MediaEval 2015: Recurrent Neural Network Approach to Emotion in Music Tack

Yu-Hao Chin and Jia-Ching Wang
Department of Computer Science and Information Engineering
National Central University, Taiwan, R.O.C.
kio19330@gmail.com, jiacwang@gmail.com

ABSTRACT

This paper describes our work for the “Emotion in Music” task of MediaEval 2015. The goal of the task is predicting affective content of a song. The affective content is presented in terms of valence and arousal criterions, which are shown in a time-continuous fashion. We adopt deep recurrent neural network (DRNN) to predict the valence and arousal for each moment of a song, and Limited-Memory-Broyden–Fletcher–Goldfarb–Shanno algorithm (LBFGS) is used to update the weights when doing back-propagation. DRNN considers the target of the previous time segments when predicting the target of the current time segment. Such time-considering manners of predictions are believed to achieve better performance in comparison of common machine learning models. We finally use the baseline feature set, adopted by the champion of last year, after comparing it with our feature set. A 10-fold cross validation evaluation is used to do the inner-experiments. The system achieves *r* values of -0.5904 for valence and 0.4195 for arousal. The *Root-Mean-Squared Error* (*RMSE*) for valence and arousal are 0.4054 and 0.3804, respectively. For the evaluation dataset, the system achieves *r* values of -0.0103+-0.3420 for valence and 0.3417+-0.2501 for arousal. The Root-Mean-Squared Error for valence and arousal are 0.3359+-0.1614 and 0.2555+-0.1255, respectively.

1. INTRODUCTION

The “Emotion in Music” task asks participants to construct a system that can automatically predict valence and arousals values for each 500ms segment of a song. The development set of the whole database consists of 431 clips, and each clip has a length of 30 seconds. The annotators are asked to slide a pointer on the monitor when annotating the valence and arousal values for the clips. The valence and arousal annotations are provided in a time-continuous manner. Please refer to [1] for more details. The time-series annotations are related to each other and we thus use a time-considering machine learning model (Deep Recurrent Neural Network, DRNN) to do the work. The rest of paper is organized as follows. Section II introduces a music information retrieval feature set. Section III introduces recurrent neural network and the Limited-Memory-Broyden–Fletcher–Goldfarb–Shanno algorithm. Section IV shows the performance of our system and makes a discussion about the experimental results. Section V provides a conclusion of our work.

2. FEATURE EXTRACTION

This section describes the feature set used in our work. Specifically, this feature set is finally dropped in our submission since the baseline feature set obtains a better performance. To illustrate our experiments in Section IV clearly, we still introduce this feature set in this paper.

We extract 10 kinds of features that are often utilized in the music emotion related research. A Matlab toolbox- MIR toolbox [2] is used to extract features from each music clip. The extracted features are beatspectrum, event density, zero-crossing rate, MFCC, roll-off, brightness, roughness, chromagram, pitch, root mean square (RMS) energy, and low energy. These features can be classified into five categories according to their properties, i.e. rhythm, timbre, tonality, pitch, and dynamics. Table I lists the class of each feature.

Table 1: Extracted features and the corresponding classes.

Feature class	Feature name
Dynamics	RMS energy, low energy
Rhythm	beatspectrum, event density
Timbre	zero-crossing rate, Roll-off, brightness, MFCC
Pitch	pitch
Tonality	chromagram

3. APPROACH

We use deep recurrent neural network to regress the valence and arousal values for a song. Different from neural network, deep recurrent neural network has at least one cyclic path of connections [3]. We set one layer to the recurrent layer, and the recurrent layer considers its nodes of the previous one time step when computing the current value of the nodes. A such model is called *L* intermediate layer deep neural network in [4].

The weights of recurrent neural network can be updated using various methods, such as back propagation through time, real-time recurrent learning, and Kalman-filtering-based weight estimation. This paper adopts back propagation through time to update the weights. Specifically, the step size of the update is estimated by a Limited-Memory-Broyden–Fletcher–Goldfarb–Shanno algorithm, which can compute the step size systemically rather than determine the step size by the multiplication of a constant rate of learning and delta values.

We adopt a multi-task architecture to predict the valence and arousal jointly. This architecture has been proved effective in various machine learning works. On the other hand, to involve the contextual information among the segments of the song, we concatenate the features of several segments together to be an input vector of the model. The size of the concatenation

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany.

is not analyzed in this paper. We just empirically set the size to three.

4. RESULTS AND DISCUSSION

This section consists of three subsections, i.e., experimental setup, experimental results, and discussion.

4.1 Experimental Setup

We adopt two feature sets, i.e., the MIR feature set mentioned in Section 2 and the baseline feature set provided by the organizers. The features are normalized by z-scores (i.e. subtracted by mean statistic and divided by the standard deviation). We train a recurrent neural network model to predict the valence and arousal values, which is implemented using a Matlab tool provided by [4]. The number of hidden layers is set to three, and only the second layer is a recurrent layer. The number of hidden nodes in each layer is set to 500. A linear function is applied to each output node, and a sigmoid function is adopted to be the activation function of each hidden node. The initialization of weights is implemented using a Xavier's weight initialization trick [5]. We train the model in a batch manner. The batch size is set to 388. The rate of learning of the back propagation is set to 2. The training process of the model is stopped after the number of iterations achieves 100. In order to avoid the over-fitting problem, we add a noise to each target when training the model. Specifically, we do not pre-train the model. The experiment of the development set is done using a 10-fold cross validation. The performances of the methods are evaluated in terms of R-Squared for valence, R-Squared for arousal, Root-Mean-Squared Error (RMS) for valence, and Root-Mean-Squared Error for arousal.

4.2 Experimental Results

Table 2 shows the performances, which are obtained using the development set, of two approaches: Approach 1) The MIR feature set is extracted from the clips. A RNN model is adopted to predict the VA values; Approach 2) The baseline feature set, provided by the MediaEval 2015 official, is extracted from the clips. A RNN model is adopted to predict the VA values as well.

Table 2: Performances of the two approaches.

Method	Valence		Arousal	
	r	RMSE	r	RMSE
Approach 1	-0.5810	0.4179	0.4079	0.3869
Approach 2	-0.5904	0.4054	0.4195	0.3804

Table 3: Performance of approach 2 for evaluation dataset.

Method	Valence		Arousal	
	r	RMSE	r	RMSE
Approach 2	-0.0103+- 0.3420	0.3359+- 0.1614	0.3417+- 0.2501	0.2555+- 0.1255

Since the MIR feature set performs worse than the baseline feature set, we only submitted Approach 2. Table 3 shows the official results of our submission.

4.3 Discussion

Apparently, our system does not obtain satisfied results in the task. Such results may come from several weakness of RNN: 1) The residual cannot be well back propagated to the nodes in the first layer; 2) The computation of the current node cannot consider its previous states by high number of time steps; 3) The parameters (e.g., batch size, number of layers, activation function, normalization method, and rate of learning.) of the model are not well set.

5. CONCLUSION

This paper presents our work of the 2015 MediaEval Emotion in Music task. Our system adopts recurrent neural network to regress the valence and arousal values. The system achieves *r* values of -0.5904 for valence and 0.4195 for arousal. The *Root-Mean-Squared Error* (*RMSE*) for valence and arousal are 0.4054 and 0.3804, respectively. For the evaluation dataset, the system achieves *r* values of -0.0103+-0.3420 for valence and 0.3417+-0.2501 for arousal. The Root-Mean-Squared Error for valence and arousal are 0.3359+-0.1614 and 0.2555+-0.1255, respectively Our systems does not perform well in the task. The unsatisfactory results may be obtained due to the lack of model tuning. A pre-training process should be involved to improve the performance.

6. REFERENCES

[1] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at mediaeval 2015. In *MediaEval 2015 Workshop*, 2015.

[2] O. Lartillot and P. Toiviainen. MIR in Matlab (II): A toolbox for musical feature extraction from audio. In *Proc. Int. Conf. Music Information Retrieval*, 2007, pages. 127–130, [Online] http://users.jyu.fi/lartillo/mirtoolbox/.

[3] H. Jaeger. 2013. *A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach*. International University Bremen.

[4] P. S. Huang, M. Kim, M. H. Johnson, and P. Smaragdis. Singing-voice separation from monaural recordings using deep recurrent neural networks. In *Proc. Int. Conf. Music Information Retrieval*, 2014, pages. 477-482.

[5] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feed forward neural networks. AISTATS 2010.