A Hybrid Approach for Video Memorability Prediction

Alexander Viola, Sejong Yoon
The College of New Jersey, USA
{violaa1,yoons}@tcnj.edu

ABSTRACT

In this working note, we present our approach and investigation on the Predicting Media Memorability Task at MediaEval 2019. We used original video frames and caption data from the provided dataset, but extracted features ourselves using selected pretrained networks. Several combinations of pretrained models and recurrent networks were attempted to conduct a comparative study. Official results, as well as our investigation on the task data are provided. The best combination yielded a Spearman's correlation score of 0.455 on short-term task and 0.218 on long-term task in the test set.

1 INTRODUCTION

MediaEval 2019 Predicting Media Memorability [2] is a continuing multimedia analysis task following up from previous years of media memorability [1] and interestingness prediction challenges [3]. The dataset and evaluation setup are identical to the previous year's. It consists of two subtasks. In the first task, the system should predict whether the viewer will remember a video in the short-term (minutes). The second subtask was for the system to predict whether the viewer will remember a video in the long-term (24-72 hours). Within the total of 10,000 videos that were annotated, 8,000 of them were provided as devset, and the remaining 2,000 videos were reserved for the test-set. Details of the annotation protocol and the prior work survey can be found in the task overview paper [2].

2 APPROACH

In our prior attempt [14], we tried to directly optimize a large neural network combining various features using early fusion strategy. Based on the lessons learned, we aim to find a good combination of existing pretrained network models that covers multiple semantic levels of the data. To achieve this goal, we investigate combinations of four building blocks: (a) a recurrent neural network [11] to capture sequential nature of the video data, (b) an image-based memorability prediction model to take high level, task-relevant semantic information into account, (c) a convolutional neural network-based pretrained model to make sure we are not missing any fundamental lower level information, and (d) video captions that readily contain the highest semantic of the data. Encouraged by the high prediction performance reported by other teams [6, 13], we examined the use of following features in the process of developing our RNN-based model. We used classic RNNs, as we could not find significant differences in using LSTM [8] potentially due to our one layer RNN model design [5, 9].

ResNet. ResNet [7], initially trained as an image classification model, generates rich penultimate-layer output pertinent to the semantic content of the image. We extracted ResNet output of three

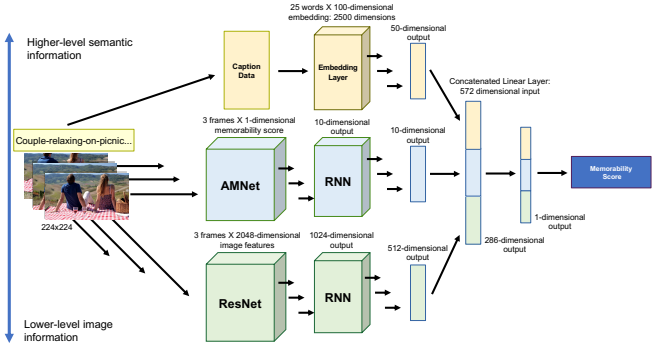

Figure 1: Proposed hybrid fusion framework for our comparative study on video memorability prediction. We used three frames from each video to extract ResNet [7] features and seven frames for AMNet [4] features.

frames (first, middle, and last of each video). These visual features are used as sequential input into a one-layer RNN, whose output is funneled through a series of fully connected layers before it either regresses to a memorability score (in models where it alone is used) or where it is part of a concatenation that contributes to a memorability score (in an ensemble model).

AMNet. Compared to the ResNet, AMNet [4] captures relatively higher-level information of the data. We extracted final image memorability scores of each frame in a given video, seven of which are supplied as sequential input to a respective RNN. Like ResNet, this feature slips through a number of linear layers before contributing to a final regression. We put one additional fully connected layer after RNN to improve generalization performance.

Caption. As a high-level semantic feature, we considered the videos caption data. As reported by multiple teams last year, caption data seems to be effective in video memorability prediction tasks [1, 6, 14]. We had technical issues while implementing RNN-based model using the caption data. For the interest of task schedule, we resorted to discarding the RNN-based approach using the caption feature in favor of creating a simple 25x100-dimensional word embedding feature that funneled through subsequent linear layers. This was functional, but results were far worse than the pretrained Word2Vec word embeddings [10] we employed last year [14].

Hybrid Fusion with Recurrent Network. To combine the features with different level of semantic information, we stacked all features into a single vector and plug it into a linear penultimate layer, that will regress to the ground truth memorability score. The training was done in an end-to-end fashion, and we found that, for feature combinations with RNN, RMSProp [12] seems to be the most effective optimization method. We also tested models without RNN for comparison, and found that Nesterov's stochastic gradient descent is the most effective method. For all experiments, we used initial learning rate of 0.001, batch size was 20, and trained for 35 epochs.

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

Table 1: Short term video memorability prediction results of all combinations evaluated on dev-set. Upper half of the table used RNN and the bottom half did not use RNN. Caption feature is treated as a static feature.

Method	5-fold Cross Validation	
	Spearman's ρ	Pearson's ρ
ResNet (run1)	0.464	0.491
AMNet	0.421	0.454
ResNet+AMNet (run2)	**0.470**	**0.503**
ResNet+Caption	0.352	0.384
AMNet+Caption	0.307	0.338
ResNet+AMNet+Caption	0.443	0.480
ResNet	0.362	0.397
AMNet	0.422	0.453
Caption	0.151	0.085
ResNet+AMNet	0.374	0.408
ResNet+Caption	0.370	0.398
AMNet+Caption	0.214	0.152
ResNet+AMNet+Caption	0.374	0.405
Caption [14] (2-fold CV)	0.450	0.472

Table 2: Long term video memorability prediction results of all combinations evaluated on dev-set. Upper half of the table used RNN and the bottom half did not use RNN. Caption feature is treated as a static feature.

Method	5-fold Cross Validation	
	Spearman's ρ	Pearson's ρ
ResNet (run1)	**0.226**	0.236
AMNet	0.213	0.231
ResNet+AMNet (run2)	0.225	**0.243**
ResNet+Caption	0.088	0.098
AMNet+Caption	0.098	0.107
ResNet+AMNet+Caption	0.217	0.232
ResNet	0.197	0.212
AMNet	0.216	0.232
Caption	0.060	0.044
ResNet+AMNet	0.201	0.217
ResNet+Caption	0.192	0.206
AMNet+Caption	0.097	0.079
ResNet+AMNet+Caption	0.198	0.215
Caption [14] (2-fold CV)	0.159	0.176

3 DISCUSSION AND OUTLOOK

Table 1 and Table 2 summarize our preliminary experiment results on dev-set. We found two observations. First, it is obvious that feature combinations with RNN significantly outperforms non-RNN counterparts in most cases. This is a clear evidence that the utilization of recurrent models is beneficial for video memorability prediction tasks. Second, ResNet features are found pretty powerful in both tasks, and often better than AMNet, that was trained for predicting memorability scores, although it is possible to expect slight synergistic improvement in cases.

Table 3: Official results on test-set. Top two rows are short term task results and bottom two rows are long term task results. It is clear that image-based memorability prediction model is not helpful for long term prediction task.

Method	Testset Official Result	
	Spearman's ρ	Pearson's ρ
ResNet (run1)	0.454	0.475
ResNet+AMNet (run2)	**0.455**	**0.493**
ResNet (run1)	**0.218**	**0.229**
ResNet+AMNet (run2)	0.177	0.197

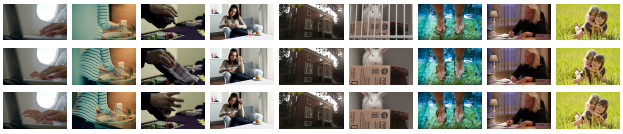

Figure 2: Three frames from each of the three videos from (left) most difficult, (middle) moderate, and (right) easiest to predict the short term memorability ranking for the best-performed combination (AMNet + ResNet w/ RNN).

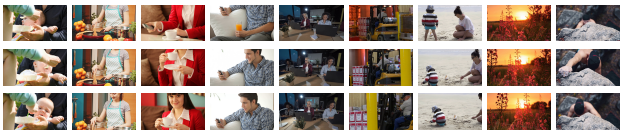

Figure 3: Three frames from each of the three videos from (left) most difficult, (middle) moderate, and (right) easiest to predict the long term memorability ranking for the best-performed combination (ResNet w/ RNN).

Table 3 shows our two best-performed models' official results on test-set. It is a well-known fact that long term memorability has different intrinsic characteristics compared to short term memorability, hence the ResNet + AMNet w/ RNN's bad performance in long term memorability task is not surprising. To obtain a better understanding of the results, we sorted all videos in dev-set, based on the absolute differences between predicted and ground truth ranking of memorability scores. Figure 2 and Figure 3 show the result of this analysis. Interestingly, it is quite clearly visible that the semantics conveyed by videos in the two tables are almost opposite, i.e. videos that have easily-predictable short term memorability scores have very similar semantic context (visible frontal human faces) as videos that have difficult-to-predict long term memorability scores, vice versa.

In the future, it would be very interesting to investigate the similarities and differences between intrinsic semantic characteristics of short and long term videos memorabilities.

ACKNOWLEDGMENTS

This work was supported in part by The College of New Jersey under Mentored Undergraduate Summer Experience (MUSE) 2019. The authors acknowledge use of the ELSA high performance computing cluster at The College of New Jersey for conducting the research reported in this paper. This cluster is funded by the National Science Foundation under grant number OAC-1828163.

REFERENCES

[1] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. MediaEval 2018: Predicting Media Memorability. In *Proc. of MediaEval 2018 Workshop, Sophia Antipolis, France, Oct. 29-31, 2018.*

[2] 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 *Proc. of MediaEval 2019 Workshop, Sophia Antipolis, France, Oct. 27-29, 2019.*

[3] Claire-Hélène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Michael Gygli, and Ngoc Q. K. Duong. Predicting Media Interestingness Task at MediaEval 2017. In *Proc. of MediaEval 2017 Workshop, Dublin, Ireland, Sept. 13-15, 2017.*

[4] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. AMNet: Memorability Estimation With Attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[5] A. Graves, A. Mohamed, and G. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

[6] Rohit Gupta and Kush Motwani. Linear Models for Video Memorability Prediction Using Visual and Semantic Features. In *Proc. of MediaEval 2018 Workshop, Sophia Antipolis, France, Oct. 29-31, 2018.*

[7] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521 (27 May 2015), 436. https://doi.org/10.1038/nature14539

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). arXiv:1301.3781 http://arxiv.org/abs/1301.3781

[11] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323, 6088 (1986), 533–536. https://doi.org/10.1038/323533a0

[12] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning. (2012).

[13] Duy-Tue Tran-Van, Le-Vu Tran, and Minh-Triet Tran. Predicting Media Memorability Using Deep Features and Recurrent Network. In *Proc. of MediaEval 2018 Workshop, Sophia Antipolis, France, Oct. 29-31, 2018.*

[14] Aaron Weiss, Benjamin Sang, and Sejong Yoon. Predicting Memorability via Early Fusion Deep Neural Network. In *Proc. of MediaEval 2018 Workshop, Sophia Antipolis, France, Oct. 29-31, 2018.*