Predicting Media Memorability Using Deep Features and Recurrent Network

Duy-Tue Tran-Van, Le-Vu Tran, Minh-Triet Tran
University of Science, Vietnam National University-Ho Chi Minh City
tvdtue@apcs.vn,tlvu@apcs.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

In the Predicting Media Memorability Task at the MediaEval Challenge 2018, our team proposes an approach that uses deep visual features and recurrent network to predict videos' memorability. Features are extracted from CNN for a number of frames in each video. We forward these through a LSTM network to model the structure of the video and predict its memorability score. Our method achieves a correlation score of 0.484 on short-term task and 0.257 on long-term task in the official test set.

1 INTRODUCTION

The Predicting Media Memorability task's main objective is to automatically predict a score which indicates how memorable a video will be [2]. Video's memorability can be affected by several factors such as: semantics, color feature, saliency, etc.

In this paper, we examine the sequential structure of videos with LSTM. We take advantage of deep convolutional neural networks to get image features as our main source of data for predicting video memorability. In our approach, there are two main stages: (i) extract image features through multiple frames of a video, (ii) predict its memorability score.

In the first stage, we sample 8 frames from each video. These frames are then fed into a pretrained Inception-v3 convolution network [10] to extract their 2048-dimension features. Once extracted, each of the video frames sequentially becomes an input of a recurrent neural network with one LSTM layer in the second stage. The memorability score corresponds to the output of the last dense layer for the last sequence's input, i.e., the video's final frame.

2 RELATED WORK

The task of predicting image memorability (IM) has made significant progress since the release of MIT's large-scale image memorability dataset and their MemNet [1]. Recently, in 2018, Fajtl et. al. [4] proposed a method, which benefits from deep learning, visual attention, and recurrent networks, and achieved nearly human consistency level in predicting memorability on this dataset. In [9], the authors' deep learning approach has even surpassed human consistency level with $\rho = 0.72$.

On the other hand, several attempts have been made in annotating and predicting video memorability (VM) [3, 5, 8]. In the latter two methods, their results both agree that video semantics, from captioning features in particular, give the best performance overall.

In our work, we explore the effect of videos' sequential aspect on memorability by using LSTM on visual features. To our knowledge, LSTM based approach in VM has only been tried in [3]. However, the results did not seem promising because of their small dataset.

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

3 MEMORABILITY PREDICTING

Feature extraction: In order to resolve the temporal factor, instead of using C3D [11], we decide to break the video into multiple frames and treat those frames as a batch representing that video. At the beginning, we extract only 3 frames (the beginning, middle, and last frames) for processing. After several tests we figure out that we can achieve higher results with more frames extracted. However, we end up with the decision of using 8 frames rather than a greater number. Indeed, the correlation was not substantially better and we want a straightforward extracting process. The length of each video in the dataset is 7 seconds. We get the very first frame of the video, then after each second, one more frame is captured, so finally for each video we have 8 frames.

We decide to use pre-trained Inception-v3 Convolutional Neural Network [10] to extract the frames' features as we want a concise network which can conduct a reasonably high accuracy. We use the publicly available model pretrained on ImageNet [7] and extract the output with a dimensionality of 2048 from the last fully connected layer with average pooling.

Predicting memorability: We consider several approaches regarding image and video memorability. In our attempts at adapting IM to VM, we simply use only the middle frame of each video and train two models with them as input data. We implemented a simple model which consists of a CNN for feature extraction and 2 fully connected (FC) layers for computing output score. We also retrain the model in [4] with those images to see if their model generalizes well to the task's dataset. We did not have enough time to try the approach in [9].

Videos' captioning features are also tested by using provided captions from the dataset. These captions accurately represent the videos in terms of semantics. Moreover, all videos are short and mostly single scene. Therefore, we use only 1 caption per video instead of generating each for every frame. A vector of 300 dimensions is extracted from each video's caption, which has been preprocessed, using the pretrained *word2vec* model [6]. We feed these vectors into our caption-only LSTM and the obtained results are shown in Table 1.

Furthermore, we propose to use a LSTM model to predict VM score using features extracted above (figure 1). Each extracted feature vector of every frame of a video is an input of a time step in our LSTM model. At the last step, a dense layer takes a 1024-dimension output vector of LSTM model and calculates the memorability score of that video.

For the short-term task, three out of five submitted runs are the results of our proposed method with three different configurations

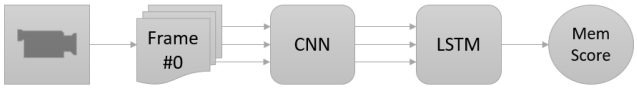

Figure 1: The proposed method for predicting memorability scores of videos using deep features and LSTM.

(512, 1024, and 2048 hidden units). The remaining two are outputs of the retrained AMNet in [4] because we also want to test its performance on the task's dataset. For the long-term task, the first run stands for our proposed method trained with long-term labels. The second run is accomplished by training our model using short-term labels and validating it by long-term labels. The next run is the result of retraining AMNet. Two final runs are actually the predicted results of two previous checkpoints in short-term task.

4 RESULTS AND DISCUSSION

In this section, we evaluate our LSTM model on the task's dataset. We present our quantitative results as well as some insight that we learned from this dataset.

Evaluation: Since we do not have the ground truth of the official test set, in order to compare these methods, we divide the development set into 3 parts: 6,000 videos for training, 1,000 videos for validating, and 1,000 videos for testing. Table 1 shows the results of different methods that we tested with our 1,000 test videos as well as the task's official test set.

With our approach of using sequential visual features of videos with LSTM, the model with 1024 hidden units achieves the best score of $\rho = 0.501$ on 1,000 test videos mentioned above and $\rho = 0.484$ on the official test set for the short-term task; while for the long-term task, the model which was trained on short-term labels and validated on long-term labels gets $\rho = 0.261$ and $\rho = 0.257$ respectively. Meanwhile, if we use only long-term labels, our method gets $\rho = 0.214$ on the official test set.

Table 1: Spearman's rank correlation results of different methods on dataset from [2].

Task	Model	ρ	
		1,000 test videos	Official test set
Short-term	Our method (2048 units)	0.532	0.480
	Our method (1024 units)	**0.511**	**0.484**
	AMNet [4] (without attention)	0.480	0.447
	AMNet [4] (with attention)	0.487	0.455
	Our method (512 units)	0.525	0.478
Long-term	Our method (long-term labels)	0.256	0.214
	Our method*	**0.261**	**0.257**
	AMNet [4] (attention + long-term labels)	0.252	0.194
	Our method (2048 units, short-term labels)	0.272	0.251
	Our method (1024 units, short-term labels)	0.266	0.252

* 1024 units, model is trained with short-term labels and validated by long-term labels.

In order to prevent overfitting while training, we apply a dropout rate of 0.5 on LSTM layer. We found that this rate gives the best results among 3 dropout rates of 0.25, 0.5, 0.75. The model also starts overfitting as it reaches its peak at around $\rho = 0.50 - 0.52$ and $\rho = 0.24 - 0.26$ on the validation subset of the short-term and long-term tasks respectively.

Discussion: The dataset from [2] on short-term memorability does follow a common trend previously stated in [1]. Videos with contents of natural scenes, landscapes, backgrounds, and exteriors tend to be less memorable. On the other hand, videos with scenes that have people, interiors, and human-made objects are easily remembered.

Figure 2: Predicted results from our models for long-term task (top) and short-term task (bottom). The images are sorted from the most memorable (left) to the least memorable (right).

On the contrary, we think predicting long-term memorability on this dataset requires more in-depth research. For all of our tried methods, the results are always better when training/validating with short-term labels. Long-term labels seem to confuse the model which leads to worse performance. One possible reason that can cause the inconsistency in this particular dataset is that there exists multiple similar videos with opposite scores about or of specific objects.

Figure 3: Similar videos can cause confusion to visual-based model in long-term memorability. Long-term scores: 0.727 (left), 0.273 (right).

As in figure 3, both videos are almost identical in terms of visual features such as color, angle, and actor. These videos might cause participants to make mistake when deciding whether they had watched it or not. Hence, their long-term labels give opposite results.

5 CONCLUSION AND FUTURE WORK

In our approach, we focus on the temporal aspect of videos by using their frames in a LSTM recurrent network. We have not tried using a combination of features in the process, hence, we will try using multiple aspects of a video to measure its performance.

Acknowledgments: We would like to thank SE-AI Lab, VNU-HCMUS for their precious support.

REFERENCES

[1] Antonio Torralba Aditya Khosla, Akhil S. Raju and Aude Oliva. 2015. Understanding and Predicting Image Memorability at a Large Scale. In *2015 International Conference on Computer Vision (ICCV)*. 2390–2398.

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

[3] Romain Cohendet, Karthik Yadati, Ngoc Q. K. Duong, and Claire-Hélène Demarty. 2018. Annotating, Understanding, and Predicting Long-term Video Memorability. In *Proceedings of the 2018 International Conference on Multimedia Retrieval, Yokohama, Japan.* 178–186.

[4] 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.

[5] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. 2015. Learning Computational Models of Video Memorability from fMRI Brain Imaging. *IEEE Trans. Cybernetics* 45, 8 (2015), 1692–1703.

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

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* (2015).

[8] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and Recall: Learning What Makes Videos Memorable. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops, Venice, Italy.* 2730–2739.

[9] Hammad Squalli-Houssaini, Ngoc Q. K. Duong, Marquant Gwenaelle, and Claire-Hélène Demarty. 2018. Deep Learning for Predicting Image Memorability. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Calgary, AB, Canada.* 2371–2375.

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Las Vegas, NV, USA.* 2818–2826.

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