Using Vision Transformers and Memorable Moments for the Prediction of Video Memorability

Mihai Gabriel Constantin[1], Bogdan Ionescu[1]
[1]University Politehnica of Bucharest, Romania
mihai.constantin84@upb.ro

ABSTRACT

This paper describes the approach taken by the AI Multimedia Lab team for the MediaEval 2021 Predicting Media Memorability task. Our approach is based on a Vision Transformer-based learning method, which is optimized by filtering the training sets for the two proposed datasets. We attempt to train the methods we propose with video segments that are more representative for the videos they are part of. We test several types of filtering architectures, and submit and test the architectures that best performed in our preliminary studies.

1 INTRODUCTION

Media Memorability has attracted the attention of researchers from different domains for a long time. This included studies that revealed that humans have an uncanny ability of memorizing large quantities of images, going so far as to correctly encode details from those images. Generally speaking, there is a certain discrepancy between the study of image and video memorability, with more attention given in the current literature to the former. In this context, the MediaEval Predicting Media Memorability task [6], now at its fourth edition, creates a common benchmarking task for predicting the short- and long-term memorability of videos. This task offers data extracted and annotated from two datasets – TRECVid [1] and Memento10k [8]; and proposes two opened task, related to direct memorability prediction and generalization prediction between the two datasets.

Our proposed method for video memorability prediction relies on the use of Vision Transformer networks for feature extraction, a dense network ending for sample regression and an important frame filtering method that attempts to use the most memorable moments from the video samples in the training process. The rest of the paper is organized as follows: Section 2 presents the works most related to our proposed approach, while our method is presented in Section 3. Section 4 presents the results both in our training and development process and on the final testing set. Finally, the main conclusions are presented in Section 5.

2 RELATED WORK

Deep Neural Networks have come a long way in addressing many machine learning problems, and for a long time, starting with the success of AlexNet [7], when it came to visual data processing in general, convolutional neural networks were the norm in getting the best results. Interestingly, some domains related to the human perception of media data did not adhere to this general trend, as concepts like fusion, data manipulation and traditional feature extractors were sometimes more important in getting good results than deep neural networks [3], indicating a need for deeply understanding the data and the way it influences human subjects.

Recently, Vision Transformers shown their usefulness for image processing, surpassing convolutional approaches in image recognition tasks [5]. To the best of our knowledge, this approach is relatively untested in the domain of media memorability. This is perhaps to be expected, as the rise of Vision Transformers is in itself a novelty at this point in time.

3 APPROACH

The general outline of our memorability prediction method is presented in Figure 1. We propose creating a three stage system. In the first stage, we theorize that not all frames might be valuable for memorability calculation and therefore propose a frame filtering method. Following this, we extract visual features by using a Vision Transformer architecture, and, in a final step, we perform regression with a dense MLP architecture.

Frame filtering. We base our frame filtering system on the assumption that not all frames are equal when trying to determine the properties of a larger video sequence. In our case, we propose using the annotations provided by the organizers for selecting the frames that may best characterize the video from a memorability standpoint. We call these frames "Memorable Moments", and while they may not represent the exact moment or the exact process of human memory retrieval, we theorize that they may represent a better approach than simply attempting to use the entire video for processing.

We test several setups for the frame filtering method as follows. First of all, we have to take into account the lag time between human memory recognition and button press. Therefore, given rt, a user's response time in milliseconds from the start of the video, we subtract the following values: 500, 1000, 1500 milliseconds from the rt value in order to define the actual time of retrieval from memory. Of course we cap the resulting value at zero in case retrieval occurred very close to the start of the video. Furthermore, we take a variable number of frames, namely $15, 30, 60$ from the resulting location and use them for analysis. We will compare our filtering method (which we call $R2$) against a default method, where all the video is taken into consideration (called $R1$).

Visual Features. For visual feature extractors, we test two popular Vision Transformer architectures, namely the DeiT [9] and the BEiT [2]. No special fusion will be employed with these two features, as at this stage we will test them separately and choose the best performing one for the final set of experiments.

MediaEval'21, December 13-15 2021, Online

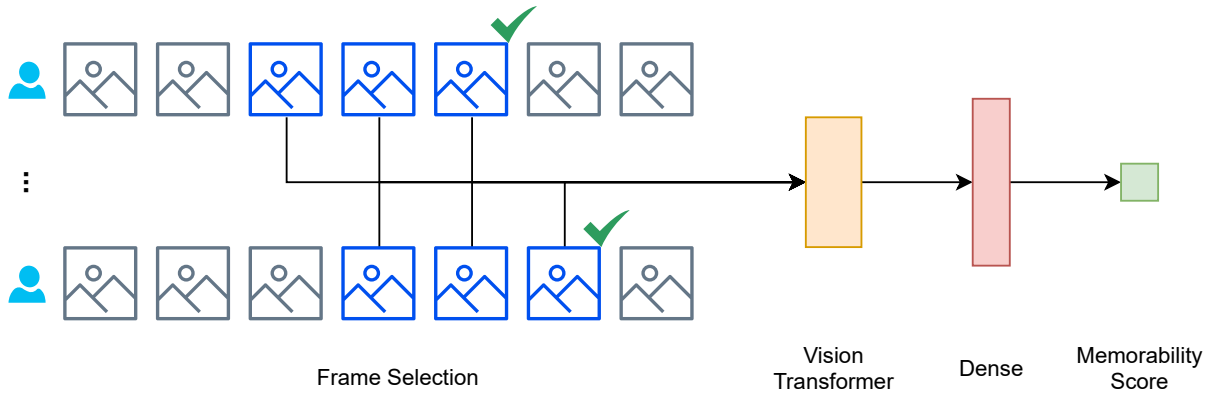

Figure 1: The diagram of the proposed frame filtering solution. The Frame Selection phase uses the most representative images from the entire set of frames in a video, that are then processed by a Vision Transformer architecture and processed by a Dense MLP head in order to obtain the final Memorability Score. Annotator picks on the training set for the Memorable Moments are presented with a green tick mark.

			R1			R2		
			Spearman	Pearson	MSE	Spearman	Pearson	MSE
Subtask 1	TRECVid	short-raw	0.293	0.312	0.01	**0.297**	0.311	0.01
		short-normalized	**0.26**	0.267	0.01	0.251	0.224	0.01
		long	0.079	0.102	0.01	**0.097**	0.114	0.04
	Memento10K	short-raw	0.407	0.409	0.01	**0.648**	0.652	0.01
		short-normalized	0.641	0.641	0.01	**0.648**	0.65	0.01
Subtask2	TRECVid	short-raw	0.089	0.11	0.02	**0.091**	0.108	0.01

Table 1: Results of the submitted systems on the two subtasks. R1 results are for the non-filtered systems, while R2 results present the filtered version of the systems.

Dense MLP. The final stage consists of classifying the chosen features extracted from the selected frames and outputting the final memorability score. This is done via a simple dense architecture with 3 hidden layers of size 1024, 512, and 256.

4 RESULTS AND ANALYSIS

In the first stage of development, we test the setups proposed in the previous Section, by training on the Memento training set and testing on its development set. With regards to the frame filtering method, we find that a setup of 1000 milliseconds delay in response time and 30 frames analyzed are the best setups, though not by a significant margin. For the Transformer architecture, we select the DeiT architecture as the best performer in these preliminary tests, though again not by a large margin.

The final results computed on the testset are presented in Table 1. It is interesting to notice that, in five out of the six $(R1, R2)$ comparison pairs, the results were better for the variant of the system which employed filtered training, via Memorable Moments, while at times even being so with a significant margin.

For the *prediction subtask* (subtask 1) we find that results for the Memento10K prediction are much better than the ones for TRECVid. This may be a result of many factors, but one of them may be represented by the lower number of video samples in the latter dataset. Also, continuing the trend recorded at the previous version of the Predicting Media Memorability task [4], we observe lower performance for long-term memorability prediction compared to short term.

Finally, regarding the *generalization subtask* (subtask 2), we find a significant drop in performance when compared to subtask 1. This may be due to differences in the types of movies in the dataset, but methods that reduce this issue must definitely be studied.

5 CONCLUSIONS

In this paper we present a media memorability prediction method that is based on the use of Vision Transformer architectures and frame filtering method we call Memorable Moments. Our experiments show good results for both these components and, for future developments we propose improving this framework by testing more feature extraction architectures, performing tests against convolutional architectures, predicting Memorable Moments on the testset, and testing this type of approach on other subjective multimedia concepts and properties.

ACKNOWLEDGMENTS

This work was funded under project AI4Media "A European Excellence Centre for Media, Society and Democracy", grant 951911, H2020 ICT-48-2020.

REFERENCES

[1] George Awad, Asad A Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, and others. 2020. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. *arXiv preprint arXiv:2009.09984* (2020).

[2] Hangbo Bao, Li Dong, and Furu Wei. 2021. BEiT: BERT Pre-Training of Image Transformers. *arXiv preprint arXiv:2106.08254* (2021).

[3] Mihai Gabriel Constantin, Liviu-Daniel Ştefan, Bogdan Ionescu, Ngoc QK Duong, Claire-Hélène Demarty, and Mats Sjöberg. 2021. Visual Interestingness Prediction: A Benchmark Framework and Literature Review. *International Journal of Computer Vision* (2021), 1–25.

[4] Alba García Seco De Herrera, Rukiye Savran Kiziltepe, Jon Chamberlain, Mihai Gabriel Constantin, Claire-Hélène Demarty, Faiyaz Doctor, Bogdan Ionescu, and Alan F Smeaton. 2020. Overview of MediaEval 2020 Predicting Media Memorability Task: What Makes a Video Memorable? *Proceedings of MediaEval'20* (2020).

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[6] Rukiye Savran Kiziltepe, Mihai Gabriel Constantin, Claire-Hélène Demarty, Graham Healy, Camilo Fosco, Alba García Seco de Herrera, Sebastian Halder, Bogdan Ionescu, Ana Matran-Fernandez, Alan F. Smeaton, and Lorin Sweeney. 2021. Overview of The MediaEval 2021 Predicting Media Memorability Task. In *Proceedings of MediaEval'21.*

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

[8] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. 2020. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16.* Springer, 223–240.

[9] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning.* PMLR, 10347–10357.