Short-term and Long-term Video memorability score prediction using Machine Learning

Tejal Nijai (Student No:19210412)
Dublin City University, Ireland
tejal.nijai2@mail.dcu.ie

ABSTRACT

Memorability is described as being easy to remember in quality or state [1]. In the present age, given the high growth rate of emerging technology and the exponential growth in Internet access, loads of images, textual information and videos have been produced every unit of time. It is beneficial to use this data for extraction of information. For processing videos, there are many features can be worked on such as semantic, visual, audio, motion of the pictures or image features. Within this paper, I worked on video (C3D and HMP), semantic (Captions) and image (InceptionV3) features with different machine learning models and neural network to predict the performance of short-term and long-term video memorability.

1 INTRODUCTION

With the dramatic surge of visual media content on platforms like Instagram, Flicker and YouTube, it is imperative that new methods for curating, annotating and organizing this content be explored [6]. There are several techniques available that examine the predictive images and texts. Noting that videos are now accessible on an immense scale and their content has valuable data should be processed to extract data. For example, using video memorability score (short term and long term) of videos can used for target marketing. Through analyzing and predicting video memorability, we will help make better videos that will draw more audience interest to improve sales. Experimenting with different methods and video features enables us identify which area to focus on which makes the video the most memorable to viewers. In this work, the main objective is to predict the memorability score of a video to indicate how memorable a video will be [4]. For better prediction, video's semantic features, visual features and image features can be processed. There are many video features such as C3D, HMP and captions or visual features (LBP, ORB, ColorHistogram, HOG and InceptionV3) can worked on for analysis and prediction. For my research, I have used C3D, HMP,InceptionV3 and captions features of video. I trained these features individually and in combination on different machine models and a neural network model. The Spearman's correlation score as a measure has been referred for evaluation of model's performance.

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 [2], effect of stimuli on encoding and later retrieval from memory [1], memory systems of the brain [7][3]etc. Considering the importance and attention to this area, there has been a lot of work is done using different techniques of machine learning for example, neural networks, CNN models, multimodal approach and transfer learning of some pre-trained models. There is a multimodal approach in which 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 [3]. In [5], three simple, linear and regularized models were chosen – L1 Regularized Logistic Regression, Linear Support Vector Regression, ElasticNet.

3 APPROACH

In this section, I will outline the approach to model media memorability using video,image features and captions.

3.1 Feature Extraction and Pre-processing

I used visual features (C3D and HMP) ,captions and image features (InceptionV3) to explore with what features could be useful for predicting good memorability value. The C3D, HMP and InceptionV3 features had been provided in .txt files. Feature lists for the corresponding C3D, HMP, and InceptionV3 functions were created by processing the respective files. The ground truth file and the respective short-term and long-term memorability scores were provided with video captions. The captions were cleaned up by removing special characters, turning captions into small cases, and deleting stop words. The collected words were used to create a bag of words. To get features this word bag was run with Scikit-learn's CountVectorizer. The method of extracting features takes a lot of time as there are about 6k development set files and 2k test set files for each C3D and HMP feature while three files have been extracted for inceptionV3 for each video so it is around 18k development set files and 6k test set files to work on. I saved these extracted features in data frame using python's numpy library to reuse it later to save time and redundant work. Extracted C3D, HMP and InceptionV3 features were combined with captions to create new feature lists such as C3D and captions, HMP and captions, and C3D, HMP and captions, as well as captions, C3D, HMP and InceptionV3. These new combinations of feature lists were fitted to traditional machine learning models such as Support vector Regression, XGBoost and Bayesian Ridge Regression. I chose two features set that gave these conventional machine learning models a better result to train them on a neural network model to try to improve performance.

3.2 Machine Models applied on features

Following Machine Learning Models have been applied:

- Support Vector Regression Model
- XGBoost Regression Model

,

,

- Bayesian Ridge Regression Model
- Neural Network Model

For each of the function combinations, above-mentioned machine learning models were used to arrive at the best one. Such models are often used for better prediction, and appear to produce good results. The tuning of these models' hyper-parameters is a major benefit in making better models. I used general purpose kernel for SVR, which is used when there is no prior knowledge of the data. I used XGBoost because of the speed of run and the flexibility of the interface it offers. Before hitting the parameters, I experimented with different parameters which gave us good results. One such parameter is Alpha which is used to reduce over-fitting patterns. The colsample_bytree hyper-parameter specifies the percentage of functions (columns) used to create each tree. The set of features is likely to be different for each tree. The target parameter was set to reg: linear for linear regression. The n estimators are used to indicate the size of the trees to be built. In the Bayesian ridge model, compute score has been set to true which calculates the marginal probability of log at any iteration for optimization.I used relu and sigmoid activation functions to construct my neural network with one of the best optimizer Adam. I tried to enhance performance by trying out different number of epochs. From which, epoch=50 gave me good results.

4 RESULTS AND ANALYSIS

The results for different types of combinations of features have been listed in Table 1 and Table 2. The ratings are determined using a correlation from Spearman.

Feature	***XGBoost***	***SVR***	***Bayesian Ridge***
Captions	0.318	0.419	0.447
C3D	0.272	0.247	0.282
HMP	0.259	0.319	0.302
InceptionV3	0.019	0.001	0.037
Captions + C3D	0.297	0.424	0.447
Captions + HMP	0.275	0.420	**0.448**
Captions + C3D + HMP	0.297	0.424	0.447
Captions + C3D+ HMP+ InceptionV3	0.321	0.424	0.447
Captions	**Neural Network Score :** 0.354		
Captions + C3D+ HMP+ InceptionV3	**Neural Network Score :** 0.348		

Table 1: Short-Term Memorability Scores

From the Table 1 and Table 2, it can be seen that, when a combined feature set of captions, C3D, HMP and InceptionV3 applied to each type of machine learning model, the respective shot-term and long-term memorability scores are higher. Noting that, I extracted

Feature	***XGBoost***	***SVR***	***Bayesian Ridge***
Captions	0.120	0.179	0.179
C3D	0.135	0.037	0.126
HMP	0.098	0.103	0.112
InceptionV3	0.012	0.037	0.066
Captions + C3D	0.110	0.190	**0.193**
Captions + HMP	0.120	0.180	0.180
Captions + C3D + HMP	0.110	0.190	**0.193**
Captions + C3D + HMP + InceptionV3	0.128	0.190	**0.193**
Captions	**Neural Network Score :** 0.166		
Captions + C3D+ HMP+ InceptionV3	**Neural Network Score :** 0.172		

Table 2: Long-Term Memorability Scores

the respective features for the test dataset and selected Bayesian Ridge model among all models that had the best predictive results. The final estimated short-term and long-term scores were then stored in **Results_Final.csv**.

5 CONCLUSION AND FUTURE WORK

Throughout this work, it was evident that better results are obtained using the semantic features(captions). Another point is that when I experimented with the combination of video, image and captions, the prediction score showed a marked increase over the individual feature implementation. It is possible to explore audio or motion characteristics to know their effect on the score for predictability.The more features used in combination, it seemed to have a high memorability ranking. Traditional machine learning approaches and a neural network were used for exploration. These features can further be experimented on deep neural network models and convolutional neural networks where different optimizers, activation functions and playing with different number of epochs runs of the model can increase the output.

REFERENCES

[1] Wilma A. Bainbridge, Daniel D. Dilks, and Aude Oliva. 2017. Memorability: A stimulus-driven perceptual neural signature distinctive from memory. *NeuroImage* 149, October 2016 (2017), 141–152. https://doi.org/10.1016/j.neuroimage.2017.01.063

[2] 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 of the United States of America* 105, 38 (2008), 14325–14329. https://doi.org/10.1073/pnas.0803390105

[3] Ritwick Chaudhry, Manoj Kilaru, and Sumit Shekhar. 2018. Show and Recall @ MediaEval 2018 ViMemNet: Predicting video memorability. *CEUR Workshop Proceedings* 2283, October (2018), 29–31.

[4] Romain Cohendet, Claire Hélène Demarty, Ngoc Q.K. Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh Toan Do. 2018. MediaEval 2018: Predicting media memorability. *CEUR Workshop Proceedings* 2283 (2018), 11–13. arXiv:arXiv:1807.01052v1

[5] Rohit Gupta and Kush Motwani. 2018. Linear models for video memorability prediction using visual and semantic features. *CEUR Workshop Proceedings* 2283 (2018), 2–4.

[6] Tanmayee Joshi, Sarath Sivaprasad, Savita Bhat, and Niranjan Pedanekar. 2018. Multimodal approach to predicting media memorability. *CEUR Workshop Proceedings* 2283, October (2018), 29–31.

[7] 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 of the United States of America* 93, 24 (1996), 13508–13514. https://doi.org/10.1073/pnas.93.24.13508