RUC at MediaEval 2018: Visual and Textual Features Exploration for Predicting Media Memorability

Shuai Wang, Weiying Wang, Shizhe Chen, Qin Jin
Renmin University of China, Beijing, China
shuaiwang@ruc.edu.cn,wy.wang@ruc.edu.cn,cszhe1@ruc.edu.cn,qjin@ruc.edu.cn

ABSTRACT

Predicting the memorability of videos has great values in various applications including content recommendation, advertisement design and so on, which can bring convenience to people in everyday life, and profit to companies. In this paper, we present our methods in the 2018 Predicting Media Memorability Task. We explored some deeply-learned visual features and textual features in regression models to predict the memorability of videos.

1 INTRODUCTION

The MediaEval 2018 Predicting Media Memorability Task [4] aims to predict what kind of media is memorable for people, which has a wide range of applications such as video retrieval, video recommendation, advertisement design and education system. We explored visual and textual representation for videos and built a regression model which can calculate a memorability score for a given video.

2 APPROACH

2.1 Framework

In general, we utilize a regressor to predict the memorability score of each video and consider late fusion to combine different features. Two kinds of fusion strategies are utilized, namely score average and second-layer regression.

Our system framework is shown in Figure 1. We firstly run regressions to get videos' memorability scores using different single features. In order to fuse multiple features, two strategies are considered and shown in Figure 1. For the score average strategy, we average the scores of different types of features from the same video, and the obtained score is the final memorability score of this video. For the second-layer regression, we concatenate the scores of different features from the same video as second-layer features, and the obtained second-layer features are fed into a second-layer regressor, which will predict the final memorability scores.

2.2 Features

The videos are soundless, so we focus on visual and textual features, especially some high-level and semantic features.

The captions of videos are short with only a few words. We argue that people may be impressed by some particular objects or their combinations. The meanings of each word should be embedded into the representations of sentences for the memorability prediction.

A pre-trained word embedding contains a large amount of semantic information, contributing to encoding the meaning of sentences.

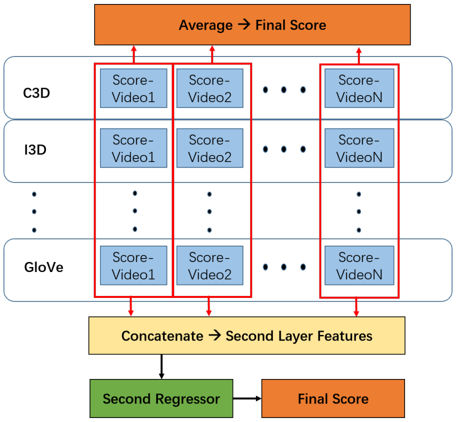

Figure 1: Two strategies of late fusion

We try the word embedding GloVe [8] as the textual feature. We combine the embedding of each word to generate the representation of sentences in different ways. Firstly, we simply add them up and take average of each dimension. Secondly, we take smooth IDF [2] as the weight for each word. Thirdly, we try the pre-trained skip-thought model [7]. And fourthly we also try ConceptNet [9]. Through these four methods we can obtain different types of video-level representations.

For visual features, we consider some deeply-learned representations and aesthetic descriptors as our features, including C3D [5], HMP [1], I3D [3] and aesthetic [6]. The C3D, HMP and aesthetic are officially provided in this task. Further more, we extract the I3D-RGB feature, which is obtained from the penultimate layer in RGB branch in I3D.

Additionally, we add some label information. If we are solving the long-term task, we firstly train a model with short-term labels and then use this model to predict the short-term scores of the test set. Then we transform the short labels of both train and test sets into 10 dimensional one-hot vectors. The 10 buckets of a one-hot vector denote the range between 0 to 1 with step 0.1. If a label is in the range of a bucket, e.g. the label is 0.56 and it is in the range of 0.5 to 0.6, we set the value of this bucket as 1 and the rest of buckets are set to 0. And then we add them to the end of each text feature. For short-term task, we map the long-term labels into one-hot vectors and use them in the same way mentioned above.

3 EXPERIMENTS AND ANALYSIS

3.1 Experimental Setup

The development set and the test set contain 8000 and 2000 videos respectively. We firstly rank the videos by their memorability scores and sample videos with a constant step value of 4. Finally we split

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

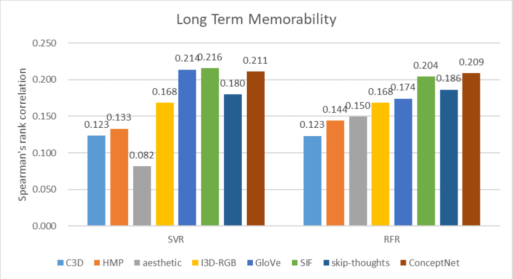

Figure 2: Results of different features for long-term memorability on the local test set

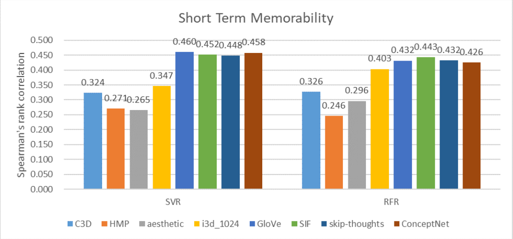

Figure 3: Results of different features for short-term memorability on the local test set

the development set into 2 parts, namely 6000 videos in train set and 2000 videos in local test set.

We simply consider two types of regressors, namely Support Vector Regression (SVR) and Random Forest Regression (RFR). The parameters were determined by grid searches. The Penalty parameter C in SVR is searched from 0.125 to 32. The parameters of n_estimators and max_depth are searched in the range [100, 1000] with step 100 and [2, 10] with step 2 respectively. The I3D model is pre-trained on ImageNet and Kinetics.

Table 1: Results of different features for long-term memorability on the test set

	all-m	visual-m	text-m	all-s	required
Spearman	0.2374	0.1875	0.2352	0.2404	0.2404
Pearson	0.2584	0.2072	0.2565	0.2621	0.2621
MSE	0.0197	0.0206	0.0198	0.0199	0.0200

Table 2: Results of different features for short-term memorability on the test set

	all-m	visual-m	text-m	all-s	required
Spearman	0.4464	0.3547	0.4383	0.4483	0.4484
Pearson	0.4957	0.3675	0.4881	0.4961	0.4961
MSE	0.0075	0.0108	0.0065	0.0080	0.0082

3.2 Results and Analysis

The results of each single feature for long-term and short-term memorability prediction are printed in Figure 2 and Figure 3 respectively. As shown in Figure 2 and Figure 3, textual representations are on the same level and textual features perform better than visual representations. We think that the captions contain more clear descriptions about the elements in the videos. If a specific object is depicted by a word, the word embeddings can describe the relations of this object and others in the whole environment. The visual features may contain some details of regions but not that intuitive. If there is no caption information, object detection and classification techniques may offer more supports.

Table 1 and Table 2 show the results on the test set, *m* and *s* means the score average and the second-layer regression strategy respectively. *all* means fusing all features, namely visual, textual and labels. *visual* denotes fusing visual representations and *text* is the fusion of all word embedding features. *required* means using average strategy and not using the label information. We used average strategy in required runs because average strategy performed generally better than second-layer regression on the local test set. We can notice that the required runs in long-term and short-term task both have the best performances. Label information helps little on local test set and does not work in official test set. We consider that maybe mapping labels into one-hot vectors is not a proper way to fully utilize the label information and it is worthy to find a proper representation format of the labels or a fusion method with other features.

We pick out a number of videos for analysis and we find that some of them depict close-ups of objects or regions, while some of them show overall scenes such as natural landscape, stories of some characters.

We draw 3 conclusions after viewing these videos and their labels.

(1) The videos with low short-term labels usually have low long-term labels.
(2) The videos with high short-term labels and low long-term labels usually depict some close-ups.
(3) There are few numbers of videos with low short-term labels and high long-term labels. These videos generally have open and wide scenes.

We find that it is difficult to predict the memorability of the videos in the second and third situations.

In sum, we consider that if a video is memorable in a long term, it is also memorable in a short term generally. Conversely, videos with high short-term labels cannot determine the long-term memorability.

4 CONCLUSION

In conclusion, we explored visual and textual representations for videos and built a regression model which can calculate a memorability score for a given video. The results show that textual representations perform better than visual features. In the future, we will focus on the visual semantic representations and object detection related works to find more interesting methods to predict memorability of videos. And how to use label information is another interesting point to be explored.

ACKNOWLEDGMENTS

This work is supported by National Key Research and Development Plan under Grant No. 2016YFB1001202. This work is partially supported by National Natural Science Foundation of China (Grant No. 61772535).

REFERENCES

[1] Jurandy Almeida, Neucimar J. Leite, and Ricardo Da S. Torres. 2011. Comparison of video sequences with histograms of motion patterns. In *IEEE International Conference on Image Processing*. 3673–3676.

[2] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*.

[3] J. Carreira and A. Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 00. 4724–4733. https://doi.org/10.1109/CVPR.2017.502

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

[5] Tran Du, Lubomir Bourdev, Rob Fergus, and Lorenzo Torresani. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *IEEE International Conference on Computer Vision*. 4489–4497.

[6] Andreas F. Haas, Marine Guibert, Anja Foerschner, Tim Co, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A. Sandin, and Jennifer E. Smith. 2015. Can we measure beauty? Computational evaluation of coral reef aesthetics. *Peerj* 3, 12 (2015), e1390.

[7] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 3294–3302.

[8] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[9] Robert Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. 4444–4451. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972