RUC at MediaEval 2019: Video Memorability Prediction Based on Visual Textual and Concept Related Features

Shuai Wang, Linli Yao, Jieting Chen, Qin Jin
School of Information, Renmin University of China, Beijing, China
shuaiwang@ruc.edu.cn,yaolinliruc@gmail.com,jietingchen1208@gmail.com,qjin@ruc.edu.cn

ABSTRACT

Memorability of videos has great values in different applications such as education system, advertising design and media recommendation. Memorability automatic prediction can make people's daily life more convenient, and bring companies profit. In this paper, we present our approaches in The Predicting Media Memorability Task at MediaEval 2019 . We explored some visual, textual and artificially designed concept related features in regression models to predict the memorability of videos.

1 INTRODUCTION

The MediaEval 2019 Predicting Media Memorability Task [2] aims to find out what type of video is memorable , namely how likely it is that the video can be remembered after people watching them. This problem has a wide range of applications such as video retrieval and recommendation, advertising design and education system. We explored some visual, textual and artificially designed concept related features in regression models to predict the memorability of videos.

2 APPROACH

Generally, we concentrate on visual features extracted from videos and textual features drew from given textual metadata. Among visual features, we consider both the visual information in a frame and the temporal factors between successive frames. In addition, we use deep network to extract high-level and semantic feature representation. Based on each individual extracted feature, we then do feature normalization. Further, we perform feature fusion to get better performance. Finally, we consider two simple but efficient regressors called Support Vector Regression (SVR) and Random Forest Regression (RFR) to get final memorability scores.

2.1 Base Features

In addition to the eight video special features provided by the official benchmark, we try to extract other new features that may be related to video memorability. We try to extract high-level representation of videos with DenseNet[9] and ResNet[8] pre-trained on ImageNet [3], respectively. Detailedly, we extract 11 frames from each video as input images and the DenseNet169 will output features with 1664 dimension. Then we combine features of 11 frames to generate video-level representation in two ways: simply taking the average, and using Gated Recurrent Unit(GRU)[1] which makes use of temporal information. The process of ResNet152 is similar and it outputs 2048 dimension features.

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

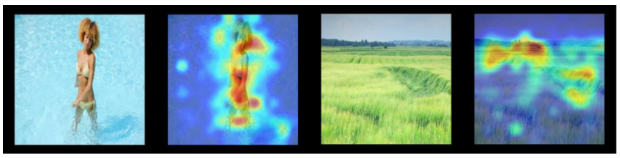

Figure 1: Examples of attention maps for high and low memorability images from videos in the dataset of MediaEval 2019. Long-term scores: left picture 1.0, right picture 0.3. Short-term scores: left picture 0.928, right picture 0.898.

The origin title of each video summarizes inclusion objects and events briefly. We try some popular word embedding models to get textual features from these captions, including GloVe[11], ConceptNet[12] and Bert[4]. We add the embedding of each word up and take average of each dimension to obtain the representation of a whole sentence.

2.2 AMNet

We find that when people watch videos, they do not pay equal attention to each region in the scene, but first focus on a certain area, which may change over time. And we learn from Baveye et al.[7] that still image regions quickly attracting us are closely related to the highly memorized areas. Therefore, we draw on the idea and directly apply the AMNet[7] to our task. AMNet is an end-to-end architecture with a Soft Attention Mechanism and a Long Short Term Memory (LSTM) recurrent neural network for memorability score regression. Moreover, AMNet uses transfer learning and is evaluated on the LaMem datasets, consequently extending our task's datasets. And this contributes to predicted memorability scores scattering in a larger scale, which is much closer to the distribution of ground truth.

Specifically, we fine-tune AMNet on the dataset of MediaEval 2019, training the long-term and short-term sub-tasks separately. Considering that AMNet is designed for still images, we extract 11 frames at a uniform time interval for each video as input. As for prediction, we take the median memorability score of 11 frames as the final result. As in figure 1, we can visually observe that the output attention maps of video frames are closely related to the highly memorable visual contents in the picture.

2.3 Concept

Generally, people have a preference for paying attention to different concepts. According to [6], most of the entities could be covered by 7 concepts: animals, building, device, furniture, nature, person, and vehicle. Among these 7 concepts, animals, person and vehicle are highly memorable. Inspired by this, we use the 7 concepts to make analyses on our caption data. We extracted meaningful entities from the captions by filtering out stop words and keeping nouns.

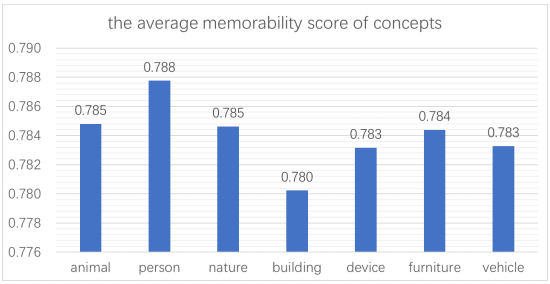

Figure 2: The average memorability scores of 7 concepts based on our caption corpus. Among these 7 concepts, person, animal and nature get higher memorability scores.

To find out whether the idea makes sense on our data, we counted the number of entities belonging to each concept. Then we take the average of the memorability scores of the videos corresponding to these concepts. The result is shown in Figure 2, showing that the preference on concepts also affects the memorability of the video to some extent.

Then, with the help of GloVe word vector pre-trained on Common Crawl data, we calculate the distance between entities and the above 7 concepts. For each entity, we can get a distance vector with 7 elements which can represent the correlation between the entity and each concept. For each caption, we take the average of the distance vectors of all the entities the caption contains, so that we get a feature vector. Then we apply a random forest regression on the feature vectors. The Spearman score on long-term memorability prediction is 0.11. This result, based solely on the textual manual features shows that the concepts of entities in videos is meaningful for predicting memorability of the videos.

We made further exploration in this direction. [10] claims that when people focus and memorize, they will pay more attention to the concepts they are familiar with. Hence we find some familiar word lists in Wikipedia and pick a list called dolch word [5] containing 156 concepts after filtering out certain parts of speech. Specifically, we replace our 7 concepts with these 156 concepts and generate the feature vectors of each caption. This time we got a Spearman score 0.15 on the long-term memorability. The result is promising for us to consider fusing concept features into the entire model.

3 RESULTS

We split the develop set into two parts, namely the training set and the validation set. We train and test on these two sets and determine the final methods according to the performances on validation set, finally the models are trained on the whole develop set and predict on the official test set. The results on validation set and official test set are shown in Table 3 and Table 2 respectively.

In Table 1, Table 3 and Table 2, "Base1" means the early fusion of DenseNet169, GloVe and C3D features, while "Base2" additionally includes ConceptNet. The "Base1" and "Base2" are the best early fusion strategies on validation set. The 'AM' is AMNet scores mentioned above and 'Dist' denotes the scores from concept distances. The plus sign means late fusion and we apply a set of weights on them empirically, which is "Base * 0.9 + AM * 0.1" and "Base * 0.6 + Dist * 0.4"

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

	Base2	AM	Dist	Base2+AM	Base2+Dist
Spearman	0.2551	0.2116	0.1534	0.2588	0.2587

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

	Base2	Base2+AM	Base1	Base1+AM	Base2+Dist
Spearman	0.196	0.213	0.198	0.216	0.211
Pearson	0.215	0.235	0.216	0.236	0.235
MSE	0.02	0.07	0.02	0.08	0.07

Table 3: Results of different features for short-term memorability on the official test set

	Base2	Base2+AM	Base1	Base1+AM	Base2+Dist
Spearman	0.436	0.466	0.446	0.472	0.470
Pearson	0.493	0.520	0.503	0.526	0.523
MSE	0.01	0.06	0.01	0.06	0.07

4 ANALYSIS AND DISCUSSION

Based on our previous experience, the deep CNN features and caption embedding features are the most effective in the memorability prediction task, such as DenseNet169 and GloVe word embeddings in our experiments. In addition, we also consider some other features to study whether there are some complementary points and pick out two combinations as "Base1" and "Base2". It's easy to remember familiar things for us, so we consider there are a fuzzy and a clear way to represent these things. AMNet can automatically pay attention to a object or an area that may attract us, and this is like a fuzzy representation, because it does not show the concept directly. The clear way is the concept distances which depict the distance map of the current video. The late fusion of these two methods and the "base" boost the performaces slightly. We suppose that the "base" namely CNN features and caption embeddings are stable, and maybe the caption embeddings have already included some information about these concepts, so the improvement of results is not very obvious.

5 CONCLUSION

In conclusion, we design a model that uses visual and textual representations to predict the memorability scores of given videos. The results show that deep CNN and caption word embeddings are effective and the attention information from AMNet and semantic distance extracted from captions can boost the performance slightly. In the future, we will focus on the concept representation and semantic representations. Also the interaction of long term and short term ground-truth is a interesting point to be explored.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Plan under Grant No. 2016YFB1001202, Research Foundation of Beijing Municipal Science Technology Commission under Grant No. Z181100008918002 and National Natural Science Foundation of China under Grant No.61772535.

REFERENCES

[1] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[2] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Hélène Demarty, Ngoc Q. K. Duong, Xavier Alameda-Pineda, and Mats Sjöberg. 2019. Predicting Media Memorability Task at MediaEval 2019. *In Proc. of MediaEval 2019 Workshop, Sophia Antipolis, France, Oct. 27-29, 2019* (2019).

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Edward W Dolch. 1936. A basic sight vocabulary. *The Elementary School Journal* 36, 6 (1936), 456–460.

[6] Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem. 2015. What makes an object memorable?. In *Proceedings of the ieee international conference on computer vision*. 1089–1097.

[7] 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*. 6363–6372.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[10] Marvin Minsky. 2006. The Emotion Machine: Commonsense Thinking. *Artificial Intelligence, and the Future of the Human Mind, Simon & Schuster* (2006), 529–551.

[11] 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

[12] Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.