

Media Memorability Prediction Based on Machine Learning

Dazhan Xu, Xiaoyu Wu, Guoquan Sun

Communication University of China, China

dzxu@cuc.edu.cn

ABSTRACT

In the context of videos today, with billions of hours of user generated video content on online platforms like social media, prediction of a cognitive measure like memorability has many potential applications including content recommendation, advertisement design and so on, which can bring convenience to people in everyday life and profit to companies. This paper describes our approach designed for the MediaEval 2020 Predicting Media Memorability Task. Our approach uses the pre-extracted features and the features extracted through a deep mutil-level encoding network along with provided textual description to predict a probability-like memorability score of videos. We use the same set of features for predicting both short-term and long-term media memorability respectively.

1 INTRODUCTION AND RELATED WORK

The Predicting Media Memorability Task [1] is proposed at the MediaEval 2020. The goal of this task is to automatically predict a memorability score for a video reflecting its probability to be remembered. For this, it is provided an available dataset composed of 1500 videos, an initial development set of 590 videos and a supplemental development set of 410 videos with short-term and long-term memorability scores along with an official test set of 500 videos without ground-truth. We build a model using the training data and utilize the trained model to predict the short-term and long-term memorability of the test data.

People have an ability to remember and recall photos and videos with a surprising amount of detail. Interestingly, not all content is stored and recalled equally well. Media Memorability has attracted research interest recently in the area of Computer Vision [2, 3]. The authors in [2, 3] used spatio-temporal features to represent video dynamics and used a regression framework for predicting memorability. Previous attempt [4] at predicting image and video memorability discuss factors affecting memorability.

In this paper, we investigate the use of various visual and semantic features to predict video memorability. Among the features provided we train models over AlexNetFC7, HOG, LBP, HSVHist, RGBHist features and C3D-Predictions semantic feature. In addition, we train models over textual features which were extracted from BERT feature extractor [5], and over features extracted from a mutil-level encoding network [6]. The models are evaluated using Spearman's rank correlation as the metric.

2 APPROACH

In this section, we discuss the task of Predicting Media Memorability using video, image and textual features. We were provided with the development set's labels only. Statically we divided the development set into the training video data (800 videos), and 200 videos for the validation set, which each data sample is associated with its visual features and the corresponding memorability score. Our approach is to develop individual models per set of features provided or extracted and to then combine some features that seem to have a better performance. In addition, the short-term memorability and long-term memorability were modelled independently using the same set of features in our experiments.

2.1 Features

We used image level features: Color Histogram in HSV and RGB space, AlexNetFC7, HOG, LBP, VGGFC7 and video-level feature: C3D. For the image-level features were extracted from 3 frames for each video, the 3 frames per each video represent the first, the middle and the last frame in the movie. C3D feature was used to represent spatio-temporal aspect of a video. We also employed two additional features:

Multi-level encoding feature. For a given video, we extract uniformly a sequence of n frames with a pre-specified interval of 0.5 second. Per frame we extract deep features using a pertained ImageNet CNN, as commonly used for video content analysis. Consequently, the video is described by a sequence of feature vectors $\{v_1, v_2, \dots, v_n\}$, where v_t indicates the deep feature vector of the t -th frame. We utilize a multi-level encoding network [6] that encodes the CNN feature vectors to get the most representative feature. By jointly exploiting multi-level encodings, the network explicitly and progressively learns to represent global, local and temporal patterns in videos.

Caption. Following a prior work [2], we considered utilizing text description information provided in the development set. Given the textual metadata per video, we generated a feature vector using BERT model [5]. This yields a 768 dimensional vector for each text description of the provided videos.

2.2 Model description

Here, we describe our proposed model and provide the training details. The development set is randomly divided into 8:2 split for training and validation respectively. We simply consider three types of regressors, namely Gradient Boosting Regression (GBR), Random Forest Regression (RFR) and a mutil-layer perceptron neural network (MLP neural network). The input feature of each model was standardized. For the MLP neural network, we input each set of features to the first input layer of it, two hidden layer along with ReLU activation and a fully connected linear layer is

Table 1: Results of different features based on three regression models for short-term and long-term memorability on the validation dataset

Features	GBR		RFR		MLP	
	short-term	long-term	short-term	long-term	short-term	long-term
AlexNetFC7	0.220	0.124	0.254	0.175	0.197	0.168
C3D	0.130	-0.110	0.023	-0.224	0.005	0.133
HOG	0.195	0.198	0.138	0.137	0.118	0.062
HSVHist	0.102	0.124	0.193	0.151	0.240	0.199
LBP	0.071	0.100	0.196	0.145	0.117	0.034
RGBHist	0.153	0.111	0.210	0.088	-0.089	-0.050
VGG	0.197	-0.027	0.207	0.132	0.038	-0.027
Caption	0.202	0.037	0.161	0.053	0.217	0.139
Multi-level	0.254	0.203	0.232	0.207	0.392	0.224
Mutil-level+Caption	0.325	0.183	0.246	0.213	0.415	0.235

applied on it to obtain a single number representing the memorability score in the range from 0 to 1. The model is trained using Adam optimizer [7] with Mean Squared Error Loss function for 50 epochs on the training set.

We trained different models for both short-term and long-term memorability scores. For each set of features we tried each of these 3 models on the validation dataset and results are presented in Table1.

Table 2: Prediction results of combining Multi-level and Caption feature for short-term and long-term on the test set

	short-term	long-term
Spearman	0.06	0.049
Pearson	0.055	0.05
MSE	0.01	0.05

2.3 Results and Analysis

The results of each single feature for short-term and long-term memorability prediction on the validation dataset are shown in Table1. From Table 1, the multi-level encoding feature preforms better than other visual representations. We think that the network explicitly and progressively learns to represent global, local and temporal patterns in videos by jointly exploiting multi-level encodings. It means that the proposed method of feature representation is effective.

In addition, 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 embedding 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. So we consider combining the two features to predict the final memorability score.

We used the simple concatenation to combine the mutil-level encoding feature and the textual feature extracted utilizing BERT feature extractor [5]. We input the combination of the features to the MLP neural network to train a model and use the model to predict the memorability score. The prediction results on the test set are shown in Table 2. But the performance of the prediction results seems poor according to the official feedback. We think that since most of the features provided or extracted are very high

dimensional and the number of videos is less, over-fitting is a major potential concern in this task. Though we drop few data in a certain ratio, it seems not to work. Another possible reason is that the division of the training set and validation set is random, leading to the contingency of the results. Overall, our results seem poor for the above reason or because of insufficient tuning of parameter settings in our experiments.

3 CONCLUSIONS

In this work, we explored visual and textual representations for videos and built a few regression models which can calculate a memorability score for a given video. From our experiments, it was clear that multi-level feature would improve the short-term and long-term video memorability prediction, but it seemed poor performance.

Due to the new dataset which is more complicated and with more actions happening. In future, we plan to conduct the Video Memorability experiment with improved features like Dense Optical Flow features, Action based features representing the sequence of actions in the video, and also aim to leverage the audio in the videos.

ACKNOWLEDGMENTS

This work was supported in part by the Natural Science Foundation of China under Grant 61801441, in part by the Project of the Beijing National Research Center for Information Science and Technology of China under Grant BNR2019TD01022, in part by the Fundamental Research Funds for the Central Universities under Grant CUC19ZD003, Grant CUC2019B066, and in part by the State Key Laboratory of Media Convergence and Communication, Communication University of China.

REFERENCES

- [1] Alba Garc ía Seco De Herrera, Rukiye Savran Kiziltepe, Jon Chamberlain, Mihai Gabriel Constantin, Claire-H é ène Demarty, Faiyaz Doctor, Bogdan Ionescu, and Alan F. Smeaton. Overview of MediaEval 2020 Predicting Media Memorability task: What Makes a Video Memorable? Working Notes Proceedings of the MediaEval 2020 Workshop. December 14-15 2020.

- [2] Romain Cohendet, Karthik Yadati, Ngoc QK Duong, and Claire-Hélène Demarty. 2018. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM, 178–186.
- [3] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision*. 2730–2739.
- [4] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. 2013. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1633–1640.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). <http://arxiv.org/abs/1810.04805>.
- [6] Jianfeng Dong; Xirong Li; Chaoxi Xu; Shouling Ji; Yuan He; Gang Yang; Xun Wang 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [7] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. (2017). [arXiv:cs.LG/1412.6980](https://arxiv.org/abs/1412.6980).