POLITECNICO DI TORINO
Repository ISTITUZIONALE

Predicting Video Memorability Using a Model Pretrained with Natural Language Supervision

Original
Predicting Video Memorability Using a Model Pretrained with Natural Language Supervision / Agarla, Mirko; Celona, Luigi; Schettini, Raimondo. - 3583:(2023). (Intervento presentato al convegno Working Notes Proceedings of the MediaEval 2022 Workshop tenutosi a Bergen (NOR) nel January 13–15, 2023).

Availability:
This version is available at: 11583/2982306 since: 2023-09-19T12:21:28Z

Publisher:
CEUR Workshop Proceedings

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

24 November 2024

Predicting Video Memorability Using a Model Pretrained with Natural Language Supervision

Mirko Agarla[1,*], Luigi Celona[1] and Raimondo Schettini[1]

[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, ITALY

Abstract

Video memorability prediction aims to quantify how much a given video content will be remembered over time. The main attributes affecting the prediction of memorability are not yet fully understood and many of the methods in the literature are based on features extracted from content recognition models. In this paper we demonstrate that features extracted from a model trained with natural language supervision are effective for estimating video memorability. The proposed method exploits a Vision Transformer pretrained using Contrastive Language-Image Pretraining (CLIP) for encoding video frames. A temporal attention mechanism is then used to select and aggregate relevant frame representations into a video-level feature vector. Finally, a multi-layer perceptron maps the video-level features into a score. We test several types of encoding and temporal aggregation modules and submit our best solution to the MediaEval 2022 Predicting Media Memorability task. We achieve a correlation of 0.707 in subtask 1 (i.e. the Memento10k dataset). In task 2 we obtain a Pearson correlation of 0.487 by training on Memento10k and testing on videoMem and of 0.529 by training on videoMem and testing on Memento10k.

1. Introduction

The exponential growth of images and videos shared on social media platforms require new ways to organize and retrieve digital contents. Like other video metrics of importance, such as quality [1], aesthetics [2, 3] or interestingness [4, 5], memorability can be regarded as a useful aspect to help make a choice between competing videos. The *Predicting Media Memorability Challenge*, hosted within the MediaEval workshop, focuses on the estimation of video memorability. In its fourth edition, the task is the same as in previous years, but it involves videos depicting in-the-wild scenes collected from social media. More information can be found in the challenge description document [6]. Most image and video memorability methods based on deep learning are usually built on top of models pre-trained on ImageNet [7, 8, 9, 10]. However, we argue that pre-trained models for semantic content classification may not model important factors for estimating memorability such as aesthetics and interestingness. Conversely to semantic categories, natural language can provide a complete description of the video. For this reason we hypothesize that models trained with natural language supervision might provide richer features useful for characterizing memorability. In this paper, we exploit the Contrastive Language-Image Pretraining (CLIP) [11] model for encoding video frames. An attention mechanism is then proposed for selecting relevant frames. Finally a Multi Layer Perceptron maps the video features into a memorability score. The experimental results for subtask 1 and subtask 2 of the MediaEval memorability task demonstrate the effectiveness of the proposed method.

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

*Corresponding author.

m.agarla@campus.unimib.it (M. Agarla); luigi.celona@unimib.it (L. Celona); raimondo.schettini@unimib.it (R. Schettini)

0000-0002-0009-8007 (M. Agarla); 0000-0002-5925-2646 (L. Celona); 0000-0001-7461-1451 (R. Schettini)

© 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)

CEUR-WS.org/Vol-3583/paper18.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

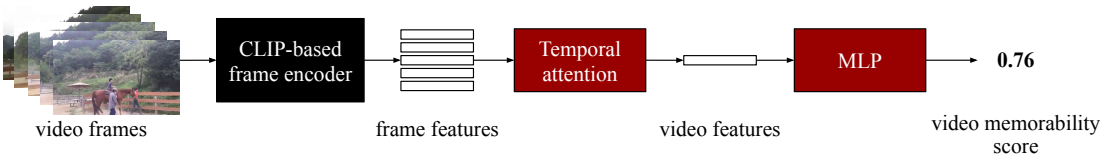


Figure 1: Overview of the approach proposed for predicting the video memorability score.

2. Approach

Figure 1 shows an overview of the proposed method that achieved the best performance among our proposals in both subtask 1 and 2 of the MediaEval Memorability challenge. Our method receives video frames as input and consists of three main modules: (i) a CLIP-based encoder [11] that extracts the features for each video frame, (ii) a temporal attention module which aggregates the frame-level features into a video-level feature vector, (iii) an MLP which maps the video-level features into a memorability score.

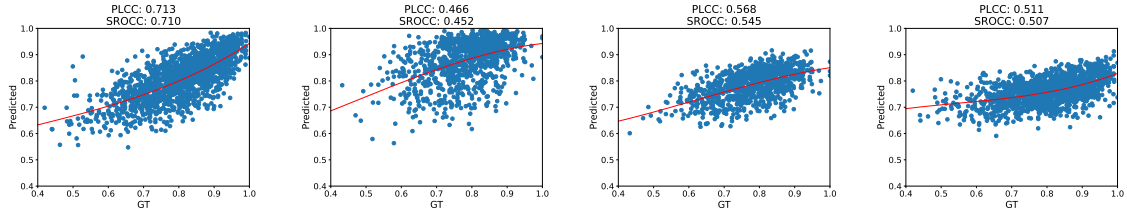
CLIP-based frame encoder. Contrastive Language-Image Pretraining (CLIP) [11] is a multimodal model that learns to represent images and text jointly in the same vector space. It consists of image and text transformer-based encoder networks [12, 13]. CLIP shows very good performance on content classification datasets [11], but it also demonstrate to be effective in perceptual tasks [14, 15]. In our method we exploit the CLIP-based Vision Transformer (ViT) encoder for video frame encoding, as known as ViT L/14. The ViT extracts 14×14 patches from an input image with size 336×336 and outputs a 1024-dimensional feature vector[1]. Given a video with T frames, we resize each frame at a resolution of 336×336 pixels and feed it into the ViT. We obtain $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_T)$ with $\mathbf{E} \in \mathbb{R}^{T \times 1024}$ representing the set of frame-level feature vectors of the input video. Each feature vector $\mathbf{e}_t$ is normalized by its L_2-norm before further processing.

Temporal attention module. The temporal attention module aims at weighting the contribution of each frame feature vector to obtain the video-level representation. A Bidirectional Gated Recurrent Unit (Bi-GRU) [16] is used for modelling temporal information among frames. The GRU consists of 6 GRU layers, each layer has an hidden state with size equal to 64 and is followed by a dropout layer with a probability of 0.2. The set of frame-level features extracted by CLIP, $\mathbf{E}$, is fed to the GRU which outputs $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_T)$ with size $T \times 128$. The matrix $\mathbf{H}$ is converted to a scaling factor $\mathbf{w} \in \mathbb{R}^{T \times 1}$ per frame throw a linear layer. The video-level feature vector $\mathbf{e}_v \in \mathbb{R}^{1024}$ is finally obtained as follows:

$$\mathbf{e}_v = \frac{1}{T} \sum_{t=1}^{T} \mathbf{e}_t * \mathbf{w}_t. \tag{1}$$

Multi-Layer Perceptron. The Multi-Layer Perceptron (MLP) estimates the video memorability score given the 1024-dimensional video-level feature vector $\mathbf{e}_v$. It consists of a stack of three linear layers. The first two linear layers reduce the size of the feature vector first to 512 and then to 128 dimensions. Each linear layer is followed by a ReLU activation function. The last linear layer outputs a scalar representing the video memorability score. The sigmoid activation function is exploited to limit the values in the range [0, 1].

[1] https://huggingface.co/openai/clip-vit-large-patch14-336



Memento10k/Memento10k Memento10k/VideoMem VideoMem/VideoMem VideoMem/Memento10k

Figure 2: Scatter plots of the GT vs. predicted memorability scores for the four train/dev combinations.

2.1. Implementation details

The method is implemented using the PyTorch [17] framework. For each video all T-frames are considered. During the training phase, the frames are shuffled for data augmentation, the optimizer Adam [18] is used with an initial learning rate equal to 1×10^{-4} which is then reduced every 5 epochs by 0.95. We train using the L_1 criterion and a batch size of 8 for a maximum of 100 epochs. However, the training process stops when there is no improvement after five consecutive epochs.

3. Results and Analysis

In this section we present the results achieved on the subtask 1 and the subtask 2. For both tasks we measure the performance in terms of Pearson's Linear Correlation Coefficient (PLCC) and Spearman's Rank Order Correlation Coefficient (SROCC). The results for the development set of subtask 1 which consists in training and testing on Memento10k [9] are depicted in Table 1. We highlight that our best approach (whose details are provided in the previous section) achieves a PLCC of 0.7132 and an SROCC of 0.7100 on the development set. The performance estimated by the organizers on the Memento10k test set corresponds to 0.707 for both correlation metrics and 0.005 of mean squared error. For subtask2 which consists in a cross-dataset scenario involving Memento10k and VideoMem [19] datasets, our best method obtains the performance reported in Table 2. As expected, in cross-dataset scenario performance decreases by about 20% for both correlations. It can also be noted that the training on VideoMem allows the method to generalize better on Ment10k with performance about 10% higher than the one obtained by training on Memento10k and testing on VideoMem. Figure 2 shows scatter plots on the four training/test combinations for the development set. The distribution of samples in the different plots reflects what was previously stated, i.e. that apart from the combination Memento10k/VideoMem the other distributions are well fit. Figure 3 shows the samples with the highest prediction errors. It is possible to notice that the proposed model tends to overestimate the memorability for such videos. Particular is the case of Memento10k/VideoMem and VideoMem/VideoMem, for which the worst error was obtained for the same video.

Ablation study. Table 1 shows the results for our less effective solutions. For the encoder, in addition to the ViT variants, we proposed several approaches that include an I3D model [20] pre-trained on Kinetics-400 [21] or Charades [22] used as a feature extractor or finetuned on memorability. We also proposed a method combining frame-level (ViT) and spatio-temporal (I3D) features. For the temporal aggregation of the features we experimented with the GRU used in different ways, a transformer and a combination of linear to reduce the dimensionality of the frame-level features followed by the temporal averaging. Finally, a simple linear layer or an MLP were tested for the memorability predictor.

Table 1
Results on the development set of Memento10k for subtask 1. The best and the second-best results on each metric are marked in boldface and underlined, respectively.

Encoder		Temporal aggregator	Predictor	PLCC	SROCC
CLIP-ViT	B/32	Linear+Avg.	Linear	0.7009	0.7000
	B/32	Bi-GRU	Linear	0.7052	0.7063
	L/14	Bi-GRU	Linear	0.7075	<u>0.7079</u>
	L/14	**Bi-GRU attention**	**MLP**	**0.7132**	**0.7100**
	L/14	Transformer	Linear	0.7033	0.7038
I3D (finetuned for memorability)		–	Linear	0.5065	0.5142
I3D (pretrained on Kinetics)		–	Linear	0.5915	0.5882
I3D (pretrained on Charades)		–	Linear	0.5724	0.5647
CLIP-ViT-B/32 + I3D (finetuned)		Linear+Avg.+Concat.	MLP	0.7059	0.7061
CLIP-ViT-B/32 + I3D (Kinetics)		Linear+Avg.+Concat.	MLP	0.7040	0.6987
CLIP-ViT-B/32 + I3D (Charades)		Linear+Avg.+Concat.	MLP	<u>0.7104</u>	0.6972

Table 2
Results of our best proposal for subtask 2 on both development and testing sets.

		Development set				Testing set			
		Memento10k		VideoMem		Memento10k		VideoMem	
		PLCC	**SROCC**	**PLCC**	**SROCC**	**PLCC**	**SROCC**	**PLCC**	**SROCC**
Training set	Memento10k	–	–	0.466	0.452	–	–	0.487	0.470
	VideoMem	0.511	0.507	–	–	0.529	0.523	–	–

Figure 3: *(Best viewed in colors and magnified.)* Visualization of worst predictions of the proposed method. For each train/dev combination, a random video frame and the predicted and GT (in parenthesis) memorability scores are reported.

4. Discussion and Outlook

Our solution involving a CLIP-ViT-L/14 + Bi-GRU attention + MLP achieves better results than the transformer, the I3D network, and the ViT+I3D. This result confirms our hypothesis that a model trained with natural language supervision can provide richer features than a model trained for action recognition. The results can also be attributed to the Bi-GRU-based temporal attention module allowing the selection of the most relevant video frames. Temporal information modeling treated as the I3D network allows the model to extract the flow relationship between frames but lacks relevant semantic information features. Moreover, the limited dataset size and the small length of the videos, approximately $3s$, make complex architectures (like the I3D and the transformer) prone to overfitting. From the cross-dataset results, we can conclude that the VideoMem dataset allows the model to generalize better on the Memento10k dataset. This is likely due to the wide variation in content, scenes, and memorability score of the Memento10k dataset. As future works, we will first exploit a frame sampling algorithm to avoid processing frames containing redundant information [1]. Secondly, we will investigate the use of spatial attention mechanisms for estimating video memorability.

References

[1] M. Agarla, L. Celona, R. Schettini, An efficient method for no-reference video quality assessment, Journal of Imaging 7 (2021) 55.

[2] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, M. Shah, Towards a comprehensive computational model for aesthetic assessment of videos, in: International Conference on Multimedia, ACM, 2013, pp. 361–364.

[3] D. V. Nieto, L. Celona, C. F. Labrador, Understanding aesthetics with language: A photo critique dataset for aesthetic assessment, in: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022.

[4] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, L. Van Gool, The interestingness of images, in: ICCV, IEEE, 2013, pp. 1633–1640.

[5] M. G. Constantin, L.-D. Ştefan, B. Ionescu, N. Q. Duong, C.-H. Demarty, M. Sjöberg, Visual interestingness prediction: a benchmark framework and literature review, International Journal of Computer Vision 129 (2021) 1526–1550.

[6] 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.

[7] M. Leonardi, L. Celona, P. Napoletano, S. Bianco, R. Schettini, F. Manessi, A. Rozza, Image memorability using diverse visual features and soft attention, in: ICIAP, Springer, 2019, pp. 171–180.

[8] S. Perera, A. Tal, L. Zelnik-Manor, Is image memorability prediction solved?, in: CVPR Workshops, IEEE/CVF, 2019, pp. 0–0.

[9] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, in: ECCV, Springer, 2020, pp. 223–240.

[10] L. Sweeney, G. Healy, A. F. Smeaton, The influence of audio on video memorability with an audio gestalt regulated video memorability system, in: International Conference on Content-Based Multimedia Indexing (CBMI), IEEE, 2021, pp. 1–6.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[12] 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).

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2020.

[14] J. Wang, K. C. Chan, C. C. Loy, Exploring clip for assessing the look and feel of images, in: AAAI Conference on Artificial Intelligence, 2023.

[15] S. Hentschel, K. Kobs, A. Hotho, Clip knows image aesthetics, Frontiers in Artificial Intelligence 5 (2022).

[16] K. Cho, B. van Merriënboer, Ç. Gulçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder–decoder for statistical machine translation, in: EMNLP, 2014, pp. 1724–1734.

[17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch (2017).

[18] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: ICLR, 2015.

[19] R. Cohendet, C.-H. Demarty, N. Q. Duong, M. Engilberge, Videomem: Constructing, analyzing, predicting short-term and long-term video memorability, in: ICCV, IEEE/CVF, 2019, pp. 2531–2540.

[20] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE CVPR, 2017, pp. 6299–6308.

[21] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, 2017.

[22] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, K. Alahari, Charades-ego: A large-scale dataset of paired third and first person videos, 2018.