Predicting Video Memorability Scores using Captions and C3D

Harshali Patil
[2]Dublin City University, Ireland
harshali.patil2@mail.dcu.ie
18212797

ABSTRACT

Today, discovering new cognitive factors are extremely important due to dramatic surge in visual content on the internet. Memorability scores are regarded as one of the prime factors for the Multimedia Evaluation. The ability to recall and remember the videos is an important human cognitive factor, which is reflected by the Memorability scores. In this paper, an approach which uses textual description provided with videos is implemented to automatically predict the Memorability scores for both short-term and long-term videos.

1 INTRODUCTION

With the growing amount of digital content on the platforms like Facebook, 9gag, YouTube and Instagram, it is important to consider videos features like interestingness, quality or memorability which benefits many real-world applications in advertising, content summarization, and recommendation. Unlike other features of cognitive features, Memorability is objectively measurable and clearly defined.

Various research had been done for modelling image memorability but only few attempts has been made to address the problem of video Memorability. This project targets video memorability as the main objective.

The primary reasons for the scarcity of studies on Video memorability is the unavailability of open datasets to test and train model. In addition, Videos are highly dimensional – sound and visual movement- as compared to Images which brings computational challenges for analyzing tasks.

The Memorability score for a video is defined as the probability of a video to be remembered till a specific time. The "MediaEval 2018: Predicting Media Memorability" challenge focuses on predicting the short-term and long-term memorability scores by training a model on a given dataset. For this challenge, total 10000 records are provided in a dataset which splits in 8000 videos for training a model and 2000 videos for the test set. As well as, pre-computed features related to videos has been provided with the datasets.

In this project, “Captions”, which are the textual description of the videos and C3D are used as features to train the model. Also the combination of both the features are used for ML algorithms. Several NLP techniques like tokenization, stop words removal, lemmatization is used for text processing of the Captions. XGBoost Model has been used to train and test the dataset. This model is evaluated using the Spearman's rank correlation coefficient.

2 RELATED WORK

With the recent studies in Image memorability, work on video memorability is also become the point of interest. [1] describe an experiment to predict sub-shot memorability of videos. They used C3D deep learning and video semantics to predict the RMSE score. [2] has used Convolutional Neural Network (CNN) on semantic features and visual features to explore the memorability. They concluded that he CNN did the acceptable performance to determining the memorability. In this paper, [3] video hashing algorithm-based memory feature is proposed. They obtain spatial histograms feature to define memory feature and adopted supervised kernel hashing to map memory feature with the hash function. [4] They train the predictor based on the global image descriptors to determine the image memorability. They analyzed the ecological and explicit measures of the image memorability. They concluded that memorability of an image is defined by the probability person detecting the pictures while going through sequence of the images.

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 [5]. Recently, in 2018, Fajtl et. al. [6] 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 [9], the authors’ deep learning approach has even surpassed human consistency level with $\rho = 0.72$.

3 APPROACH

3.1 Feature Extraction

As the part of text processing all the captions are tokenized. From those token list, all the stop words are removed. Further lemmatization and part of speech tagging is done on the captions. This resulted in a cleaned caption which can be further used in the machine learning algorithms to train the model Semantic features and aesthetic feature, which are captions and C3D respectively, are used as features to train the model. The processed textual description of videos i.e. Captions are then fed into a Term Frequency—Inverse Data Frequency (TF-IDF) vector. It gives a frequency of each word in the document. It is directly proportional to the Number of occurrences of that word in the document. IDF considers the weight of rare words present in the captions.

For aesthetic features, C3D features is one of the important features which is already provided. It supports the 3- dimensional convolutional networks.

3.2 Model

The features were high in dimensional and high variance which can result in overfitting. Also, to get higher model performance and computational speed, XGBoost model is used.

Initially, the data is split in 80% train and 20% test data. After this, video features are fed into the XGBoost model which gave the RMSE score for both short-term and long-term memorability.

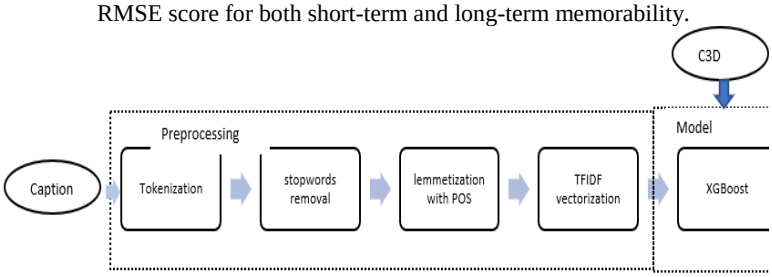

Figure 1: Flow chart for the apprach

4 RESULTS AND ANALYSIS

Spearman's correlation coefficient.is used to evaluate the predictions made by the XGBoost model with the ground truth values provided in the dataset.

From the evaluated result present in table 1 and table 2: It is concluded that the Model used with the Caption provides better results than the model used with Caption and C3D together.

Also, from the results, it is clearly depicted that the short-term memorability is more predictable as compared to the long-term memorability.

Features	RMSE	Spearman's corr. Coeff.
Captions	0.0720	0.401
Captions+C3D	0.0717	0.382

Table 1: Short-term memorability

Features	RMSE	Spearman's corr. Coeff.
Captions	0.1406	0.193
Captions+C3D	0.1422	0.172

Table 2: long-term memorability

Like the previous studies, it is observed that the Captions gives superior performance over the subjective features of the videos.

5 CONCLUSIONS & FUTURE WORK

This Paper presents a methodology using Machine Learning to predict the video memorability. The findings are done on the publicly available video dataset. The main aim of this research is to determine the effect of embedded features of the videos on the memorability of the videos. The results obtained revealed that the both semantic and aesthetic features can determine the video memorability with a good performance. However, model with semantic features outperformed C3D.

In Future work, more complicated mode and the combination of the features can be used to check the performance and predict more better results for the Video Memorability.

REFERENCES

[1] S. Shekhar, "Show and Recall: Learning What Makes Videos Memorable," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[2] A. Kar, "What Makes a Video Memorable?," in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2017.

[3] W. Wang, "A memorability based method for video hashing," in *2015 IEEE 16th International Conference on Communication*, 2015.

[4] P. Isola, "What Makes a Photograph Memorable?," in *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2014.

[5] "Hammad Squalli-Houssaini, Ngoc Q. K. Duong, Marquant Gwenaelle, and Claire-Hélène Demarty . 2018. Deep Learning for Predicting Image Memorability. In 2018 IEEE International Conference on Acoustics,".

[6] "Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. AMNet: Memorability Estimation with Attention.".

[7] " Antonio Torralba Aditya Khosla, Akhil S. Raju and Aude Oliva. 2015. Understanding and Predicting Image Memorability at a Large Scale. In 2015 International Conference on Computer Vision (ICCV). 2390–2398. [2] Romain Cohendet, Claire-Hélène Demarty, Ngoc".