THUHCSI in MediaEval 2017 Emotional Impact of Movies Task

Zitong Jin, Yuqi Yao, Ye Ma, 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
{jzt15,yaoyq15,ma-y17}@mails.tsinghua.edu.cn,xumx@tsinghua.edu.cn

ABSTRACT

In this paper we describe our team's approach to MediaEval 2017 Challenge *Emotional Impact of Movies*. Except for the baseline features, we use OpenSMILE toolbox to extract audio features eGeMAPS from video clips. We also aim at the continuous flow of emotion, where using time-sequential models such as LSTM will be useful and effective. Fusion methods are also considered and discussed in this paper. The evaluation results of our experiments show that our features and models are competitive in both valence / arousal and fear prediction, indicating our approaches' effectiveness.

1 INTRODUCTION

The MediaEval 2017 Challenge *Emotional Impact of Movies* consists of two subtasks. Subtask 1 aims at Valence/Arousal prediction while subtask 2 aims at Fear prediction. Long movies are considered for both cases and prediction needs to be given every 5 seconds for the consecutive ten seconds' segment. LIRIS-ACCEDE [1, 2] dataset is used for training and testing, including both discrete and continuous sections of data. For more details, please refer to [5].

Video affective analysis and prediction is an important and challenging issue, which has drawn the attention of many researchers recently. The *Emotional Impact of Movies* task has been held for three years, so there are many participants who took part in the challenge in 2015 and 2016 [4, 9].

2 APPROACH

In this section, we will describe the main approaches for the subtasks, including feature extraction, pre-processing, prediction models, fusion and post-processing methods.

2.1 Subtask 1: valence / arousal prediction

Feature extraction. Except for the baseline features provided by the organizers, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [6] is extracted from audio channel, which contains 88 features and has been proved effective in the same task of last year [8]. In our experiments, we extract them with the OpenSMILE toolkit [7] from 5-second-long segments which are cut from original videos in advance.

As for the visual features, the general purpose visual features provided by the organizers (except CNN features) are merged into one large feature. This is mainly on account of the fact that these features are short and complementary, and that combining them can greatly reduce the training workload to try every one of them.

MediaEval'17, 13-15 September 2017, Dublin, Ireland

All input features are scaled into vectors of zero mean and unit variance for normalization.

Prediction models. Two aspects of models are adopted in our experiments, which are traditional machine learning models and time-sequential models. Specifically, the traditional models consist of Support Vector Regression (SVR) and AdaBoost while the time-sequential ones are Long-Short Term Memory (LSTM) models. The LSTM models may capture the emotional flow of video and enhance the performance. We take the problem as a Sequence-to-One regression problem and the input features of LSTM models are segmented in a 10-second-long sliding window of 5 seconds overlapping.

All models are trained separately for valence and arousal.

Fusion methods. To combine features of different modalities, except for the early fusion method which simply concatenates different features, late fusion method is also considered. As for the traditional prediction models, average fusion is used to avoid overfitting. As for the LSTM models, the hidden vectors of several LSTM models taking different inputs are fused using an one-layer fully-connected network to obtain final prediction, which is trained with LSTM models simultaneously.

After fusion, to reduce the fluctuation of output and smooth out the random noise, a 25-frame-long triangle filter is applied to each video.

2.2 Subtask 2: Fear prediction

Feature extraction. We use the same feature sets as Subtask 1. However, the main problem and the biggest challenge in Subtask 2 is that the samples are so unbalanced that simply predicting "zero" obtains the accuracy score of 84.34% in the test set (see Run 4). Therefore, to solve the unbalanced problem, SMOTE (Synthetic Minority Over-sampling TEchnique [3]) method is adopted after feature extraction to re-sample. The main idea of SMOTE algorithm is to generate new samples for minorities using interpolation, which will make it more balanced.

Prediction models. Random Forest model is adopted in fear prediction, which may behave better than Support Vector Machine (SVM) in unbalanced problem. We first use Random Forest model to obtain the probability of predicting fear ("one") for each video clip. Then we set up the decision threshold p, and predict fear when the probability is larger than p. The value of p are adjusted according to the validation set's results. Due to the time constaints, we didn't try the LSTM model for Subtask 2.

Fusion methods. Similar to Subtask 1, both early and late fusion are used. In late fusion, the probability of different models are averaged to get one probability.

Table 1: Results of Subtask 1 on test set

Runs	Valence		Arousal	
	MSE	*r*	MSE	*r*
Run 1	0.2230	-0.0985	0.1577	0.2261
Run 2	**0.1670**	-0.0990	0.1269	-0.0122
Run 3	0.1833	**0.3707**	**0.1166**	**0.3213**
Run 4	0.2074	-0.0111	0.1318	0.2708
Run 5	0.2046	0.0122	0.1300	0.2750

3 EXPERIMENTS AND RESULTS

In this section, we will describe our specific runs in more detail and show the results. Note that all the hyper-parameters are selected due to the results of validation set, and the ratio of training data and validation data is 4:1.

3.1 Subtask 1: valence / arousal prediction

We've submitted 5 runs for valence / arousal prediction, where the first two use LSTM and the other ones use SVR and AdaBoost, all listed below:

Run 1: For valence, 2-layer LSTM model of hidden size 500 taking eGeMAPS as input; For arousal, 3-layer LSTM model of hidden size 500 taking VGG as input.

Run 2: For valence, late fusion of three 2-layer LSTM models of hidden size 1000 taking eGeMAPS, VGG and other visual features as input respectively; For arousal, the input features are Emobase, eGeMAPS and CEDD respectively.

Run 3: For both valence and arousal, SVR model taking VGG as input.

Run 4: For valence, AdaBoost model taking eGeMAPS as input; For arousal, AdaBoost model taking other visual features as input.

Run 5: For both valence and arousal, late fusion of Run 3 and Run 4.

In detail, the "other visual features" in Run 2 and 4 means the concatenation of all the visual features except the CNN feature. CEDD means Color and Edge Directivity Descriptor, which is one of the visual feature provided. VGG means CNN features extracted using VGG16 fc6 layer.

From Table 1 we can see that, the best run of valence MSE is Run 2, using late fusion of LSTM models. Run 3 achieves the best results on other metrics, using SVR model and VGG feature. Notice that Run 2, the LSTM late fusion method, is better at MSE than Run 1, the single LSTM model, which means late fusion of three models utilizes different information in different features and enhances the performance to some extent. However, LSTM models perform worse in Pearson's r, compared to traditional machine learning models. This could be because LSTM models tend to predict similar values of all time, and thus obtain lower MSE and lower Pearson's r.

Taken together, Run 3 using SVR and VGG achieves best results, which means CNN features may contain useful information for emotion analysis, and traditional model could behave well when trained properly.

Table 2: Results of Subtask 2 on test set

Runs	Accuracy	Precision	Recall	f1
Run 1	0.7352	0.0206	0.0530	0.0239
Run 2	0.8153	0.2318	**0.2781**	**0.2352**
Run 3	0.8461	0.2035	0.0208	0.0371
Run 4	0.8434	0.0000	0.0000	0.0000
Run 5	**0.8469**	**0.2383**	0.2186	0.2165

3.2 Subtask 2: fear prediction

We've submitted 5 runs for fear prediction, all using Random Forest model, listed below:

Run 1: Random Forest + other visual features.

Run 2: Random Forest + VGG.

Run 3: Random Forest + all visual features.

Run 4: All predicting "zero" (just for test)

Run 5: Late fusion of Run 1 and Run 2.

From Table 2 we can see that, Run 2 using VGG features achieve best results on recall and f1, while Run 5 using late fusion achieve best results on accuracy and precision. As mentioned before, the problem of subtask 2 is very unbalanced, and the fear samples are much fewer. Therefore, there is no surprise that accuracy and precision are one pair while recall and f1 are the other pair. Predicting more "zeros" will lead to higher accuracy while lower recall, and vice versa.

When considering f1 score, which is the harmonic mean of both precision and recall, Run 2 using VGG feature performs best, which confronts with the result of subtask 1 that CNN features contain useful information for emotion analysis.

4 CONCLUSION AND DISCUSSION

In this paper, we illustrate our approach to the MediaEval 2017 Challenge "Emotional Impact of Movies" task. In valence / arousal prediction subtask, both LSTM and SVR models are trained and compared. In fear prediction subtask, Random Forest model using different features are compared. Besides, early fusion and late fusion are adopted in experiments, which shows promising results in some aspects.

However, some problems have not been solved yet. For instance, some of the LSTM models tend to predict similar values of all time, leading to a very low Pearson's r, which may be caused by inappropriate experiment configuration. Unbalanced problem in subtask 2 still exists even using SMOTE algorithm, which means changing models or features could make no big difference, and all predicting "zero" can still obtain a very high accuracy. These problems remain to be solved in the future.

ACKNOWLEDGMENTS

This work was partially 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] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, and Liming Chen. 2015. Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 77–83.

[2] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 43–55.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[4] Emmanuel Dellandréa, Liming Chen, Yoann Baveye, Mats Sjöberg, and Christel Chamaret. 2016. The MediaEval 2016 Emotional Impact of Movies Task. In *Proceedings of MediaEval 2016 Workshop*. Hilversum, Netherlands.

[5] Emmanuel Dellandréa, Martijn Huigsloot, Liming Chen, Yoann Baveye, and Mats Sjöberg. 2017. The MediaEval 2017 Emotional Impact of Movies Task. In *Proceedings of MediaEval 2017 Workshop*. Dublin, Ireland.

[6] Florian Eyben, Klaus Scherer, Khiet Truong, Bjorn Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Devillers, Julien Epps, and Petri Laukka. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 12, 2 (2016), 190–202.

[7] 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.

[8] Ye Ma, Zipeng Ye, and Mingxing Xu. 2016. THU-HCSI at MediaEval 2016: Emotional Impact of Movies Task. In *Proceedings of MediaEval 2016 Workshop*. Hilversum, Netherlands.

[9] Mats Sjöberg, Yoann Baveye, Hanli Wang, Vu Lam Quang, Bogdan Ionescu, Emmanuel Dellandréa, Markus Schedl, Claire-Hélène Demarty, and Liming Chen. 2015. The MediaEval 2015 Affective Impact of Movies Task.. In *Proceedings of MediaEval 2015 Workshop*. Wurzen, Germany.