMOTION PREDICTION

Xinxing Li, Haishu Xianyu, Jiashen Tian,Wenxiao Chen, Fanhang Meng, Mingxing Xu, Lianhong Cai

Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing, China
xumx@tsinghua.edu.cn, {lixinxin14,xyhs14}@mails.tsinghua.edu.cn

ABSTRACT

Music Dynamic Emotion Prediction is a challenging and significant task. In this paper, We adopt the dimensional valence-arousal (V-A) emotion model to represent the dynamic emotion in music. Considering the high context correlation among the music feature sequence and the advantage of Bidirectional Long Short-Term Memory (BLSTM) in capturing sequence information, we propose a multi-scale approach, Deep BLSTM (DBLSTM) based multi-scale regression and fusion with Extreme Learning Machine (ELM), to predict the V-A values in music. We achieved the best performance on the database of *Emotion in Music* task in *MediaEval 2015* compared with other submitted results. The experimental results demonstrated the effectiveness of our novel proposed multi-scale DBLSTM-ELM model.

Index Terms— Music dynamic emotion prediction, multi-scale, deep bidirectional long short-term memory, extreme learning machine

1. INTRODUCTION

Music is an essential part of human life. In today's digital era, people can have access to music through network and electronic products expediently. Emotion, as one of crucial element in music, is significantly meaningful in real-world applications. It can be applied to broad areas that include massive music data management, personalized music recommendation and psycho-music therapy.

There are several psychological emotion models. They can be categorized into two classes: 1) discrete emotion states; 2) dimensional continuous emotion space [1]. Among the second class, the valence-arousal (V-A) emotion model proposed by Russel in [2] to depict the emotion in the 2D plane is widely used, by which way the music dynamic emotion prediction can be translated into the regression of V-A value in music.

Music dynamic emotion prediction in the dimensional emotion model is a big challenge. Many efforts have been done in this field recently. The traditional regression methods, such as Multiple Linear Regression (MLR) [3], Support Vector Regression (SVR) [4], have been applied in solving this problem. However, the effectiveness of these traditional regression methods are not prominent [5, 6, 7], partly because these regression methods just get information from the single annotated frame and do not take the context information in the sequence into account.

In recent work, Long Short-Term Memory (LSTM) [8] model has made a breakthrough in 'Emotion in Music Task' in MediaEval 2014 [9] with its ability of accessing long range previous context. Considering the process of composing, performing and annotating music, the emotion within music is not only associated with the previous context but also with the future ones. In order to make use of context in both time-based directions, Bidirectional Long Short-Term Memory (BLSTM) [8] appears to be a good choice. This can be seen by its performance in numerous research involving sequence modelling, such as large-vocabulary speech recognition [10] and handwriting recognition [11]. The choice in this paper is to choose Deep BLSTM (DBLSTM) as the regression model. For all we know, it is the first time that DBLSTM is adapted to the task of emotion in music, and experimental results demonstrate DBLSTM's better performance in sequence regression.

Although a BLSTM model has the ability to capture both the previous and future contexts over a long period of time, the information obtained by using BLSTM is still limited by the length of the sequence. Therefore we have proposed a multi-scale fusion approach based on Extreme Learning Machine (ELM) to promote the performance of the BLSTM model.

This paper is organized as follows. Section 2 describes the multi-scale DBLSTM-ELM model. Section 3 provides the setting and the process of the experiment. Section 4 gives the experimental results and analysis. Section 5 is the conclusions drawn from the experiment.

This work is supported by the 973 Program (2012CB316401) of China, and is partially supported by the National Natural Science Foundation of China (61171116, 61433018, 61375027) and 863 Program (2015AA016305).

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

2. DBLSTM-ELM MODEL

The section begins with a brief introduction to the DBLSTM and ELM models, followed by a description of the structure of our model.

2.1. Deep BLSTM

Neural networks have developing over time. Recurrent neural networks (RNNs) is a significant branch of neural networks that has shown a clear ability for the processing of sequence and time. LSTM is a redesign of the RNN model around the memory cell [12]. Compared with RNN, LSTM is better at exploiting and storing information for long periods of time, which is a benefit achieved from the use of special purpose-built memory cells units [13]. Each memory block consists of three gate units, which are input, output, and forget gates which each separately have the ability to write, read and reset the functionality of the cell [14]. Particular among the three gates, forget gates are shown to be essential for very long input sequence [15].

Although LSTM can have access to context for long periods of time, both LSTM and RNN can only get information from the previous context, they can not make use of the future context. As with the prediction of sequence, there is richer purpose if this can be extended to exploit information in both directions. Bidirectional RNNs (BRNN) perform this work by processing the sequence with two separate hidden layers in both the forward and backward direction [12, 16]. Fig. 1 illustrates the architecture of BRNNs. As can be seen in Fig. 1, the Inputs are fed forwards separately to the Forward Layer from t=1 to T and the Backward Layer from t=T to 1 and then the Forward and Backward Layer are fed forward to the same Output Layer.

BLSTM is a combination of LSTM and BRNN [14]. Thus the BLSTM not only can exploit context for long periods of time, but also can have access to the context in both previous and future directions. This idea was developed from considering the deep feedforward network, in order to get better representation of data, multiple recurrent hidden layers were stacked on the top of each other, and so the final DBLSTM model used in this paper was reached.

2.2. ELM

Extreme Learning Machine (ELM) [17] is a learning algorithm for single-hidden layer feedforward neural networks (SLFN). The input weights and hidden layer biases of SLFNs are randomly assigned, and the output weights are analytically determined. For N arbitrary distinct samples ($\mathbf{x}_i$, $\mathbf{t}_i$) where $\mathbf{x}_i$ = [x_{i1}, x_{i2},..., x_{in}]^T and $\mathbf{t}_i$ = [t_{i1}, t_{i2},..., t_{im}]^T, we have a SLFN with $\tilde{N}$ hidden nodes and activation function $g(x)$. The input weights $\mathbf{w}_i$ and bias b_i is randomly assigned. To decide the output weights, the hidden layer output matrix is calculated in the following way [17]:

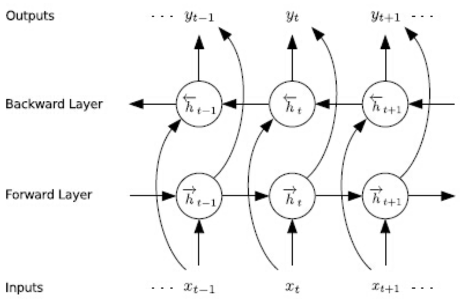

Fig. 1. Bidirectional RNN[13]

$$\mathbf{H}(\mathbf{w}_1, ..., \mathbf{w}_{\tilde{N}}, b_1, ..., b_{\tilde{N}}, \mathbf{x}_1, ..., \mathbf{x}_N) \\ = \begin{bmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & g(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_1 + b_{\tilde{N}}) \\ \vdots & \cdots & \vdots \\ g(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1) & \cdots & g(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}}$$

where $\mathbf{w}_i = [w_{i1}, w_{i2}, ..., w_{in}]^T$ is the weight vector connecting the ith hidden node and the input nodes, $\beta_i = [\beta_{i1}, \beta_{i2}, ..., \beta_{im}]^T$is the weight vector connecting the ith hidden node and the output nodes, b_i is the threshold of the ith hidden node, and $g(x)$ is the activation function.

The output weight β is calculated: $\beta = \hat{\mathbf{H}}\mathbf{T}$, where $\hat{\mathbf{H}}$ is the Morre-Penrose generalize inverse of matrix $\mathbf{H}$ and $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ...\mathbf{t}_N]^T$, $\mathbf{T}$ is a $N \times m$ matrix.

2.3. Model Structure

The proposed model was reached by the combination of D-BLSTM and ELM models. The DBLSTM-ELM model structure was as is presented in Fig. 2. In Fig. 2, m$_1$,...m$_k$ represent the DBLSTM models with different sequence lengths. The DBLSTM models with different sequence length provide the temporary predictions first, then the ELM model accomplishes the fusion of multi-scale results.

$\mathbf{o}_i = [o_1, ..., o_t]^T$, is the output of the DBLSTM model m$_i$, t is the sequence length. $\mathbf{d}_i$ is the differential of $\mathbf{o}_i$ while $\mathbf{s}_i$ denotes the value of $\mathbf{o}_i$ after smoothing. Combining $\mathbf{o}_i$, $\mathbf{d}_i$ and $\mathbf{s}_i$, produced a supervector $\mathbf{W} = [\mathbf{o}_1, \mathbf{d}_1, \mathbf{s}_1, ..., \mathbf{o}_k, \mathbf{d}_k, \mathbf{s}_k]$ as the input to ELM. The output of ELM is the final result.

3. EXPERIMENT SETTINGS

3.1. Data and Feature

The data used was sourced from the *Emotion in Music* task in *MediaEval 2015* [18]. The organizers provided the development set which consisted of 431 clips of 30 seconds annotated with precise V/A values every 500 ms from different songs, with the songs themselves appended. The evaluation set consisted of 58 full-length songs. There was a need to predict both A and V values every 500 ms for the 58 complete songs.

The organizers also provided a baseline feature set of the development set and the test set. The baseline feature set

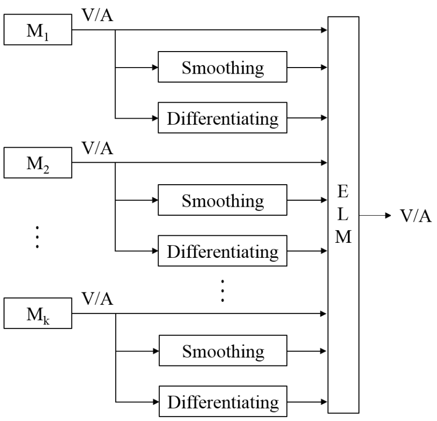

Fig. 2. Framework of DBLSTM-ELM

was extracted with *openSMILE*, and consisted of 260 low-level features including the mean and standard deviation of 65 low-level acoustic descriptors, and their first-order derivatives [18].

3.2. Model Training

3.2.1. Data Partition

In total there were five different data partitions, 411 clips as training data and 20 clips as validation data. The 20 validation data were randomly selected according to the genre distribution of the test data (the 58 complete songs), there was no overlap between the five sets of validation data.

3.2.2. DBLSTM Training

Separately the DBLSTM model was trained for valence and arousal regression. In order to find the best network structure, the performance of DBLSTM models were compared with different number of layers and units on the validation data. The number of layers that were trialed were altered from one to six. With respect to the number of units, these were trialed for 100, 150, 200, 250 and 300. Finally the DBLSTM model with five layers and 250 units was used.

The first two layers were pre-trained with features of 431 full-length songs in the way similar to the training of an autoencoder. The weights in the other three layers were initialized using random numbers with zero mean and standard deviation 0.1. Training with learning rate 5E-6 and momentum 0.9 was stopped after a maximum of 100 iterations or after 20 iterations if there was no new best error on the validation data. To alleviate over-fitting, Gaussian noise with zero mean and standard deviation of 0.6 was added to the input sequence, and sequences were presented in random order during training. All DBLSTM models were trained with CURRENNT.[1]

Four kinds of DBLSTM models were trained with different sequence scales of 10, 20, 30 and 60, respectively. Each kind of model was trained with the five data partitions for three trails, finally producing 15 models for each sequence scale.

3.3. Model Selection

In order to select four models with different sequence scales for fusion, two different criteria were applied separately to become two groups of four models. The first criterion was Root Mean Square Error (RMSE) which initially selected the model with the best RMSE for each sequence scale, while the second one considered both the RMSE and the data partition in order to guarantee the training sets of the selected models would be different from each other. This was done for valence and arousal separately. In the experiments, for both valence and arousal, there were two models shared by the two groups. In summary, there were six unique models for fusion.

3.4. ELM Training

ELMs were trained for valence and arousal separately. The output of 20 DBLSTMs of the four scales on the five validation sets were used for the training. Each of the 20 DBLSTMs chosen was the best of the 3 trials performed with the same scale and on the same train set. The outputs of the DBLSTM models were used with their delta derivatives and smoothed values to compose the supervectors, which were then input to the ELM for training.

3.5. Emotion Prediction for Test Data

The 58 complete songs in the test set were cut into four different lengths: 10, 20, 30 and 60 half-seconds. Then the test set data, now in different lengths, was used in order to predict using the corresponding DBLSTM models in the two model groups (selected by the two model selection criteria, as outlined in 3.3 above). Then, in order to fuse the predictions of selected DBLSTMs of different scales, the DBLSTMs' output together with their delta derivatives and smoothed results were composed into a supervector, which was input to the trained ELM model. Then the two sets of fused predictions of the two model groups were averaged to produce the final prediction of V/A values. In addition to this fusion policy, another fusion was created by averaging the predictions produced by all six models for comparison.

4. RESULTS AND DISCUSSION

To evaluate the performance of the DBLSTM-ELM model, two objective experiments were conducted. The prediction accuracy was evaluated with RMSE.

[1] http://sourceforge.net/p/currennt

4.1. Performance of DBLSTM with Different Length

Table 1 presents the result for the first experiment using DBLSTM with different sequence lengths on the validation set and test set. We note that the error on the test data was much higher than the one on the validation data. For one thing there was a mismatch between the development and test sets in terms of the data sources (FMA versus medleyDB and jamendo); for another, the development data was selected from *MediaEval2014* dataset to maximize inter-annotator agreement (the Cronbach's α is 0.76 ± 0.12 for arousal, and 0.73 ± 0.12 for valence), while on test set, that is 0.65 ± 0.28 for arousal, and 0.29 ± 0.94 for valence.

The results on the validation set show the DBLSTM models' effectiveness in music affective computing. Table 1 suggests that the regression accuracy of DBLSTM was related to the sequence length. With valence values on test set, larger scales tended to produce more accurate predictions; whereas for arousal values smaller scales gave a better performance on the test set. On test set, the model with length of 60 gave the best results on valence, whereas for arousal the model with length of 10 performed best. The performance of models with length of 20 and 30 was about equal on both valence and arousal.

4.2. Results after Multi-Scale Fusion

Table 2 presents the RMSE after the multi-scale fusion in the second experiment. The results in the first and third row of Table 2 were from the submission of the *Emotion in Music* task for which different pre-training data were used. The pre-training data included the 431 clips of 30 seconds from the development set, together with the 58 full songs from the test set. It can be seen that when the test set is included during pre-training process, the accuracy of prediction improved marginally.

The results in the second and fourth row of Table 2 were the fusion of the regression results in Table 1. Compared with Table 1, it can be observed that the fusion methods of both average (AVG) and ELM performed better overall than predictions given by a single scale. In detail, both fusion methods outperform the four kinds of DBLSTM with valence, as for arousal, the results fusion methods were only marginally worse than for DBLSTM with 10 sequence-length and better than the other three variants. It can be also seen that ELM gave better results than AVG in regards to valence and that AVG outperformed ELM with respect to arousal.

4.3. Comparison with Other Works

The results presented in Table 3 each came from the *Emotion in Music* task using the same baseline features as outlined in 3.1 above. For both valence and arousal, DBLSTM-based fusion method gets the best results. Compared with Table 1, DBLSTM models with any sequence length performs better than LSTM model on valence, and on arousal all except sequence length 60 get better results than LSTM model, which demonstrates the effectiveness of capturing information in both previous and future directions for music dynamic emotion prediction.

Table 1. RMSE of DBLSTM models with different sequence length

Sequence Length	Validation Data		Test Data	
	V-RMSE	A-RMSE	V-RMSE	A-RMSE
10	0.181	0.157	0.364	**0.231**
20	0.188	0.152	0.360	0.241
30	0.180	**0.148**	0.363	0.240
60	**0.162**	0.151	**0.333**	0.279

Table 2. RMSE of V/A after multi-scale fusion
The content in the parenthesis is the data used in pre-train, '431x30' represents the 431 clips with length of 30 seconds, '58-full' represents 58 full-length songs in test set, '431-full' represents 431 full-length songs in the training set

Fusion(Pre-train)	V-RMSE	A-RMSE
AVG(431x30, 58-full)	0.331	**0.230**
AVG(431-full)	0.331	0.233
ELM(431x30, 58-full)	**0.308**	0.234
ELM(431-full)	0.318	0.239

Table 3. RMSE of different regression models

Model	V-RMSE	A-RMSE
MLR [19]	0.366	0.270
SVR [20]	0.366	0.255
RNN+smooth [21]	0.365	0.247
LSTM [22]	0.373	0.242

5. CONCLUSIONS

This paper presents a proposal to apply DBLSTM-based multi-scale regression and fusion with ELM in order to predict the value of V/A in music. The experimental DBLSTM model used presented an advantageous position over that of other regression models in sequence processing, and multi-scale fusion was seen to further enhance its performance. On further investigation, it was also discovered that the performance of DBLSTM was related to the sequence scale.

In future work, we would intend to explore the effect of different sequence lengths on the performance of the model and the performance of the DBLSTM when there is a mismatching of the length of training sequences and testing sequences.

6. REFERENCES

[1] Kai Sun, Junqing Yu, Yue Huang, and Xiaoqiang Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 566–569.

[2] James A Russell, "A circumplex model of affect.," in *Journal of personality and social psychology*. 1980, vol. 39, pp. 1161–1178, American Psychological Association.

[3] Shean-Tsong Chiu, "Regression analysis: Theory, methods, and applications," 1991, vol. 33, pp. 479–480, Taylor & Francis Group.

[4] Alex J Smola and Bernhard Schölkopf, "A tutorial on support vector regression," in *Statistics and computing*. 2004, vol. 14, pp. 199–222, Springer.

[5] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen, "A regression approach to music emotion recognition," in *Audio, Speech, and Language Processing, IEEE Transactions on*. 2008, vol. 16, pp. 448–457, IEEE.

[6] Byeong-jun Han, Seungmin Rho, Roger B. Dannenberg, and Eenjun Hwang, "Smers: Music emotion recognition using support vector regression," in *Proceedings of the 10th International Society for Music Information Retrieval Conference*, 2009, pp. 651–656.

[7] Erik M Schmidt, Douglas Turnbull, and Youngmoo E Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proceedings of the international conference on Multimedia information retrieval*. ACM, 2010, pp. 267–274.

[8] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," in *Neural computation*. 1997, vol. 9, pp. 1735–1780, MIT Press.

[9] Eduardo Coutinho, Felix Weninger, Björn Schuller, and Klaus R Scherer, "The munich lstm-rnn approach to the mediaeval 2014 emotion in music task," in *Working Notes Proceedings of the MediaEval 2014 Workshop*, October 2014.

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

[11] Youssouf Chherawala, Partha Pratim Roy, and M Chenet, "Context-dependent blstm models. application to offline handwriting recognition," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2565–2569.

[12] Alex Graves et al., *Supervised sequence labelling with recurrent neural networks*, vol. 385, Springer, 2012.

[13] Alan Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013.

[14] Lifa Sun, Shiyin Kang, Kun Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.

[15] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins, "Learning to forget: Continual prediction with lstm," in *Neural computation*. 2000, vol. 12, pp. 2451–2471, MIT Press.

[16] Mike Schuster and Kuldip K Paliwal, "Bidirectional recurrent neural networks," in *Signal Processing, IEEE Transactions on*. 1997, vol. 45, pp. 2673–2681, IEEE.

[17] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew, "Extreme learning machine: theory and applications," in *Neurocomputing*. 2006, vol. 70, pp. 489–501, Elsevier.

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

[19] Liu Yang, Liu Yan, and Gu Zhonglei, "Affective feature extraction for music emotion prediction," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.

[20] Chmulik Michal, Guoth Igor, Malik Miroslav, and Jarina Roman, "Uniza system for the "emotion in music" task at mediaeval 2015," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.

[21] Pellegrini Thomas and Barriere Valentin, "Time-continuous estimation of emotion in music with recurrent neural networks," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.

[22] Coutinho1 Eduardo, Trigeorgis George, and Zafeiriou1 Stefanos, "Automatically estimating emotion in music with deep long-short term memory recurrent neural networks," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.