AIMultimediaLab at MediaEval 2023: Studying the Generalization of Media Memorability Methods.

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

[1]University Politehnica of Bucharest, Romania

Abstract

Video memorability is one of the vital aspects of subjective multimedia perception and, as such, is closely and thoroughly studied in the computer vision literature. This paper presents the methods proposed by AIMultimediaLab for the generalization subtask of the 2023 edition of the Predicting Video Memorability task. We explore several methods for augmenting the training process for a video Vision Transformer network, aiming to increase the number of hard-to-predict samples in the training set in order to increase the robustness of the targeted AI model. Starting from our previous works, we analyze several visual features that define "hard-to-predict" samples, and based on these features, we augment the training data of our models to target those specific videos that pose problems for memorability prediction.

1. Introduction

The prediction of video memorability is an essential aspect in the subjective analysis of multimedia content, with the MediaEval Predicting Video Memorability[1] series of benchmarking tasks playing an important role in bringing attention in the computer vision community to the study of this concept. While previous editions of this benchmarking task have focused on memorability prediction in videos extracted, annotated and processed in similar conditions, this edition [1] focuses on the generalization task. Concretely, the organizers ask participants to train on data extracted from one memorability dataset, namely the Memento10k [2], and test their trained systems on data extracted from the VideoMem [3] dataset. This allows for an interesting setup where AI models are exposed to different types of videos, annotated by different people, and extracted from different sources, thus creating a testing scenario that better simulates real-world conditions.

As we will show throughout the paper, this work represents the continuation of some of our previous works on memorability, particularly those targeting the use of vision transformers in the prediction of subjective concepts, and a sample-based analysis of videos we defined as "hard-to-predict" from a memorability standpoint. We continue this work by applying training augmentation, particularly for the problematic videos for our vision transformer networks. The rest of the paper is structured as follows. Section 2 presents previous works our methods are based on. Following this, the methods employed by our team are presented in Section 3, while our results are presented in Section 4. Finally, the paper concludes with Section 5.

2. Related work

Our proposed method is built upon two of our previous works. The first one, published in the previous edition of the MediaEval memorability task [4], uses vision transformers to predict

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

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

[1]https://multimediaeval.github.io/editions/2023/tasks/memorability/

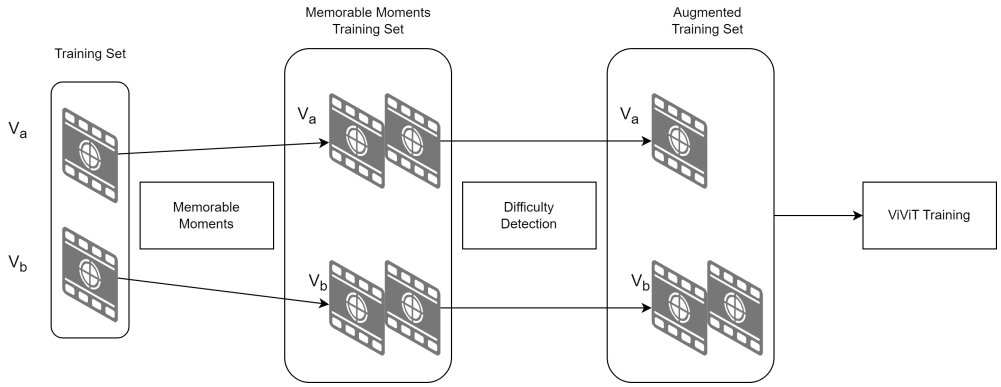

Figure 1: The diagram of the proposed solution. Starting from the training set, as proposed by the organizers, we apply the Memorable Moments method for detecting the two most representative segments of the original video. Following this, based on the difficulty detection methods we will present, the top most representative segment is kept for the videos detected as easy, and both segments for the videos detected as hard-to-classify, obtaining an augmented training set.

video memorability. More precisely, it is represented by a vision transformer model, derived from the popular ViViT [5] neural network. We will also use the "Memorable Moments" video segment selection method to select the most representative segments from the training videos and use only those segments during the training phase. The second work is represented by a feature-based analysis of all the runs submitted during the previous edition of the memorability task [6]. In this paper, we showed the emergence of some video samples that are significantly harder to classify by all participants, regardless of the systems they used or their pre-processing methods.

Starting from these two works, we select two segments as representatives for each video in the training set, based on the Memorable Moments approach presented in the first work [4]. We then employ several methods, similar to our second work [6], for detecting the videos in the training set that may be hard to classify by the proposed ViViT-derived model. While [6] presents just an analysis of submitted runs and launches some interesting hypotheses concerning the features that make a video hard to classify, this paper seeks to test these hypotheses and apply them directly to media memorability prediction.

3. Method

A general diagram of the training method we propose is presented in Figure 1. We propose using the Memorable Moments selection scheme, as presented in [4], in order to select the two most representative video segments in each video from the training set. Following this, we analyze several methods of determining which videos are challenging and which are easy to classify regarding their memorability score, using a set of features and visual descriptors. In the last step, we keep only the most representative video segment for easy-to-classify videos, and keep both segments for the hard-to-classify ones. We theorize that this imbalance we create in the dataset may allow the ViViT-based model to better learn the ground truth of samples that it may otherwise mispredict.

3.1. Memorable Moments

We use the Memorable Moments selection scheme, as presented in [4]. Concretely, given a video clip composed of N frames: $V = f_1, f_2, ...f_N$. Using the annotations provided by the competition organizers, we provide a score of 1 for the frame corresponding to the moment of recall while accounting for a 500 milliseconds delay in response and extending this score to a window of 15 frames around the moment of recall. We gather all the annotations and thus obtain a frame-level score of recall for each video as follows: $R_V = [s_1, s_2, ...s_N]$. Finally, taking the top two recall scores, we obtain the top two most significant Memorable Moments for each particular video. We then extract two video segments around these frames, each of them 15 frames long.

3.2. Prediction difficulty assessment

In [6] we presented a set of methods for determining which features are the most discriminative when analyzing the difficulty of media memorability prediction. Perhaps unsurprisingly, we found that videos with average memorability ground truth scores are more challenging to predict accurately than videos that are either very memorable or have low memorability. Other discriminative features are as follows: sharpness computed via the Laplacian operator [7] (sharper videos are harder to classify with regards to memorability), contrast computed in RGB space [8] (higher contrast videos are harder to classify), and dynamism computed via the Farnebäck method [9].

We use these four methods to determine which videos are more challenging to classify by AI models. As presented in the previous section, we will only keep two Memorable Moments video segments only for the videos deemed as "hard-to-predict". We will therefore compute the values for each of the four features (ground truth score, sharpness, contrast, and dynamism), and split the training set into four quartiles, according to the value of each feature, with the top quartile, Q_1, representing the videos that theoretically should be easier to predict according to each feature, and the bottom quartile, Q_4, representing those that would be hard to predict. The entire training set will thus be divided, for each feature f, as follows: $T = Q_{1,f} \cup Q_{2,f} \cup Q_{3,f} \cup Q_{4,f}$. We will then keep two Memorable Moments segments only for the videos that belong to the bottom quartile, Q_4.

3.3. Vision transformer network

We apply these training augmentation schemes to a vision transformer deep neural network that is based upon the ViViT architecture [5]. Specifically, we used the tubelet embedding that encodes spatio-temporal information as 3-dimensional tubes and feeds them to the network for training and inference. This architecture handles the 3-dimensional input by passing it through a series of repeatable spatio-temporal attention blocks. Based on the conclusions of our previous work [4], we design the network so that it can handle 15 frames at input, and use 8 parallel self-attention heads in each block, and a number of 8 repeatable transformer blocks.

This network is then trained in five different setups. In the original setup, only one segment per video is fed into the network at training time. We consider this setup as the baseline for our approach. The following four setups will contain augmented samples, represented by one additional video segment for each video in the Q_4 quartile for each of the selected discriminative features.

	SRCC	
Augmentation Method	devset	testset
ViViT - baseline	0.651	0.361
ViViT + GT score	0.668	**0.382**
ViViT + sharpness	0.628	0.291
ViViT + contrast	0.631	0.265
ViViT + dynamism	0.680	0.380

Table 1
Results of the proposed augmentation methods, according to Spearman's Rank Correlation Coefficient (SRCC) metric. We compare the baseline training results with the four augmentation categories, while also comparing the results on the Memento10k devset, used for method validation, with the results on the official testset composed of VideoMem videos.

4. Results

We present the results in Table 1. Firstly, we analyze the results on the validation set of this task, which is composed of the Memento10k devset. While the un-augmented baseline system results are good enough with a Spearman's Rank Correlation Coefficient (SRCC) of 0.651, two of the proposed augmentation methods outscore this for the validation experiments, even if by a small margin. These two methods are represented by the ground-truth score-based quartile augmentation with an SRCC value of 0.668 (ViViT + GT score), and the dynamism-based augmentation (ViViT + dynamism), with an SRCC value of 0.680. On the other hand, the two other methods score lower on the devset when compared with the baseline method.

Similar trends are noticeable when looking at the official testset results. The ViViT + GT score method has the highest performance, with an SRCC value of 0.382, closely followed by the ViViT + dynamism approach with 0.380. The baseline method scores 0.361, while the sharpness and contrast methods have even lower scores.

When comparing the direct prediction performance on the Memento10k devset with the official generalization scores on the VideoMem dataset, we notice a sharp decline in performance. This is an indication of the significant level of difficulty generalization tasks pose. On the other hand, we are pleased to report that at least two of the proposed feature-based augmentation methods scored better than the baseline method. The GT score-based method achieved a 5.81% increase over the baseline run, with similar performance from the dynamism-based method.

5. Conclusions

We presented a sample-based augmentation method for media memorability prediction, in a generalization setup, where the training and the testing data came from different datasets, which involved different video sources and annotators. Our best performing method augmented the samples at training time based on their ground truth scores. Concretely, videos that have ground truth memorability values close to the average were augmented, thus resulting in a schema that increases the number of hard-to-predict videos, allowing the AI model to learn more details about these videos.

Acknowledgements

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

References

[1] 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, L. Sweeney, Overview of the MediaEval 2023 predicting video memorability task, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[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: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16, Springer, 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/CVF International Conference on Computer Vision, 2019, pp. 2531–2540.

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

[5] 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.

[6] M. G. Constantin, M. Dogariu, A. C. Jitaru, B. Ionescu, Assessing the difficulty of predicting media memorability, in: 20th International Conference on Content-based Multimedia Indexing, 2023, pp. 188–192.

[7] J. Wan, X. He, P. Shi, An iris image quality assessment method based on laplacian of gaussian operation., in: MVA, 2007, pp. 248–251.

[8] Y. Ke, X. Tang, F. Jing, The design of high-level features for photo quality assessment, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 1, IEEE, 2006, pp. 419–426.

[9] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13, Springer, 2003, pp. 363–370.