Combining Textual and Visual Modeling for Predicting Media Memorability

Alison Reboud[*], Ismail Harrando[*], Jorma Laaksonen[+],
Danny Francis [*], Raphaël Troncy[*], Héctor Laria Mantecón[+]

[*]EURECOM, Sophia Antipolis, France
[+]Aalto University, Espoo, Finland
{alison.reboud,ismail.harrando,danny.francis,raphael.troncy}@eurecom.fr
{jorma.laaksonen,hector.lariamantecon}@aalto.fi

ABSTRACT

This paper describes a multimodal approach proposed by the MeMAD team for the MediaEval 2019 "Predicting Media memorability" task. Our best approach is a weighted average method combining predictions made separately from visual and textual representations of videos. In particular, we augmented the provided textual descriptions with automatically generated deep captions. For long term memorability, we obtained better scores using the short term predictions rather than the long term ones. Our best model achieves Spearman scores of 0.522 and 0.277 respectively for the short and long term predictions tasks.

1 INTRODUCTION

Considering video memorability as a useful tool for digital content retrieval as well as for sorting and recommending an ever growing number of videos, the Predicting Media Memorability Task aims at fostering the research in the field by asking its participants to automatically predict both a short and long term memorability score for a given set of annotated videos. The full description for this task is provided in [2]. Last year's best approaches for both the long term[5] and short term tasks [14] indicated that high level representations extracted from deep convolutional models performed the best in terms of visual features. Furthermore, the best long term model [5] was a weighted average method including Bag-of-Words features extracted from the provided captions. Following this approach, we created multimodal weighted average models with visual deep features and textual features extracted from both the provided video titles, as well as from automatically generated deep captions.

2 APPROACH

2.1 Visual Approaches

VisualScore. Our visual-only memorability prediction scores are based on using a feed-forward neural network with visual features in the input, one hidden layer of 430 units and one unit in the output layer. The best performance was obtained with 6938-dimensional features consisting of the concatenation of I3D [1] video features, ResNet-152 and ResNet-101 [6] image features and two versions of SUN-397 [15] concept features. The image and concept features were extracted from the middle frames of the videos. The hidden layer uses ReLU activations and dropout during the training phase, while the output unit is sigmoidal. We trained separate models for the short and long term predictions with the Adam optimizer. The number of training epochs was selected with 10-fold cross-validation with 6000 training and 2000 testing samples.

CaptionsA. Our first captioning model uses the DeepCaption software[1] and is quite similar to the best-performing model of the PicSOM Group of Aalto University's submissions in TRECVID 2018 VTT task [13]. The model was trained with COCO [10] and TGIF [9] datasets using the concatenation of ResNet-152 and ResNet-101 [6] features as the image encoding. The embed size of the LSTM network [7] was 256 and its hidden state size 512. The training used cross-entropy loss.

CaptionsB. Our second model has been trained on the TGIF [9] and MSR-VTT [16] datasets. First, 30 frames have been extracted for each video of these datasets. Then, these frames have been processed by a ResNet-152 [6] that had been pretrained on ImageNet-1000: we keep local features after the last convolutional layer of the ResNet-152 to obtain features maps of dimensions 7x7x2048. At that point, videos have been converted into 30x7x7x2048-dimensional tensors. A model based on the L-STAP method [4] has been trained on MSR-VTT and TGIF: all videos from TGIF, and training and testing videos from MSR-VTT have been used for training, and validation has been performed throughout training with the usual validation set of MSR-VTT, containing 497 videos. Cross-entropy has been used as the training loss function. The L-STAP method has been used to pool frame-level local embeddings together to obtain 7x7x1024-dimensional tensors: each video is eventually represented by 7x7 local embeddings of dimension 1024. These have been used to generate captions as in [4].

VisualEmbeddings. The local embeddings used for CaptionsB have also been used to derive global video embeddings, by averaging the mentioned 7x7 local feature embeddings. These global video embeddings have then been fed to a model of two hidden layers, the first one and the second one having respectively 100 and 50 units, and ReLU activation function. The number of training epochs is 200 with an early stopping monitor.

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

[1]https://github.com/aalto-cbir/DeepCaption

2.2 Textual Approaches

Through initial experiments and from last year's results on this task, the descriptive titles provided with each video prove to be an important modality for predicting the memorability scores. In order to build on this observation, we generate captions for each video using the two visual models described above (**CaptionsA** and **CaptionsB**). While the generated captions are not always accurate, they seem to noticeably help the model disambiguate some titles and use some of the vocabulary already seen on the training set (e.g. the title contains words such as *couple"* or *"cat"* while the generated caption would say *"a man and a woman"* or *"an animal"*, respectively, which are more common words in the training set and thus help the model generalize better on inference time). The models described in this section use a concatenation of the original provided title and the generated captions as their input.

Multiple techniques for generating a numerical score from this input sequence were considered (in ascending order of their performance on cross-validation).

Recurrent Neural Network. We use an LSTM [7] to go through the GloVe embeddings [12] of the input and predict the scores at the last token. This model performed consistently the worst, probably due to the length of the input sequence at times, and the empirical observation that word order doesn't seem to matter for this task.

Convolutional Neural Network. We use the same model as [8] except for a regression head instead of a classifier trained on top of the CNN, and GloVe embeddings as input. This model leaks less information thanks to max-pooling, and performs much better than its recurrent counterpart.

Self-attention. Similar to the previous methods, we feed our input text to a self-attentive bi-LSTM [11] to generate a sentence embedding that we use to predict the memorability scores. This model performs on par with the CNN method.

BERT. We used a pre-trained BERT model [3] to generate a sentence embedding for the input by max-pooling the last hidden states and reducing their dimension through PCA (from 768 to 250). This model performs better than the previous ones but it is more computationally demanding.

Bag of Words. We vectorize the input string by counting the number of instances of each token (and frequent n-grams) after removing the stop words and the least frequent tokens. The score is predicted by training a linear model on the counts vector. This simple model performs the best on our cross-validation, which can be justified by the lack of linguistic or grammatical structure in the titles and generated captions that would justify the use of a more sophisticated model.

For all the models considered, the addition of the generated captions improves the prediction score on the validation set considerably. It also should be noted that the use of short-term scores for long-term evaluation yields substantially better results throughout all of our experiments.

3 RESULTS AND ANALYSIS

During the evaluation process, we created four test folds of 2000 videos and therefore four models trained on 6000 videos. For the VisualScore approach, we decided to use predictions from a model trained on the entire set of 8000 videos (VisualScore8k), as well as the mean predictions from the combinations of the four models trained on 6000 videos (VisualScore6k). For the Long Term task, all models except from the WA3lt exclusively use short-term scores.

Table 1: Results on test set for short term memorability

Method	Spearman	Pearson	MSE
Textual	0.441	0.464	0.01
VisualScore	0.495	0.543	0
WA1	0.512	0.552	0
WA2	**0.522**	**0.559**	0
WA3	0.520	0.557	0

Table 2: Results on test set for long term memorability

Method	Spearman	Pearson	MSE
Textual	0.239	0.25	0.03
VisualScore	0.268	0.289	0.03
WA2	**0.277**	**0.296**	0.03
WA3	0.275	0.295	0.03
WA3lt	0.260	0.285	0.02

- WA1 = 0.5Textual+0.5VisualScore
- WA2 = 0.25Textual+0.25VisualEmb+0.5VisualScore8k
- WA3 = 0.25Textual+0.25VisualEmb+0.5VisualScore6k
- WA3lt = WA3 with long-term scores

We observe that the weighted average method which was trained on the whole training set and included our two visual approaches and our textual approach works the best for short term predictions. For long term prediction, one of the key observations to make is that WA3lt got the second worst results. This is consistent with our early observation that short-term scores for long-term evaluation yields substantially better results.

4 DISCUSSION AND OUTLOOK

This paper describes a multimodal weighted average method outperforming the best results of the Predicting Media Memorability Task 2018. One of the key contribution of this paper is to have demonstrated that using deep captions helped improving the predictions. We also conclude that, quite surprisingly, a simple n-gram frequency count was more efficient at modelling memorability than more sophisticated textual models. Finally, the fact that long term memorability was better predicted using short term predictions indicates that we failed at capturing the memorability decay of a scene from a few minutes to a few days. In the future, we would like to focus more on this aspect of the task.

ACKNOWLEDGEMENTS

This work has been partially supported by the European Union's Horizon 2020 research and innovation programme via the project MeMAD (GA 780069).

REFERENCES

[1] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4724–4733.

[2] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Helene Demarty, Ngoc Q. K. Duong, Xavier Alameda-Pineda, and Mats Sjoberg. 2019. The Predicting Media Memorability Task at MediaEval 2019. *Proc. MediaEval workshop* (2019).

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805

[4] Huet B. Francis, D. 2019. L-STAP : Learned Spatio-Temporal Adaptive Pooling for Video Captioning. In *First International Workshop on AI for Smart TV Content Production (AI4TV)*.

[5] Rohit Gupta and Kush Motwani. 2018. Linear Models for Video Memorability Prediction Using Visual and Semantic Features. In *MediaEval*.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[8] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014).

[9] Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. TGIF: A New Dataset and Benchmark on Animated GIF Description. *CoRR* abs/1604.02748 (2016). http://arxiv.org/abs/1604.02748

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*.

[11] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-attentive Sentence Embedding. *CoRR* abs/1703.03130 (2017). http://arxiv.org/abs/1703.03130

[12] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

[13] Mats Sjöberg, Hamed R. Tavakoli, Zhicun Xu, Héctor Laria Mantecón, and Jorma Laaksonen. 2018. PicSOM Experiments in TRECVID 2018. In *Proceedings of the TRECVID 2018 Workshop*. Gaithersburg, MD, USA.

[14] Duy-Tue Tran-Van, Le-Vu Tran, and Minh-Triet Tran. 2018. Predicting Media Memorability Using Deep Features and Recurrent Network. In *MediaEval*.

[15] Jianxiong Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Computer Vision and Pattern Recognition (CVPR)*. 3485–3492.

[16] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5288–5296.