AIMultimediaLab at MediaEval 2022: Predicting Media Memorability Using Video Vision Transformers and Augmented Memorable Moments

Mihai Gabriel Constantin[1,*], Bogdan Ionescu[1]

[1]University Politehnica of Bucharest, Romania

Abstract

This paper describes AIMultimediaLab's approach and results achieved during the 2022 MediaEval Predicting Video Memorability task. The proposed approach represents a continuation of last year's work, using, updating and better analysing the concept of Memorable Moments. This is done by improving the scheme we use for selecting Memorable Moments, and allowing for the possibility that more than one video segment is representative for the entire video clip from a memorability standpoint. Furthermore, we propose studying a new architecture for processing the selected Memorable Moments, by implementing a variant of the popular ViViT architecture, that is more suited to analysing pure video content.

1. Introduction

Media Memorability is one of the domains that lately gained considerable traction in the research community, thanks to the need of novel and better methods of classifying the huge quantities of data associated with social media and video content sharing platforms. While previous work focused more on the prediction of image-based content, lately a significant push for video-based processing can be noticed in the multimedia research environment. In this context, the MediaEval 2022 Predicting Video Memorability task [1], one of the drivers of this tendency and now at its fifth edition, proposes three subtasks, based on the prediction of short-term video memorability: a video-based prediction task, a generalization task and an EEG-based task. The data offered by the organizers of this task is extracted from two popular datasets, namely the Memento10k dataset [2] for the first two tasks and the VideoMem [3] dataset for the second task, that are enhanced by the addition of EEG data for the third task.

As we will show, this paper represents a continuation and an ongoing work on the study of Memorable Moments in particular, and of the way video segments can represent or can be interpreted as representative of an entire video in general. For this edition of the MediaEval Memorability task, we will only be participating to the first subtask. The rest of the paper is organized as follows: Section 2 presents our approach, while Section 3 presents and analyzes the results, and the paper concludes with Section 4.

2. Proposed Approach

Our proposed method represents a continuation of last year's work [4], where we studied the use of two popular Vision Transformer architectures, namely DeiT [5] and the BEiT [6]

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online

*Corresponding author.

mihai.constantin84@upb.ro (M. G. Constantin)

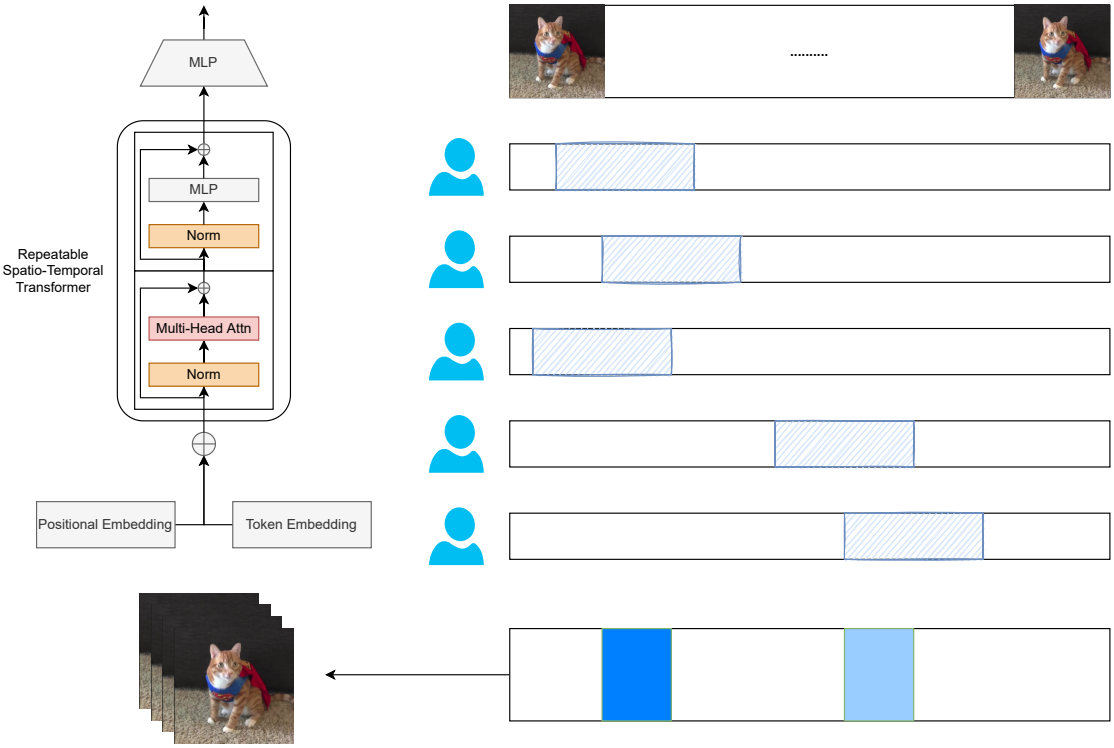

Figure 1: The diagram of the proposed solution. The Memorable Moments Frame Selection phase computes the most representative video segments, represented by the blue segments. These frames, along with the ground truth memorability score for the entire video clip are then passed to the ViViT architecture. Annotator picks on the training set for the Memorable Moments are presented with blue sketched segments.

networks for extracting visual features from selected frames, and a frame selection method we called Memorable Moments. We proposed and proven that having a frame selection method is a positive addition to the overall performance, as this would spare the network from processing frames that perhaps represent noise or are unimportant to the overall memorability score of the video in itself. The first of the changes we propose in this paper is represented by the replacement of the two architectures with an architecture dedicated to video processing, namely the popular transformer-based ViViT [7] model. Furthermore, we propose augmenting the previously developed Memorable Moments. While in the past, we only selected one region per video, that corresponded to the region most indicated by the annotators as memorable, now we allow several regions per video, as we theorize that this will take into account more representative video segments. This approach is shown in Figure 1, with the neural transformer architecture shown on the left and the frame selection scheme shown on the right side of the figure.

2.1. Neural Network Architecture

As already stated, we use the ViViT architecture [7] for processing the frames. Specifically, we deploy the spatio-temporal multi-headed self-attention ViViT model. This model uses a tubelet embedding that takes 3-dimensional tubes and feeds them to the network, created from the two spatial and one temporal dimension, in order to ensure that the network has access directly to

spatio-temporal information. We format this network to take 15 frames as the input, therefore we fix the window of frames from the beginning of our experiments. This architecture takes the input and passes it through a variable number of repeatable spatio-temporal transformer blocks. Each of these blocks are composed of a self-attention block and a MLP block.

We vary several parameters of the network in order to search for a robust architecture that would best fit our experiments. First of all, we test several values for the number of parallel self-attention heads in each block, using the values 4, 8, 16, 32. Secondly, we vary the number of repeatable blocks, using the values 4, 8, 16, 32.

2.2. Memorable Moments

Regarding the Memorable Moments scheme, we propose several variations for deploying this frame filtering method. Theoretically, given a video clip that is composed of a set of N frames denoted $V = [f_1, f_2, ..., f_N]$, and a set of M annotators $A = [a_1, a_2, ...a_M]$, each annotator that will go watch the video will press a button whenever they recognise the target video. Given a delay in response time of 500 milliseconds, corresponding to approximately 15 frames, that we determined and used in the previous version of the Memorable Moments scheme, we can calculate the central frame f_i that corresponds to each annotator's a_j moment of recognition.

Furthermore, we can allocate a score of 1 to each frame in a video that corresponds to an annotation, and can even extend that to the window of 15 frames that we chose at the begging of the experiments. Therefore if a central frame f_i gets a score of 1 from annotator a_j, the entire window composed of $[f_{i-7}, ..., f_i, ...f_{i+7}]$ gets that score. Finally, summing up all the annotations and given s_i^j the score for frame i from annotator j for a video we get the formula $S_i = \sum_{j=1}^{M} s_i^j$

In the next step we propose three methods for selecting the frames, called *Single*, *Double* and *Multi*. The *Single* method consists of basically selecting the highest value of S_i and using a 15-frame window around it as the single representative segment of the video. The *Double* method consists of selecting the two highest values of S_i, and using them as two central frames that generate two representative segments of the video. The final method, the *Multi*, uses a threshold value $\alpha \in (0, 1)$. We select the highest value of S_i, and all values that are higher than $\alpha \times S_j$, therefore getting a variable number of representative segments for each video. In cases of equality between S_i values we choose to take the frame with the lowest i value. Finally, we test several values for α, namely 0.95, 0.90, 0.85, 0.75.

3. Results and Analysis

3.1. Results on the development set

We conduct a set of preliminary experiments in the training phase, consisting of finding the optimal values for the ViViT size and for the α parameter. The results of these experiments are presented in Table 1. Each variation of the method has only one variable parameter, the other ones being in default mode for the experiment. Also, when varying the number of heads and the number of repeats the *Single* method for Memorable Moments is applied. Considering these experiments are done on the development set, 7000 movies are used for training (the training-set) and 1500 for validation (the development-set).

Nr. Heads	Spearman
4	0.5831
8	*0.5942*
16	0.5876
32	0.5879

Nr. Repeats	Spearman
4	0.5831
8	*0.6054*
16	0.5980
32	0.5601

α	Spearman
0.95	0.5974
0.90	0.6376
0.85	*0.6410*
0.75	0.6248

Table 1
Study performed on the development set, concerning the three variable parameters of the proposed method, namely the number of heads, the number of repeats and the α parameter.

Method	Spearman	Pearson	MSE
AIMultimediaLab-subtask1-Single	0.618	0.622	0.007
AIMultimediaLab-subtask1-Double	0.648	0.650	0.006
AIMultimediaLab-subtask1-Multi	**0.665**	**0.669**	**0.006**

Table 2
Final results for the proposed method, under the *Single*, *Double* and *Multi* Memorable Moments configuration.

3.2. Results on the testing set

Following this, we use the three values determined in the previous experiments, namely 8 multi-attention heads, 8 repeating blocks and an α value of 0.85 and submit three systems for evaluation by the Memorability task organizers. These three systems are represented by the *Single*, *Double* and *Multi* variations of the Memorable Moments selection scheme. The results are presented in Table 2, where the best performing method is shown to be the *Multi* configuration, with a Spearman value of 0.665. Wihle we can observe a significant growth in performance even when comparing the *Single* with the *Double* methods, an even better performance is recorded by the *Multi* approach, with almost 2% growth over the *Double* approach.

We theorize at this moment that this type of performance was to be expected, as each additional Memorable Moments configuration progressively adds more videos as representative of the video clips in the collection, therefore creating more training data. We propose that it may be interesting to research this problem on a different dataset, that perhaps contains more actions. Our reason for proposing this is that, in the current Memento dataset the video clips have 3 seconds and generally the actions shown in the clips do not change. It is possible that in longer clips the changes in actions or angles may be more significant and having more representatives for each video may improve the results even more.

4. Conclusions

In this paper we presented our approach for the MediaEval 2022 Predicting Video Memorability, consisting of an updated frame selection method called Memorable Moments, that has the role of selecting one or more representatives from each video clip for processing and training, and a video vision transformer ViViT architecture. Results show that selecting more than one representative for each video improves overall performance.

Acknowledgements

Financial support provided under project AI4Media, a European Excellence Centre for Media, Society and Democracy, H2020 ICT-48-2020, grant #951911.

References

[1] L. Sweeney, M. G. Constantin, C.-H. Demarty, C. Fosco, A. G. S. de Herrera, S. Halder, G. Healy, B. Ionescu, A. Matran-Fernandez, A. F. Smeaton, M. Sultana, Overview of the MediaEval 2022 predicting video memorability task, in: Working Notes Proceedings of the MediaEval 2022 Workshop, 2023.

[2] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 223–240.

[3] R. Cohendet, C.-H. Demarty, N. Q. Duong, M. Engilberge, Videomem: Constructing, analyzing, predicting short-term and long-term video memorability, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2531–2540.

[4] M. G. Constantin, B. Ionescu, Using vision transformers and memorable moments for the prediction of video memorability, in: MediaEval Multimedia Benchmark Workshop Working Notes, 2021. URL: http://ceur-ws.org/Vol-3181/.

[5] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.

[6] H. Bao, L. Dong, F. Wei, Beit: Bert pre-training of image transformers, arXiv preprint arXiv:2106.08254 (2021).

[7] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: A video vision transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6836–6846.