SELAB-HCMUS at MediaEval 2023: A cross-domain and subject-centric approach towards the memorability prediction task

Minh-Quang Nguyen[1,3,†], Minh-Huy Trinh[1,3,†], Huy-Giap Bui[1,3,†], Khac-Trieu Vo[1,3,†], Minh-Triet Tran[1,2,3], Thien-Phuc Tran[1,3,*,†] and Hai-Dang Nguyen[2,3]

[1]*Faculty of Information Technology, University of Science - VNU-HCM, Ho Chi Minh City, Vietnam*
[2]*Software Engineering Lab, University of Science - VNU-HCM, Ho Chi Minh City, Vietnam*
[3]*Viet Nam National University, Ho Chi Minh City, Vietnam*

Abstract

The captivating field of Human Memorability extends across various dimensions, each offering distinctive insights. In this paper, we undertake a comprehensive exploration, examining the unique impact of individual domains on the complex fabric of human memorability. Within this inquiry, we introduce two innovative yet straightforward methodologies crafted not only to spark the reader's interest but also to achieve remarkable results in our analytical pursuits. The intricacies of human memorability are delved into, presenting fresh perspectives and inventive approaches to enrich the discourse on this compelling subject.

1. Introduction

Due to the explosive quantities of data from social media and short content platforms in recent years, Media Memorability has attracted more research on the retention of users and their cognitive reactions towards the content. Whereas multiple previous works emphasized the visual domain of the media, we reckon that humans perceive the media in a multimodal fashion with a deeper understanding of the content.

Event-Related Potentials (ERPs) and Event-Related Spectral Perturbations (ERSPs) are crucial tools in neuroscience, offering precise insights into brain processing timing and location. They are pivotal in various applications, such as understanding learning disorders and enabling thought-controlled devices. In educational and clinical settings, ERPs and ERSPs provide invaluable insights into the intricate workings of the mind, advancing our understanding of brain function, one brainwave at a time.

The study of video memorability has diverse applications, including education, content retrieval, summarization, storytelling, and advertising. This motivates the organization of Video Memorability Prediction in MediaEval [1]. Participants in this task will need to predict memorability scores for videos using a dataset containing annotations, visual features, and EEG recordings to gauge their memorability over both short and long periods.

The authors introduce efficient and promising methods spanning diverse domains for predicting video memorability (for Subtasks 1 and 2), emphasizing their lightweight nature and

Proc. of the MediaEval 2023 Workshop, Amsterdam, The Netherlands, 2024.
*Corresponding author.
†These authors contributed equally.
nmquang21@apcs.fitus.edu.vn (M. Nguyen); tmhuy23@apcs.fitus.edu.vn (M. Trinh); bhgiap23@clc.fitus.edu.vn (H. Bui); vktrieu23@apcs.fitus.edu.vn (K. Vo); tmtriet@hcmus.edu.vn (M. Tran); ttphuc21@apcs.fitus.edu.vn (T. Tran); nhdang@selab.hcmus.edu.vn (H. Nguyen)
0009-0008-3520-4624 (M. Nguyen); 0009-0007-5721-3751 (M. Trinh); 0009-0006-7411-8804 (H. Bui); 0009-0001-7049-8532 (K. Vo); 0000-0003-3046-3041 (M. Tran); 0009-0008-0800-3884 (T. Tran); 0000-0003-0888-8908 (H. Nguyen)

straightforward structure. Additionally, a novel technique in Subtask 2 is presented and thoroughly assessed for its robustness.

2. Related Work

Various factors, especially media features, influence memorability. Models using visual or texture features have been proposed, and combining diverse models often improves performance. For example, Guinaudeau *et al.* [2] merged visual models (ResNet and DenseNet) with a texture model (Sentence-BERT), achieving commendable results. Insights from this study [3] suggest that highly memorable videos tend to have saturated colors, people and faces, manipulable objects, and man-made environments, while less memorable videos often feature darker settings, clutter, or inanimate scenes.

Contemporary computer vision faces challenges in generality and usability, requiring additional labeled data. An alternative involves learning directly from raw text associated with images. Radford *et al.* [4] show caption prediction's efficacy, learning image representations from a dataset of 400 million image-text pairs. This enables natural language use in downstream tasks. Alec *et al.* [5] verify CLIP's [6] transferability across 30+ datasets, showcasing its performance without specific data training.

3. Experimental setup and results

Motivations

For Subtask 1, we are convinced that an ordinary subject would perceive the video as a multimodal object – the video is embedded with content in various forms, including visual, semantic, and temporal content. The user may require multiple content domains to assemble an 'impression' of the video to memorize it thoroughly. Therefore, we utilize the nature of the videos to extract their corresponding features and use them to predict the memorability of each video.

For Subtask 2, we recognize that memorability is personalized and varies among individuals based on their preferences for different aspects of video content. Instead of focusing on the content, we aim to understand and identify the most memorable moments for each person in each corresponding video, emphasizing a user-centric approach.

3.1. Subtask 1: Using cross-domain features to estimate memorability score

Method	Spearman	Pearson	MSE
Resnet + EfficientNet + CLIP Text	0.313	0.326	**0.006**
CLIP (Text + Vision)	**0.445**	**0.452**	0.008
NGram	0.336	0.350	0.008

Table 1. Final results from the proposed method in Subtask 1.

N-gram, ResNet, and EfficientNet

Initially, we used a simple yet efficient statistical method to analyze words from captions in the Memento10k dataset. We followed the N-gram approach, aligning with previous research. This method served as our starting point. Our results show how words are expressed plays a crucial role in making video content more memorable. Significantly, our approach outperformed the use of isolated features from ResNet [7] and EfficientNet [8].

In the provided Memento10k dataset, each video is paired with five lowercase, stop-word-removed, punctuation-free, and lemmatized captions. The text underwent N-gram processing (unigrams, bigrams, trigrams), with memorability scores assigned based on the methodology in [9]. This method involves computing mean memorability scores, scaled by N-gram frequency. The predicted memorability score for each video in the testing set is the mean score of its

N-gram. The final predicted memorability score is derived from a weighted sum of N-gram components: 80% from unigrams, 12% from bigrams, and 8% from trigrams.

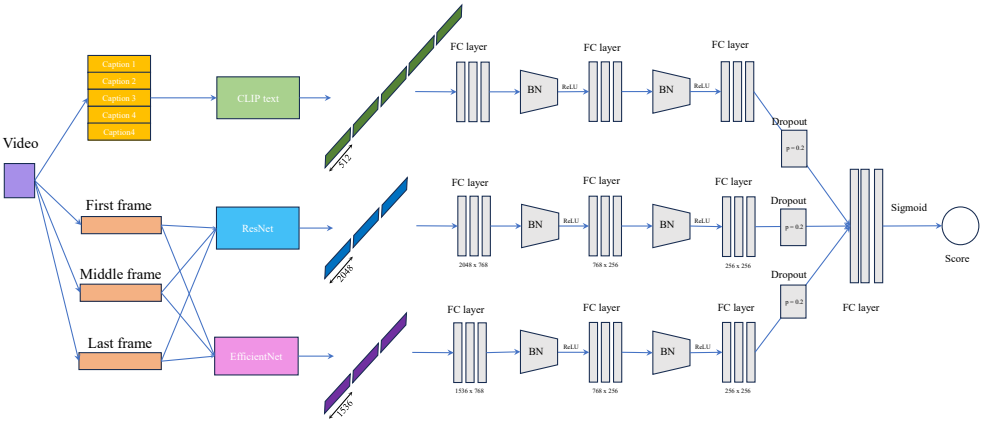

Figure 1: Structure of the proposed model. Only the first, middle, and last frames are taken to extract visual-based features. The text-based features are extracted from the image captioning method. BN is the abbreviation of batch normalization.

Comparing CLIP and N-gram, CLIP outperformed as a text feature extractor, leading us to select CLIP for the next experiment. For CLIP, five captions yield five feature vectors concatenated into a single 2560-length vector. Visual-based features are extracted from three frames (first, middle, last), with ResNet and EfficientNet generating vectors of lengths 2048 and 1536, respectively. Integration involves concatenating the three vectors from each model.

Fusing CLIP text and image encoding

Previous results using CLIP for textual feature extraction have proven effective for the task. This motivates us to use CLIP for visual feature extraction. Because CLIP is trained on text-image pairs, it can learn the nuanced connections between them. And because both the text and image are encoded into the same representation space, we can use a much simpler architecture for fusing them, which might reduce over-fitting and generalize better to the test dataset.

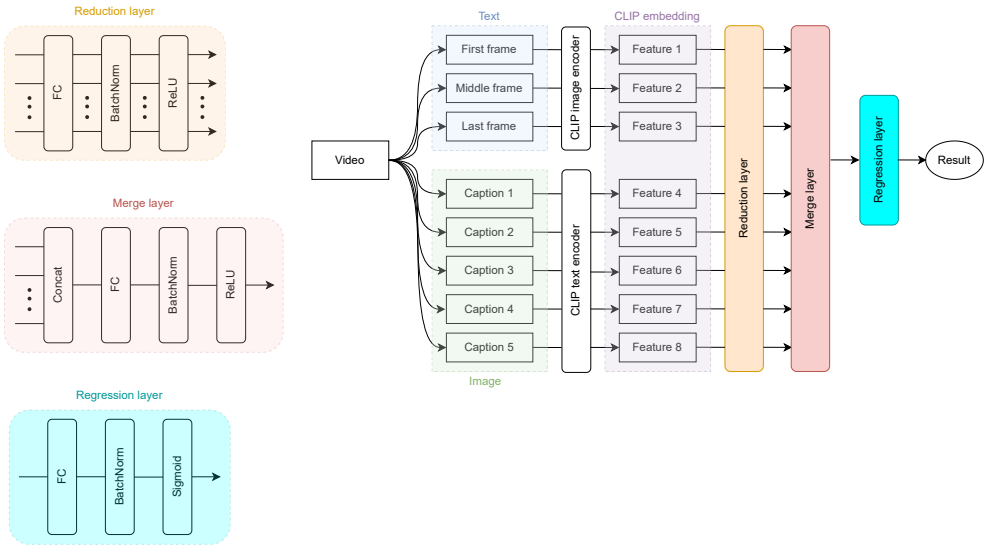

Figure 2: Overall structure of our clip text-image fusion model. Because the text and image embedding are in the same representation space, we can use a single model to process them.

Instead of training two separate networks for text and visual features, we trained a single network on the combined features. We reduced dimensionality on the features before concatenating them, as shown in Figure 2. Since CLIP features are noticeably smaller than ResNet's and EfficientNet's features, we only used a single linear layer to reduce the dimension of each feature from 512 to 128, which is shared across both textual and visual features of CLIP. This drastically reduces the number of parameters of the model. We suspect the model might have difficulty learning but can generalize better.

Similarly to the previous approach, the first, last, and middle frames are extracted and encoded into CLIP's embedding, along with five captions. Batch normalization is applied to help the model converge faster.

3.2. Subtask 2: Regression on neural signals from ERP and ERSP

Method	AOC
ERP one sensor (FC6) signals	0.536
ERP all 28 sensors signals	0.540
ERP all 28 sensors signals (with subject one-hot encoding)	**0.657**

Table 2. Final results from the proposed method in Subtask 2 (different regression models on ERP signals). ERP one sensor (FC6) signals are the result of our regression model taking 30 input signals from the FC6 sensor. ERP all 28 sensors' signals are the result of our regression model taking 28×30 input signals from the 28 sensors. ERP, all 28 sensors' signals (with subject one-hot encoding) are the result of our regression model taking 28 x 30 + 12 input signals from the 28 sensors and subject one-hot encoding.

We introduce two regression methodologies applied to the provided dataset within this specific subtask. The initial approach involves employing a straightforward linear regression step to assimilate the features inherent in ERPs (28 x 30 input features to 1 output tensor) and ERSPs (28 x 30 x 30 input features to 1 output tensor) on the collected data from different sensors. For ERSPs, an additional normalization step is undertaken before the commencement of the training process. The results of the method are shown in the table 2.

In the second method, we adopt a personalized approach, training models for each individual due to unique neural signals observed in the dataset. We believe that individual cognitive responses vary when interacting with video content. Using one-hot encoding, we create individualized vectors for each subject, adding 12 more features (for 12 test subjects) to the original 28×30 ERP features. The results in Table 2 highlight the method's high efficiency.

4. Discussion and Outlook

The main findings from Subtask 1 highlight that text-based features are more effective than visual-based features in determining a video's impact on human-memorability. This emphasizes the importance of providing diverse captions for each video to enhance predictability, as well as the need to consider multiple factors from various perspectives. It's crucial to acknowledge that the disparities in backgrounds and content between the training set and test set raise questions about the robustness of the findings.

In Subtask 2, distinct patterns in ERP and ERSP diagrams suggest that video memorization is encoded in neural signals. These unique patterns call for individualized investigation to achieve optimal results, highlighting the importance of conducting a detailed analysis of the neural responses of each participant to improve predictive performance.

Acknowledgments

This research was funded by Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19.

References

[1] M. G. Constantin, C.-H. Demarty, C. Fosco, A. G. Seco de Herrera, S. Halder, G. Healy, B. Ionescu, A. Matran-Fernandez, R. S. Kiziltepe, A. F. Smeaton, L. Sweeney, Overview of The MediaEval 2023 Predicting Video Memorability Task, in: Proceedings of the MediaEval 2023 Workshop, Amsterdam, The Netherlands, 2024.

[2] C. Guinaudeau, A. Girbau Xalabarder, Textual Analysis for Video Memorability Prediction, in: the 13th MediaEval Multimedia Benchmark Workshop, Bergen, Norway, 2023. URL: https://universite-paris-saclay.hal.science/hal-04091024.

[3] A. Newman, C. Fosco, V. Casser, A. Lee, B. A. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, CoRR abs/2009.02568 (2020). URL: https://arxiv.org/abs/2009.02568. arXiv:2009.02568.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR139, 2021.

[5] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller, Multi-view convolutional neural networks for 3d shape recognition, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA, 2015, pp. 945–953. URL: https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.114. doi:10.1109/ICCV.2015.114.

[6] R. A. Kim, J. W. Hallacy, C. Ramesh, A. G. G. Agarwal, S. Sastry, G. Askell, A. Mishkin, P. Clark, J. Krueger, G. I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR139, 2021.

[7] 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, pp. 770–778.

[8] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 6105–6114.

[9] M. M. A. Usmani, S. Zahid, M. A. Tahir, Quest for insight: Predicting memorability based on frequency of n-grams (2022).