CEUR-WS.org/Vol-2283/MediaEval_18_paper_41.pdf

Multimodal Approach to Predicting Media Memorability

Tanmayee Joshi, Sarath Sivaprasad, Savita Bhat, Niranjan Pedanekar
TCS Research, Pune, India
tanmayee.joshi@tcs.com,sarath.s7@tcs.com

ABSTRACT

In this paper, we present a multimodal approach to modelling media memorability for the "Predicting Media memorability" task at MediaEval 2018. Our approach uses video and image based features along with provided textual description to predict a probability-like memorability score for each of the seven second audioless video clips. We use the same set of features for predicting both short-term and long-term media memorability.

1 INTRODUCTION

With the dramatic surge of visual media content on platforms like *Instagram*, *Flickr* and *YouTube*, it is imperative that new methods for curating, annotating and organizing this content be explored. To this effect, non-traditional metrics for tagging media content have been examined. Previous works have used metrics such as *aesthetics*[2], *interestingness*[5], *memorability*[9] to annotate and rank images. The "MediaEval 2018: Predicting Media Memorability" task [1] focuses on predicting ‘short-term’ and ‘long-term’ memorability for videos.

An important aspect of human cognition is the ability to remember and recall photos and videos with a surprising amount of detail. Interestingly, not all content is stored and recalled equally well [8]. Previous attempts [5, 8] at predicting image and video memorability discuss factors affecting memorability. The experiment stated in [1] showed people, content not related to them personally, and recorded a varying probability of detecting a repetition of a given video after a short/long delay. Deep learning models have given promising predictions over image memorability [9, 12]. We propose an ensemble of deep learning models that takes into account various properties that are correlated with memorability. We capture these aspects by deriving respective features through text embedding, frames and video.

2 APPROACH

In this section, we outline our multimodal approach to model media memorability using video, image and text features. The visual features are inspired from different properties of images such as *saliency* and *aesthetics*. We assumed that memorability of the video is affected by properties of images comprising the video. We also hypothesized that captions provide additional cues for understanding semantics of videos. The short-term memorability and long-term memorability were modelled independently using the same set of features explained in this section.

2.1 Visual Features

We used image level features based on: *color*, *saliency*, *aesthetics*, *memorability* and presence of human *faces*. Image features were calculated on frame number *0*, *56* and *112*. Except for *aesthetics*, we computed mean and standard deviation of these features across the three frames to give feature representation for a video clip. C3D features were used to represent spatiotemporal aspect of a video.

Color: Color and its distribution have a significant influence on human cognition [6, 15]. Color information was captured using 3D HSV feature [4] and colorfulness [7]. The statistics over these vectors provided a *128* dimensional vector per video.

Saliency: [3, 11] observed that saliency feature is relevant for predicting memorability. For every image, saliency map was created using pre-trained image saliency net (Salnet) [10]. We hypothesized that the intensity distribution of saliency inside the image and its change across frames contribute more towards memorability than the spatial spread and orientation of salient pixels. We created bins from saliency maps based on the intensity of pixels, with histogram boundaries at variable lengths to accommodate the variance of pixel distribution.

Aesthetics: Aesthetics and human judgements of memorability are highly correlated [5, 8]. We used median value of aesthetic visual features across frames, provided with the dataset.

Face-based Feature: Using a state-of-the-art deep learning method [13], we computed the number of faces per keyframe. Using this information, the dataset was divided into two parts: with faces and without faces. Running "Mann-Whitney U" test over the memorability of two populations, we found that two populations are significantly different (*p-value 1.06e-31*).

Image Memorability: We hypothesized that memorability of the video is affected by the memorability of the images comprising the video. We used MemNet proposed by [9] to get a memorability score per image. This score was used directly as a part of ensemble.

C3D: We used fc7 activations of the C3D network (provided with the dataset) as a feature vector to capture activity in the video. It captures spatiotemporal information [14].

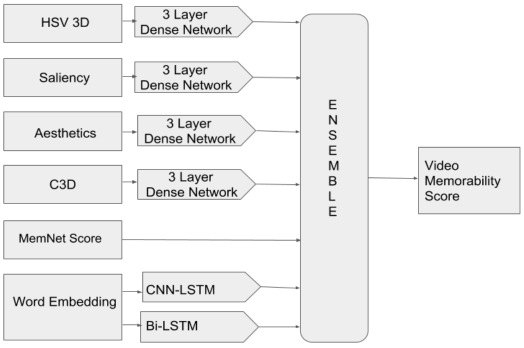

Figure 1: Model Architecture

Copyright held by the owner/author(s).
MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

2.2 Text-based Features

To analyse the data with respect to captions, we divided the train data into 4 bins. Each bin consisted of captions corresponding to 4 quartiles in memorability annotation. We defined a metric *'word relevance'* inspired from term frequency−inverse document frequency (*Tf-Idf*) statistic typically used in *Information Retrieval.* We define *word relevance* for a word w_i in a bin j as WR_{ij}. Let total number of bins be N, total number of words in a bin j be W_j, frequency of a word w_i in a bin j be w_{ij}, frequency of bins where the word w_i appears be b_{wi}, and frequency of word w_i in other bins be $w_{i\hat{j}}$.

$$WR_{ij} = \frac{w_{ij}}{W_j}(1 + \log\frac{N}{b_{wi}})\frac{1}{w_{i\hat{j}}} \qquad (1)$$

We created a wordlist of all unique words from the video captions. After stemming and lemmatizing, we removed all stopwords from the list. Words with WR value above a threshold were shortlisted as candidate words and their frequency in captions was considered as a feature. We believe that higher value of WR quantifies the word's association to a particular range of memorability. We hypothesized that WR increases with the relative higher frequency of a word in a particular bin with respect to its frequency in other bins. It was observed that words related to topics like food and toddlers fall in higher memorability range and generic words related to topics such as landscape and scenery fall in the lower memorability range.

We also used pre-trained GloVe embeddings[1] of words to capture more information from textual description. We preprocessed the caption data by removing stopwords. We created a 100 dimensional word-embedding vector for each word.

3 EXPERIMENTS

We ranked videos by assigning a probability like score to each video clip, treating it as a regression problem. The annotations for short term and long term memorability were skewed towards higher values with mean of *0.86* (short term) and *0.78* (long term). All input features were normalized and the ground-truth was kept unwhitened so that the model captured the skewed output distribution. We divided the given dataset into train and validation sets in the ratio 3:1 such that the annotations in two sets have similar distribution. We explored different combinations of features for predicting memorability.

Experiment 1: Low level features namely, colorfulness, blur[11] value, HSV histogram were concatenated and SVR was used over this vector.

Experiment 2: We concatenated the face based feature to the 3D HSV. The resultant 130 dimension vector represents the color spread and facial information in an image. We passed this vector and features for C3D , aesthetics and saliency through dense fully connected layers indepedently. We ensembled these models using their normalized correlation values on validation as coefficients for weighted average.

Experiment 3: We used word embeddings to train different neural network architectures. CNN-LSTM and Bi-LSTM models give best correlation on training and validation data. We ensembled models trained on image features, video features and text embeddings using weighted average and SVR.

Table 1: Correlation for Ensemble model

Ensemble Model	Short-Term Memorability		Long-Term Memorability	
	Pearson	Spearman	Pearson	Spearman
Weighted Average	0.50	0.46	0.25	0.22
SVR	0.48	0.44	0.25	0.23

We used sigmoid activation in the last layer for all networks so as to restrict the output to the range of 0 to 1. ReLU activation was used for all other layers. We also fine-tuned the model over validation data before predicting on test data. Final predictions submitted for evaluation were based on models from experiments 2 and 3.

4 RESULTS AND ANALYSIS

In experiment 1, the model performed poorly with a near zero Spearman's rank correlation over the validation set. This shows that only low level image features are not sufficient to understand media memorability. As part of the challenge we submitted results of five runs based on experiments 2 and 3. Table 1 lists correlations obtained from two best performing models. As per the evaluation on an unseen test set, best performing model gives Spearman's and Pearson's correlation for short-term memorability as *0.46* and *0.50* respectively. The correlations for long term memorability are *0.23* and *0.25* respectively. In the run with experiment 2, we obtained Spearman's rank correlation of *0.39* and *0.17* for short term and long term memorability respectively. Our best submission based on experiment 3 gives an improvement of *7%*.

As mentioned earlier, this model from experiment 3 used textual information along with visual features. The improvement shows that the words from captions are contributing to predicting memorability scores. We believe that additional textual features such as location cues, emotion cues may be useful in further improvements. Mean values of our predictions of short term and long term memorability over validation data are *0.83* and *0.78* respectively. The values are close to the mean values of their respective annotations in the training data. This shows that our model succeeds in capturing the skewed distribution of training data.

5 CONCLUSION AND FUTURE WORK

This paper presents the ensemble model by team *AREA66* for predicting media memorability. We use visual features based on image and video along with textual features from given captions. The results show that better results are obtained by combining visual and textual features. On the other hand, only visual features give lowest values for prediction correlation. We also noticed that experiments with only low level image features give poor results. In future, we plan to explore effects of textual information on video memorability. Secondly, we aim to explore more sophisticated methods for utilizing low level image features to improve prediction performance.

[1] https://nlp.stanford.edu/projects/glove/

REFERENCES

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

[2] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1657–1664.

[3] Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem. 2015. What makes an object memorable?. In *Proceedings of IEEE International Conference on Computer Vision*. 1089–1097.

[4] Ankit Goyal, Naveen Kumar, Tanaya Guha, and Shrikanth S Narayanan. 2016. A multimodal mixture-of-experts model for dynamic emotion prediction in movies. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2822–2826.

[5] 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.

[6] Alan Hanjalic. 2006. Extracting moods from pictures and sounds: Towards truly personalized TV. In *IEEE Signal Processing Magazine*, Vol. 23. IEEE, 90–100.

[7] David Hasler and Sabine E Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, Vol. 5007. International Society for Optics and Photonics, 87–96.

[8] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What makes a photograph memorable?. In *IEEE transactions on pattern analysis and machine intelligence*, Vol. 36. IEEE, 1469–1482.

[9] 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.

[10] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. 2016. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 598–606.

[11] José Luis Pech-Pacheco, Gabriel Cristóbal, Jesús Chamorro-Martinez, and Joaquín Fernández-Valdivia. 2000. Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings of 15th International Conference on Pattern Recognition*, Vol. 3. IEEE, 314–317.

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

[13] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1701–1708.

[14] 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.

[15] Patricia Valdez and Albert Mehrabian. 1994. Effects of color on emotions.. In *Journal of experimental psychology: General*, Vol. 123. American Psychological Association, 394.