Using Aesthetics and Action Recognition-based Networks for the Prediction of Media Memorability

Mihai Gabriel Constantin[1], Chen Kang[2], Gabriela Dinu[1], Frédéric Dufaux[2], Giuseppe Valenzise[2], Bogdan Ionescu[1]
[1]CAMPUS, University Politehnica of Bucharest, Romania
[2]Laboratoire des Signaux et Systèmes, Université Paris-Sud-CNRS-CentraleSupélec, Université Paris-Saclay, France
mgconstantin@imag.pub.ro

ABSTRACT

In this working note paper we present the contribution and results of the participation of the UPB-L2S team to the MediaEval 2019 Predicting Media Memorability Task. The task requires participants to develop machine learning systems able to predict automatically whether a video will be memorable for the viewer, and for how long (e.g., hours, or days). To solve the task, we investigated several aesthetics and action recognition-based deep neural networks, either by fine-tuning models or by using them as pre-trained feature extractors. Results from different systems were aggregated in various fusion schemes. Experimental results are positive showing the potential of transfer learning for this tasks.

1 INTRODUCTION

Media Memorability was studied extensively in recent years, playing an important role in the analysis of human perception and understanding of media content. This domain was approached by numerous scientists from different perspectives and fields of study, including psychology [1, 13] and computer vision [3, 12], while several works analyzed the correlation between memorability and other visual perception concepts like interestingness and aesthetics [6, 8]. In this context, the MediaEval 2019 Predicting Media Memorability task requires participants to create systems that can predict the short-term and long-term memorability of a set of soundless videos. The dataset, annotation protocol, precomputed features, and ground truth data are described in the task overview paper [5].

2 APPROACH

For our approach, we used several deep neural network models based on image aesthetics and action recognition. For the first category, we fine-tuned the aesthetic deep model presented in [9]. It is based on the ResNet-101 architecture [7]. For the action recognition networks, we used features extracted from the I3D [2] and TSN [15] networks and attempted to augment these features with the C3D features provided by the task organizers. Finally, we performed some late fusion experiments to further improve the results of these individual runs. Figure 1 summarizes and presents these approaches. The approaches are detailed in the following.

2.1 Aesthetics networks

The aesthetic-based approach modifies the ResNet-101 architecture [7], trained on the AVA dataset [11] for the prediction of image aesthetic value, following the approach described in [9]. This approach generates a deep neural model that can process single image aesthetics and must be fine-tuned to process the short and long term memorability of videos. To generate a training dataset that will support the fine-tuning process, we extracted key-frames in two ways: (i) key frames from the 4th, 5th, and 6th second of each sample; (ii) one key frame every two seconds to test multi-frame training. In the retraining stage of the network for the memorability task, the provided devset is randomly split into three parts, with 65% of the samples representing the training set, 25% the test set and 10% the validation set. We adapted the last layer for this task by creating a fully connected layer with 2,048 inputs and 1 output. During the fine-tuning process, we applied mean square error as loss function, using an initial learning rate of 0.0001. We ran the training process for 15 epochs, with a batch size of 32.

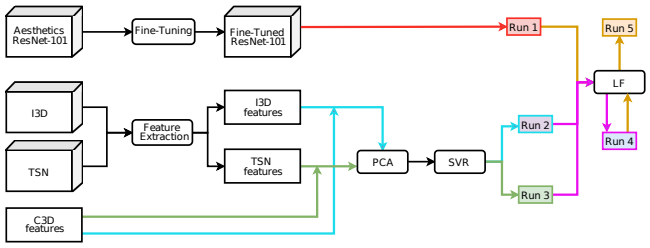

Figure 1: The diagram of the proposed solution.

2.2 Action recognition networks

Apart from the precomputed C3D features, we extracted the "Mixed_5" layer from the I3D network [2], trained on the Kinetics dataset [10] and the "Inception_5" layer of the TSN network [15], trained on the UCF101 dataset [14]. These features were used as inputs for a Support Vector Regression algorithm that generates the final memorability scores. We conducted preliminary early fusion tests with combinations of these features in order to select the best possible combinations, testing both each feature vector individually and all possible combinations of two feature vectors. We also employed a PCA dimensionality reduction, reducing the size of each vector to 128 elements. Finally, to train the SVR system, we used a random 4-fold approach, with 75% of the data representing the training set and 25% representing the validation set. We used parameter tuning for the SVR model, via a RBF kernel and performing a grid search with two parameters: the C parameter and the gamma parameter (taking values 10^k, where $k \in [-4, ..., 4]$).

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

Table 1: Results of the proposed runs (preliminary experiments on *devset*, and official results on *testset*).

		Devset - Spearman's ρ		Testset - Spearman's ρ	
Run	*System description*	*Short-term*	*Long-term*	*Short-term*	*Long-term*
run1	Aesthetic-based	0.448	0.230	0.401	0.203
run2	Action-based (TSN+I3D)	0.473	0.259	0.45	0.228
run3	Action-based (C3D+I3D)	0.433	0.204	0.386	0.184
run4	Late Fusion Action-based (run2 + run3)	0.466	0.200	0.439	0.218
run5	Late Fusion Aesthetic and Action (run1 + run2)	**0.494**	**0.265**	**0.477**	**0.232**

2.3 Late fusion

We employed several late fusion schemes on the best performing systems, trying to benefit from their combined strengths. We used three different strategies for combining these scores, namely: (i) LFMax, where we took the maximum score for each media sample; (ii) LFMin, where we took the minimum score; (iii) LFWeight, where each score from different samples was multiplied with a weight w. We assigned each weight varying values according to the formula $w = 1-r/c$, where the rank r had the value 0 for the best performing system, 1 for the second best and so on, and c represents a coefficient that dictates rank influence on the weights.

3 EXPERIMENTAL RESULTS

The development dataset consists of 8,000 videos, annotated with short and long term memory scores, while the test dataset consists of 2,000 videos. The official metric used in the task is Spearman's rank correlation (ρ). The best performing systems in the development phase are selected, retrained on the whole devset by using the optimal parameters and lastly run on the testset data.

3.1 Results on the devset

During the tests performed on the devset, several systems and combinations of parameters stood out as best performers. Table 1 shows the performances recorded by the best performing aesthetic, action-based, and late fusion systems.

We used several dataset variations in retraining the aesthetic-based deep network. More precisely, we found that, for the short-term memorability, the best performing systems were the ones trained with keyframes extracted from the 5th second and the ones extracted from the multi-frame approach. The results were both similar with a Spearman's ρ of 0.45. On the other hand, in the long-term memorability subtask we found that the best performing systems were the ones trained with keyframes from the 5th frame. Although this may seem somewhat surprising, giving that bigger data sets usually account for better results, we believe that the reason behind this is that each video contains only one scene. Therefore not much additional information is given to the system when more frames are extracted because the frames are very similar. However, we would also like to point out that the results for the other frame extraction schemes were not much lower than these.

Regarding the 3D action-recognition based systems, we noticed that individual systems, based on only one feature vector (TSN, I3D or C3D) had a low performance, with a Spearman's ρ score of under 0.42. This performance further dropped when we used the original vectors, without applying PCA reduction, therefore demonstrating the positive influence that dimensionality reduction has on the final results. Therefore we decided to apply an early fusion scheme, where we tested all the possible combinations of the feature vectors, by concatenating them. The best performing combinations were TSN + I3D and C3D + I3D.

Finally, in the late fusion part of the experiment, we generally decided to test late fusion schemes between the two action-recognition based systems and between the best performing action-recognition system (TSN + I3D) and the aesthetic-based system. In general, results for the LFMin systems were underperforming, while the LFMax systems were better than their components, but without bringing a significant increase in results. The best performing late fusion schemes proved to be based on LFWeight, more precisely using a c value of 5. This was an expected result, as it confirms some of our previous work in other MediaEval tasks [4].

3.2 Results on the testset

For the final phase, we retrained all the systems on the entire set of videos from devset, using the parameters computed in the previous phases and tested them on the videos from the testset. Table 1 presents also the results for this phase.

As expected, the best performance comes from a late fusion system using both aesthetic and action-based components (short-term $\rho = 0.477$ and long-term $\rho = 0.232$). Generally, we observe that the system ranking for the submitted systems is consistent with the one we observed during the development phase, however, the results are lower than those predicted then, with significant drops in performance for the aesthetic-based system and the action-based (C3D + I3D) approaches. In terms of single-system performance, the action-based TSN + I3D system performs best, followed by the aesthetic-based system.

4 CONCLUSIONS

In this paper we presented the UPB-L2S approach for predicting media memorability at MediaEval. We created a framework that uses aesthetic and action recognition based systems and some late fusion combinations of these systems, that predict short-term and long-term memorability scores for soundless video samples. The results show that these systems are able to individually predict these scores, while the best results are achieved via late fusion weighted schemes. This enforces the idea of better exploiting transfer learning to tasks where labeled data are in particular hard to obtain.

ACKNOWLEDGMENTS

This work was partially supported by the Romanian Ministry of Innovation and Research (UEFISCDI, project SPIA-VA, agreement 2SOL/2017, grant PN-III-P2-2.1-SOL-2016-02-0002).

REFERENCES

[1] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. 2008. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences* 105, 38 (2008), 14325–14329.

[2] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[3] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, and Martin Engilberge. 2019. VideoMem: Constructing, Analyzing, Predicting Short-term and Long-term Video Memorability. In *International Conference on Computer Vision (ICCV)*.

[4] Mihai Gabriel Constantin, Bogdan Andrei Boteanu, and Bogdan Ionescu. 2017. LAPI at MediaEval 2017-Predicting Media Interestingness.. In *MediaEval*.

[5] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Hélène Demarty, Ngoc Q. K. Duong, Xavier Alameda-Pineda, and Mats Sjöberg. 2019. Predicting Media Memorability Task at MediaEval 2019. *In Proc. of MediaEval 2019 Workshop, Sophia Antipolis, France, Oct. 27-29, 2019* (2019).

[6] Mihai Gabriel Constantin, Miriam Redi, Gloria Zen, and Bogdan Ionescu. 2019. Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates. *ACM Computing Surveys (CSUR)* 52, 2 (2019), 25.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2013. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence* 36, 7 (2013), 1469–1482.

[9] Chen Kang, Giuseppe Valenzise, and Frédéric Dufaux. 2019. Predicting Subjectivity in Image Aesthetics Assessment. In *IEEE 21st International Workshop on Multimedia Signal Processing, 27-29 Sept 2019, Kuala Lumpur, Malaysia*.

[10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and others. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[11] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2408–2415.

[12] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision*. 2730–2739.

[13] Roger N Shepard. 1967. Recognition memory for words, sentences, and pictures. *Journal of verbal Learning and verbal Behavior* 6, 1 (1967), 156–163.

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[15] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.