Predicting Media Memorability Using Deep Features with Attention and Recurrent Network

Le-Vu Tran, Vinh-Loc Huynh, Minh-Triet Tran
Faculty of Information Technology, University of Science, Vietnam National University-Ho Chi Minh City
tlvu@apcs.vn,hvloc15@apcs.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

In the Predicting Media Memorability Task at the MediaEval Challenge 2019, our team proposes an approach that uses deep visual features with attention, and recurrent network to predict video memorability. For several frames in each video, attentive regions are marked by utilizing AMNet . Features are then extracted from those preprocessed frames. We next forward these through an LSTM network to model the structure of the video and predict its memorability score.

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 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 three main stages: (i) determine which regions of multiple frames of a video are more remarkable, (ii) extract image features from those remarkable video frames, (iii) predict each video's memorability score.

In the first stage, we sample 8 frames from each video, each frame is fed through AMNet [3] to determine which regions are remarkable. For each frame, 3 activation maps are generated to mark attentive regions. As a result, for each video, we increase from 8 frames to $8 \times 4 = 32$ frames (1 original frame + 3 attention frames).

In the second stage, all 32 frames are concatenated in the following order O_1, M_{11}, $M12$, M_{13}, O_2, M_{21}, M_{22}, M_{23},... where O_i is the i^{th} original frame, and M_{ij} is the j^{th} masked frame of the i^{th} original frame; are then fed into a pre-trained Inception-v3 convolution network [7] to extract their 2048-dimension features.

Once extracted, each of the video features sequentially becomes an input of a recurrent neural network concatenating with a dense layer in the third stage. The memorability score corresponds to the output of the dense layer mentioned earlier.

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 [4]. Recently, in 2018, Fajtl et. al. [3] 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 [6], the authors' deep learning approach has even surpassed human consistency level with $\rho = 0.72$.

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 [1]. However, the results did not seem promising because of their small dataset.

3 MEMORABILITY PREDICTING

Attention: For each frame in a particular video, we fed it through AMNet, by default, it iteratively generates 3 attention maps that linked to the image regions correlated with the memorability. Then we multiply those heat maps with the original frame to remove completely regions that we don't want to appear in the frame. Figure 1 gives a better point of view of what we have done in this stage. As a result, after this stage, each frame of a video becomes a batch of 4 frames (1 original frame + 3 masked frames). We consider that batch of 4 frames as the input for the next stage.

Feature extraction: To resolve the temporal factor, instead of using C3D [8], we decide to break the video into multiple frames and treat those frames as a batch representing that video. At the beginning, we extracted only 3 frames (the beginning, middle, and last frames) for processing. After several tests, we figured out that we can achieve higher results with more frames extracted. However, we ended up with the decision of using 8 frames rather than a greater number. Indeed, the correlation was not substantially better

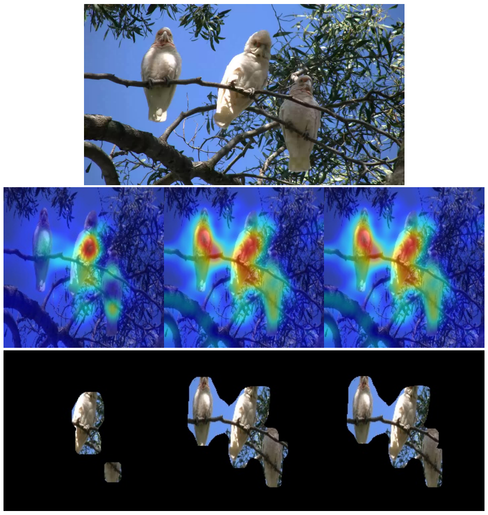

Figure 1: Original frame, its three activation maps (second row), and its masked frames (third row).

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

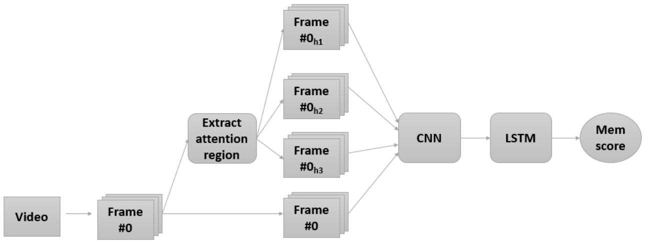

Figure 2: The proposed method.

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. At this stage, for each video, we have 8 original frames. Including the attention stage described above, finally, for each video, we have the total number of 32 frames. We then use pre-trained Inception-v3 Convolutional Neural Network [7] to extract the frames' features as we want a concise network which can conduct a reasonably high accuracy. We use the publicly available model pre-trained on ImageNet [5] and extract the output with a dimensionality of 2048 from the last fully connected layer with average pooling.

Predicting memorability: We considered several approaches regarding image and video memorability. In our attempts at adapting IM to VM, we simply used 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 the output score. We also retrained the model in [3] with those images to see if their model generalizes well to the task's dataset. Furthermore, we propose to use an LSTM model to predict VM score using features extracted above (figure 2). 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 the 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 (1024, 2048, and 4096 hidden units). The remaining two are the results of the captioning mechanism from [9] (we use the mechanism from [9] to generate attention heat maps similarly to the AMNet mechanism mentioned earlier) with two different configurations (2048 and 4096 hidden units). For the long-term task, we repeat the same configurations but trained on different data from the 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.

Since we do not have the ground truth of the official test set, 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.

With our approach, the very same model with 1024 hidden units achieved the best result for both subtasks.

Table 1: Spearman's rank correlation results

Task	Model	ρ	
		1,000 test videos	Official test set
Short-term	Region Attention (1024 units)	0.496	0.445
	Region Attention (2048 units)	0.481	0.434
	Region Attention (4096 units)	0.468	0.436
	Caption Attention (2048 units)	0.431	0.414
	Caption Attention (4096 units)	0.365	0.384
Long-term	Region Attention (1024 units)	0.249	0.208
	Region Attention (2048 units)	0.221	0.202
	Region Attention (4096 units)	0.245	0.187
	Caption Attention (2048 units)	0.171	0.097
	Caption Attention (4096 units)	0.168	0.124

To prevent overfitting while training, we apply a dropout rate of 0.5 on the LSTM layer. We found that this rate gives the best results among 3 dropout rates of 0.25, 0.5, 0.75.

Discussion: According to the groundtruth, the dataset on short-term memorability does follow a common trend previously stated in [4]. 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.

On the contrary, we think predicting long-term memorability on this dataset requires more in-depth research. For all of our 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 of the inconsistency in this particular dataset is that there exist multiple similar videos with opposite scores about or of specific objects.

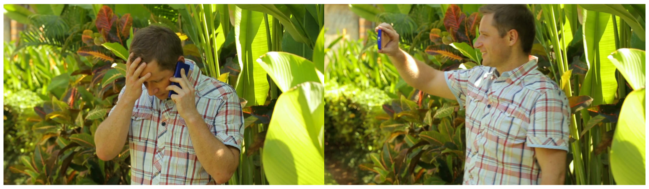

Figure 3: Similar videos can cause confusion to visual-based model in long-term memorability. Long-term scores: 0.727 (left), 0.273 (right).

As in Figure 3, both videos are almost identical in terms of visual features such as color, angle, actor, etc. These videos might cause participants to make mistakes when deciding whether they 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 an 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

Research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19. We would like to thank AIOZ Pte Ltd for supporting our team with computing infrastructure.

REFERENCES

[1] 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.

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

[3] 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.

[4] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*. 2390–2398.

[5] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

[6] Hammad Squalli-Houssaini, Ngoc QK Duong, Marquant Gwenaëlle, and Claire-Hélène Demarty. 2018. Deep learning for predicting image memorability. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2371–2375.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[8] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[9] Viet-Khoa Vo-Ho, Quoc-An Luong, Duy-Tam Nguyen, Mai-Khiem Tran, and Minh-Triet Tran. 2018. Personal diary generation from wearable cameras with concept augmented image captioning and wide trail strategy. In *Proceedings of the Ninth International Symposium on Information and Communication Technology*. ACM, 367–374.