THU-HCSI at MediaEval 2016: Emotional Impact of Movies Task

Ye Ma, Zipeng Ye, Mingxing Xu
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
{y-ma13, yezp13}@mails.tsinghua.edu.cn, xumx@tsinghua.edu.cn

ABSTRACT

In this paper we describe our team's approach to MediaEval 2016 Challenge "Emotional Impact of Movies". Except for the baseline features, we extract audio features and image features from video clips. We deploy Convolutional Neural Network (CNN) to extract image features and use OpenSMILE toolbox to extract audio ones. We also study multi-scale approach at different levels aiming at the continuous prediction task, using Long-short Term Memory (LSTM) and Bi-directional Long-short Term Memory (BLSTM) models. Fusion methods are also considered and discussed in this paper. The evaluation results show our approaches' effectiveness.

1. INTRODUCTION

The MediaEval 2016 Challenge "Emotional Impact of Movies" consists of two subtasks: Global emotion prediction of a short video clip (around 10 seconds) and continuous emotion prediction of a complete movie. LIRIS-ACCEDE [2, 1] dataset is used in the challenge. A brief introduction to the dataset for training and testing as well as the details of these two subtasks has been given in [3]. In this paper, we mainly discuss the approach employed by our system.

2. APPROACH

2.1 Subtask 1: global emotion prediction

2.1.1 Feature Extraction

Except for the baseline features provided by the organizers, there are two types of features used in our experiments, which are audio features and image features. Audio features only utilize the audio wave files extracted from video files, and image features only utilize the static frames extracted from videos.

As to the audio features, we use the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which consists of 88 features and has been used in many emotion recognition tasks for their potential and theoretical significance [4]. In our experiments, we extract these features from each video clip with the OpenSMILE toolkit [5].

The image features were extracted by using a fine-tuned Convolutional Neural Network (CNN). We adopt a 19-layer CNN model on the ILSVRC-2014 dataset pretrained by VGG team [10] and replaced the last softmax layer (which was used for classification) with a fully-connected layer and an Euclidean loss layer (which was used for regression). The input images are static frames extracted from video clips at the rate of 2 Hz and the output labels are valence and arousal annotations. We trained the valence and arousal CNNs with Caffe [8] separately and used the first fully-connected layer of CNN models as the output features which were reduced by using Principal Component Analysis (PCA) in our experiments.

2.1.2 Prediction Models

Support Vector Regression (SVR) models with RBF kernel were trained for valence and arousal separately. We've tried both early fusion and late fusion for audio- and visual-features, which will be elaborated in Section 3.

2.2 Subtask 2: continuous emotion prediction

2.2.1 Feature Extraction

For audio features, we used a set of features provided by INTERSPEECH 2013 Computational Paralinguistics Challenge [11] which consists of 130 dimensions. For image features, the same CNN feature as Subtask 1 is chosen and reduced by PCA to 256 dimensions.

2.2.2 Prediction Model

We applied Long-Short Term Memory (LSTM) [7] to model the context information in movies. Since the emotion evoked by a video clip is not only associated with the previous content but also the future one, Bidirectional Long Short-Term Memory (BLSTM) [6] is considered as a better choice because of its ability to use both previous and future information.

In our experiments, two types of models with three layers were used. Type 1 has three LSTM layers and type 2 is the same except the middle layer is BLSTM. The dimensions of the two hidden layers are as listed in Table 1.

2.2.3 Multi-scale Fusion and Post-processing

Similar to [9], total five models of different scales were trained with different sequence lengths, i.e., 8,16,32,64 and 128, respectively. For each scale, we selected one appropriate model from 3 trails.

We divided the whole dataset into three parts: 70% for training, 20% for validation and 10% as the test set for fusion and post-processing.

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

Table 1: Dimensions of the 2 hidden layers

Feature	Model	Layer1	Layer2
Audio	Type 1	128	64
Audio	Type 2	128	32
Audio + Image	Type 1	256	128
Audio + Image	Type 2	256	64

Table 2: Global result of test set

Runs	Valence		Arousal	
	MSE	r	MSE	r
Run 1	0.2188	0.2680	**1.4674**	0.2725
Run 2	0.2170	0.2740	1.5910	**0.3444**
Run 3	**0.2140**	**0.2955**	1.5312	0.2667

Finally, we applied a post-processing with a sliding triangular filter to smooth the final results. In our experiments, the filter window size is 9.

3. EXPERIMENTS AND RESULTS

In this section, we will describe our methods and experiments in more detail and show the results.

3.1 Subtask 1: global emotion prediction

We've submitted three runs for global prediction task in total, listed below:

Run 1: (Baseline + eGeMAPS + CNN) features + SVR + early fusion

Run 2: (Baseline + eGeMAPS) features + SVR + early fusion

Run 3: (Baseline + eGeMAPS + CNN) features + SVR + late fusion

In detail, CNN features in Run 1 and Run 3 are compressed using PCA algorithm, which is 512 dimensions for arousal and 128 dimensions for valence. These dimensions are decided upon the results of 5-fold cross-validation on training set. Besides, the weight of late fusion in Run 3 is also determined on validation.

From Table 2 we can see that, the best run of valence is Run 3 while the best of arousal is Run 1, which are late fusion and early fusion respectively. Notice that runs using CNN features performs better on arousal than those who don't, indicating that image features may contain more information about emotion's polarity than audio ones. Besides, it is worth mentioning that the arousal's Pearson r of Run 2 is the highest among all runs, implying that higher relevance may lead to higher MSE loss to some content.

3.2 Subtask 2: continuous emotion prediction

In order to select the best model for each scale and fusion, we designed a series of experiments. We have five different scales, two feature sets and two types of models. For each possible combination, we trained 3 trials with randomized initial weights. Therefore, there are total 60 ($5 \times 2 \times 2$) experiments.

Run 1: Only audio features were used. The sequence length of LSTM was 16 for valence while 64 for arousal.

Table 3: Continuous result of test set

Runs	Valence		Arousal	
	MSE	r	MSE	r
Run 1	0.1086	**0.017**	0.1601	**0.054**
Run 2	0.1276	-0.023	**0.1244**	-0.023
Run 3	**0.1016**	-0.002	0.1354	0.030
Run 4	0.1029	-0.003	0.1294	0.026
Run 5	0.1018	0.000	0.1376	0.052

The model was type 2 for valence and type 1 for arousal.

Run 2: Audio and video feature vectors were concatenate as a multi-modality feature vector. The scale was 16 for valence and 128 for arousal. The model was type 2 for valence and type 1 for arousal.

Run 3: We used the same features as Run 1. Multi-scale models were trained and fused by using simple average to generate the final results.

Run 4: We used the same features as Run 2. Multi-scale models were trained and fused by using simple average to generate the final results.

Run 5: Same as Run 3 except the fusion in which we weighted those models' results with different weights, i.e. 0.4, 0.3, 0.2 and 0.1 from the low loss to high loss, respectively.

4. CONCLUSION

In this paper, we illustrate our approach to the MediaEval 2016 Challenge "Emotional Impact of Movies" task. As to global emotion prediction subtask, combining the features learnt from video by using CNN enhances the regression performance of arousal with early fusion as well as the performance of valence with late fusion.

As to continuous prediction, the best result obtained in this paper is Run 3 for valence and Run 2 for arousal. Fusion by multi-scale has a good performance for valence. For arousal, the Run 3 is better than Run 1 but Run 4 is worse than Run 2, so fusion could not always make it better.

5. ACKNOWLEDGMENTS

This work was partially supported by the 863 Program of China (2015AA016305), the National Natural Science Foundation of China (61171116, 61433018) and the Major Project of the National Social Science Foundation of China (13&ZD189).

6. REFERENCES

[1] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *A ective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 77–83. IEEE, 2015.

[2] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on A ective Computing*, 6(1):43–55, 2015.

[3] E. Dellandréa, L. Chen, Y. Baveye, M. Sjöberg, and C. Chamaret. The mediaeval 2016 emotional impact of movies task. In *Proceedings of MediaEval 2016 Workshop*, Hilversum, Netherlands, 2016.

[4] F. Eyben, K. Scherer, K. Truong, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, and P. Laukka. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on A ective Computing*, 12(2):190–202, 2016.

[5] 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*, pages 835–838. ACM, 2013.

[6] A. Graves and J. r. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602610, 2005.

[7] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[9] X. Li, J. Tian, M. Xu, Y. Ning, and L. Cai. Dblstm-based multi-scale fusion for dynamic emotion prediction in music. pages 1–6, 2016.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] 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(2):292, 2013.