Predicting Media Memorability Using Regression Algorithms and Ensemble Techniques

Ajit Kumar
[1]Dublin City University, Ireland
ajit.kumar7@mail.dcu.ie

ABSTRACT

Generally, various visual media are unequally memorable by the human brain [1]. The media Memorability Challenge focuses on determining and how unforgettable a video is. This paper explains the method developed to forecast video memorability scores in the short and long term by using different supervised machine learning regression algorithms and ensemble weighted average techniques based on a textual feature like caption, aesthetic feature and visual feature of a video.

1 INTRODUCTION

With the exponential development of research in the field of image memorability, the movement towards video memorability has now been expanded, and is an important area in computer vision science. [2]. Since the internet has transformed the planet into a digital village, billions of videos are posted every day to the popular sites, such as Twitter, YouTube, Instagram and Flickr, and millions of people watch such videos every hour [3]. This is increasing at an alarming rate which has caused problem to the education sector, entertainment industries, and many more who are inclined towards digital media for their growth and development.

So, this is an important motivation for video memorability prediction as there is an utmost need to develop new techniques so that these digital media can be store based on their long-term and short-term memorability score and they can be retrieve in an optimize way which is of utmost importance.

2 RELATED WORK

New strategies for modeling the memorability of video clips and automatically predicting how memorable such videos are by using brain functional magnetic resonance imaging (fMRI) were described in [1] by the author. Visual features, audio features, etc. have been included in this author to accomplish their purpose. The purpose of this paper was to build a model for predicting video memorability using algorithmically extracted features.

Using natural language processing (NLP) techniques and a recurrent network (RNN), this approach uses the scenario semantics derived from the video names. The efficiency of semantic-based methods is compared with those of aesthetic feature-based methods using SVR and ANN models, and the likelihood of predicting the highly subjective media memorability is examined using simple features. The model using RNN and semantics is the best among all the five models [4].

In [5] the author has discussed the use of different visual and semantic attributes in the video memorability prediction models. CNN model is used for semantic features and the author has concluded that the caption-based model performed better than C3D based model.
In [6] multimodal approach has been used for the predictions of video memorability and author concluded that the combination of both visual and textual features gave better result.

3 APPROACH

Going through the previous research in the same domain has helped me a lot in pursuing a standardized methodology and choosing the right features so that I can get the best possible outcome and I will address the same in the sub-section listed below.

3.1 Feature Selection and Data Preprocessing

Based on the specific studies, I have selected text feature like caption and visual features like C3D and aesthetic to move forward in this research and among them Caption received the best results for short-term and long-term forecasts of memorability.

I cannot go with raw text fitting a machine learning or deep learning model. I have stated with first normalizing case means the entire text has been converted to one case and then followed by removing the punctuation marks, stop words and stemming the words. After that performed vectorization on the cleaned text to transform the text to feature vector that can be used as an estimator using TfidfVectorizer form Scikit-Learn [6].

Dependent variables are normalized in various to achieve better predictions in various models like LSTM and MultiTaskLassoCV. k-Fold Cross-Validation which is defined as a re-sampling method used to validate the machine learning models. For C3D feature model k-Fold Cross-Validation is used along with Random Forest.

3.2 Model Selection

In this study, six different machine learning algorithms like Random Forest, Bayesian Ridge, Linear Regression, Support Vector Regressor, MultiTaskLassoCV, LSTM, and along with ensemble weighted average.

For caption three models like Random Forest, Bayesian Ridge, and Long short-term memory (LSTM) has been used. In which Bayesian Ridge has outperformed both Random Forest and LSTM in both short-term video memorability and long-term memorability.

For C3D Support Vector Regressor (SVR), Random Forest and MultiTaskLassoCV have been used. In which Random Forest has shown promising result for short-term memorability and MultiTaskLassoCV have shown good result for long-term video memorability.

For aesthetic feature, Random Forest, Bayesian Ridge, Linear Regression has been used. In which Random Forest has performed well for short-term memorability and linear regression has performed better than both Random Forest and Bayesian Ridge for long-term memorability scores.

Based on the long-term and short term score one model form each feature has been selected like for caption Bayesian Ridge perform better, for C3D Random Forest performs better in short term and MultiTaskLassoCV has shown good score for long-term and for aesthetic Random forest and linear regression have shown great result in short-term and long term respectively. All these models mentioned above model have been combined for ensemble techniques.

3.2 An Ensemble of Top Three Models

The Weighted average ensemble is a method in which multiple model predictions are used to get new model predictions which are proportional to the estimated performance of the combined model [7].In this method, the top three models have been chosen based on the Spearman's correlation coefficient score and are assigned different weights which sum up to 1.

For short term memorability, Bayesian Ridge for (caption) and Random Forest for (C3D and Aesthetic Feature) is used.
For Long term memorability, Bayesian Ridge for (Caption) and for C3D MultiTaskLassoCV is used and Linear Regression is used for Aesthetic Feature.

4 RESULTS AND ANALYSIS

Below mentioned Tables 1 and 2 show the comparison of different methods used based on short term and long-term memorability scores. The results are evaluated based on Spearman's Correlation Coefficient. For short-term memorability, the Bayesian Ridge model has shown the best result, whereas MultiTaskLassoCV has shown promising results in long-term memorability. At last ensemble, techniques are used in which the best performing model in each feature is used to calculate the spearman score and the same have been used to predict the final short-term and long-term memorability. Below mentioned combination of weighted average are used to get better results.

Short-term ensemble
(0.5 * caption) + (0.25 * C3D) + (0.25 * Aesthetic)
Long-term ensemble
(0.5 * caption) + (0.25 * C3D) + (0.25 * Aesthetic)

Short-Term Memorability		
Features Name	**Model Used**	**Spearman Score**
Caption	Bayesian Ridge	**0.417**
	LSTM	0.412
	Random Forest	0.398
C3D	Random Forest	0.323
	MultiTaskLassoCV	0.313
	SVR	0.307
Aesthetic Feature (Median)	Random Forest	0.302
	Linear Regression	0.272
	Random Forest	0.251
Caption + C3D + Aesthetic (Median)	**Ensemble =** (Bayesian Ridge + Random Forest + Random Forest)	**0.445**

Table 1. Short Term Memorability Using Spearman Correlation

Long-Term Memorability		
Features Name	**Model Used**	**Spearman Score**
Caption	Bayesian Ridge	0.154
	Random Forest	0.153
	LSTM	0.151
C3D	MultiTaskLassoCV	**0.184**
	Random Forest	0.143
	SVR	0.141
Aesthetic Feature (Median)	Linear Regression	0.126
	Random Forest	0.125
	Bayesian Ridge	0.092
Caption + C3D + Aesthetic (Median)	**Ensemble** (Bayesian Ridge + Linear Regression + MultiTaskLassoCV)	**0.163**

Table 2. Long Term Memorability Using Spearman Correlation

5 CONCLUSIONS

In this work, various supervised machine learning algorithm and LSTM is used for predicting the short term and long-term video memorability. Visual features along with the textual feature of the video are used. Captions based model performed better in terms of short-term video memorability and C3D performed better in long-term video memorability. On the other hand, only visual features provide the lowest predictive correlation values which are significantly lower than textual features. Based on the above-shown tables better results can be obtained by combining visual and textual features.
For future work, I would like to explore other features like HMP, HOG, and color histogram, etc. To improve the performance of both long term and short-term video memorability.

REFERENCES

[1] J. Han, J. Han, C. Chen, L. Shao, X. Hu, and T. Liu, "Learning Computational Models of Video Memorability from fMRI Brain Imaging," IEEE Trans. Cybern., vol. 45, no. 8, pp. 1692–1703, 2015.
[2] R. Cohendet, C. H. Demarty, N. Q. K. Duong, M. Sjöberg, B. Ionescu, and T. T. Do, "MediaEval 2018: Predicting media memorability," CEUR Workshop Proc., vol. 2283, no. July 2018
[3] S. Shekhar, D. Singal, H. Singh, M. Kedia, and A. Shetty, "Show and recall: Learning what makes videos memorable," Proc. - 2017 IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017, vol. 2018-January, pp. 2730–2739, 2017.
[4] W. Sun and X. Zhang, "Video memorability prediction with recurrent neural networks and video titles at the 2018 MediaEval predicting media memorability task," CEUR Workshop Proc., vol. 2283, pp. 4–6, 2018.

[5] R. Gupta and K. Motwani, “Linear models for video memorability prediction using visual and semantic features,” CEUR Workshop Proc., vol. 2283, pp. 2–4, 2018.
[6] T. Joshi, S. Sivaprasad, S. Bhat, and N. Pedanekar, “Multimodal approach to predicting media memorability,” CEUR Workshop Proc., vol. 2283, no. October, pp. 29–31, 2018.
[7] A. F. Smeaton et al., “Dublin’s participation in the predicting media memorability task at MediaEval 2018,” CEUR Workshop Proc., vol. 2283, pp. 7–9, 2018.
[8] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, “Understanding and predicting image memorability at a large scale,” Proc. IEEE Int. Conf. Comput. Vis., vol. 2015 International Conference on Computer Vision, ICCV 2015, pp. 2390–2398, 2015.