THAU-UPM at MediaEval 2021: From Video Semantics To Memorability Using Pretrained Transformers

Ricardo Kleinlein, Cristina Luna-Jiménez, Fernando Fernández-Martínez
Grupo de Tecnología del Habla y Aprendizaje Automático (THAU Group), Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, Avda. Complutense 30, 28040 Madrid, Spain
{ricardo.kleinlein,cristina.lunaj,fernando.fernandezm}@upm.es

ABSTRACT

This paper reports on our experience after participating at the MediaEval 2021: Predicting Media Memorability challenge. The memorability of a video is defined as the proportion of people that successfully remembered having watched a video on a second viewing during a memory game. Given this setup, teams were requested to provide systems able to predict the degree of memorability for individual videos from two different datasets: TRECVid and Memento10k. Our proposal builds upon previous work in which we find that non-adapted features extracted from Transformer architectures can be closely tied to semantic differences between samples, which in turn point to the overall memorability degree within different semantic units, or topics. We feed these precomputed features to linear regressors, showing that even without adapting the input representation competitive prediction rates can be achieved.

1 INTRODUCTION

Scientific modelling of cognitive variables of human perception of multimedia productions has eluded a mathematical formulation until the last decades, leaving it as a discipline within psychology[1]. Although usually perceived to be largely dependent on the subjective appraisals experienced by an individual, the analysis of group-level data sets points to the existence of patterns most humans attach at least to some degree when faced before multimedia content. One such instance is the problem of media memorability.

The MediaEval workshop, and in particular the *Predicting Media Memorability* challenge, provides now for the 4^{th} consecutive edition with reliable data that researchers can use to further understand media memorability. A detailed description of the challenge, as well as the data sources used in this task can be found in [9].

2 RELATED WORK

From the seminal work of Isola et al.[7], researchers have investigated whether the prediction of media memorability depends primarily on visual descriptors such as image colour; brightness; and hue, as opposed to other approaches, which suggested that high-level, data-driven representations (e.g., image composition, scene recognition, and image classification features) are best suited to the task.

Our hypothesis, supported by studies from both neuroscience and psychology, is that there are certain topics (particularly those related to people) that are inherently better remembered, than other themes such as nature, war-like scenes and open spaces [8, 11]. Moreover, it seems that a major principle in creating new memories comes from the fact that the brain deals with scene and object representations at the same level of abstraction [12]. This highlights the need for global descriptors of the media content if the goal is to predict its likelihood to be remembered.

In recent times, the Transformer family of models has been proposed as an alternative to other neural architectures, with promising results until now [4, 5, 14, 17]. This is largely due to the inner representation of input features these models are able to compute, which tend to show a high degree of robustness to previously unseen data. Because of their success, we use them as either text of image encoders, in order to transform text descriptions or video frames into meaningful, semantically-rich vector embeddings.

3 APPROACH

In a previous study[10], we found that a Transformer trained on a sentence similarity task yielded features closely aligned to the automatic detection of topics within the set of available video descriptions. We also observed that some semantic units like human, baby, girl, or man showed a higher average memorability than other topics closer to nature views, open spaces or war-like contents. One of the pillars of our analysis relied on the fact that the model used to encode sentences into embeddings was not fine-tuned or adapted to our task. Therefore, here we extend our methodology to other pretrained Transformer architectures.[1]

Here we explore a wider range of models, covering systems able to deal not only with text inputs, but also with visual information. The main distinction between different runs (shown in Table 1) lies in the model combinations used to encode the textual and visual features. These embeddings are then fed as input to linear regression models that constitute the only part of the pipeline specifically trained on the task of predicting media memorability. Every video is represented by a single embedding, computed as the mean value of that video's individual sentence or frame-level embeddings.

3.1 Text Transformers

Language is a natural way to describe to others what we see, and hence through it, we can encapsulate the semantics of a video in a succinct and readable way. We choose three different architectures, SBERT [16]; GPT-2[15]; and CLIP[14], each covering a different aspect of language modelling. SBERT is a variation of the popular

[1] All the models used here can be downloaded from https://huggingface.co.

MediaEval'21, December 13-15 2021, Online

Run	Dataset	SBERT	GPT-2	CLIP (text)	CLIP (visual)	ViT	BEIT	PCA dims.	Method
1	TRECVid	x	x	x				64	Bayes
	Memento10k							256	LR
2	TRECVid				x			32	Bayes
	Memento10k							512	
3	Both				x	x	x	2048	Bayes
4	Both	x	x	x	x	x	x	4096	LR
5	Both	x	x	x	x	x	x	4096	Bayes

Table 1: Overview of the runs submitted. Different runs solve the task using different sets of precomputed features. Within every dataset, the same solution is proposed despite labels be raw or normalised.

BERT language model[4]; the embeddings computed using SBERT are targeted at telling apart pairs of sentences with similar or dissimilar meaning, which is beneficial when looking for topics in texts. We use the *all-mpnet-base-v2* implementation. GPT-2 set a remarkable milestone in the path of automatic text generation[15], since it is able to synthesize texts coherent both in structure, use of language and grammar. Features extracted using this model build a general-purpose language representation. CLIP was announced as a model able to combine information from both visual and textual sources in order to perform image classification and image synthesis simultaneously [14]. Its text-encoder is considered separately from the rest of the model to encode sentences describing videos with emphasis on the content of the video.

3.2 Visual Transformers

Although text descriptions can convey most of the semantic units within a video clip, many aspects of the clip itself are missed. For instance, a text such as "two people walking" can evoke a unending amount of different images. However, extracting the semantics from images is a process far more complex to interpret and analyze. Fortunately, Transformers have also been applied to computer vision tasks. Hence, we can proceed analogously and elaborate on the embedding representations extracted from video frames (extracted at 1 FPS) using pretrained models. We use the visual branch of a CLIP model, plus two additional systems designed under the same guiding principles of the original BERT. In particular, we use BEiT [2] and ViT [18] as additional visual encoders. Both were trained on image classification over the ImageNet-21k dataset [3], at a resolution of 224x224 pixels, though following different approaches.

3.3 Predictive models

We limit our setup to simple linear predictors: linear regression and Naïve Bayes regression. Both are simple enough conceptually, yet different in their inner working[2], allowing us to concentrate our efforts on the predictive power of the non-adapted input features. Also, *Principal Component Analysis* (PCA) is used to project the input vectors on spaces with lower dimensionality [6]. Each run was submitted according to the learning method and PCA dimensions that performed the best over the development set of data on each dataset.

[2]We used the default implementations from *sklearn* library [13].

Dataset	Labels	run 1	run 2	run 3	run 4	run 5
TRECVid	short-raw	0.204	0.265	**0.291**	0.205	0.198
	short-norm	0.193	0.272	**0.293**	0.193	0.198
	long	**0.125**	0.102	0.077	0.009	0.01
Memento10k	raw	0.596	0.601	**0.656**	0.651	0.651
	norm	0.598	0.606	**0.657**	0.652	0.651

Table 2: Spearman's rank correlation coefficient of our proposed models when evaluated over the official test set. In bold, the run that performed the best on each set of labels.

4 RESULTS AND OUTLINE

Table 2 shows Spearman's rank correlation coefficient values over the test set of data for each run submitted. First, it is noticeable that the combination of all visual embeddings outperforms any other approach, except in the long-term set of labels. In that case, it points to the possibility that text-based representations may encode better the semantics needed to predict long-term media memorability. Another interesting observation can be made about the relative difference in performance shown by the same approaches depending on which dataset is considered. In fact, the worsening in the predictions made over TRECVid data is likely to be related to its smaller size, as well as the fact that we have noticed that most videos within TRECVid fall in a narrow range of the topics detected in Memento10k.

ACKNOWLEDGMENTS

The work leading to these results was supported by the Spanish Ministry of Scienceand Innovation through the GOMINOLA (PID2020-118112RB-C21 and PID2020-118112RB-C22,funded by MCIN/AEI/10.13039/501100011033), CAVIAR (TEC2017-84593-C2-1-R, funded by MCIN/AEI/10.13039/501100011033/FEDER "Una manera de hacer Europa"), and AMIC (TIN2017-85854-C4-4-R, funded by MCIN/AEI/10.13039/501100011033/FEDER "Una manera de hacer Europa") projects. This research also received funding from the European Union's Horizon2020 research and innovation program under grant agreement N°823907 (http://menhir-project.eu, accessedon17 November 2021). Furthermore, R.K.'s research was supported by the Spanish Ministry ofEducation (FPI grant PRE2018-083225).

REFERENCES

[1] Rudolf Arnheim. 1954. *Art and visual perception: a psychology of the creative eye.* University of California Press.

[2] Hangbo Bao, Li Dong, and Furu Wei. 2021. BEiT: BERT Pre-Training of Image Transformers. (2021). arXiv:2106.08254 https://arxiv.org/abs/2106.08254

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv* abs/2010.11929 (2021).

[6] Karl Pearson F.R.S. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, 11 (1901), 559–572. https://doi.org/10.1080/14786440109462720

[7] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1469–1482. https://doi.org/10.1109/TPAMI.2013.200

[8] Andrew Jaegle, Vahid Mehrpour, Yalda Mohsenzadeh, Travis Meyer, Aude Oliva, and Nicole Rust. 2019. Population response magnitude variation in inferotemporal cortex predicts image memorability. *eLife* 8 (aug 2019), e47596. https://doi.org/10.7554/eLife.47596

[9] Rukiye Savran Kiziltepe, Mihai Gabriel Constantin, Claire-Hélène Demarty, Graham Healy, Camilo Fosco, Alba García Seco de Herrera, Sebastian Halder, Bogdan Ionescu, Ana Matran-Fernandez, Alan F. Smeaton, and Lorin Sweeney. 2021. Overview of The MediaEval 2021 Predicting Media Memorability Task. In *Working Notes Proceedings of the MediaEval 2021 Workshop*.

[10] Ricardo Kleinlein, Cristina Luna-Jiménez, David Arias-Cuadrado, Javier Ferreiros, and Fernando Fernández-Martínez. 2021. Topic-Oriented Text Features Can Match Visual Deep Models of Video Memorability. *Applied Sciences* 11, 16 (2021). https://doi.org/10.3390/app11167406

[11] T. Konkle, T. F. Brady, G.A. Alvarez, and A. Oliva. 2010. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General* 139, 3 (2010), 558–578.

[12] Talia Konkle, Timothy F. Brady, George A. Alvarez, and Aude Oliva. 2010. Scene Memory Is More Detailed Than You Think: The Role of Categories in Visual Long-Term Memory. *Psychological Science* 21, 11 (2010), 1551–1556. https://doi.org/10.1177/0956797610385359 arXiv:https://doi.org/10.1177/0956797610385359 PMID: 20921574.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

[15] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[16] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[18] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual Transformers: Token-based Image Representation and Processing for Computer Vision. (2020). arXiv:cs.CV/2006.03677