# Predicting Video Memorability

Santanu Bhattacharjee
santanu.bhattacharjee2@mail.dcu.ie
Dublin City University
Dublin, Leinster

## ABSTRACT

In the era of social media awareness, videos are becoming more and more acceptable form of communication among us. Thus, impact of such videos whether it is made in the context of advertisement or Learning, is a critical factor to be considered while making. Human cognitive factors like memorability can be considered as one such metric to do the impact analysis. In this paper, I present an approach for solving the MediaEval 2018 Predicting Media Memorability Task. In the beginning, I explore what are the features most contributing to the final prediction. In addition, I show how other features generated from existing ones can be critical for achieving better result.

## KEYWORDS

MediaEval 2018, Video Memorability, Multimedia Analysis, Neural Network, Deep Learning Framework

## 1 INTRODUCTION

Human memorability is a research topic over long time. Previous work [1] in this domain suggests that human can process and store hundreds of images that they come across every day. However, this metric can be highly influenced by personal attributes like individual interests, attention etc. [2] When it comes to short-term and long-term memorability, there is very few literatures available which can differentiate the attributing factors for these phenomena.

In recent past, there are quite few research works have already been taken place on image memorability. [3] [4] In contrast, very few works are there on video memorability. It can still be considered as a new field of research. The task of Predicting media memorability at MediaEval 2018 is a kind of benchmarking initiative for such Multimedia Analysis.

In this paper, it is depicted that how Machine Learning along with Feature Engineering can produce some convincing results in predicting memorability of videos.

## 2 RELATED WORK

Memorability of images is quite explored [3] [4] and few of these solutions provided good results on benchmark test. [5] Another interesting work has shown uncorrelation between memorability and interestingness of the image. [6]

Memorability of videos is comparatively new in multimedia analysis. However, there are few notable works that showed promising results. One such work [7] introduced ensemble approach where features are used in different models before aggregating. Another very recent research work [8] shows how visual features along with salience maps of sampled frames from video can be combined with captions to predict decent memorability score.

Where few of these approaches [9] are based on machine learning algorithms, other approaches [7] [8] are using neural network architecture for their implementation.

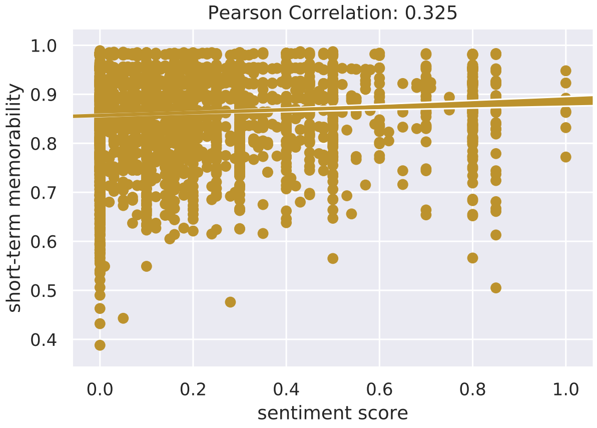

**Figure 1: Correlation Between Sentiment and Short Term Memorability**

## 3 METHODOLOGY

I propose a deep learning-based computational model architecture for predicting such memorability scores. However, before showing model design I want to give a brief on feature engineering that is done on different features available in dataset.

### 3.1 Feature Engineering

In the dataset, I find the following features and I try to explore and use few of these features in my experiments.

**Features:** *Video Captions, C3D, Color Histogram, Histogram of motion patterns (HMP), Histograms of Oriented Gradients (HOG), InceptionV3, LBP, ORB*

- Video Caption turns out to be one of the important features in my model. I remove all stop words from each caption and tokenize the text to sequence.
- I extract sentiment score against each caption and found a decent correlation between that score and memorability levels [Fig 1]. I use this feature in my model as well.
- I get the word counts of different captions and try to understand possible correlation with memorability score. Although, it is shown that the end model is not highly improved in presence of this feature.

- Most of the other features are available for every video file in the dataset. I use different functions to read these features by the order of file-names given in dev-set_video-captions text file to make the train-test split easy.
- Shape of these visual features are changed according to the need of the model.

## 3.2 Model Design

Multiple features need to be passed in the system as inputs to the model. In addition, there are two different scores (short-term and long-term memorability) need to be calculated from model. Considering these two requirements, I design a multi-input/multi-output neural network [Fig 2] for this problem to solve. I use Keras Functional API (with TensorFlow backend) for the implementation.

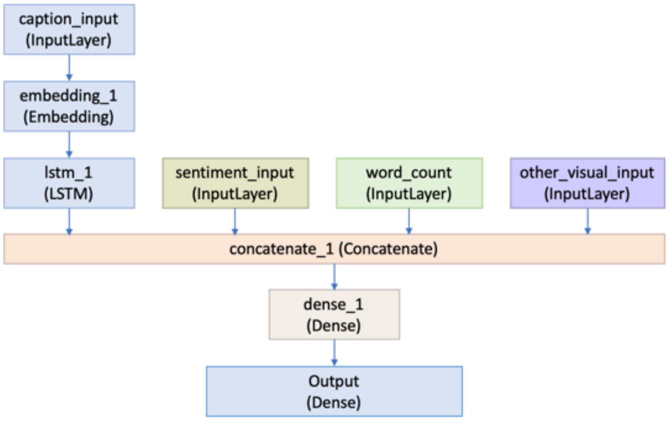
caption_input (InputLayer)
embedding_1 (Embedding)
lstm_1 (LSTM)
sentiment_input (InputLayer)
word_count (InputLayer)
other_visual_input (InputLayer)
concatenate_1 (Concatenate)
dense_1 (Dense)
Output (Dense)

**Figure 2: Neural Network Model Design**

*3.2.1 Input Layers.* As part of the experiment, I try different combination of features as inputs to the model. In broad way I can divide the input layers in four categories:

- Category 1: Captions as input
- Category 2: Sentiment scores as input
- Category 3: Words count as input
- Category 4: Other visual features as input

*3.2.2 Embedding Layer.* Each caption is converted to a sequence of tokens. Embedding layers is having input dimension of 5,870 and output dimension of 15 with input fixed length of 37. The sequence of tokens goes through the next LSTM layer.

*3.2.3 LSTM Layer.* The sequence of tokens generated in embedding layer is passed through a single layer of LSTM of 32 hidden dimensions. The final representation of the captions is combined with other features from different input layers.

*3.2.4 Concatenation Layer.* In this network layer, all input feature values are concatenated to get a single representation of multiple inputs.

*3.2.5 Dense Layers.* One dense layer is placed just after previous concatenation layer with 10 neurons. In this layer, both L1 and L2 regularization are applied to control weights in the network. “Relu” activation function is used.

**Table 1: Spearman‘s Correlation and MAE on test and validation sets for Short Term memorability scores**

| Features | Training | | Validation | |
|---|---|---|---|---|
| | Corr. | MAE | Corr. | MAE |
| Caption, Sentiment | 0.73 | 0.051 | 0.39 | 0.1 |
| Caption, Sentiment, Word count | 0.41 | 0.063 | 0.23 | 0.1 |
| Caption, Sentiment, Visual Features | 0.53 | 0.057 | 0.25 | 0.1 |

**Table 2: Spearman‘s Correlation and MAE on test and validation sets for Long Term memorability scores**

| Features | Training | | Validation | |
|---|---|---|---|---|
| | Corr. | MAE | Corr. | MAE |
| Caption, Sentiment | 0.78 | 0.051 | 0.14 | 0.1 |
| Caption, Sentiment, Wordcount | 0.74 | 0.055 | 0.14 | 0.1 |
| Caption, Sentiment, Visual Features | 0.68 | 0.063 | 0.12 | 0.1 |

Output layer is another dense layer just after the previous layer and it is consisting of only two neurons representing two scores model is predicting. “Sigmoid” activation function is used in output layer.

The final model is compiled with RMSprop optimizer. Mean Absolute Error (mae) is defined as loss function. “Accuracy” is another metric monitored during model training.

## 3.3 Model Training

The final model is trained on different set of inputs as indicated in previous sections in this paper. The configuration parameters are optimized to their best possible values during experiments. Batch size was set to 32 and 50 epochs were considered.

# 4 RESULTS

Performance of model prediction is measured by Spearman‘s correlation between ground-truth values and predicted values of memorability [Table 1 and 2]. However, to understand the model performance in more detail, I monitor “mae” loss and accuracy while training and validating the model. It is empirically evident that model with features caption and sentiment scores did well than model with other visual features for predicting short term memorability. In contrast, long term memorability scores from different models are very closely comparable.

# 5 CONCLUSION

In this work, I have explored importance of available/generated features in prediction of memorability, as well as, I propose a neural model framework for this analytical problem. The performance of the predictive model is quite comparable with the performance of other state-of-art models published in recent past. Although there is further scope of improvement especially considering I did not take any salient features from the videos. In addition, action oriented features from video along with audio can lead to a better model.

# REFERENCES

[1] TimothyFBrady,TaliaKonkle,GeorgeAAlvarez,andAudeOliva. 2008. Visual long-term memory has a massive storage capacity for object details. Proceedings of the National Academy of Sciences 105, 38 (2008), 14325âĂŞ14329.

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

[3] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, âĂIJUnderstanding and predicting image memorability at a large scale,âĂİ in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2015, pp. 2390âĂŞ2398.

[4] Y. Baveye, R. Cohendet, M. Perreira Da Silva, and P. Le Callet, âĂIJDeeplearning for image memorability prediction: the emotional bias,âĂİ in Proc. ACM Int. Conf. on Multimedia (ACMM), 2016, pp. 491âĂŞ495.

[5] J. Fajtl, V. Argyriou, D. Monekosso, and P. Remagnino, âĂIJAmnet: Memorability estimation with attention,âĂİ in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6363âĂŞ6372.

[6] Hammad Squalli-Houssaini, Ngoc Duong, Marquant GwenaÃńlle, Claire-HÃľlÃĺne Demarty. Deep learning for predicting image memorability. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr 2018, Calgary, Canada.

[7] R. Gupta and K. Motwani, âĂIJLinear Models for Video Memorability Prediction Using Visual and Semantic Features,âĂİ In The Proceedings of MediaEval 2018 Workshop, 29-31 October 2018, p. 31.

[8] R. Chaudhry, M. Kilaru, and S. Shekhar, âĂIJShow and Recall @ MediaEval 2018 ViMemNet: Predicting Video Memorability,âĂİ In The Proceedings of MediaEval 2018 Workshop, 29-31 October 2018, p. 15.

[9] Y. Liu, Z. Gu, and T. H. Ko, âĂIJLearning Memorability Preserving Subspace for Predicting Media Memorability,âĂİ In The Proceedings of MediaEval 2018 Workshop, 29-31 October 2018, p. 33.