Predicting Memorability via Early Fusion Deep Neural Network

Aaron Weiss*, Benjamin Sang*, Sejong Yoon
The College of New Jersey, USA
{weissa7,sangb1,yoons}@tcnj.edu

ABSTRACT

In this working note, we present our approach and investigation on the MedialEval 2018 Predicting Media Memorability Task. We used a portion of the features provided, while also employed additional features. Two different training approaches were attempted to train a deep neural network architecture, fusing multiple features we used. Official results, as well as our investigation on the task data are provided.

1 INTRODUCTION

MediaEval 2018 Predicting Media Memorability [4] is a new multimedia analysis task following up from previous years of media interestingness prediction challenges [6]. 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 dev-set, 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 [4].

2 APPROACH

In this section, we first describe the features we employed and then present our method.

2.1 Features

We used many of the provided features, including Aesthetic visual features[7], the final classification layer of the C3D[15] model, Color Histogram in HSV space, Histogram of Motion Patterns[1], and the outputs of the $fc7$ layer of the InceptionV3[14] deep neural network. We also employed two additional features:

Image Memorability Prediction. Extracted three frames from every video, at the time stamps 0.5, 3.0, and 5.5 seconds. For 7 second videos, this results in good coverage of the entire video in the case of rapidly changing scenes. Then, we used MemNet [10] image memorability prediction model to extract image memorability scores of the three frames. Finally, the three prediction scores were averaged as a memorability score prediction for the entire video.

Caption. Following a prior work [5], we considered utilizing caption data provided in the dataset. Given the textual metadata per video, we generated a feature vector using Google's Word2Vec [12] model. This yields a 300-dimensional vector for each word within the provided video caption. Then, the vectors in each video were averaged to create one vector per video as a feature.

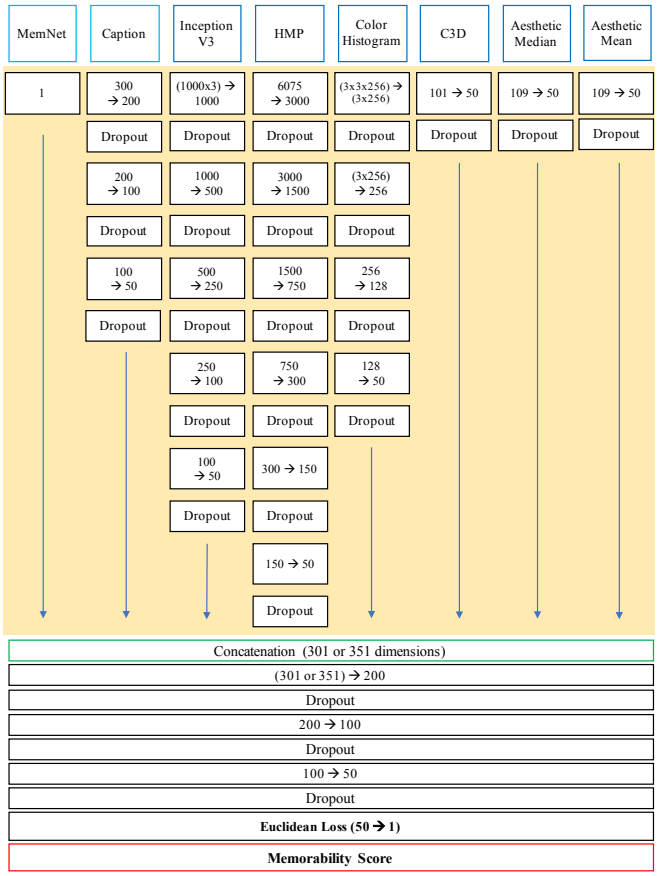

Figure 1: Our deep neural network structure. Each fully connected layer used ReLU [13] for the activation function. All dropout layers used 0.5 for the drop rate. Concatenation will make the fused features into a 301 (without caption) or 351 (with caption) dimensional vector. To speed up the training, we attempted to pre-train the lower layers of the network (shaded in yellow). Please refer the text for the details.

2.2 Feature Fusion via Concatenation

Given the described features, the key task is to find the best combination/subset of the features that correlates well to the video memorability score. In this work, we tried deep neural network stacking multiple fully connected layers with modern regularization techniques. Fig. 1 depicts our network structure.

Our network design focused on two aspects: (a) include sufficient number of layers for input features with high dimensions

*A. Weiss and B. Sang equally contributed to this work.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

Table 1: Result of all submissions and additional experiments.

Subtask	Method	Initial Learning Rate	Epochs	2-fold Cross Validation			Testset Official Result		
				Spearman's ρ	Pearson's ρ	MSE	Spearman's ρ	Pearson's ρ	MSE
Short	A (run1)	0.05	100	0.421	0.448	0.010	0.284	0.190	0.007
	A (run2)	0.001	300	0.400	0.410	0.010	0.287	0.130	0.010
	B (run3)	0.001	100	0.285	0.247	0.010	0.310	0.316	0.006
	B (run4)	0.001	200	0.378	0.392	0.010	**0.340**	0.363	0.006
	B (run5)	0.001	300	0.375	0.391	0.010	0.338	0.345	0.006
	C	0.001	200	0.450	0.472	0.010	-	-	-
Long	A (run1)	0.05	100	0.131	0.125	0.020	0.074	0.039	0.025
	A (run2)	0.001	300	0.109	0.105	0.030	0.078	0.026	0.025
	B (run3)	0.001	100	0.078	0.081	0.030	0.086	0.090	0.022
	B (run4)	0.001	200	0.081	0.085	0.030	0.090	0.095	0.024
	B (run5)	0.001	300	0.085	0.094	0.020	**0.093**	0.097	0.023
	C	0.001	300	0.159	0.176	0.020	-	-	-

Table 2: Results using each feature we employed (5-fold cross validation on the dev-set). MemNet features are already memorability scores, so we calculated Spearman's ρ directly using the extracted MemNet scores and the ground truth video memorability scores over the whole dev-set.

Feature	Spearman's ρ	
	Short	Long
Aesthetic Mean	0.2754	0.1287
Aesthetic Median	0.2782	0.1191
C3D	0.2960	0.1269
HMP	0.2212	0.0744
Color Histogram (HSV)	0.3146	0.1078
InceptionV3	0.0960	0.0354
Caption	0.4638	0.2020
MemNet	0.4029	0.2022

so that subtle but important variations are not ignored and (b) all features are equally treated, and important but small dimensional features are not overwhelmed by the other large features. Each linear weight followed by a ReLU [13] activation function and a dropout regularization [8]. We used 0.5 for all dropout rates. The network hyperparameters were determined by preliminary experiments and Table 2 summarizes the results using each feature individually. Several methods have been proposed for feature fusion in deep neural networks, particularly for convolutional neural nets [2, 3]. After some preliminary trials, we decided to use the simple concatenation as no significant difference was found.

2.3 Pre-training Layers

One of the well-known issues of the deep neural network training is the vanishing gradient problem [9]. While we used ReLU to alleviate the problem, we found that our network in Fig. 1 easily get stuck during the training. To speed up the training, borrowing the idea from transfer learning, we pre-trained the lower layers before the concatenation. We denote the network without pre-training as model A and the one with pre-training as model B. As evident from our final result in Table 1, this improved the performance of the model in the test-set. We ran 100 epochs for the pre-training.

3 DISCUSSION AND OUTLOOK

Overall results on our submissions are summarized in Table 1. We used ADAM for the optimization [11] and for most of the cases, we used the default learning rates of 0.001. Due to schedule constraints, we did not include Caption features in the submitted methods A and B. We report cross validation result including the Caption feature with the best-performing configuration (with pre-trained layers) as method C. It is clear from the result, that the pre-training approach B showed more balanced generalization performance, regardless of dev-set/test-set split. Moreover, B shows consistent performance improvement over increasing training epoch, indicating that the model is being trained in the right direction.

On the downside, several challenges were identified. First, the performance of the feature-fused network did not improve much over individual features. Only when high level information, e.g. caption, is involved, we reached the baseline performance. As reported in [5], it is clear that high level pre-processing is essential to achieve a reasonable performance. One may consider late fusion instead of early fusion, for some of the features we considered, e.g. MemNet. Second, long-term video memorability is more difficult to predict than the short-term one. From our experiments, it was unclear which feature, would improve the long-term video memorability prediction as all of them yielded poor performance. Even the high level semantic features struggled in this case. This is not surprising given the true long-term memorability scores are 1 (memorable for all annotators) in many cases. More robust prediction model, that can distinguish subtle differences might be needed for this subtask.

ACKNOWLEDGMENTS

This work was supported in part by The College of New Jersey under Support Of Scholarly Activity (SOSA) 2017-2019 grant. 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] J. Almeida, N. J. Leite, and R. da S. Torres. 2011. Comparison of video sequences with histograms of motion patterns. In *2011 18th IEEE International Conference on Image Processing*. 3673–3676. https://doi.org/10.1109/ICIP.2011.6116516

[2] E. Boyaci and M. Sert. 2017. Feature-level fusion of deep convolutional neural networks for sketch recognition on smartphones. In *2017 IEEE International Conference on Consumer Electronics (ICCE)*. 466–467. https://doi.org/10.1109/ICCE.2017.7889398

[3] Chen, Shi-Qi, Zhan, Rong-Hui, Hu, Jie-Min, and Zhang, Jun. 2017. Feature Fusion Based on Convolutional Neural Network for SAR ATR. *ITM Web Conf.* 12 (2017), 05001. https://doi.org/10.1051/itmconf/20171205001

[4] 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*.

[5] Romain Cohendet, Karthik Yadati, Ngoc Q.K. Duong, and Claire-Hélène Demarty. 2018. Annotating, understanding, and predicting long-term video memorability. In *Proc. of the ICMR 2018 Workshop, Yokohama, Japan, June 11-14*.

[6] 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*.

[7] Andreas F Haas, Marine Guibert, Anja Foerschner, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A Sandin, Jennifer E Smith, Mark JA Vermeij, and others. 2015. Can we measure beauty? Computational evaluation of coral reef aesthetics. *PeerJ* 3, e1390 (2015).

[8] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR* abs/1207.0580 (2012). arXiv:1207.0580 http://arxiv.org/abs/1207.0580

[9] Sepp Hochreiter. 1991. Untersuchungen zu dynamischen neuronalen Netzen. *Diploma, Technische Universität München* 91, 1 (1991).

[10] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and Predicting Image Memorability at a Large Scale. In *International Conference on Computer Vision (ICCV)*.

[11] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980 http://arxiv.org/abs/1412.6980

[12] 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

[13] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10)*. Omnipress, USA, 807–814. http://dl.acm.org/citation.cfm?id=3104322.3104425

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. IEEE Computer Society, Washington, DC, USA, 4489–4497. https://doi.org/10.1109/ICCV.2015.510