Media Memorability Prediction using Deep Neural Networks

Anup Bolli
Dublin City University, Ireland
anup.bolli2@dcu.mail.ie

ABSTRACT

In this paper, an attempt has been made to predict the memorability scores of the video, probability of video being recognized, using different machine learning algorithms on semantic features such as Captions – video description. Different experiments were carried out to predict memorability of videos using basic machine learning algorithms to Recurrent Neural Networks (RNN) with optimized parameters.

1 INTRODUCTION

With the technological development and exponential growth of content data being generated on daily basis, it is impossible to consume the entire data in one's individual life and even its become a tough job for media platforms such as social media, search engines and recommender systems to provide right content to right individual at right time [1].

Human Cognition has huge capacity for recognizing lots of video in great detail after single view and forget videos even after watching several times. This shows that there are some intrinsic features that make video more memorable such as humor, emotions, salient events and actions etc and less memorable such as natural landscapes, animal videos, less interesting videos etc. [2].Interesting things that come out from this observations are "Can we build a Human alike computational model to predict memorability of Video" and "What are the factors that affect the memorability of the video". [3]

As a part of MediaEval competition 2018, an attempt was made to predict how memorable a video is, in short term and long term based on certain precomputed features. In this competition data was collected by showing a series of 9-sec video clips to the participants, few minutes later they were shown with random videos. if they recalled, they were told to press button. The memorability scores were calculated based on the average number people who remembered the video. This experiment was repeated again after 24 -72 hours later, so each video comprised of two scores -Short term memorability and long-term memorability scores. The precomputed features of the Video such as HMP, interestingness, inception, aesthetics, captions were provided, which can be useful to make competing choice between competing videos.

2 RELATED WORK

The pioneer work of Image memorability (IM) prediction by Isola et al [4], become a baseline for computational understanding of video memorability (VM) which attracted increasing attention. In subsequent work , several features were found to affecting memorability of video such as Saliency [5], interestingness and aesthetics [4] or emotions [6].The best results were obtained from emotions based features by applying fine-tuned deep learning model. [6] The model managed to get predictability score of 0.64 which is closer to human memorability score of 0.68.

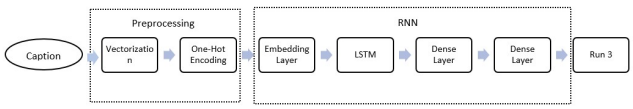
In [3], The Authors have proposed a novel approach which combines power of audio-visual and fMRI-derived features. They built a primary model learned on fMRI-features which mirrors the brain activity of memorizing videos. In second step they were able to predict VM without fMRI-scans. The extensive feature engineering was performed on several computed features before building the final predictor. The analysis included C3D deep learning features, color features, semantic features obtained from video captions, dense trajectories and saliency features. Caption feature turned out to be best parameter for predicting VM.

Application of deep learning models on VM has overcome the challenges of IM prediction and achieved good results. The Authors in [1] have attempted to make long term VM, have used new protocol to measure very long-term memory performances in the duration from weeks to years to collect good quality ground truth data and made them available for research community. The features used for building computational model were Spatio-temporal visual features (features extracted from C3D model, a 3D convolution network proposed for generic video analysis), Audio features, Emotional features (SentiBank and Affect) and Image captions. Image captions alone was able to predict best results with the scores of 0.31 and 0.39.

3 APPROACH

3.1 Semantic-based Model

From the literature review, it was clear that predicting memorability scores using captions resulted in best results. First attempt was made to use a basic algorithm with optimized parameters: Support Vector Regression which has given good results.



As a part of data pre-processing, words in captions of video were tokenized to integers using vectorization and padded with 0's to transform into uniform dimension of 50. This uniform vector was

further applied to One-hot encoding method to convert into 0's and 1's, which make sure that the algorithm is not biased towards words with high integer value. The dataset has divided into 80% training data and 20% test data, out of 80% of training data, 20% was used for validation.

Initially a 3-layer basic RNN layer architecture was built for predicting VM, the first layer comprised of Long Short-Term Memory (LSTM) layer with 128 units and return sequences as True. The tokenized data was transformed into 3-dimensional vector, to make it available for LSTM layer. The output of LSTM layer was further fed into 64 node dense layer with rectangular linear unit (ReLu) activation function and drop out factor of 0.4. The final layer is 2 node layers with sigmoid activation function. The model was trained with Adam optimizer, Mean Squared Error (MSE) as loss function and trained for 20 iterations.

To increase the prediction accuracy another model was designed. The main model corresponds to 3-layer RNN model (Run3) as depicted in the above pic. The tokenized captions of videos were fed to Embedding layer with the input dimension of 5191(max words count in caption file) output dimension of 20, with no regularization factor. Further the output of Embedding layer which is a vector of 5191 into 50 matrixes was passed on to bidirectional LSTM with 64 units with the dropout factor of 0.5. The last layer is 2- node dense layer for predicting the long term and short-term memorability scores using a sigmoid function using a Sigmoid activation function and dropout of 0.5 factor. The model is compiled with Adam optimizer, Mean square error as a loss function and accuracy for metrics. [7]

Another Neural network was built to improve accuracy. The first 3 layers consisted of 220 units with ReLU activation function and dropout factor 0.5. The last layer of is a 2-node dense layer predicting short term and long-term memorability scores with ReLU activation function. The model is trained for 20 epochs with a batch size of 50 and Nadam optimizer with a learning rate of 0.002 and beta values of 0.9 to 0.999.

Table 1:Test Results: Spearman's Rank Correlation

Run	Method	Short-term	Long-Term
1	SVR + Captions	0.334	0.141
2	ANN + Captions	0.440	0.145
3	RNN + Captions	0.165	0.114
4	LR + Captions	0.283	0.114

4 RESULTS AND ANALYSIS

Spearman's Rank correlation was used to compare the predictions made by models with the ground truth values provided in the dataset. According to the Spearman's correlation rank, Artificial neural network applied on captions of video is the best among all the models for short term memorability scores and ANN applied on LBP is the best for long term memorability scores.

In long term memorability predict, all the models performed similarly and whereas in short term memorability RNN preformed less than half times of the other models. Since all the models predicted high scores in short term than long term memorability, it can be inferred that short term VM is more predictable than long term VM.

In RNN model, initially a simple LSTM model was constructed which gave a small memorability scores, to enhance the accuracy complex model was build for this project but it took more computational power and 40- 45 mins duration to complete even one epoch.

5 CONCLUSIONS

In this paper, I have tried to experiment with different neural network algorithms on semantic features. In the process of this task, I have observed that semantic features were the good predictors of memorability scores in both long term and short term. The future work would be devoted to enhancing both short term and long term memorability with the focus on different features of videos such as Inception, Histogram of Motion Pictures, C3D along with ensemble neural networks.

REFERENCES

[1] J. Almeida, N. J. Leite and R. d. S. Torres, "Annotating, Understanding, and Predicting Long-term VideoMemorability," *International Conference on Multimedia Retrieval,* pp. 178-186, 2011.

[2] S. Shekhar, D. Singal, H. Singh, M. Kedia and A. Shetty, "Show and Recall: Learning What Makes Videos Memorable," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* p. 2730–2739, 2017.

[3] H. Junwei, C. Changyuan, L. Shao, X. Hu, J. Han and T. Liu, "Learning Computational Models of Video Memorability from fMRI Brain Imaging," *Cybernetics,* pp. 1692 - 1703, 2015.

[4] P. Isola, J. Xiao, A. Torralba and A. Oliva, "What makes an image memorable?," *Computer Vision andPattern Recognition (CVPR).,* p. 145–152, 2011.

[5] O. Le, M. Mancas and Meur, "Memorabilty of natural scenes: The role of attention," *In Proc. IEEE Int. Conf. on Image Processing (ICIP),* p. 196–200, 2013.

[6] A. Khosla, A. S. Raju, A. Torralba and A. Oliva, "Understanding and Predicting Image Memorability at a Large Scale," *Computer Vision(CV),* pp. 2380-7504, 2015.

[7] W. Sun and X. Zhang, "Video Memorability Prediction with Recurrent Neural Networksand Video Titles at the 2018 MediaEval Predicting MediaMemorability Task," *MediaEval'18,* 2018.