Frame-based Evaluation with Deep Features to Predict Emotional Impact of Movies

Khanh-An C.Quan[1], Vinh-Tiep Nguyen[1], Minh-Triet Tran[2]
[1]University of Information Technology, Vietnam National University-Ho Chi Minh city
[2]University of Science, Vietnam National University-Ho Chi Minh city
15520006@gm.uit.edu.vn,tiepnv@uit.edu.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

In this paper, we describe our approach for the Emotional Impact of Movies Task at the MediaEval 2018 Challenge. Specifically, we employ features extracted from ResNet-50 from image frames. Then, a fully connected neural network is used for learning the prediction models. Later, we applied the Window Sliding Technique for post-processing the results. The experimental results show the effectiveness of our approach.

1 INTRODUCTION

Analysing the emotional impact of a video clip to viewers can be utilized to enhance or control psychological effects of media to people [2, 7], to boost user engagement to media content [6], or to generate personalized media content [5].

The MediaEval 2018 Emotional Impact of Movies Task consists of two subtasks. The first subtask is to predict the score of induced valence and induced arousal every second along movies. The other is fear prediction, but we have not worked on it. Both subtasks are evaluated by Mean Squared Error and Pearson's Correlation Coefficient. The dataset used for both is the LIRIS-ACCEDE [1] dataset. Full details of the challenge tasks and database can be found in [3].

There are various sources of information that can be exploited to predict the emotional impact of a movie clip. Although visual content is an essential source to infer viewers' emotion, audio and text are also potential components for this task. Frame-based and sequence-based approaches can be applied to analyse video frames to evaluate emotional impact.

In our method, we follow the frame-based approach to predict video emotional impact. From the training dataset, we extract deep features of each frame and train two models to predict valence and arousal properties of a video frame. Then we apply the two trained models to evaluate each frame in the test set independently. Finally, we employ the sliding window technique to smooth the final results.

2 APPROACH

In this section, we will describe how we approach the valence-arousal prediction subtasks. The proposed method contains four stages: frame extraction, features extraction, prediction models and post-processing methods. Our system pipeline is shown in Figure 1 below.

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

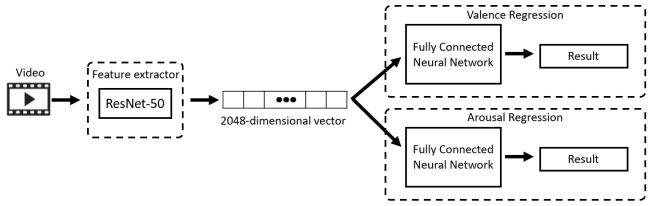

Figure 1: Overview of Frame-based Prediction Models

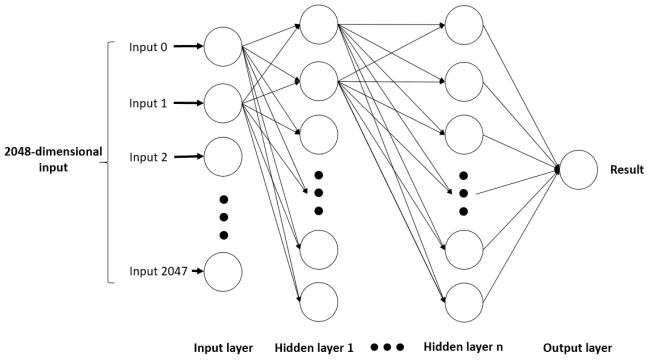

Figure 2: Valence/Arousal prediction models

2.1 Frame extraction

Firstly, we extracted one frame per second of all movies on the training and test set. For frame extraction, we use ffmpeg the framework and the extract command provided by the organizers to extract frames.

2.2 Features extraction

For the frame extraction, we use pre-trained 50-layer Residual Network (ResNet-50) [4] for ImageNet. The ResNet-50 used as a feature extractor and 2048-dim features vector are extracted from each frame of the movies. In our experiments, we used the Keras ResNet-50 pre-trained model on ImageNet dataset and calculate the features vector from the global average pooling that applied to the output of the last convolutional layer.

2.3 Prediction models

We apportion the training set provided by the organizer into training and validation sets with a ratio of 80:20. An overview of the prediction models is shown in Figure 2.

We employ 2-layer fully connected neural network to learn the emotional models. The models take 2048-dim features vectors extracted from ResNet-50 as input. We experimented with varying the number of the nodes for the first and the second layer with 128, 256, 512 and training epochs as 10, 15, 20 . We use Root Mean Square Propagation (RMSProp) with the learning rate 10^{-4}.
All prediction models are trained separately for valence and arousal.

2.4 Post-processing

After get the valence/arousal results, we applied the Average Window Sliding Technique to smooth out the random noise. We tested the window size of the algorithm with 3, 5, 7.

3 RESULTS AND ANALYSIS

In this section, we will describe in detail the experimental specification, five runs that we have summitted for the valence-arousal subtask and the result.

3.1 Experimental specification

The experiments are processed on Google Compute Engine with 2 vCPU, 7.5 GB RAM and Nvidia Tesla K80 GPU. The average times for extracting 93406 frames on the training set about 1 hour, 40 minutes for extracting features by ResNet and 3 minutes for training each models.

3.2 Submitted runs

We tested all trained models with the validation set. After we got the results of all models on the validation set, we sorted descending by the mean square error and selected from Top-1 to Top-4 model to submit. The details of each run are listed below.

All runs take ResNet-50 features as input. From Run 2 to Run 5, the results take the Window Sliding Technique with the window size = 7.

- **Run 1:** For both valence and arousal, 2-layer fully connected neural network with 128 nodes on the first layer, 512 nodes on the second layer trained on 20 epochs.
- **Run 2:** The same models with **Run 1** but we also take the Window Sliding Technique with the window size = 7 to smooth out the random noise.
- **Run 3:** For valence, 2-layer fully connected neural network with 256 nodes on the first layer, 512 nodes on the second layer trained on 10 epochs. For arousal, 2-layer fully connected neural network with 512 nodes on the first layer, 512 nodes on the second layer trained on 15 epochs.
- **Run 4:** For valence, 2-layer fully connected neural network with 256 nodes on the first layer, 512 nodes on the second layer trained on 15 epochs. For arousal, 2-layer fully connected neural network with 128 nodes on the first layer, 512 nodes on the second layer trained on 10 epochs.
- **Run 5:** For valence, 2-layer fully connected neural network with 512 nodes on the first layer, 512 nodes on the second layer trained on 10 epochs. For arousal, 2-layer fully connected neural network with 512 nodes on the first layer, 512 nodes on the second layer trained on 10 epochs.

3.3 Results and Analysis

Table 1: Results of the valence-arousal subtask.

Runs	Valence		Arousal	
	MSE	r	MSE	r
Run 1	0.11936	0.10665	0.17448	0.05282
Run 2	**0.11504**	**0.14565**	**0.17055**	0.07525
Run 3	0.11943	0.14513	0.17443	0.06978
Run 4	0.11731	0.14097	0.17901	0.01877
Run 5	0.11526	0.14306	0.17282	**0.09123**

Figure 3: Examples of the frame similarity between the tranning set and test set on Valence

As shown in the Table 1, Run 2 obtains the best result for the valence-arousal subtask. But in general, there is a slight difference in the results. Comparing Run 1 with Run 2, applying Window Sliding Technique provides better results. As shown in the Figure 3, there is the similarity in frames between the training set and test set on valence.

4 CONCLUSION

We propose a simple method to evaluate the emotional impact, i.e. valence and arousal properties, of a video frame. We study several settings of classification modules with 1 to 2 fully connected layers and different numbers of nodes in each layer to select an appropriate model for each property. Experimental results demonstrate that although our method is simple, it achieves promising results for this task. This is the initial step to develop better method to utilize temporal information of frame sequences, and other media types, such as audio and text components.

ACKNOWLEDGMENTS

We would like to express our appreciation to Multimedia Communications Laboratory, University of Information Technology, VNU-HCM, Vietnam, and Software Engineering Laboratory, University of Science, VNU-HCM, Vietnam.

REFERENCES

[1] Y. Baveye, E. Dellandrea, C. Chamaret, and Liming Chen. 2015. LIRIS-ACCEDE: A Video Database for Affective Content Analysis. *IEEE Transactions on Affective Computing* 6, 1 (Jan.-March 2015), 43–55.

[2] L. Canini, S. Benini, and R. Leonardi. 2013. Affective Recommendation of Movies Based on Selected Connotative Features. *IEEE Transactions on Circuits and Systems for Video Technology* 23, 4 (April 2013), 636–647. https://doi.org/10.1109/TCSVT.2012.2211935

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

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.

[5] Rajiv Ratn Shah, Yi Yu, and Roger Zimmermann. 2014. ADVISOR: Personalized Video Soundtrack Recommendation by Late Fusion with Heuristic Rankings. In *Proceedings of the 22Nd ACM International Conference on Multimedia (MM '14)*. ACM, New York, NY, USA, 607–616. https://doi.org/10.1145/2647868.2654919

[6] K. Yadati, H. Katti, and M. Kankanhalli. 2014. CAVVA: Computational Affective Video-in-Video Advertising. *IEEE Transactions on Multimedia* 16, 1 (Jan 2014), 15–23. https://doi.org/10.1109/TMM.2013.2282128

[7] Sicheng Zhao, Hongxun Yao, Xiaoshuai Sun, Xiaolei Jiang, and Pengfei Xu. 2013. Flexible Presentation of Videos Based on Affective Content Analysis. In *Advances in Multimedia Modeling*, Shipeng Li, Abdulmotaleb El Saddik, Meng Wang, Tao Mei, Nicu Sebe, Shuicheng Yan, Richang Hong, and Cathal Gurrin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 368–379.