THUHCSI in MediaEval 2018 Emotional Impact of Movies Task

Ye Ma[1], Xihao Liang[1], Mingxing Xu[1]
[1]Department of Computer Science & Technology, Tsinghua University, Beijing, China
ma-y17@mails.tsinghua.edu.cn,liangxh16@mails.tsinghua.edu.cn,xumx@tsinghua.edu.cn

ABSTRACT

In this paper we describe our team's approach to the MediaEval 2018 Challenge *Emotional Impact of Movies*. We extract several sets of audio and visual features, and then apply the time-sequential models such as LSTM and BLSTM to model the continuous flow of emotion in movies. Different fusion methods are also considered and discussed. The results show that our methods achieve promising performance, indicating the effectiveness of the features and the models we choose.

1 INTRODUCTION

The Challenge *Emotional Impact of Movies* of MediaEval has been held since 2015[1, 2, 9]. This challenge mainly focuses on the emotion aroused from the movies and how to predict it. This year's task consists of two subtasks. Subtask 1 aims at Valence / Arousal prediction and Subtask 2 aims at Fear prediction. Details of both subtasks could be found in [3].

2 APPROACH

In this section, we describe in detail our team's main approach, including feature extraction, prediction models, fusion methods, pre-processing and post-processing.

2.1 Feature extraction

Audio features. Previous results[6, 8] have showed the great potential of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [4]. This feature set contains 23 low level descriptors (llds), which is proved effective in acoustic tasks such as speech emotion recognition. In our experiments, we extract the low level descriptors of eGeMAPS using the OpenSMILE toolbox [5]. Then we compute the mean and standard deviation in a centered 5-second-long sliding window of all 23 features to obtain the feature of 46 dimension for each second of the movie clip.

Besides, baseline features provided by the organizer are also considered, which is the Emobase 2010 feature set (1582 dimensions).

Visual features. Baseline features consist of multiple general-purpose visual features. Following last year's experiments, we concatenate all the visual features to one big feature except the CNN feature, which is of 1271 dimensions. The CNN feature is treated separately from other features because it is much larger (4096 dimensions) and has the different source from others.

In order to utilize more visual information, we try using SentiBank for feature extraction. We apply the MVSO detectors[7] on image frames extracted every one second from the movies to obtain the final layer of Inception net, which can be referred as the composition ratio of different concepts (4342 dimensions).

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

2.2 Prediction models

Last year's results [6] showed that the Support Vector Machines (SVM) are better than Long Short-Term Memory models (LSTM). However, as the size of the training dataset is larger than that from last year and time sequential models should perform better on bigger dataset, this year we adopt LSTM as the prediction model to predict the emotional flow. In detail, we take the problem as a Sequence-to-Sequence problem and the time length of input sequences is determined by the validation set.

This year, we also use the Bidirectional LSTM, which is mainly for these two reasons: First, the ground truth of emotion is labelled while the annotators are watching the movies, so the latency and mismatch of ground truth and movie content must be considered. Second, the emotional flow in movies is changing smoothly, where the Bidirectional LSTM could be less affected by the fluctuation of input features.

Besides, another difference from last year is that we train models for valence and arousal together. Considering that both valence and arousal share similar emotion concept, it is reasonable to use the same underlying structure. Therefore, every regression model is trained to predict a two dimensional vector which represents both valence and arousal.

As for the Subtask 2, the experiments are done in two steps for simplicity: First, we train a classification model to predict the label for every second. Second, we identify a segment as "Fear" according the labels of every seconds within it. Specifically, we filter out the seconds whose probability of evoking fear is lower than the threshold we set and only keep the sequences whose length is longer than certain threshold, which could remove noise from the sequence.

2.3 Fusion methods

In our experiments, we apply multiple fusion methods, which are shown as follows.

Early fusion: We concatenate features from different modalities and different sources to one bigger vector. This method is simple and straightforward while sometimes very effective.

Late fusion: We trained several LSTM models simultaneously. The output of the last layer of these LSTM models are merged together and used as the input of the next fully-connected layer.

Average fusion: To avoid over-fitting and reduce noise, we compute the average of several models' prediction.

In addition, we apply a triangle filter of 25 seconds to reduce the noise of the outputs.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

Table 1: Results of Subtask 1 on test set

Runs	Valence		Arousal	
	MSE	*r*	MSE	*r*
Run 1	0.1021	0.1714	0.1414	**0.0870**
Run 2	0.1036	0.1820	0.1399	-0.0181
Run 3	**0.0924**	**0.3048**	0.1399	0.0761
Run 4	0.0980	0.2422	**0.1396**	0.0612
Run 5	0.0944	0.2511	0.1460	-0.0667

3 EXPERIMENTS AND RESULTS

In this section, we will elaborate our specific experiment settings and show the results. Note that all hyper-parameters below such as sequence length, hidden size, number of layers are all determined by the validation set. The ratio of training and validation data is 4:1.

3.1 Subtask 1

In our experiments on the validation set, it shows that BLSTM models perform better than LSTM models, which verifies our assumption. And we also find that BLSTM performs best when the sequence length is 100. As for the features, we have tested multiple early fusion combinations and early fusion of Emobase, visual features (except CNN) and eGeMAPS performs the best. Thus, we have submitted 5 runs for subtask 1 all using BLSTM models whose sequence length is 100, and the input features of them are all the same. The first three runs only differ in the number of BLSTM layers, which is 4, 2 and 3 respectively. Run 4 is the average fusion of the first three runs. Run 5 is the late fusion of two BLSTM models, of which the inputs are Emobase and visual features (except CNN) respectively. All runs are trained using a dropout probability of 0.5 to avoid over-fitting.

From Table 1 we can see that the best run of valence is Run 3, which is a 2-layer BLSTM model using Emobase, visual features (except CNN) and eGeMAPS as inputs. As for arousal, Run 4 achieves best performance in MSE, which indicates average fusion sometimes enhances the performance to some extent. The result of valence prediction is remarkably better than that of arousal prediction. This is probably because arousal is harder to predict than valence.

3.2 Subtask 2

As for subtask 2, we try to use the method discussed in Section 2.2. However, it performs much worse than expected. Due to the problem of imbalanced dataset, the prediction probability of fear is very low and only a few segments of consecutive seconds are predicted as "fear". Some movies in development set even have no "fear" segments. It shows that LSTM models may not be proper for imbalanced problem. We've also tried to use techniques for imbalanced problem, such as down-sampling movies and adding more weight for positive samples. Nevertheless, these methods hardly work. Owing to time constraints, we didn't submit runs for this subtask finally, and we will continue researching in future work.

4 DISCUSSION AND OUTLOOK

In summary, this year we've further studied the *Emotional Impact of Movies* task and discovered some useful insights. Firstly, temporal models such as LSTM and BLSTM can capture more information in time sequential problems, when given enough training data. And BLSTM models could be less affected by the latency and mismatch between annotations and movies, which perform better than single directional LSTM. As for fusion methods, early fusion and average fusion are both simple and intuitive, but they usually have a good performance.

Still, some problems remain to be solved. SentiBank features are not so useful as expected in this task. More and more CNN related features should be extracted and tested. Arousal is much harder to predict than valence in our experiments, which needs further investigation. For subtask 2, the problem of imbalanced dataset still remains unsolved this year, even though the evaluation metric has been changed to intersection over union. In addition, some novel techniques from other domains such as object segmentation and voice activity detection could be applied to this subtask to handle this new metric. Moreover, adding more fear related movies to dataset could be another effective approach to alleviate the imbalanced problem.

In conclusion, this paper illustrates our approach to the MediaEval 2018 Challenge *Emotional Impact of Movies* task. We've trained BLSTM models using multi-modality features and several fusion methods, which achieves promising performance in valence and arousal prediction task. Fear prediction task is not fully solved and remains to be further investigated.

ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (61433018, 61171116) and the National High Technology Research and Development Program of China (863 program) (2015AA016305).

REFERENCES

[1] 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.

[2] 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.

[3] Emmanuel Dellandréa, Martijn Huigsloot, Liming Chen, Yoann Baveye, Zhongzhe Xiao, and Mats Sjöberg. 2018. The MediaEval 2018 Emotional Impact of Movies Task. In *Proceedings of MediaEval 2018 Workshop*. Sophia Antipolis, France.

[4] 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.

[5] 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.

[6] Zitong Jin, Yuqi Yao, Ye Ma, and Mingxing Xu. 2017. THUHCSI in MediaEval 2017 Emotional Impact of Movies Task. In *Proceedings of MediaEval 2017 Workshop*. Dublin, Ireland.

[7] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. 2015. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 159–168.

[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.