Investigating Memorability of Dynamic Media

Phuc H. Le-Khac, Ayush K. Rai, Graham Healy, Alan F. Smeaton, Noel E. O'Connor
Dublin City University, Ireland
{khac.le2,ayush.rai3}@mail.dcu.ie
{graham.healy,alan.smeaton,noel.oconnor}@dcu.ie

ABSTRACT

The Predicting Media Memorability task in MediaEval'20 has some challenging aspects compared to previous years. In this paper we identify the high-dynamic content in videos and dataset of limited size as the core challenges for the task, we propose directions to overcome some of these challenges and we present our initial result in these directions.

1 INTRODUCTION

Since the seminal paper on image memorability by Isola *et. al.* [10], there has been growing interest in building computational models to understand and predict the intrinsic memorability of media, as well as other subjective perceptions of media [7]. The Predicting Media Memorability task at MediaEval is designed to facilitate and promote research in this area by requiring task participants to develop computational approaches to generate measures of media memorability, which in turn also helps to better understand the subjective memorability of human cognition. Having a prediction model of a media's memorability also allows for interesting applications to be developed, such as techniques to enhance memorability of images through the use of neural style transfer [14].

Continuing from the success of the previous two years [4, 5], the MediaEval'20 Predicting Media Memorability challenge [8] remains the same, where teams need to build prediction models for the memorability scores from a dataset of videos with captions. There are two types of memorability scores: short-term scores for videos that are shown for a second time within a short timespan of the initial viewing, and long-term scores for videos that are shown again two to three days later. The videos making up this year's training dataset contain 590 videos with sound from the TRECVid 2019 Video-to-text dataset [1], compared to 8,000 soundless videos from VideoMem [3] dataset.

2 RELATED WORK

In previous challenges, a variety of methods that utilised different data modalities were explored [2, 11, 17]. As previous works have shown [6, 13], utilising high-level semantic features, either extracted using deep networks or provided by human annotators via text captions, are among the most effective methods to predict memorability. However, the modalities and features that are most predictive of memorability remain unclear i.e. the best approach on last year's challenge [2] used an ensemble of models trained on a variety of modalities.

MediaEval'20, 14-15 December 2020, Online

3 APPROACH

3.1 Motivation

High dynamic video. The short videos used in previous years' memorability challenges were extracted from raw professional footage that is used for example in creating high-quality film and commercials [3]. Consequently, the majority of videos therefore contain only one scene and are mostly static. Conversely, the data for this year's 2020 challenge is extracted from the TRECVid 2019 [1] Video-to-Text dataset. These videos were collected from social media posts and TV shows. They contain multiple scene changes in one video, and thus are more dynamic and with more complex movement and activity levels, as illustrated in Fig.1.

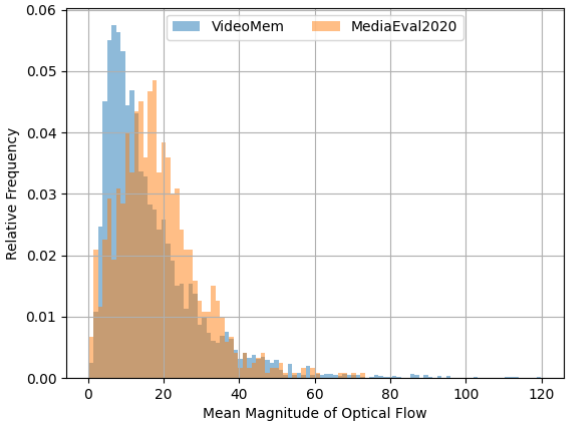

Figure 1: Histogram showing distributional differences in Mean Magnitude of Optical Flow (amount of motion) between the VideoMem dataset used in previous years and the MediaEval2020 dataset. The Mean Magnitude of Optical Flow of a video was obtained by averaging the optical flow features over all pixels in a frame and across all frames.

While previous approaches to memorability computation use higher-semantic features, such as those extracted from deep neural networks, text captions provided by human-annotators and human-centric features such as emotions or aesthetics, most of these features were extracted using frame-based models. In contrast, this year's dataset provided an unexplored challenge for methods that can capture semantic features from highly dynamic videos.

Small size dataset. Another challenge for this year's data set is the limited number of annotated videos with memorability scores. Compared to previous years, this is an order of magnitude smaller

in size, with 590 videos in the training set and 500 videos in the test set. A development set with an additional 500 videos was released in the later stage of the benchmark.

Challenging benchmark. The changes to this year's dataset made the task considerably more challenging, as the videos used were more dynamic thus making the extraction of high-level semantic features more difficult, and further compounded by the limited size of the dataset.

Motivated by these insights, we focus our work on using video-level features to capture the dynamic in videos, and attempt to pre-train a memorability model on a larger dataset before fine-tuning it on this year MediaEval's data.

3.2 Spatiotemporal baseline

While the provided captions for each video can be used as high-level semantic-rich features for predicting media memorability. However, in our approach we focus on a method to extract high-level features from raw video input, without any additional annotation.

Motivated from the fact that this year's video is highly dynamic as discussed above, we chose to use features extracted from C3D [16], the only video-based method instead of frame-based model. We used the C3D pre-extracted features provided by the task organisers as the input for our memorability regression model.

C3D model learned spatio-temporal representation. C3D [17], short for 3D Convolution, is among the earliest approaches to learning generic representation from videos. Extending the 2D Convolution operations common in most image-processing deep learning models, C3D's homogeneous architecture composed of small $3 \times 3 \times 3$ convolution kernels, expanding in all height, width and time dimension of videos. Trained on a generic action recognition dataset, each video passed through C3D returns a 4096-dimension feature vector that extracts high-level semantic features from a video segment.

Memorability Regression Model. With the extracted features via C3D, we used a multi-layer perceptron (MLP) with two hidden layers of 512 units followed by a Rectified Linear Unit (ReLU) non-linearity layer. To mitigate overfitting, we used Dropout [15] with a probability of 0.1. Batch Normalisation [9] was used to normalise the features in each batch after each layer to help optimisation. The final model is a sequential application 3 building blocks composed from BatchNorm, Dropout, fully-connected and ReLU activation layers. For the final output layer, the ReLU activation is replaced with a Sigmoid to obtain the final memorability score in the range from 0 to 1.

There was a total of more than 2 million parameters in the model, most belonging to the first fully-connected layer due to the large C3D's feature size of 4096. To minimise the effect of outliers in ranking score, we use L1 loss instead of the standard L2 Mean-Squared Error (MSE) and saw a slight improvement in convergence speed. The model is trained for 100 epochs on the training set and the final model was picked based on the performance on the validation set.

3.3 Large scale Memorability Pre-training

To overcome the limited size of this year's dataset, we used the recently released Memento10K [12] for pre-training our model before fine tuning on the MediaEval challenge's dataset. Containing 10,000 videos, the Memento10K dataset is roughly the same size as the VideoMem [3] dataset. Moreover, videos in the Memento10K contains more action and are more similar to the new data of this year's challenge. Therefore we decided to focus our large-scale pre-training approach on the Memento10K dataset and replicate the accompanying SemNet model [12] instead of pre-training on VideoMem [3] from previous years. The SemNet model contains three separate sub-networks to process three different input streams: *image*, *optical flow* and *video* stream.

After pre-training SemNet on Memento10k data, we only retain the video stream sub-network to fine tune on the MediaEval dataset and discarded all other components.

4 RESULTS AND DISCUSSION

The results of our runs together with this year's mean and variance are reported in Table 1. Due to technical issues and time constraints, we only submitted the baseline regression model based on C3D features and did not have the result for fine tuning SemNet on the Memento10K [12] dataset for this year's challenge.

Table 1: Results of our approach compared with the challenge's statistics

	Short-term			Long-term		
Run	**Spearman**	Pearson	MSE	**Spearman**	Pearson	MSE
Mean	0.058	0.066	0.013	0.036	0.043	0.051
Variance	0.002	0.002	0.000	0.002	0.001	0.000
C3D	0.034	0.078	0.1	-0.01	0.022	0.09

Even with the simplest method possible, the result from our regression baseline with pre-extracted C3D features is not very far from the mean. This support our hypothesis that spatio-temporal representation is important for this year's dataset and the performance can be increased with other complementary high-level features. Future work can explore this hypothesis further by fine tuning the entire C3D model instead of just using the pre-extracted features, or by using different models that also capture spatio-temporal features.

On the other hand, compared to the state-of-the-art result of 0.528 [2] in last year's challenge, the mean of the Spearman rank correlation for all runs this year (0.058) is an order of magnitude lower. This highlights the importance of large scale pre-training and transfer learning techniques, and methods that can extract high-level features from dynamic videos effectively.

We believe the direction of our solution, given the challenges in this year's challenge, not only shows promise but also indicates the importance of spatio-temporal models to capture high-level semantics of videos for memorability prediction.

ACKNOWLEDGMENTS

This work was co-funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning (18/CRT/6183) and the Insight Centre for Data Analytics (SFI/12/RC/2289_P2), co-funded by the European Regional Development Fund. A.K. Rai also acknowledges support from FotoNation Ltd.

REFERENCES

[1] George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, and Georges Quenot. 2020. TRECVID 2019: An Evaluation Campaign to Benchmark Video Activity Detection, Video Captioning and Matching, and Video Search & Retrieval. In *TREC Video Retrieval Evaluation Notebook Papers and Slides*. https://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.19.org.html

[2] David Azcona, Enric Moreu, Feiyan Hu, Tomás E Ward, and Alan F Smeaton. 2019. Predicting Media Memorability Using Ensemble Models. In *Proceedings of MediaEval 2019, Sophia Antipolis, France*. CEUR Workshop Proceedings, 3. http://ceur-ws.org/Vol-2670/

[3] Romain Cohendet, Claire-Helene Demarty, Ngoc Duong, and Martin Engilberge. VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019). IEEE, 2531–2540. https://doi.org/10.1109/ICCV.2019.00262

[4] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q K Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. MediaEval 2018: Predicting Media Memorability. In *Proceedings of MediaEval 2018, CEUR Workshop Proceedings* (2018). 3. http://ceur-ws.org/Vol-2283/

[5] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Hélène Demarty, Ngoc Q K Duong, Xavier Alameda-Pineda, and Mats Sjöberg. The Predicting Media Memorability Task at MediaEval 2019. In *Proceedings of MediaEval 2019, Sophia Antipolis, France* (2019). CEUR Workshop Proceedings, 3. http://ceur-ws.org/Vol-2670/

[6] Mihai Gabriel Constantin, Chen Kang, G. Dinu, Frédéric Dufaux, Giuseppe Valenzise, and B. Ionescu. Using Aesthetics and Action Recognition-Based Networks for the Prediction of Media Memorability. In *Proceedings of MediaEval 2019, Sophia Antipolis, France* (2019). CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2670/

[7] Mihai Gabriel Constantin, Miriam Redi, Gloria Zen, and Bogdan Ionescu. 2019. Computational Understanding of Visual Interestingness Beyond Semantics: Literature Survey and Analysis of Covariates. *ACM Comput. Surv.* 52, 2, Article 25 (2019), 37 pages. https://doi.org/10.1145/3301299

[8] Alba García Seco de Herrera, Rukiye Savran Kiziltepe, Jon Chamberlain, Mihai Gabriel Constantin, Claire-Hélène Demarty, Faiyaz Doctor, Bogdan Ionescu, and Alan F. Smeaton. 2020. Overview of MediaEval 2020 Predicting Media Memorability task: What Makes a Video Memorable?. In *Working Notes Proceedings of the MediaEval 2020 Workshop*.

[9] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15)*. JMLR.org, 448–456.

[10] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. 2011. Understanding the Intrinsic Memorability of Images. In *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.), Vol. 24. Curran Associates, Inc., 2429–2437. https://proceedings.neurips.cc/paper/2011/file/286674e3082feb7e5afb92777e48821f-Paper.pdf

[11] Roberto Leyva and Faiyaz Doctor. Multimodal Deep Features Fusion For Video Memorability Prediction. In *Proceedings of MediaEval 2019, Sophia Antipolis, France* (2019). CEUR Workshop Proceedings, 3. http://ceur-ws.org/Vol-2670/

[12] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. 2020. Multimodal Memorability: Modeling Effects of Semantics and Decay on Video Memorability. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 223–240.

[13] Alison Reboud, Ismail Harrando, Jorma T. Laaksonen, D. Francis, Raphaël Troncy, and H. Mantecón. Combining Textual and Visual Modeling for Predicting Media Memorability. In *Proceedings of MediaEval 2019, Sophia Antipolis, France* (2019). CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2670/

[14] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. 2019-06-14. Increasing Image Memorability with Neural Style Transfer. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 2 (2019-06-14), 1–22. https://doi.org/10.1145/3311781

[15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)* (2015-12). 4489–4497. https://doi.org/10.1109/ICCV.2015.510

[17] Le-Vu Tran, Vinh-Loc Huynh, and Minh-Triet Tran. Predicting Media Memorability Using Deep Features with Attention and Recurrent Network. In *Proceedings of MediaEval 2019, Sophia Antipolis, France* (2019). CEUR Workshop Proceedings, 3. http://ceur-ws.org/Vol-2670/