The Munich LSTM-RNN Approach to the MediaEval 2014 "Emotion in Music" Task

Eduardo Coutinho[1,2,3], Felix Weninger[1], Björn Schuller[3,1,4], Klaus R. Scherer[3,5]

[1]Machine Intelligence & Signal Processing Group, Technische Universität München, Munich, Germany
[2]School of Music, University of Liverpool, Liverpool, United Kingdom
[3]Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland
[4]Department of Computing, Imperial College London, London, United Kingdom
[5]Department of Psychology, Ludwig-Maximilians-Universitẗ München, Munich, Germany
{e.coutinho,weninger,schuller}@tum.de, klaus.scherer@unige.ch

ABSTRACT

In this paper we describe TUM's approach for the MediaEval's "Emotion in Music" task. The goal of this task is to automatically estimate the emotions expressed by music (in terms of Arousal and Valence) in a time-continuous fashion. Our system consists of Long-Short Term Memory Recurrent Neural Networks (LSTM-RNN) for dynamic Arousal and Valence regression. We used two different sets of acoustic and psychoacoustic features that have been previously proven as effective for emotion prediction in music and speech. The best model yielded an average Pearson's correlation coefficient of 0.354 (Arousal) and 0.198 (Valence), and an average Root Mean Squared Error of 0.102 (Arousal) and 0.079 (Valence).

1. INTRODUCTION

The MediaEval 2014 "Emotion in Music" task comprises two subtasks. The first subtask (Subtask 1), pertains to the development of new features to automatically estimate Arousal and Valence for music excerpts. The second subtask (Subtask 2), consists of the estimation of Arousal and Valence scores continuously in time for a second long segments of the same music excerpts. In both tasks, the development set consists of 744 instances and the evaluation set comprises 1000 instances. For more details, please refer to [1]. The TUM-MISP team participated only in Subtask 2.

2. METHODOLOGY

We used two features sets in our experiments. The first feature set (*FS1*) consists of the official set of low-level audio descriptors (LLDs) used in the 2013 INTERSPEECH Computational Paralinguistics Challenge (ComParE; see [6] for full details). It comprises 65 LLDs as well as their first order derivates (130 LLDs, in total). LLDs related to voice were computed using 60 ms long time frames and Gaussian windows ($\sigma = 0.4$). LLDs related to all other features were were calculated using 25 ms long time frames and Hamming window functions. In both cases, overlapping windows were used with a step size of 10 ms (17% and 40% overlaps, respectively). Finally, for the purpose of this work, we also computed the mean and standard deviation functionals of each feature over 1 s time windows with 50% overlap (step size of 0.5 s). This resulted in 260 features extracted at a rate of 2 Hz. All features were extracted using the open-source feature extractor openSMILE ([3]). The second feature set (*FS2*) consists of the same features included in *FS1*, plus four new features - Sensory Dissonance (SDiss), Roughness (R), Tempo (T) and Event Density (ED). These features correspond to two psychoacoustic dimensions consistently associated with the communication of emotion in music and speech ([2]) - Roughness (SDiss and R) and Duration (T and ED) - which are absent from *FS1*. The four features were extracted with the MIR Toolbox [5], using the *mirroughness* (SDiss - with Sethares formula; and R - with Vassilakis algorithm), *mirtempo* (T) and *mireventdensity* (ED) functions.

As regressors, and given the importance of the temporal context in emotional responses to music (e.g., [2]), we considered LSTM-RNN as defined in [4]. LSTM networks make use of special memory blocks, which endow the model with the capacity of accessing a long-range temporal context and predicting the outputs based on such information. An LSTM network is similar to an RNN except that the nonlinear hidden units are replaced by a special kind of memory blocks. Each memory block comprises one or more self-connected memory cells and three multiplicative units – input, output and forget gates – which provide the cells with analogues of write, read and reset operations. The multiplicative gates allow LSTM memory cells to store and access information over long sequences (and corresponding periods of time).

2.1 Models training

We used a multi-task learning framework for the joint learning of Arousal and Valence time-continuous values. A cross-validation procedure was used in the development phase, where we created an extra fold to estimate the performance of our approaches during the development phase. The fold subdivision followed a modulus based scheme (instance ID modulus 11). The instances yielding a remainder of 10 were left out to create a small test set for performance estimation. On the remaining instances, a 10-fold cross-validation was performed. We computed 5 trials of the same model each with randomized initial weights in the range [-0.1,0.1]. Our basic architecture consisted of deep LSTM-RNN with 2 hidden layers. We optimised the number of LSTM blocks in each hidden layer, as well as the learning rate (a momentum

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

of 0.9 was used for all tests), and the standard deviation of the Gaussian noise applied to the input activations (used to alleviate the effects of over-fitting). An early stopping strategy was also used to avoid overfitting the training data – training was stopped after 20 iterations without improvement of the validation set performance (sum of squared errors). The instances in the 10 training sets were presented in random order to the model during training. The input (acoustic features) and output (emotion features) data were standardised to zero mean and unit variance on the correspondent training sets used in each cross-validation fold.

In four of our five runs (see next subsection) we pre-trained the first hidden layer. Our unsupervised pre-training strategy consisted of de-noising LSTM-RNN auto-encoders. We first created a LSTM-RNN with a single hidden layer trained to predict the input features ($y(t) = x(t)$). Both the development and test set instances were used to train the DAE. In order to avoid over-fitting, in each training epoch and timestep t, we added a noise vector n to $x(t)$, sampled from a Gaussian distribution with zero mean and variance n. After determining the auto-encoder weights a second hidden layer was added. In two of the runs, all of the weights were trained using the regression targets and keeping the first layer weights constant. In the other two, the first layer weights were retrained.

2.2 Runs

We submitted five runs for Subtask 2. All runs consisted of LSTM-RNNs using two hidden layers in order to attempt modeling high-level abstractions in the data (Deep Learning). The specifics of each run are as follows: Run 1) The basic architecture was directly trained using the regression targets and *FS1*; Run 2) We pre-trained the first layer, added a second one, and all weights (with the exception of the first layer weights that were kept constant) were trained using the regression targets and *FS1*; Run 3) Same as Run 2, but all weights (including the first layer weights) were trained using the regression targets and *FS1*; Run 4) Same as Run 2, but using *FS2*; Run 5) Same as Run 3, but using *FS2*; The submitted results for each test run consisted of the average outputs of the five best models (across all folds and trials) as estimated using the method described in Section 2.1.

3. RESULTS AND EVALUATION

In Table 1, we report the official challenge metrics (r - Pearson's linear correlation coefficient; and $RMSE$ - Root Mean Squared Error) calculated individually for each music piece and averaged across all pieces (standard deviations also shown) of the test set. In short, we observe that *Run 4* lead to the best results. Individual two-tailed t-tests revealed that: a) $r(Arousal)$ was significantly higher for *Run 4* compared to *Run 1*, *Run 2*, *Run 5* ($p < 0.0001$), and *Run 3* ($p < 0.01$); b) $r(Valence)$ was higher for *Run 4* compared to all other runs, but only significantly higher than *Run 3 (*$p < 0.05$*)*; c) $RMSE(Arousal)$ was significantly lower for *Run 4* compared to all other runs ($p < 0.0001$); d) $RMSE(Valence)$ was significantly lower for *Run 4* compared to all other runs ($p < 0.0001$) except *Run 5*.

Run 4 consisted of a LSTM-RNN with two layers, including a pre-trained first layer (with weights kept constant while training using the regression targets) and *FS2* as input. The optimised architecture consisted of 200 and 5 LSTM blocks (first and second layers, respectively), trained with a learning rate of 10^{-6} and Gaussian noise with a variance of 0.5 applied to the inputs during development (no noise added when processing the test set).

Table 1: Official results of the MISP-TUM team for the five runs submitted.

		Arousal	Valence
r	Run 1	0.247±0.456	0.170±0.458
	Run 2	0.246±0.458	0.181±0.503
	Run 3	0.291±0.479	0.152±0.503
	Run 4	**0.354±0.455**	**0.198±0.492**
	Run 5	0.232±0.434	0.172±0.450
$RMSE$	Run 1	0.134±0.062	0.096±0.056
	Run 2	0.121±0.058	0.090±0.055
	Run 3	0.120±0.059	0.090±0.056
	Run 4	**0.102±0.052**	**0.079±0.048**
	Run 5	0.112±0.055	0.082±0.050

4. CONCLUSIONS

The LSTM-RNN approaches to the 2014 MediaEval "Emotion in Music" task all delivered consistent improvements over the baselines. The results reveal the importance of fine-tuning the feature set and the deep learning strategy, which could be attributed to the relatively small training set.

5. ACKNOWLEDGMENTS

This work was partially supported by the ERC in the European Community's 7th Framework Program under grant agreements No. 338164 (Starting Grant iHEARu to Björn Schuller) and 230331 (Advanced Grant PROPEREMO to Klaus Scherer).

6. REFERENCES

[1] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at mediaeval 2014. In *MediaEval 2014 Workshop*, Barcelona, Spain, October 16-17 2014.

[2] E. Coutinho and N. Dibben. Psychoacoustic cues to emotion in speech prosody and music. *Cognition & emotion*, 27(4):658–684, 2013.

[3] F. Eyben, F. Weninger, F. Groß, and B. Schuller. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proceedings of the 21st ACM International Conference on Multimedia, MM 2013*, pages 835–838, Barcelona, Spain, October 2013.

[4] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[5] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[6] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer. On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common. *Frontiers in Psychology*, 4(Article ID 292):1–12, May 2013.