Understanding Media Memorability From Event-Related Potential Records And Visual Semantics

Ricardo Kleinlein[1,*], Enrique R. Sebastián[2] and Fernando Fernández-Martínez[1]

[1]Grupo de Tecnología del Habla y Aprendizaje Automático, Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid 28040 Madrid, Spain

[2]Instituto Cajal, CSIC, Madrid, Spain

Abstract

The memorability of a video has been defined in the literature as an intrinsic property of its visual features, expressed as the proportion of an audience that successfully remembers having watched that video on a subsequent viewing. Hence our brains must cope not only with information about pixel statistics and scene semantics, but also to encode whether it is worth keeping information about them in memory for future retrieval. These are the hypothesis behind the 5^{th} edition of the Predicting Media Memorability challenge, which we tackle from a two-fold perspective: first we pursue a semantics-based approach, using both pre-trained and fine-tuned visual and textual Transformers; on the other hand, we process Event-Related Potential (ERP) data according to two feature extraction methods to obtain a representation compatible with cross-subject predictive models of media memorability, namely: (1) to extract sample-level functionals and feed them as input features to a random forest classifier, and (2) to compute coherence maps between sensor recordings at four frequency bands, training a shallow neural classifier from them. Ultimately, we seek to further comprehend why, whereas some of our visual models display performances that rival that of the current state-of-the-art predictive systems, ERP-based approaches pose a far more complex challenge.

1. Introduction

A detailed scientific modelling of the factors by which some visual memories remain attached to us for a long time while others fade shortly after has eluded a mathematical formulation for decades. Recent studies point to the possibility that all the visual information that reaches our eyes carry along a measure that would account for its likelihood to be remembered in subsequent viewings, i.e., its intrinsic memorability [1, 2, 3]. With the rise of social media, an automatic system able to classify a video on these terms is of the utmost interest, both from a commercial and a scientific perspective. In this paper, we report on our experience during the 5^{th} edition of the Predicting Media Memorability Challenge [4]. The availability of Electroencephalography (EEG) data enables us not only to study the link between visual features and memorability but also to explore possible mechanisms by which human brain stores that information, building predictive models of media memorability accordingly.

2. Related Work

Although studies on the issue date back to R.N. Shepard (1967) and Standing (1973) [5, 6], it has not been until the work of Isola et al.[3] that researchers began to think of media memorability as

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online
*Corresponding author.
ricardo.kleinlein@upm.es (R. Kleinlein); enrique.rodsebastian@gmail.com (E. R. Sebastián); fernando.fernandezm@upm.es (F. Fernández-Martínez)

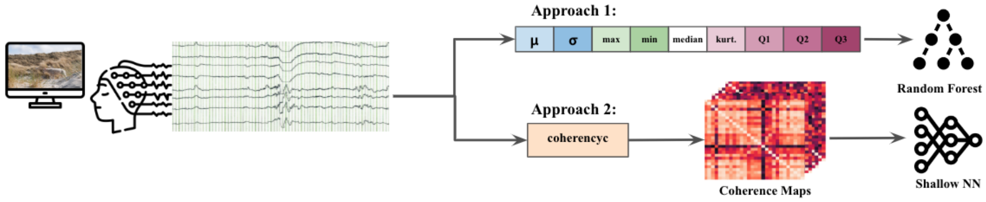

Figure 1: In order to predict memorability from EEG data we developed two different approaches, (1) based on extracting statistical functionals of the record for each subject and video pair, and (2) computing coherence maps between sensors at 4 frequency bands during the first second of exposition of a subject to a given video.

Run#	Model description	MSE		PCC		SRCC	
		Val.*	Test	Val.*	Test	Val.*	Test
1	VisualCLIP (adapted)	0.009	0.009	0.430	0.401	0.427	0.395
2	TextCLIP (adapted)	0.007	0.008	0.597	0.557	0.6	0.556
3	Mean late-fusion (1) & (2)	0.007	0.007	0.601	0.595	0.599	0.592
4	Pretrained VisualCLIP	0.008	0.006	0.547	0.647	0.549	0.64
5	Mean late-fusion (2) & (4)	0.007	0.006	0.628	0.664	0.629	0.658

Table 1
Prediction rates both at validation and testing time for the models submitted to the subtask 1. MSE: Mean Squared Error; PCC: Pearson's Correlation Coefficient; SPCC: Spearman's Rank Correlation Coefficient. *Validation is carried out using a 5-fold cross-validation scheme over both train and dev data partitions.

a deterministic function of fundamental visual properties (such as image colour or its brightness) and/or the high-level semantic features of a multimedia clip [7, 8, 9]. We use Transfomers, highly successful in an array of different tasks [10, 11], either as visual and textual feature extractors or fine-tuning them as predictive models of media memorability (Section 3.1).

EEG data open the path for further understanding of the mechanisms underpinning the encoding of media memorability by the human brain. Much of the difficulty lies in the entanglement between different brain regions operating simultaneously along the process [12, 13, 14]. However, coherence between different brain areas (a measure of the strength of the coupling between the signal recorded by two sensors at specific frequency bands) has been found to relate to memory impairment in Alzheimer's disease [15, 16] and other dementia-related health disorders [17]. Furthermore, techniques based on similar functional connectivity between EEG channels has been demonstrated to correlate with long-term semantic memory [18]. Therefore in Section 3.2 we propose two alternative preprocessing methods for ERP data, both outlined in Figure 1.

3. Experimental setup and results

A detailed description of both the requirements and the data resources available for each subtask can be consulted at [4]. During the experimental phase we placed a special emphasis not only on accurately predicting memorability but also on explaining the decisions made by our models.

System description	AUC	
	Val.*	Test
Statistical Functionals	0.529	0.501
Delta channel only	0.490	0.500
Beta channel only	0.514	0.509
Late-fusion of all channels (Median)	0.534	0.509
Late-fusion of all channels (Max.)	0.529	0.509

Table 2
Prediction rates for validation and test sets for the model predictions for subtask 3. AUC: Area Under Curve score. *Validation rates are computed using a 5-fold cross-validation strategy with a Leave-One-Subject-Out (LOSO) scheme.

3.1. Subtask 1: Predicting memorability rates from visual features

Our fundamental hypothesis, supported by previous experiences [9, 19], is that video-level semantic features are robust indicators of video memorability, given the strong correlation found between certain topics and the average memorability rates of videos depicting them. Here we elaborate on this idea: either keeping a frame-wise (extracted at 1FPS) pre-trained CLIP Visual Transformer (ViT) as a feature extractor upon which a linear regressor is trained on the task of media memorability (run #4), or fine-tuning a ViT and its textual counterpart on Memento10K data [8] (run #1, run #2). We also investigate the degree to which both modalities can help each other in making a prediction, and hence the output of the run #3 is the average between the prediction made by run #1 and run #2, while run #5 is the analogous for run #2 and run #4. In all cases, fine-tuning is performed optimizing the mean square loss between predicted labels and the ground-truth memorability scores for 10 epochs. Prediction rates at both validation and testing are shown in Table 1.

3.2. Subtask 3: Memorability classification from ERP

We propose two different processing pipelines, illustrated in Figure 1, aimed both at computing useful numerical representations for the final task of predicting whether a video will be remembered, irrespective of the subject data comes from. This is an inherently complex scenario, since two subjects can respond very differently to the same video. Validation and testing classification Area Under the Curve (AUC) rates are shown in Table 2. Our first approach consists on concatenating statistical functionals - mean value, standard deviation, median, maximum and minimum values, kurtosis index and the first three quartiles of a sample - to describe each trial (subject-video pair). As predictive algorithm, we train a random forest model. For our second approach, for each subject and video we compute the coherency between each ERP channel pairwise. We used the function “coherencyc” from Matlab’s® third party toolbox Chronux[1] to compute the mean coherency value for different power bands: delta (0.5-4Hz), theta (4-8Hz), alpha (8-14Hz) and beta (14-30Hz). This yields a 28x28x4 matrix of coherencies between channels in specific spectral bands. These values, once arranged as a single vector embedding, conform to the input features for a shallow neural network whose hidden layer has 256 neurons, with a ReLU activation function [20] and Adam optimizer [21] and adaptive learning rate.

[1]https://doi.org/10.1016/j.jneumeth.2010.06.020

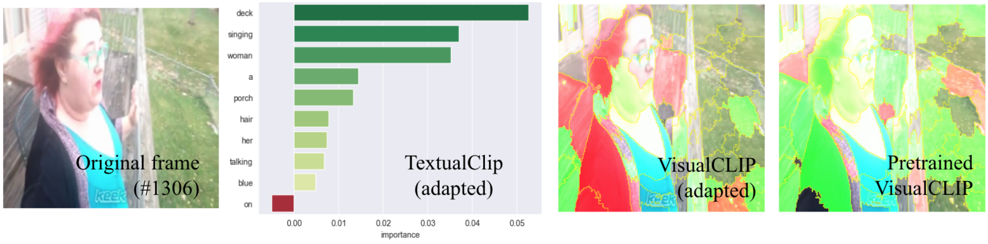

Figure 2: From left to right: Original frame, and LIME explanations for predictions made by runs (2), (1) and (4) from Table 1, respectively. Green indicates areas that contribute positively to greater memorability scores while red regions denote the opposite.

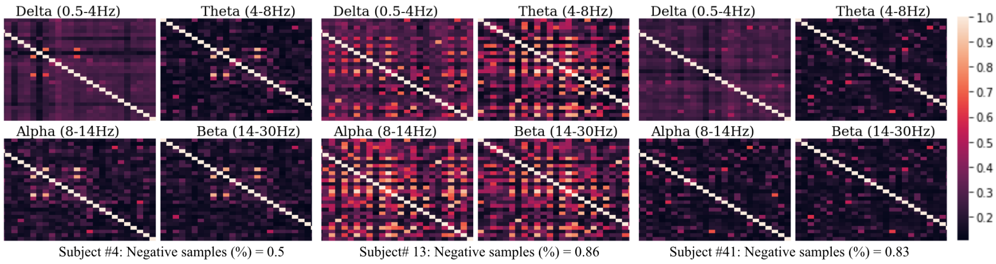

Figure 3: Average coherence maps at each power band for 3 subjects in the training set. Each point in these matrices represents the pair-wise average coherence between two sensors at a given frequency band, coloured according to the strength of their coupling. We found significant differences amidst these features between participants, even when their success rates are similar.

4. Discussion and outlook

Interestingly enough, a fine-tuned ViT performs worse than a simpler linear regressor trained from the features obtained by a pretrained version of the full model, even though the same does not seem to happen in the case of text. Computing explanations using a custom version of LIME [22], a popular post-hoc local surrogate method [23], we notice that while the text-based model bases its predictions on concepts that we know correlate well with memorability [9], our fine-tuned ViT (run #1) might be generalising worse due to overfitting (Fig. 2). As illustrated in Figure 3, it is hard to notice a clear pattern of neural activity amidst the subjects when using ERP data to predict memorability. Different people show high memorability rates (subjects 4 and 9), yet the rest fail about 80% of the time, hence leaving an extremely unbalanced dataset that adds up to the overall complexity of the task. As a future line of research, we believe it would be particularly interesting to explore multimodal EEG-visual-textual models, in order to further develop scientific knowledge on what information from a video clip is actually leaving a lasting footprint on our brains.

Acknowledgements

Our work has been supported by the Spanish Ministry of Science and Innovation: projects GOMINOLA (PID2020-118112RB-C21, PID2020-118112RB-C22, funded by MCIN/AEI/10.13039/501100011033), AMIC-PoC (PDC2021-120846-C42, funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR"), and the Spanish Ministry of Education (FPI grant PRE2018-083225).

References

[1] P. Isola, D. Parikh, A. Torralba, A. Oliva, Understanding the intrinsic memorability of images, in: Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11), Curran Associates Inc., Red Hook, NY, USA, 2011, p. 2429–2437.

[2] P. Isola, J. Xiao, A. Torralba, A. Oliva, What makes an image memorable?, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 145–152.

[3] P. Isola, J. Xiao, D. Parikh, A. Torralba, A. Oliva, What makes a photograph memorable?, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014) 1469–1482.

[4] 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, 2023.

[5] R. N. Shepard, Recognition memory for words, sentences, and pictures, Journal of Verbal Learning and Verbal Behavior 6 (1967) 156–163.

[6] L. Standing, Learning 10000 pictures, Quarterly Journal of Experimental Psychology 25 (1973) 207–222.

[7] Z. Bylinskii, L. Goetschalckx, A. Newman, A. Oliva, Memorability: An image-computable measure of information utility, 2021. arXiv:2104.00805.

[8] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, 2020. arXiv:2009.02568.

[9] 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).

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[11] 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: International Conference on Machine Learning (ICML), 2021.

[12] J. Han, C. Chen, L. Shao, X. Hu, J. Han, T. Liu, Learning computational models of video memorability from fmri brain imaging, IEEE Transactions on Cybernetics 45 (2015) 1692–1703.

[13] R. F. Thompson, J. J. Kim, Memory systems in the brain and localization of a memory, Proceedings of the National Academy of Sciences 93 (1996) 13438–13444.

[14] A. Jaegle, V. Mehrpour, Y. Mohsenzadeh, T. Meyer, A. Oliva, N. Rust, Population response magnitude variation in inferotemporal cortex predicts image memorability, eLife 8 (2019).

[15] G. Adler, S. Brassen, A. Jajcevic, Eeg coherence in alzheimer's dementia, Journal of Neural Transmission 110 (2003) 1051 – 1058.

[16] M. J. Hogan, G. Swanwick, J. Kaiser, M. Rowan, B. Lawlor, Memory-related eeg power and coherence reductions in mild alzheimer's disease, International Journal of Psychophysiology 49 (2003) 147–163.

[17] D. Laptinskaya, P. Fissler, O. C. Küster, J. Wischniowski, F. Thurm, T. Elbert, C. A. F. von Arnim, I.-T. Kolassa, Global eeg coherence as a marker for cognition in older adults at risk for dementia, Psychophysiology 57 (2020).

[18] S. Hanouneh, H. U. Amin, N. M. Saad, A. S. Malik, Eeg power and functional connectivity correlates with semantic long-term memory retrieval, IEEE Access 6 (2018) 8695–8703.

[19] R. Kleinlein, C. Luna-Jiménez, F. Fernández-Martínez, Thau-upm at mediaeval 2021: From video semantics to memorability using pretrained transformers, in: MediaEval'21 Online, 2021.

[20] A. F. Agarap, Deep learning using rectified linear units (relu), ArXiv abs/1803.08375 (2018).

[21] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980 (2014).

[22] R. Kleinlein, A. Hepburn, R. Santos-Rodríguez, F. Fernández-Martínez, Sampling based on natural image statistics improve local surrogate explainers, in: The 33rd British Machine Vision Conference, 2022.

[23] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM International Conference on Knowledge Discovery and

Data Mining (ICKDD), 2016.