Predicting Media Memorability from a Multimodal Late Fusion of Self-Attention and LSTM Models

Ricardo Kleinlein[1], Cristina Luna-Jiménez[1], Zoraida Callejas[2], Fernando Fernández-Martínez[1]
[1]Speech Technology Group, Center for Information Processing and Telecommunications, E.T.S.I. de Telecomunicación,
Universidad Politécnica de Madrid, Spain
[2] Department of Languages and Computer Systems, University of Granada, Spain
ricardo.kleinlein@upm.es

ABSTRACT

This paper reports on the GTH-UPM team experience in the Predicting Media Memorability task at MediaEval 2020. Teams were requested to predict memorability scores at both short-term and long-term, understanding such score as a measure of whether a video was perdurable in a viewer's memory or not. Our proposed system relies on a late fusion of the scores predicted by three sequential models, each trained over a different modality: video captions, aural embeddings and visual optical flow-based vectors. Whereas single-modality models show a low or zero Spearman correlation coefficient value, their combination considerably boosts performance over development data up to 0.2 in the short-term memorability prediction subtask and 0.19 in the long-term subtask. However, performance over test data drops to 0.016 and -0.041, respectively.

1 INTRODUCTION

The improvement in computational capabilities is progressively allowing researchers to tackle problems long though to be out of reach due to the subjective nature of the phenomena involved. One good instance is memorability prediction. The seminal work of Isola *et al.* set the ground for later work on computational modelling of image memorability [11]. Since 2018 the *Predicting Media Memorability Challenge*, hosted within the MediaEval workshop, has pushed forward the extent of the original problem to encompass memorability prediction over multimedia sources of information [3, 4]. In its current edition the goal of the task holds the same as previous years, yet video clips now cover a kind of material resembling short videos commonly found in social media. Further information can be found in the challenge description paper [7].

Several multimodal late fusion strategies have been proposed regarding the image and video memorability prediction problem [5]. Additionally, attention mechanisms have been successfully applied to problems in which data come naturally in a sequential form [16]. In particular, self-attention layers have been proved to boost performance when tackling the computational modelling of media memorability [6].

2 APPROACH AND EXPERIMENTS

Every video sample in the dataset presents the following sources of information: between 2 and 5 text captions that roughly describe the content of the video, the video audio signal and its visual frames. As stated before, multimodal systems are able to learn modality-wise data representations, and combine their predictive power in order to make a final, unique memorability prediction. We hypothesize that a late fusion scheme will benefit from incorporating a self-attention mechanism that learns to focus on what it is particularly relevant on a given sample's prediction.

We propose a system based on the late fusion by a Support Vector Regressor (SVR) of the predictions made by three single-modality models whose architecture is depicted in Figure 2. In all cases the biLSTM encoders have 75 units, with all the learners sharing the same architecture but trained independently. Prediction comes as the outcome of the last sigmoid layer. Learned layers suffer from a dropout rate fixed at 0.3. For every single-modality learner the training pipeline holds the same; batch size is set at 128, with initial learning rate 0.001 and Adam optimizer [12]. Figure 1 shows the general prediction pipeline from these models. Results shown in this paper are obtained following a 5-fold cross-validation procedure over the 1000 videos of the development data. Training is stopped after 5 epochs with no improvement over the Spearman correlation coefficient, computed over the fold's validation data. Experimental results are summarized in Table 1. Next we introduce in greater detail the feature extraction processing carried out for every modality.

2.1 Text captions

We merge all the captions of a sample into a single one in a Bag-Of-Words fashion. Afterwards, we extract the lemma of every word in the text using NLTK's WordNet-based Lemmatizer [1, 14]. Finally, the input of the text modality is made by the sequence of fasttext 300-dimensional word embeddings corresponding to every word in the sample's BOW-text [2]. At training time, random noise with $\mu = 0$ and $\sigma = 0.15$ is added to the niput embeddings in order to improve learning robustness.

2.2 Audio signal

Based on previous experience, we hypothesize that event detection-oriented embeddings provide a robust basis to study multimedia perceptual variables such as attention or memorability [13]. Therefore we compute aural embeddings using the default VGGish configuration, which is pretrained on Audioset, a large audio event-detection database [8, 9]. That way every video audio signal is defined by a sequence of 128-dimensional embeddings, each spanning 960 ms of audio and without overlap between them.

2.3 Video image

Videos in the dataset are no longer than a few seconds, characterized by an event happening quickly and conforming the most relevant

MediaEval'20, December 14-15 2020, Online

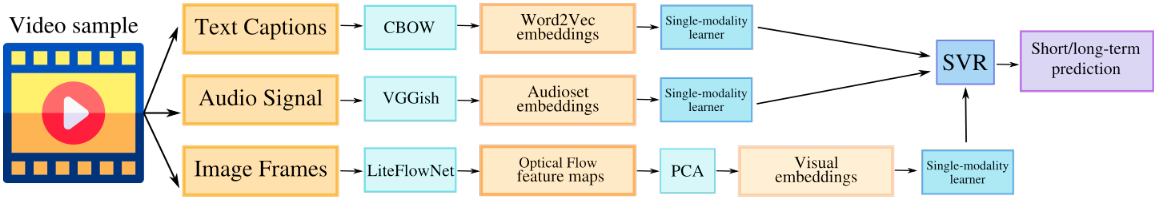

Figure 1: Proposed video memorability prediction pipeline. The system is the same when dealing with both short- and long-term memorability scores, but single-modality learners are trained independently for every time interval and modality.

		Spearman coeff. for fold # – Development Set						Test Set		
Time range	Model	1	2	3	4	5	AVG	Spearman	Pearson	MSE
Short-term	Word2Vec Captions	0.00	0.05	0.13	-0.03	-0.06	0.02	–	–	–
	Audioset embeddings	-0.06	-0.04	0.07	0.02	0.01	0.00	–	–	–
	Optical Flow + PCA(128)	0.11	0.01	0.07	-0.1	0.08	0.03	–	–	–
	Prediction ensemble + SVR	**0.22**	**0.20**	**0.20**	**0.23**	**0.17**	**0.20**	0.016	0.011	0.01
Long-term	Word2Vec Captions	0.08	0.06	0.06	0.12	0.13	0.09	–	–	–
	Audioset embeddings	0.07	0.05	-0.10	0.12	0.17	0.06	–	–	–
	Optical Flow + PCA(128)	-0.02	0.13	-0.05	0.10	0.19	0.07	–	–	–
	Prediction ensemble + SVR	**0.19**	**0.19**	**0.19**	**0.23**	**0.18**	**0.19**	-0.041	-0.028	0.05

Table 1: Spearman correlation coefficient scores computed for every validation fold in the dataset, as well as the overall average and official test results. Both short- and long-term scores are shown for every predictive model studied.

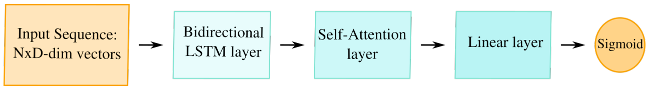

Figure 2: Architecture of the single-modality learners.

part of the clip. Because of that, videos are expected to display quick changes in pixel values between consecutive frames due to visual events taking place. In order to capture the degree of visual change along a clip, we compute optical feature maps for its frames, extracted at 3 FPS, using a LiteFlowNet model [10]. We further reduce optical flow features' dimensionality by projecting them into a 128-dimensional subspace computed by PCA [15]. A sample is represented by a temporally-sorted sequence of 128-dimensional features that retains most of the information regarding the optical flow features maps.

2.4 Ensemble of modality-wise models

We independently train single-modality models from the features explained in the sections above. Thereafter, a memorability prediction is computed for every sample in the dataset. The combination of the three memorability scores is the input for a SVR that makes a final prediction that reflects the knowledge extracted from the different the modalities.

As it can be seen from Table 1, individual learners are not able to fully characterize a video sample and learn the relationship with its memorability score. However, the ensemble of the three of them achieves a Spearman correlation coefficient value of 0.2 in the short-term problem and 0.19 in the long-term one over development data. However, we notice that the performance on the test data significantly drops, achieving much lower scores on both subtasks.

3 DISCUSSION AND OUTLOOK

Despite individual learners showing very low or even zero coefficient values, a SVR based on their posteriors seems to weakly capture the relationship between media content and its memorability score, with similar correlation values obtained at both short-term and long-term subtasks. This might be partially caused by the limited amount of data available, which is likely to be dragging the learning process, and therefore making the SVR to learn the development dataset's score distribution. Prediction's distribution suggests that the system might be learning to approximate every sample to the mean memorability score, rather than exploiting the knowledge extracted from the computed features. Future work includes extending the amount of training data with similar datasets. It is also left for future studies to explore different data encodings, with special emphasis on smaller, more compact data representations that might better suited for cases where large datasets are not available.

ACKNOWLEDGMENTS

The work leading to these results has been supported by the Spanish Ministry of Economy, Industry and Competitiveness through CAVIAR (MINECO, TEC2017-84593-C2-1-R) and AMIC (MINECO, TIN2017-85854-C4-4-R) projects (AEI/FEDER, UE). Ricardo Kleinlein's research was supported by the Spanish Ministry of Education (FPI grant PRE2018-083225).

REFERENCES

[1] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python.* O'Reilly Media.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[3] Romain Cohendet, Claire-Hélène Demarty, Ngoc Duong, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, and France Rennes. 2018. MediaEval 2018: Predicting Media Memorability Task. (2018). arXiv:cs.CV/1807.01052

[4] Mihai-Gabriel Constantin, Bogdan Ionescu, Claire-Hélène Demarty, Ngoc Duong, Xavier Alameda-Pineda, and Mats Sjöberg. 2019. The Predicting Media Memorability Task at MediaEval 2019.

[5] Mihai Gabriel Constantin, Chen Kang, Gabriela Dinu, Frédéric Dufaux, Giuseppe Valenzise, and Bogdan Ionescu. 2019. Using Aesthetics and Action Recognition-Based Networks for the Prediction of Media Memorability. In *Working Notes Proceedings of the MediaEval 2019 Workshop, Sophia Antipolis, France, 27-30 October 2019 (CEUR Workshop Proceedings)*, Martha A. Larson, Steven Alexander Hicks, Mihai Gabriel Constantin, Benjamin Bischke, Alastair Porter, Peijian Zhao, Mathias Lux, Laura Cabrera Quiros, Jordan Calandre, and Gareth Jones (Eds.), Vol. 2670. CEUR-WS.org. http://ceur-ws.org/Vol-2670/MediaEval_19_paper_60.pdf

[6] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. AMNet: Memorability Estimation with Attention. (2018). arXiv:cs.AI/1804.03115

[7] 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*.

[8] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

[9] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. (2017). arXiv:cs.SD/1609.09430

[10] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. 2018. LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[11] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What makes a photograph memorable? *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36, 7 (2014), 1469–1482.

[12] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. (2017). arXiv:cs.LG/1412.6980

[13] Ricardo Kleinlein, Cristina Luna Jiménez, Juan Manuel Montero, Zoraida Callejas, and Fernando Fernández-Martínez. 2019. Predicting Group-Level Skin Attention to Short Movies from Audio-Based LSTM-Mixture of Experts Models. In *Proc. Interspeech 2019*. 61–65. https://doi.org/10.21437/Interspeech.2019-2799

[14] George A. Miller. 1995. WordNet: A Lexical Database for English. *COMMUNICATIONS OF THE ACM* 38 (1995), 39–41.

[15] Karl Pearson. 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.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. (2017). arXiv:cs.CL/1706.03762