Textual Analysis for Video Memorability Prediction

Camille Guinaudeau[1,2], Andreu Girbau Xalabarder[2]

[1]Japanese-French Laboratory for Informatics, Tokyo, Japan
[2]National Institute of Informatics, Tokyo, Japan

Abstract

This article presents the analysis carried out by the Japanese French Laboratory for Informatics (JFLI) and the National Institute of Informatics (NII) to understand what makes a video memorable. To do so, we first propose an analysis of the results obtained by two sequential models applied on visual and textual representations. We then study the manual descriptions and automatic captions in order to identify specificities in the textual representations of videos associated with a high memorability score. We show that they are described by longer and more precise texts (manual *and* automatic) than the videos associated with lower memorability scores, opening the way to research on the correlation between textual vagueness and video memorability.

1. Introduction

Efficiently predicting the memorability of a video could be very useful in the context of multimedia document analysis, e.g. automatic summarization. To progress on this multimodal task, the mediaeval evaluation benchmark organizes for the fifth year the *Predicting Video Memorability* task [1]. Previous work on video memorability (VM) showed that semantic information plays a major role for VM prediction, either through manual descriptions ([2, 3, 4]), manual or automatic captioning ([5], [3], [6]) or semantic concepts [7].

While they all agree that semantic information is important in distinguishing memorable videos from less memorable ones, they struggle to identify the semantic cue that helps make this distinction. [8] analyzed the vocabulary terms corresponding to the most positive and negative correlation coefficients of their caption-based models and showed that the highest negative coefficients are all dominated by terms related to natural scenery; whereas the highest positive coefficients are dominated by terms related to people. Similarly, [4] showed that videos that are the closest related to nature and landscapes generally exhibit worse average memorability ratios, but found that in the *Memory10k* dataset most documents deal with people, and couldn't find a clear distinction between topics that are memorable and topics that are not. Finally, following the idea that people pay more attention to the concepts they are familiar with, [7] used 156 familiar concepts [9] to generate ConceptNet feature vectors to represent the videos. However, the fusion of this semantic information failed to improve the results.

In this paper, we present an analysis of the multimodal characteristics that can be used to predict video memorability. First, we describe the characteristics and models used for the first subtask of the *Prediction Video Memorability* task on the *Memento10k* dataset [5]. We then analyze the results and study the videos' textual representations with respect to their memorability scores. Since manual textual representations may include a translation of whether humans find the videos memorable, we conduct these analyzes on both manual and automatic textual representations.

MediaEval'22, 12-13 January 2023, Bergen, Norway and Online

Table 1
Spearman (Sp.) and Pearson (P.) correlation and Mean Square Error (MSE) values on development and test sets. The parameters column gives the weight values for the linear combination of scores, optimized on the development set.

Runs	Param.	Development			Test		
		Sp.	P.	MSE	Sp.	P.	MSE
Desc.	-	0.599	0.599	0.008	0.571	0.575	0.007
Captions	-	0.511	0.501	0.008	-	-	-
ResNet (RN)	-	0.557	0.550	0.008	-	-	-
DenseNet (DN)	-	0.561	0.550	0.008	-	-	-
RN + DN	0.5	0.587	0.576	0.008	0.546	0.531	0.008
Captions + (RN + DN)	0.3	0.622	0.610	0.007	-	-	-
Desc. + Captions	0.6	0.629	0.620	0.007	0.589	0.594	0.007
Desc. + (RN + DN)	0.5	0.668	0.666	0.007	0.626	0.628	0.006
Desc. + Captions + (RN + DN)	0.7, 0.6	**0.674**	**0.667**	0.007	**0.629**	**0.633**	0.006

2. Video Memorability Prediction

Features To represent the videos, two different kinds of textual features were used. First, the manual descriptions provided by the organizers were represented by a 384 dimensional dense vector using Sentence-BERT [10]. Second, two automatic captions for each video were computed using the ClipCap Image captioning model [11]. To do so, one keyframe was automatically extracted for each video using the Katna tool[1], and the ClipCap model was applied using the model trained either on the MS-COCO dataset [12] or on the Conceptual Caption dataset [13]. Captions from both models were concatenated and, similarly to the manual descriptions, were represented by 384 dimensional dense vectors using Sentence-BERT. In addition to these textual features, two visual features provided by the organizers were also employed: Densenet [14] and Resnet [15].

Models and scores fusion Two models were defined for memorability scores prediction. The first one, used for prediction based on textual representations, is a 3-layer sequential model. In order to avoid overfitting a dropout regularization is added on the visible layer. We use Huber loss as the loss function and a learning rate of 1×10^{-5}. The second model uses video features extracted from deep learning models provided by [1], over 3 frames -first, middle, and last-. We concatenate the features of the 3 video frames and input the resulting feature vector to a 3-layer sequential model. We use dropout on each layer, as loss function the Mean Squared Error (MSE), and a learning rate of 2×10^{-3}. To keep the output values scaled from 0 to 1, we use a sigmoid function in the output layer for both models. Also, for both models, we make use of the Adam optimizer [16]. To combine textual and visual features for memorability scores prediction, the embeddings are first fed to the models and the fusion scores S_F are then computed thanks to $S_F(i) = \alpha * S_M(i) + (1 - \alpha) * S_N(i)$, with $\alpha \in [0, 1]$ and where $S_M(i)$ is the score for the i^{th} video according to modality M and $S_N(i)$ the score for the i^{th} video according to modality N.

3. Results and Analysis

Table 1 summarizes the results obtained on development and test sets of the *Memento10k* dataset.

[1] https://katna.readthedocs.io/en/stable/index.html

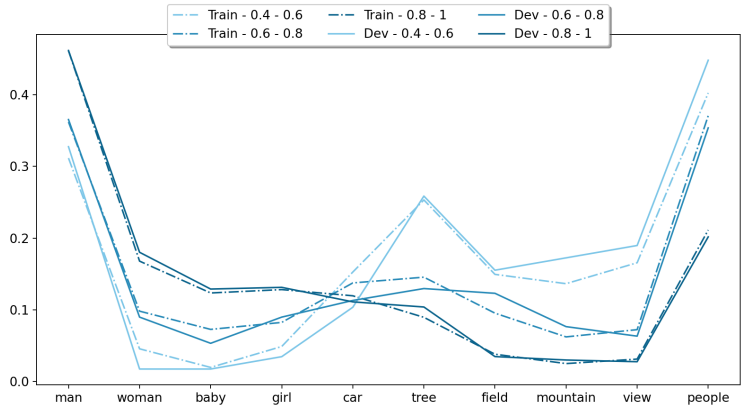


Figure 1: Normalized frequency of words regarding memorability scores.

Concerning the scores obtained from textual representations, although the manual descriptions provide better scores, automatic captions give encouraging results, meaning that automatic captions carry information, automatically extracted from images, related to memorability. Combination of both textual features further improves the results, which suggests that the information carried by manual descriptions and automatic captions is not strictly redundant.

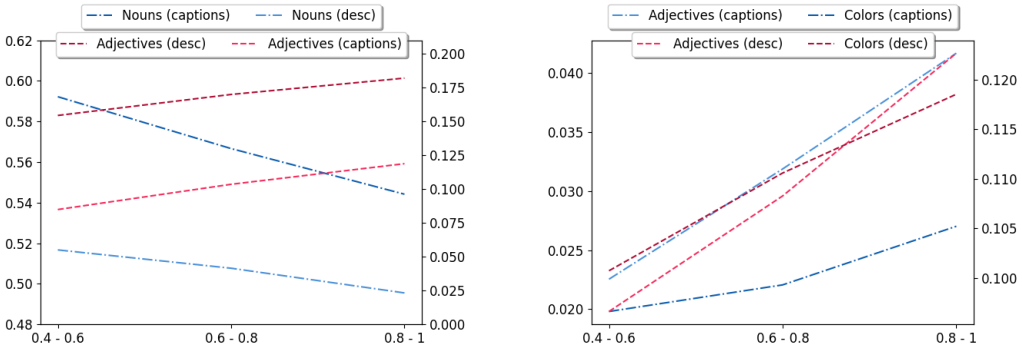
Visual characteristics provide correlation scores almost on par with those obtained from manual descriptions, and the combination of the two visual features (DenseNet and ResNet) increases the results. As before, the combination of visual features and automatic captions further improves the correlation scores. Finally, from this Table, we can see that, although the combination of automatically calculated representations (captions and visual features) provides encouraging results (0.62), the best results are still obtained when manual descriptions are used, suggesting that humans have the ability to introduce information related to memorability into their representations, that is not carried by visual characteristics or automatic captions.

Once again, these experiments show that semantic information is a mainstay in the prediction of video memorability. In order to understand what makes video memorable, we have analyzed the textual information attached with each video. First, we dig into topic analysis to verify if some concepts are more memorable than others, as claimed by [7]. To do so, we identified the 10 words in the manual descriptions that have the most important differences in normalized frequency between videos associated with the higher memorability scores and the lower memorability scores. We then computed the normalized frequency of this 10 words for 3 splits of the development and training set regarding the values of the memorability scores: between 0.4 and 0.6[2], between 0.6 and 0.8 and between 0.8 and 1.

Figure 1 shows the values of these normalized frequencies. Similarly to [4], we couldn't see a clear separation in the topics addressed by the memorable videos and by the other videos. If, as presented in [8], the vocabulary associated with nature is more frequent in videos associated with a lower similarity scores, the *human* topic is addressed by all the videos, whatever their score. However, the *way* this topic is described in the manual descriptions is different between memorable videos (where the vocabulary is more precise *man*, *woman*, *baby*, *girl*) and non memorable ones, where the term used is more vague (*people*).

A possible explanation for this phenomenon is that annotators, remembering the video better, use more precise vocabulary in video with high memorability scores when writing the manual description. Another possibility is that the video is more memorable *because* the elements

[2]There is almost no video with scores lower than 0.4.



(a) Nouns and adjectives

(b) Color and physical adjectives

Figure 2: Proportion of adjectives and nouns in video descriptions w.r.t. memorability scores.

within it are more easily identifiable by viewers. This hypothesis is in line with the notion of familiarity mentioned by [7]. To confirm this hypothesis, we carried out textual analyzes – trough part-of-speech tagging and dependency parsing using the *nltk* tool [17] – of the manual descriptions and the automatic captions, to check whether the sentences used to describe the most memorable videos are more precise than those used for the less memorable ones.

First, we found that the length of both manual descriptions and *automatic* captions were longer for memorable videos than for non-memorable videos, suggesting that either 1) memorable videos have more to describe or 2) elements present in the video are described more extensively for memorable videos. Moreover, we observed, as showed in the left part of Figure 2, that the proportion of nouns used in the texts (descriptions or captions) associated with the videos decreases when the memorability score increases while the proportion of adjectives grows. This observation leads us to believe that textual representations tend to be more accurate for the most memorable videos. To corroborate this assumption, the proportion of 8 frequent adjectives for physical descriptions and 8 common English colors[3] were computed on manual descriptions and automatic captions. The right side of Figure 2 shows the evolution of this proportion and confirms that the proportion of colors and physical adjectives usage improves with the values of memorability scores on both manual descriptions and *automatic* captions. Finally, similar analyzes were conducted on the *VideoMem* dataset and the automatic captions associated with the videos follow the same pattern than those of *Memento10k*: length and adjectives proportion growing and nouns proportion decreasing while the memorability scores rise.

4. Future work

In this paper, we analyze the results of two sequential models based on visual and textual features to understand what makes a video memorable. We show that automatic captions and visual features can provide encouraging results but still miss memorability related information carried by manual descriptions. We also conduct an analysis of the manual and automatic textual representation of videos, showing that more memorable videos are associated with more precise descriptions *even in automatic captions*. In order to confirm this finding, it would be interesting to estimate the vagueness of the descriptions associated with videos, with a tool such as VAGO [18], and use this vagueness prediction as a clue for memorability prediction.

[3] small, large, long, tall, little, big, young and old / white, yellow, green, blue, purple, red, orange and black.

References

[1] 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: Proceedings of the MediaEval Multimedia Benchmark Workshop Working Notes, 2023.

[2] 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.

[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.

[4] R. Kleinlein, C. Luna-Jiménez, D. Arias-Cuadrado, J. Ferreiros, F. Fernández-Martínez, Topic-Oriented Text Features Can Match Visual Deep Models of Video Memorability, Applied Sciences 11 (2021).

[5] 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, 2020.

[6] S. Shekhar, D. Singal, H. Singh, M. Kedia, A. Shetty, Show and recall: Learning what makes videos memorable, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017.

[7] S. Wang, L. Yao, J. Chen, Q. Jin, RUC at MediaEval 2019: Video Memorability Prediction Based on Visual Textual and Concept Related Features, in: Proceedings of the MediaEval 2019 Workshop, 2019.

[8] R. Gupta, K. Motwani, Linear Models for Video Memorability Prediction Using Visual and Semantic Features, in: Proceedings of the MediaEval 2018 Workshop, 2018.

[9] E. W. Dolch, A basic sight vocabulary, The Elementary School Journal 36 (1936).

[10] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019.

[11] R. Mokady, A. Hertz, A. H. Bermano, ClipCap: CLIP Prefix for Image Captioning, arXiv preprint arXiv:2111.09734 (2021).

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, 2014.

[13] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018.

[14] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[16] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: Proceedings of the 3rd International Conference for Learning Representations, 2015.

[17] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.", 2009.

[18] P. Guélorget, B. Icard, G. Gadek, S. Gahbiche, S. Gatepaille, G. Atemezing, P. Égré, Combining vagueness detection with deep learning to identify fake news, in: 2021 IEEE 24th International Conference on Information Fusion, 2021.