Show and Recall @ MediaEval 2018 ViMemNet: Predicting Video Memorability

Ritwick Chaudhry, Manoj Kilaru, Sumit Shekhar
Adobe Research
{rchaudhr,kilaru,sushekha}@adobe.com

ABSTRACT

In the current age of expanding access to the Internet, there has been a flood of videos on the web. Studying the human cognitive factors that affect the consumption of these videos is becoming increasingly important, to be able to effectively organize and curate them. One such important cognitive factor is Video Memorability, which is the ability to recall a video's content after watching it. In this paper, we present our approach to solving the MediaEval 2018 Predicting Media Memorability Task. We develop a 3-forked pipeline for predicting Memorability Scores, which leverages the visual image features (both low-level and high-level), the image saliency in different video frames, and the information present in the captions. We also explore the relevance of other features such as image memorability scores of the different frames in the video, and present a detailed analysis of the results.

1 INTRODUCTION

With the explosion of visual content on the Internet, it is becoming increasingly important to discover new cognitive metrics to analyze the content. Memorability of visual content is one such metric. Previous studies on memorability [3] suggest that even though we come across a plethora of photos and images each day, our long-term memory is capable of storing massive number of objects with details, from images that we have come across. Although memorability of visual content is affected by personal context and subjective consumption [9], it has been shown [2, 11] that there is a high degree of consistency amongst people in the ability to retain information. This makes memorability an objective target.

Recent efforts in trying to predict the memorability of images have been successful, with the development of a large scale dataset on image memorability [13]. In [13], near human consistency rank correlation for image memorability is achieved, thereby establishing that human cognitive abilities are within reach for the field of computer vision. Despite these efforts in the realm of images, there has been limited work in predicting memorability of videos, given the added complexities that videos bring in.

Therefore, we seek to analyze the task of predicting memorability scores for videos in the context of MediaEval 2018.

2 RELATED WORK

The concept of memorability has been studied in psychology and neuroscience studies. They mostly focused on visual memory, studying for instance the human capacity of remembering object details [3], effect of stimuli on encoding and later retrieval from memory [1], memory systems of the brain [18, 19] etc. Broadly, prior work on recall of information about viewed visual content can be divided into the following categories:

Image Memorability: Isola *et al.* [12] started out the computational study revolving around the cognitive metric, memorability of images. The authors showed that across various subjects and under wide range of contexts, memorability of an image is consistent, which indicates that image memorability is an intrinsic property of images. Since then many prior works have explored this problem [10, 11, 14, 15, 17]. Khosla et al. [13] introduced largest annotated image memorability dataset (containing 60,000 images from diverse sources) and showed that fine-tuned deep features outperform all other features by a large margin. Fajtl *et al.* [7] used a visual attention mechanism and designed an end-to-end trainable deep neural network for estimating memorability. Siarohin *et al.* [21] adopted a deep architecture for generating a memorable picture from a given input image and a style seed.

Video Memorability: Han et al. [8] commenced computational studies on memorability of videos by learning from brain functional magnetic resonance imaging. As the method used fMRI measurements of the users for learning the model, it would be difficult to generalize. The authors in [5, 20] used spatio-temporal features to represent video dynamics and used a regression framework for predicting memorability.

We extend their work by proposing a trainable deep learning framework for predicting Video Memorability scores.

3 APPROACH

In this section, we discuss the task of predicting Video Memorability. The feature extraction from videos is described in Section 3.1 and an analysis of features for memorability prediction is discussed in Section 3.2

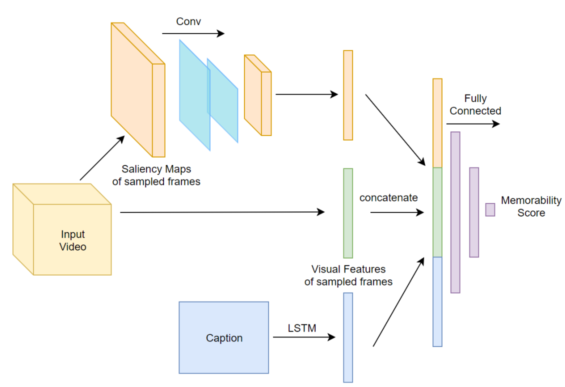

Figure 1: Our proposed model architecture

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

3.1 Feature Extraction

We used following features, divided into 3 groups:

Group 1 (G1)

- **C3D:** C3D features are outputs of the final classification layer of deep 3D convolutional networks trained on a large scale supervised video dataset;
- **Color Histogram:** This is computed in the HSV space using 64 bins in each color space for 3 key-frames (first, middle and last frames) for each video;
- **InceptionV3:** This corresponds to the final class activations of the InceptionV3 deep network for object detection, trained on the ImageNet dataset;
- **Saliency:** The aspect of visual content which grabs human attention has shown to be useful in predicting memorability [6, 11]. We used the highest ranking saliency prediction model in the MIT Saliency Benchmark on the MIT300 dataset (AUC, sAUC), DeepGaze II [16] and generated saliency maps for all 3 key frames;
- **Captions:** We used textual captions present in the dataset which were generated manually for describing the videos. Captions can be a compact source of representing the video content, and thus can be useful for predictions.

Group 2 (G2)

- **Image Memorability:** We divided the video into 10 frames and used image memorability scores for each frame predicted using a pre-trained model [13].

Group 3 (G3)

- **HMP:** Histogram of motion patterns is computed for each video and Principal Component Analysis (PCA) (with 128 principal components) is applied on them to obtain a reduced dimensional encoding;
- **HoG:** HoG descriptors (Histograms of Oriented Gradients) are calculated on 32x32 windows of each key frame and 256 principal components are extracted from each feature.

3.2 Model Description and Prediction Analysis

Here, we describe our proposed model and provide the training details. The dataset [4] consists of 8000 training videos and 2000 test videos, with each video being 7 seconds long. The train data is randomly split into 80:20 split for training the model and validation respectively. We describe our 3-forked pipeline architecture (see Figure 1) for predicting Memorability Scores, which leverages the aforementioned visual features, image saliency and captions.

Saliency: The saliency maps extracted from the video frames are down scaled to 120 by 68. A 2 layer CNN is applied on these maps with each layer consisting of a 2D convolution, batch normalization, relu activation and a max pool operation. Finally they are vectorized and a fully connected linear is applied on it.

Table 1: Rank correlation and MSE on test train and validation sets For Short Term memorability scores

Features	Train		Validation		Test	
	Spearman	MSE	Spearman	MSE	Spearman	MSE
G1,G3	0.5897	0.0039	0.3959	-	0.3554	0.0060
G1,G2	0.5815	0.0042	0.341	-	0.3149	0.0062
G1	0.7401	0.0027	0.3222	-	0.3096	0.0068

Table 2: Rank correlation and MSE on test train and validation sets For Long Term memorability scores

Features	Train		Validation		Test	
	Spearman	MSE	Spearman	MSE	Spearman	MSE
G1,G3	0.8941	0.0065	0.2034	-	0.0878	0.0313
G1,G2	0.2882	0.0197	0.153	-	0.1399	0.0198
G1	0.2975	0.0296	0.1437	-	0.1499	0.0286

Captions: Each word in the captions is represented using pre-trained 100 dimensional Glove embeddings and each embedding is passed through single layered LSTM of hidden dimension 100. The final representation of this caption is appended with rest of features as shown in the Figure 1.

Other Visual Features: C3D, Color Histogram, HMP, HoG, InceptionV3, Image Memorability features are concatenated and then combined with saliency and caption representations. Then a five layered dense fully connected linear neural is applied to obtain a single number representing the memorability score. The model is trained using Stochastic Gradient Descent with Mean Squared Error Loss function.

We trained different models for both Long Term and Short Term memorability scores using the aforementioned architecture. Results are presented in Table 1 and Table 2

We believe that all higher level G1 features are required for memorability prediction. To test whether low level features (Group G3) will help in prediction, we ran experiments, including and excluding the G3 features (Table 1). We also experimented with using the Image Memorability scores of sampled frames from the video.

4 DISCUSSION AND OUTLOOK

In this work, we have described a robust way to model and compute Video Memorability. It is empirically clear that using G3 features (low level image features of keyframes) help. Also, including Image memorability scores of key frames didn't lead to any improvement in performance, hinting to the fact that videos are much more than just a set of frames, and that temporal features matter. 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.

REFERENCES

[1] Wilma A Bainbridge, Daniel D Dilks, and Aude Oliva. 2017. Memorability: A stimulus-driven perceptual neural signature distinctive from memory. *NeuroImage* 149 (2017), 141–152.

[2] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. 2013. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General* 142, 4 (2013), 1323.

[3] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. 2008. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences* 105, 38 (2008), 14325–14329.

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

[5] 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.

[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] 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 transactions on Cybernetics* 45, 8 (2015), 1692–1703.

[9] R Reed Hunt and James B Worthen. 2006. *Distinctiveness and memory*. Oxford University Press.

[10] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. 2011. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*. 2429–2437.

[11] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2013. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis & Machine Intelligence* 1 (2013), 1.

[12] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable? (2011).

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

[14] Aditya Khosla, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2012. Memorability of image regions. In *Advances in Neural Information Processing Systems*. 296–304.

[15] Jongpil Kim, Sejong Yoon, and Vladimir Pavlovic. 2013. Relative spatial features for image memorability. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 761–764.

[16] Matthias Kummerer, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. 2017. Understanding Low- and High-Level Contributions to Fixation Prediction. In *The IEEE International Conference on Computer Vision (ICCV)*.

[17] Matei Mancas and Olivier Le Meur. 2013. Memorability of natural scenes: The role of attention. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 196–200.

[18] James L McGaugh, Larry Cahill, and Benno Roozendaal. 1996. Involvement of the amygdala in memory storage: interaction with other brain systems. *Proceedings of the National Academy of Sciences* 93, 24 (1996), 13508–13514.

[19] James L McGaugh, Ines B Introini-Collison, Larry F Cahill, Claudio Castellano, Carla Dalmaz, Marise B Parent, and Cedric L Williams. 1993. Neuromodulatory systems and memory storage: role of the amygdala. *Behavioural brain research* 58, 1-2 (1993), 81–90.

[20] 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 Conference on Computer Vision and Pattern Recognition*. 2730–2739.

[21] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. 2017. How to Make an Image More Memorable?: A Deep Style Transfer Approach. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 322–329.