Exploring Video Transformers and Automatic Segment Selection for Memorability Prediction

Iván Martín-Fernández[1,*], Sergio Esteban-Romero[1], Jaime Bellver-Soler[1], Manuel Gil-Martín[1] and Fernando Fernández-Martínez[1]

[1]*Grupo de Tecnología del Habla y Aprendizaje Automático (THAU Group), Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid (UPM)*

Abstract

This paper summarises THAU-UPM's approach and results from the MediaEval 2023 Predicting Video Memorability task. Focused on the generalisation subtask, our work leverages a pre-trained Video Vision Transformer (ViViT), fine-tuned on memorability-related data, to model temporal and spatial relationships in videos. We propose novel, annotator-independent automatic segment selection methods grounded in visual saliency. These methods identify the most relevant video frames prior to conducting memorability score estimation. This selection process is implemented during both training and evaluation phases. Our study demonstrates the effectiveness of fine-tuning the ViViT model compared to a scratch-trained baseline, emphasising the importance of pre-training for predicting memorability. However, the model shows comparable sensitivity to both saliency-based and naive segment selection methods, suggesting that fine-tuning may harness similar benefits from various video segments. These results underscore the robustness of our approach but also signal the need for ongoing research.

1. Introduction and Motivating Work

Memorability is an aspect of human perception that has attracted the interest of researchers in psychology, neuroscience and computer science alike due to its relevance to areas as diverse as disease diagnosis, marketing and education. Taking advantage of the burgeoning advances in artificial intelligence architectures for media retrieval, classification and analysis as a proxy for modelling the connections between human senses and our understanding of the world through cognitive processes is particularly appealing, which explains the steady stream of work on the subject in recent years.

The MediaEval Predicting Video Memorability task, currently in its sixth edition [1], plays an important role in this effort. This contribution focuses on the generalisation subtask, focused on training systems that are able to learn general knowledge about the task that can be tested using different datasets.

To the best of our knowledge, most recent tackles on the Predicting Video Memorability task rely on using image-level architectures to extract knowledge from a handful of frames and then performing some sort of fusion strategy to obtain a single representation for the entire video, using powerful image-only backbone models such as the Vision Transformer but neglecting architectures that use video itself as input [2, 3, 4]. A notable exception comes from Constantin

MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online

*Corresponding author.

ivan.martinf@upm.es (I. Martín-Fernández); sergio.estebanro@upm.es (S. Esteban-Romero); jaime.bellver@upm.es (J. Bellver-Soler); manuel.gilmartin@upm.es (M. Gil-Martín); fernando.fernandezm@upm.es (F. Fernández-Martínez)

0009-0004-2769-9752 (I. Martín-Fernández); 0009-0008-6336-7877 (S. Esteban-Romero); 0009-0006-7973-4913 (J. Bellver-Soler); 0000-0002-4285-6224 (M. Gil-Martín); 0000-0003-3877-0089 (F. Fernández-Martínez)

and Ionescu [5], who train a Video Vision Transformer (ViViT) [6] to predict memorability from video segments, thus integrating the temporal aspect of videos into the core architecture of the system. They also present a technique for selecting which video segments are used to train and evaluate the model, based on the time it took annotators to recall watching a video. Although the authors prove the effectiveness of this method, we aim to develop an alternative that is based purely on input data and can therefore be used in the absence of this time-specific annotation. Furthermore, our strategy can be used for both training and evaluation, which we argue is an advantage over annotation-based approaches where an arbitrary segment selection method has to be designed for the testing phase in order to avoid data leakage.

In the spirit of transfer learning and generalisation, we propose to fine-tune a Video Vision Transformer, pre-trained on a generic video classification task, on memorability related data. Furthermore, we evaluate different selection strategies, where video segments are fed into the model in both training and evaluation steps.

2. Approach

We hypothesise that the ViViT architecture has the potential to be a robust, data agnostic model for memorability prediction, and therefore perform well in the generalisation task scenario. With this in mind, our approach is based on incorporating generic knowledge into the training process using two complementary strategies: a) fine-tuning a pre-trained ViViT model instead of starting training from scratch, and b) proposing automatic segment selection methods that do not rely on annotator data.

2.1. Fine-tuning Video Transformers

The ViViT Transformer is an adaptation of the original Vision Transformer that is able to process and model temporal relationships between frames as well as the spatial relationships that appear in each image by including a three-dimensional *Tubelet Embedding* encoder before the Transformer input. We start our training from the official ViViT checkpoint available at *huggingface*[1]. Its training data, Kinetics 400 [7], consists of 10-second clips extracted from YouTube videos and depicting one of 400 possible human actions, with a minimum of 400 clips per action class. We believe that modelling the subtleties of anthropoid imagery with this vast amount of content is key in understanding media memorability, as there is a direct relationship between both concepts [8]. Our regression head consists of a linear layer followed by a sigmoid activation function, which is appended to the last hidden state of the final encoder. This design operates under the hypothesis that this representation is inherently meaningful, requiring no further transformations.

We train on a single 32-frame long segment extracted from each video from the training set, using one of the segment selection methods that will be described next. The frame number selection is imposed by the architecture of the model that we wish to fine-tune. In order to compare our fine-tuning proposal, we train a baseline ViViT model from scratch, using the implementation proposed in [5] (i.e., 15 frames per segment, 8 attention heads per Transformer encoder, and 8 encoders). This baseline is trained on every possible 15 frame segment that can be extracted from each of the videos in the training set, so as to maximise the amount of information that is used for learning. We aim to test whether this simpler architecture can trade the lack of pre-training data with the ability of generating more meaningful representations of memorability-related videos.

[1]https://huggingface.co/google/vivit-b-16x2-kinetics400

Figure 1: Saliency maps for a sample frame. Whitest pixels are the predicted most salient.

2.2. Designing an automatic segment selection method

Using [3] as reference, we elaborate on the idea of selecting the most representative segment for the video and propose a novel method that is annotator-independent and selects the most relevant set of frames only using visual information, instead of relying on label related data. Based on the existing conception that saliency, defined as the prominence of features within an image that naturally attract human attention, is closesly related to memorability [9, 10, 11], we propose a method that automatically selects the most salient segment of a video and use it as input. We compare two different methods for computing image saliency. The first one, based on [12] and denoted **Fine Grained** according to the OpenCV implementation [13], involves analysing localised variations in the image to identify salient regions. The second method, **Spectral Residual** [14], identifies areas that stand out in the spectral domain of an image. By comparing these approaches, we aim to determine if the nuanced detail detection of the Fine Grained method or the global anomaly identification of the Spectral Residual approach is more effective in isolating memorable segments in videos. To identify the most representative video segment, we calculate the total pixel saliency across all frames, sum the saliency within a sliding window of $n = 32$ frames, and normalsze these values. The frame with the highest normalized window saliency and its adjacent n frames are then selected.

To test our approach, we compare it to two image-agnostic baseline methods: **Uniform Sampling** of n frames from the entire clip, and extracting the n frames from the **Center Segment** of the video.

3. Results and Discussion

As a preliminary study, we compare our fine-tuning approach with the *from scratch* baseline in order to analyse the effect of progressively unfreezing the weights of the Transformer encoders, starting from the one next to the regression head and going towards the input. We resort to the Uniform Sampling method for fine-tuning in this step. The results in term of Spearman Rank Correlation Coefficient (SRCC), the official metric for the task, are shown in Table 1, where we observe that our fine-tuning proposal significantly outperforms the baseline with just a single unfrozen encoder. This supports our idea that the ViViT model is greatly benefitted by a pre-training step in which general knowledge is acquired, and that it can translate this learnt relationships into the memorability problem. On the other hand, the fact that our best result comes from unfreezing all the model weights and letting it update as a whole leads us to think that the specific visual and semantic language related to the task still plays a crucial role in its solving, and therefore this aforementioned generic knowledge must be conditioned to it. This synergy between broad and specific expertise encourage us to use the fine-tuning approach for our runs, and to explore whether an automatic segment selection can enhance the adaptation process.

With this in mind, we show our final testing set results for our runs in Table 2, where we

Table 1
Results on the Memento10k dev set, where the baseline is compared with different fine-tuning strategies.

# Unfrozen encoders	*Baseline*	1	3	5	All
SRCC	*0.4119*	0.5573	0.5663	0.6274	**0.6529**

Table 2
SRCC results for the different segment selection strategies.

Segment Selection Strategy	Memento10k dev set	VideoMem test set
Uniform Sampling	0.653	0.437
Center Segment	0.651	0.440
Salient Segment - Fine Grained	**0.657**	**0.441**
Salient Segment - Spectral Residual	0.640	0.433

compare the different segment selection methods. We perceive that there is no significant difference between the saliency based methods and the naive approaches used for comparison, neither on the Memento10k developing set nor in the VideoMem test set, apart from a slight drop in performance when using the Spectral Residual method, indicating that the relationship between the spectral characteristics of an image and its memorability is somewhat weaker than a more nuanced approach. As can be seen in Figure 1, the Fine-Grained saliency maps are more detailed, in contrast with the less defined aspect of the Spectral Residual, which may influence on the selected segment. However, it seems that fine-tuned method benefits equally from segments across the whole video, independently of which part of it is used as input. Although we believe this is a sign of the robustness of our proposal, a more in-depth analysis of the relationship between image saliency and annotators response in terms of memorability could possibly further enhance the capabilities of this type of architecture.

4. Conclusions

In this paper we outline our contribution to the MediaEval 2023 Predicting Video Memorability task. We propose to leverage pre-trained Video Transformers in order to create robust memorability predictors that take sequences of frames as input. We also explore automatic segment selection methods based on saliency. Our results show that fine-tuning significantly outperforms training from scratch on our setup, but that the model is not specially sensible to automatic selection methods. We aim to deepen our exploration on the matter by developing advanced methods based on saliency and other perceptual features that output multiple candidate segments in order to broaden the training information, as well as evaluating the potential benefits of these methods on models trained from scratch.

Acknowledgments

We would like to thank M. Gabriel Constantin for his insights on his work, which have been greatly helpful for our research. I.M.-F.'s research was supported by the UPM (Programa Propio I+D+i). This work was funded by Project ASTOUND (101071191 — HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) of the European Commission and by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C22) and BeWord (PID2021-126061OB-C43), funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR"

References

[1] M. G. Constantin, C.-H. Demarty, C. Fosco, A. García Seco de Herrera, S. Halder, G. Healy, B. Ionescu, A. Matran-Fernandez, R. Savran Kiziltepe, A. F. Smeaton, L. Sweeney, Overview of the mediaeval 2023 predicting video memorability task, in: Proc. of the MediaEval 2023 Workshop, Amsterdam, The Netherlands and Online, 2024.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[3] M. G. Constantin, B. Ionescu, Using vision transformers and memorable moments for the prediction of video memorability, in: MediaEval 2021 workshop, 2021.

[4] M. Agarla, L. Celona, R. Schettini, et al., Predicting video memorability using a model pretrained with natural language supervision, in: MediaEval Multimedia Benchmark Workshop 2022 Working Notes, 2023.

[5] M. G. Constantin, B. Ionescu, Aimultimedialab at mediaeval 2022: Predicting media memorability using video vision transformers and augmented memorable moments, Working Notes Proceedings of the MediaEval 2022 Workshop (2023).

[6] 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.

[7] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, arXiv preprint arXiv:1705.06950 (2017).

[8] P. Isola, D. Parikh, A. Torralba, A. Oliva, Understanding the intrinsic memorability of images, Advances in neural information processing systems 24 (2011).

[9] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, B. Ghanem, What makes an object memorable?, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[10] M. Mancas, O. Le Meur, Memorability of natural scenes: The role of attention, in: 2013 IEEE International Conference on Image Processing, 2013, pp. 196–200. doi:10.1109/ICIP.2013.6738041.

[11] V. Mudgal, Q. Wang, L. Sweeney, A. F. Smeaton, Using saliency and cropping to improve video memorability, arXiv preprint arXiv:2309.11881 (2023).

[12] S. Montabone, A. Soto, Human detection using a mobile platform and novel features derived from a visual saliency mechanism, Image and Vision Computing 28 (2010) 391–402.

[13] G. Bradski, The OpenCV Library, Dr. Dobb's Journal of Software Tools (2000).

[14] X. Hou, L. Zhang, Saliency detection: A spectral residual approach, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8. doi:10.1109/CVPR.2007.383267.