AMNetVid – Predicting Video Memorability Using Attention Maps

Andrey Totev
20212042
Dublin City University
Dublin, Ireland
andrey.totev2@mail.dcu.ie

ABSTRACT

This working note presents an approach to predicting the short- and long-term memorability of a video as specified by the MediaEval Challenge 2018. Our solution does not rely on any of the features provided with the Challenge, nor it trains on video frames directly. We utilize an existing memorability model – AMNet – to create a sequence of *attention maps* which do not contain any visual elements from the original frames. Our RNN-based model treats the attention maps as signal – it learns their spatiotemporal aspects in order to provide predictions.

INTRODUCTION AND RELATED WORK

The Predicting Media Memorability task launched in 2018 as part of the MediaEval Benchmarking Initiative for Multimedia Evaluation [1]. Some 21 participant papers were published following the two editions of the challenge in 2018 and 2019.

Notably, half of the 2019 proceedings utilized AMNet [2] – a novel attention-based memorability model developed in 2018. Some of the participating teams [3] fine-tuned AMNet and used it as an alternative to off-the-shelf architectures such as ResNet. Others [4] applied the model directly to a sequence of frames feeding its predictions to an RNN. A third team [5] used a byproduct of the model called "attention maps" in a preprocessing step – to determine memorability-significant regions of the image and zero out pixels that fall outside of those regions.

Our model, AMNetVid, is a video extension of AMNet that operates on its attention maps. The computer vision part of the task is entirely delegated to the still image model. In fact, we specifically alter AMNet's code base, such that its attention maps are drawn on black background rather overlayed on top of the original frame. AMNetVid is inspired by signal processing. It feeds the attention map sequences into a Convolutional LSTM [6, 7] network which learns the graphical-temporal patterns. This approach provides more than 4% improvement of the prediction score compared to directly using AMNet for regression.

APPROACH

1 Used Hardware

All experiments were performed on a workstation with the following specification:

- CPU: Intel Core i7-9750H
- RAM: 32GB
- HDD: 1TB SSD
- GPU: Nvidia RTX 2070 8GB

2 Attention Maps

For each frame, "AMNet iteratively generates three attention maps linked with image regions correlated to memorability" [2]. As AMNetVid operates as a signal processing extension to AMNet, we adjusted AMNet's code base[1] to output the attention maps on black background rather than overlaying them on top of the original frame. This way we achieve separation of concerns where the computer vision model (AMNet) generates the signal, and the sequence model is responsible for learning spatiotemporal patterns in the signal with regards to memorability.

On the hardware system used, it took approximately 10 seconds to generate attention maps for 28 frames per video, resulting in approximately 24 hours to process all 8000 videos.

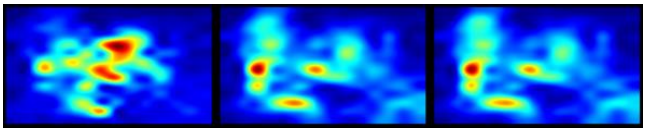

Figure 1: A video frame's attention map triplet

3 Evaluated Models

In our experiments we evaluated three models:

- AMNetVid - operates on attention maps and AMNet prediction scores for a sequence of four video frames
- Mean score (baseline) - calculates the mean AMNet score among 28 frames used from each video
- RNN - predicts video memorability by inputting a sequence of AMNet frame scores into an LSTM-based model

Mean score. This is the baseline – calculates the mean of the AMNet scores across the frames of a video.

Still image score in RNN. This model learns temporal patterns within the frames' AMNet scores without using the attention maps.

[1] AMNetVid GitHub repository, DOI 10.5281/zenodo.4723988

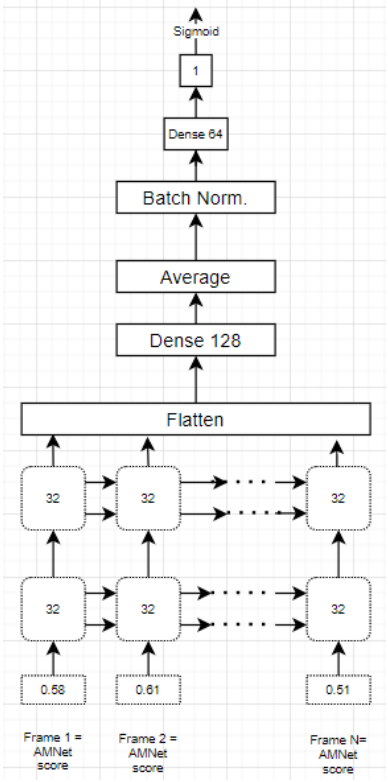

Figure 2: RNN network operating on the memorability score predictions for a sequence of frames

AMNetVid. In this work we devise a Convolutional LSTM (C-LSTM) architecture operating on a sequence of frames, where each frame is represented by a 9-channel "image" formed by stacking the red, green, and blue channels of the three attention maps. The same model architecture is used for short-term and long-term memorability predictions but is trained separately for each of the tasks.

Convolutional LSTM (C-LSTM) architectures are known to capture the input's spatiotemporal patterns better than fully connected LSTMs [7]. C-LSTMs tend to score lower than 3D convolutional networks [6] on data sets like something-something. However, C-LSTMs are more capable of modelling non-linear transitions in time whereas 3D convolutions treat time as a third spatial dimension, are able to detect shorter events, and place their spatial focus on larger contiguous image areas [6].

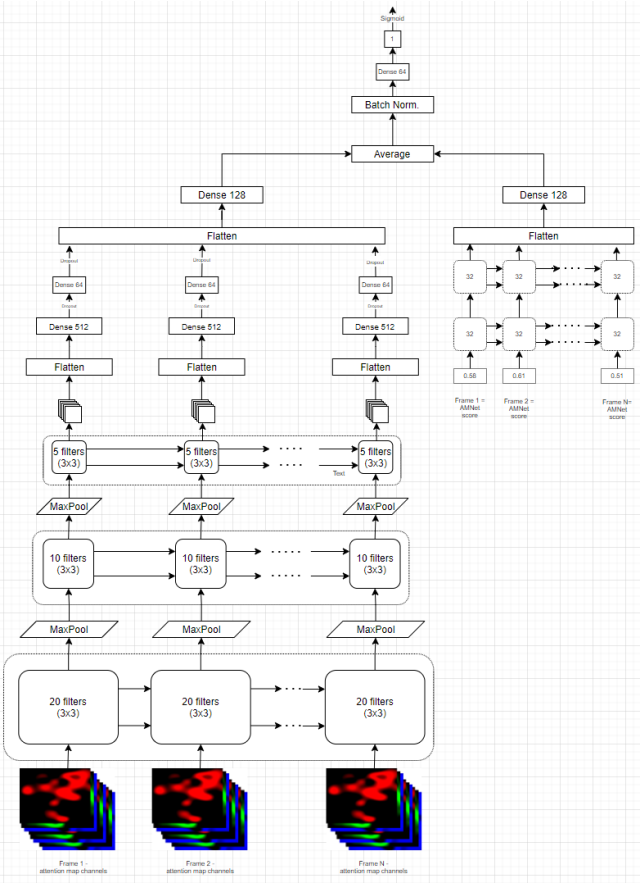

Figure 3: AMNetVid – video memorability regression based on AMNet score and attention maps for a sequence of frames

RESULTS AND ANALYSIS

	AMNet Mean (28 frames)	**AMNet RNN (4 frames)**	**AMNet Vid (4 frames)**
Short-term	**0.4214**	**0.4230 (+0.4%)**	**0.4384 (+4.0%)**
Long-term	**0.1867**	**0.1866 (0.0%)**	**0.2015 (+7.9%)**

Table 1: Comparison of baseline model, AMNet RNN score, and AMNetVid

We trained each model in three rounds, consisting of 20 epochs. From each of the three rounds, we took the best performing model, evaluated against our test data set, and calculated the mean

Spearman correlation coefficient across the three best models. Our findings show that AMNetVid performs approximately 4% better than the baseline model for short term memorability and 8% better for long term memorability.

The significantly higher improvement in the long-term memorability score could be explained with the memorability data sets that AMNet was trained on [8, 9]. That is, the tasks could be a more closely aligned with MediaEval's short-term memorability problem. An indication in this direction could be seen in the density plot for the three memorability scores where the AMNet curve has a very similar shape compared to short-term memorability (although slightly "shifted" to the left.)

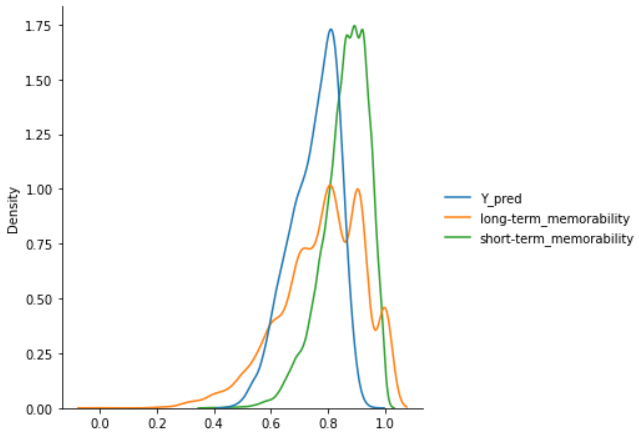

Figure 4: Density curve of the AMNet predictions (1st column in Table 1) along with ground truth density curves for long- and short-term memorability

DISCUSSION AND OUTLOOK

In this working note we presented AMNetVid - a signal-processing inspired extension of AMNet for videos which provides 4% and 8% improvement in the predictions for short- and long-term memorability respectivelt.

AMNetVid operates solely on attention map sequences. Not utilizing additional features like aesthetics, emotion, color histograms, etc. makes the model a good potential fit for incorporating in larger ensemble models.

Another area of interest is stacking the model on top of AMNet, so that it is fit directly on the internal state which is more information rich than attention maps.

Quantitatively expanding the existing architecture, although more computationally intensive, is also possible:

a) Longer input sequences, multiple sequences per video
b) Deeper neural network architecture
c) Increased input image resolution (up to 224x224 supported by ResNet/AMNet)
d) Data augmentation during training
e) Comparison with a 3D convolutional model

REFERENCES

[1] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Hélène Demarty, Ngoc Q. K. Duong, Xavier Alameda-Pineda, and Mats Sjöberg, 2019. The Predicting Media Memorability Task at MediaEval 2019. In *Proc. of the MediaEval 2019 Workshop, Sophia Antipolis, France,* 2019.

[2] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino, 2018. AMNet: Memorability Estimation with Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pages 6363-6372.

[3] Shuai Wang, Linli Yao, Jieting Chen, and Qin Jin. 2019. RUC at MediaEval 2019: Video Memorability Prediction Based on Visual Textual and Concept Related Features. In *Proc. of the MediaEval 2019 Workshop, Sophia Antipolis, France,* 2019.

[4] Alexander Viola and Sejong Yoon. 2019. A Hybrid Approach for Video Memorability Prediction. In *Proc. of the MediaEval 2019 Workshop, Sophia Antipolis, France*, 2019.

[5] Le-Vu Tran, Vinh-Loc Huynh, and Minh-Triet Tran. 2019. Predicting Media Memorability Using Deep Features with Attention and Recurrent Network. In *Proc. of the MediaEval 2019 Workshop, Sophia Antipolis, France*, 2019.

[6] Joonatan Mänttäri, Sofia Broomé, John Folkesson, and Hedvig Kjellström. 2020. Interpreting Video Features: A Comparison of 3D Convolutional Networks and Convolutional LSTM Networks. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020.

[7] Xingjian Shi, Zhourong Chen, Hao Wang, and Dit-Yan Yeung. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, Vol. 1, 2015, pages 802-810.

[8] A. Khosla, A. S. Raju, A. Torralba and A. Oliva, 2015. Understanding and Predicting Image Memorability at a Large Scale. *International Conference on Computer Vision (ICCV)*, 2015. DOI 10.1109/ICCV.2015.275

[9] Isola, P., Xiao, J., Torralba, A., and Oliva, A, 2011. What makes an image memorable? *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. Pages 145-152