

Video Memorability Prediction using Deep Features and Loss-based Memorability Distribution Estimation

Safaa Azzakhnini^{1,*†}, Olfa Ben-Ahmed^{1,†} and Christine Fernandez-Maloigne^{1,†}

¹XLIM Research Institute, URM CNRS 7252, University of Poitiers, France

Abstract

In this paper, we address the video memorability prediction problem, which was a part of MediaEval 2022. This problem involves learning a model that maps the video content to a global memorability score. However, this value is still related to uncertain factors, including time and human subjectivity. Hence, we focus in this work mainly on modeling the subjective nature of the memorability experiments. Our approach relied on predicting the video memorability based on the maximum likelihood estimation of the Bernoulli distribution of the success variable of each video. To evaluate our approach, preliminary experiments were performed on the Memento10k dataset. First promising results were achieved by using a simple fusion of the visual and textual pre-trained features as input.

1. Introduction

Understanding content memorability is a task that has gained psychologists' interest for over a decade [1, 2]. Most of the performed studies showed that mapping image content to image memorability could be measured and, therefore, predictable [1]. With advances in machine learning and deep learning approaches, this task has also gained the attention of the computer vision community to investigate these algorithms to build models able to predict the memorability of a given multimedia content [3]. The first works started with studying image memorability and proposed models that have achieved promising results [4, 5, 6]. These results have encouraged researchers to extend this challenge to videos and release public datasets to tackle this task, mainly the Memento10k [7] and VideoMem [8] datasets used in predicting video memorability in the current MediaEval challenge [9].

Related works in video memorability prediction have mainly focused on extracting the visual and temporal information from videos and the semantic information from textual data. From the previous challenge, authors in [10] combined a set of pre-trained visual and textual features using transformers fed into a regressor to perform the regression task. In [11], the authors proposed a fusion approach based on a weighted sum of visual and textual segment features followed by a linear neural network. Another work [12] focused on filtering the frames that are more informative in the video and extracting the features from them using vision transformers. From another perspective, the authors who released the Memento10k dataset [7], proposed a new mathematical formulation of memorability decay, resulting in a model that can produce a quantitative estimation of how a video decays in memory over time. Although this model investigated the correlation between memorability and decay, it ignored modeling the memorability experiment's subjectivity.

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


*Corresponding author.

†These authors contributed equally.

✉ safaa.azzakhnini@univ-poitiers.fr (S. Azzakhnini); olfa.ben.ahmed@univ-poitiers.fr (O. Ben-Ahmed); christine.fernandez@univ-poitiers.fr (C. Fernandez-Maloigne)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In order to design a suitable model for memorability prediction, it is necessary to understand what factors influence a video memorability score. In the Memento10K dataset [7], video memorability is computed using a memory test called the memory game. In this test, participants are given a stream of videos and are asked whether or not they have seen them before. The score computed for each video corresponds to an average among the total number of successes obtained by changing participants and video positions in the stream (lags). A global score is then associated with a video as a memorability score that characterizes the probability that a person will be able to identify this video when seen before. In this dataset, the authors showed that the memorability score is linearly correlated with lags and decay. They also provided the response time taken in each experiment in which we noticed a correlation with the participants' answers. Therefore, given the subjective nature of the memory test, we propose to model video memorability following a subjective methodology. Hence, we assume that each memorability score related to each video follows a specific distribution that can be estimated instead of estimating the global average. Therefore, a loss function based on maximum likelihood estimation was proposed to estimate the distribution parameter. We consider this approach as a first attempt to model video memorability by considering the relative obtained successes for each video. This paper is organized as follows. A more detailed description of the approach is provided in section 2. The obtained results on the validation and test sets are provided and analyzed in section 3. Finally, we conclude the paper in section 4

2. Proposed approach

Deep neural networks have shown remarkable success in many classification problems where the goal of the model is to find a function that directly maps the input space to an output value. However, in problems where subjectivity is present, the characterization of uncertainties is desired. In the memorability estimation experiment, several factors are involved, including the positions of the video the first and second time in the stream and the time spent since the first time looking at the video. For each video, these settings vary and lead to a different outcome, where this latter can be either a correct response or not.

Statistically, a single experiment for two possible outcomes is called a Bernoulli trial, where the probability of success is p and failure is $1 - p$. Therefore, each video memorability may be regarded as binomial distribution with n trials, where in each Bernoulli trial, the outcome variable may be a correct response or not.

We formulate the problem as the following: Let's $V = \{v_1, v_2, \dots, v_N\}$ be a set of N videos and $M = \{m_1, m_2, \dots, m_N\}$ the set of the N corresponding memorability scores. For each video, a number of n independent settings were performed to compute the global score, where each is characterized by a success or failure outcome. Hence, the memorability m_j of a video v_j can be modeled as a set of Bernoulli trials with the same parameter p , where trials are independent of one another, $m_j \sim \text{Bernoulli}(p)$. The main quantity of interest is the proportion of subjects who respond favorably (the proportion of successful trials, which is the total number of successful trials divided by the total number of subjects).

Every video v_j is represented by a single embedding X_j . This embedding corresponds to a simple concatenation of visual and textual features. The visual features were computed as the mean of extracted pre-trained features from Resnet50 and Efficient-Net models on the frame level (features provided with the dataset), while the textual features were obtained using pre-trained features from the Bert transformer model [13].

The proposed approach consists, therefore, of designing a neural network that estimates p , the probability of 'success' (i.e., the proportion of observed successes) given the input features

X. In order to estimate the parameter p that is most likely to have given rise to the observation X , we use as a loss function the log-likelihood of p based on the observation X with n outputs x_i with $i = 1 \cdots n$. Hence, the goal of the model is to find p^* that minimizes $l(p | X)$ such that:

$$l(p | X) = - \sum_{i=1}^n (x_i \log p + (1 - x_i) \log (1 - p)) \quad (1)$$

3. Results and Analysis

Preliminary analysis of the average memorability scores The histogram in figure 1b shows the distribution of the memorability scores on the training set. Indeed, the memorability score over the training examples follows a skewed and asymmetric distribution with a skew of -0.529 . The nature of this distribution makes learning the low values a difficult task. Therefore, a transformation should be applied to the memorability variable. We have applied the square transform to the output variable, reducing the distribution's skewness (Figure 1b). Furthermore, we have used the inverse transform to the predictions not to impact the regression analysis results.

Preliminary analysis of the response time Figure 1a shows the distributions of the response time (rt) for the examples where the correct variable is equal to 1 (success) and equal to 0 (failures). We notice that rt follows different distributions for the two cases. Therefore, this shows that the rt feature is highly correlated with the memorability outcome. Based on this observation, we investigated adding a regularization term to the loss function based on maximizing the correlation between rt and the predicted memorability as a simple approach.

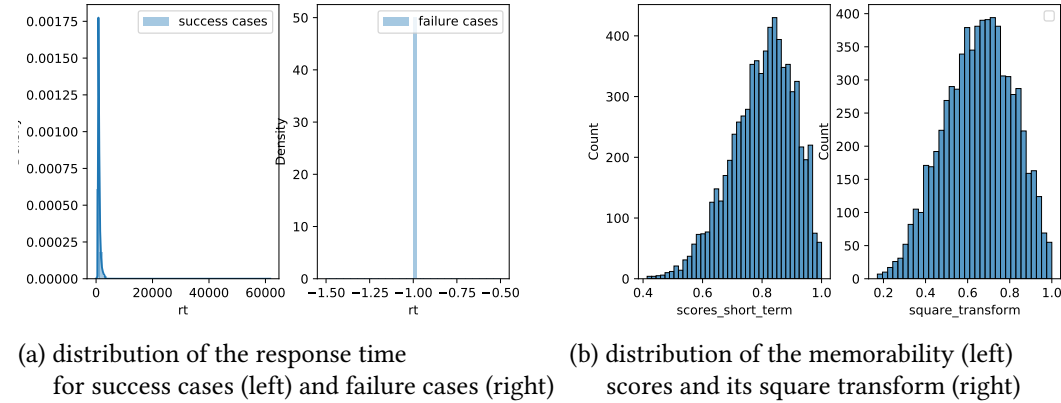


Figure 1: The distribution of the response time and memorability score on the training set of the Memento10K dataset

Obtained results on the validation set Table 1 shows obtained Mean Squared Error (MSE), Mean Absolute Error (MAE), Spearman and Pearson's rank correlation coefficient values, and the coefficient of determination R^2 . We have compared the approach with using a neural network where the loss function corresponds to minimizing the MSE of the average memorability scores. These findings show that the proposed approach results in a better coefficient of determination (from 0.38 to 0.42). The Spearman and Person correlation coefficients have also been improved compared to using an MSE as a loss function.

Table 1: Obtained performance on the validation set of the Memento10k dataset

	Our approach	Without the proposed loss
Mean squared error (MSE)	0.0158	0.007
Mean absolute error (MAE)	0.1012	0.0651
Spearman correlation	0.6511	0.6406
Pearson correlation	0.6502	0.6325
Coefficient of determination (R^2)	0.4226	0.3895

Obtained results on the test set In this section, we present the obtained results on the test set. We evaluated four independent runs to gain more insights about the approach. We started by evaluating the model using visual and textual features in the first run. We included the temporal dependencies between the frames in the second run using the LSTM model and visual information only. In the third run, we have added the proposed regularization term based on the response time (reg1). Finally, we added another simple regularizer to maximize the R^2 between predicted and true values (reg2). The obtained results are shown in table 2.

Table 2: Obtained results on the test set of the Memento10K dataset

	Spearman correlation	Pearson correlation	MSE
Proposed approach	0.597	0.608	0.007
LSTM using images only	0.537	0.547	0.008
The approach with reg1	0.574	0.589	0.01
The approach with reg2	0.576	0.587	0.007

Based on the obtained values, the best run corresponds to the model predictions using the proposed loss only (with a spearman correlation of 0.597). Although the performance has decreased compared to the values obtained on the validation set, but it still provides promising results. This can be improved by more investigation on the feature level as well as by considering more frames rather than using three frames. Furthermore, we observe that the added regularization terms could not result in a better performance. Therefore, we consider including the response time factor in a more suitable way in our future experiments.

4. Conclusion

In this paper, we investigated the subjectivity of the memorability experiments by considering each memorability score as a Bernoulli distribution with the parameter p . We proposed a neural network that estimates p based on using the maximum likelihood as a loss function. The obtained performance on the validation set shows promising results. We consider this approach to be early work on the subjectivity modeling of video memorability. An interesting next step would be to examine the memorability distribution estimation by considering lags and video decays. Concerning the neural network, it might be beneficial to use more sophisticated approaches such as video transformers to capture the spatiotemporal relationships in the data.

5. Acknowledgement

Support for this research was provided by a grant from La Région Nouvelle Aquitaine (CPER-FEDER P-2019-2022), in partnership with the European Union (FEDER/ ERDF, European Regional Development Fund)

References

- [1] N. C. Rust, V. Mehrpour, Understanding image memorability, *Trends in cognitive sciences* 24 (2020) 557–568.
- [2] Z. Bylinskii, L. Goetschalckx, A. Newman, A. Oliva, Memorability: An image-computable measure of information utility, in: *Human Perception of Visual Information*, Springer, 2022, pp. 207–239.
- [3] S. Lahrache, R. El Ouazzani, A survey on image memorability prediction: From traditional to deep learning models, in: *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, IEEE, 2022, pp. 1–10.
- [4] H. Squalli-Houssaini, N. Q. Duong, M. Gwenaëlle, C.-H. Demarty, Deep learning for predicting image memorability, in: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 2371–2375.
- [5] A. Khosla, A. S. Raju, A. Torralba, A. Oliva, Understanding and predicting image memorability at a large scale, in: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2390–2398.
- [6] Q. Xu, F. Fang, A. Molino, V. Subbaraju, J.-H. Lim, Predicting event memorability from contextual visual semantics, *Advances in Neural Information Processing Systems* 34 (2021) 22431–22442.
- [7] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, in: *European Conference on Computer Vision*, Springer, 2020, pp. 223–240.
- [8] R. Cohendet, K. Yadati, N. Q. Duong, C.-H. Demarty, Annotating, understanding, and predicting long-term video memorability, in: *Proceedings of the 2018 ACM on international conference on multimedia retrieval*, 2018, pp. 178–186.
- [9] L. Sweeney, M. G. Constantin, C.-H. Demarty, C. Fosco, A. García Seco 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: *MediaEval Multimedia Benchmark Workshop Working Notes*, 2022.
- [10] R. Kleinlein, C. Luna-Jiménez, F. Fernández-Martínez, Thau-upm at mediaeval 2021: From video semantics to memorability using pretrained transformers (2021).
- [11] Y. Lu, X. Wu, Cross-modal interaction for video memorability prediction (2021).
- [12] M. G. Constantin, B. Ionescu, Using vision transformers and memorable moments for the prediction of video memorability (2021).
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).

Author Agreement to Publish a Contribution as Open-Access on CEUR-WS.org

Herewith I/we (the author(s) resp. the copyright holders) **agree** that my/our contribution:

authored by: Safaa Azzakhmini, Ofa Bey-Ahmed and Christine Fernandez-Mabigne

with corresponding author Safaa Azzakhmini and Ofa Bey-Ahmed

Name: XLIM, University of Poitiers

Affiliation: 11 Bd Marie et Pierre Curie, 86360 - Chasseneuil du Poitou
France

Address: safae.azza khmini@gmail.com

Email: shall be made available as an open-access publication under the **Creative Commons License Attribution 4.0 International (CC BY 4.0)**, available at <https://creativecommons.org/licenses/by/4.0/legalcode>, and be published as part of the proceedings volume of the event

Name and year of the event: MediaEval'22

Editors of the proceedings (editors):

I/we agree that my/our contribution is made available publicly under the aforementioned license on the servers of CEUR Workshop Proceedings (CEUR-WS). I/we grant the editors, RWTH Aachen, CEUR-WS, and its archiving partners the non-exclusive and irrevocable **right to archive** my/our contribution and to **make it accessible** (online and free of charge) for **public distribution**. This granted right extends to any associated metadata of my/our contribution. Specifically, I/we license the associated metadata under a Creative Commons CC0 1.0 Universal license (public domain). I/we agree that our author names and affiliations is part of the associated metadata and may be stored on the servers of CEUR-WS and made available under the CC0 license. I/we acknowledge that the editors hold the copyright for the proceedings volume of the aforementioned event as the official collection of contributions to the event.

I/we have not included any copyrighted third-party material such as figures, code, data sets and others in the contribution to be published.

I/we warrant that my/our contribution (including any accompanying material such as data sets) does not infringe any rights of third parties, for example trademark rights, privacy rights, and intellectual property rights. I/we understand that I/we retain the copyright to my/our contribution. I/we understand that the dedication of my/our contribution under the CC BY 4.0 license is irrevocable. I/we understand and agree that the **full responsibility/liability** for the content of the contribution rests upon me/us as the authors of the contribution. I/we **release and relieve** the aforementioned editors, RWTH Aachen, individuals providing the CEUR-WS service, and the archiving partners of CEUR-WS **of liability caused by ordinary/simple negligence** in the publication or archiving of my/our aforementioned contribution via the servers used by CEUR-WS.

I/we have read the conditions of the Creative Commons License Attribution 4.0 International (CC BY 4.0), and agree to apply this license to my/our contribution.

Location, Date, Signature of the corresponding author representing all authors
(Signature must be handwritten with a pen on paper)

Safaa Azzakhmini 