Predicting Long-Term and Short-TermVideo Memorability using Machine Learning

Divya Aren
Student Number 20210762
MCM Computing
Dublin City University
Dublin, Ireland
divya.aren@gmail.com

***Abstract*—With the immense increase in the generation of video content on the Internet, video research and analysis has proved to be a budding area in research for which human cognition is required to be taken into account. Video memorability is one such cognitive measure which is the ability to recall the content after watching it. Memorability is an important area of study for industries like Marketing and Advertising, Entertainment, Online Study Camps, and Edu-Techs as it is directly proportional to the impact the video content can create. In this study, we have attempted to predict the short-term and long-term memorability using classic machine learning models. We use multiple features and their combinations to identify the contrast different models and recognize the model which is best at predicting long-term and short-term memorability.**

Keywords—video memorability, video annotation, features, machine learning, random forest, support vector regression.

I. Introduction

Videos are exploding on the Internet today with so many websites and applications like TikTok, YouTube, Facebook, Instagram, etc. which are heavily focused on creating new digital content that is being forwarded multiple times in a day. The concept of a video being viral is not new to hear in our everyday lives. Thus, it is in the interest of these multimedia organizations to incubate strategies for curation of such content which can trend worldwide using advanced technologies like machine learning. Advances in the study of human-computer interaction have enabled us to take cognizance of the psychological and cognitive factors in designing an effective system. Like other cues of video importance, such as aesthetics or interestingness, memorability can be regarded as useful to help make a choice between otherwise comparable videos. [6] To define, memorability is the ability to recall content after viewing it. The significance of Short-Term and Long-Term Memorability is that Short-Term memorability is the ability to recall information after just a short time of viewing and Long-Term memorability defines the metric of recalling ability after a long time of viewing the content. The study discusses the relevance of memorability as a component of video analytics and is helpful in identifying the best of the classical machine learning algorithms to predict the long-term and short-term memorability. Spearman's Correlation Scores are used to evaluate the classical machine learning models at this task.

II. Literature Review

Survey-based recall experiments were the earlier approach to model video memorability. Han et al [2] in his study discusses the approach in detail where they deploy a survey for about 20 participants who were initially made to watch several videos played together in a sequence. This was then followed by a recall task after 48 hours or even a week, where they were asked if they remember the videos being shown. Video Memorability has gained a lot of traction with the analytics community interested to investigate the use of various low- and high-level visual features [4], Convolution, Spatio-temporal 3D features [1], movements in a video by identifying the patterns of motion and using memory colour maps. In general, consistently identifying which videos and which parts of a video are memorable or forgettable could be used intrinsically for identifying visual data useful for people, concisely representing information, and allowing people to consume information more efficiently [3] with the help of machine learning.

III. Approach

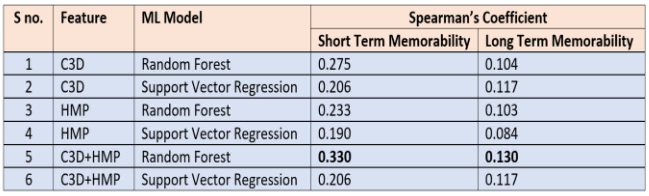
In this study, we investigate the various visual features to predict video memorability. The paper is an attempt to successfully conduct an analysis of the available features and to utilize principles of feature engineering to develop an accurate prediction of long-term and short-term video memorability. The dataset which is published as part of the Media Eval's 2018 challenge comprises 8,000 videos. It has been divided into an official test set of 2,000 videos and a development set of 6,000 videos. The corresponding ground truth for the development set can be found in file ground-truth_dev-set.csv. It contains one line per video, which consists of the video's name, its short-term memorability score, the number of annotations that were used to calculate the short-term memorability score, its long-term memorability score, the number of annotations that was used to calculate the long-term memorability score.

It was important to understand the different features of videos. The challenge deals with some pre-computed features like C3D and HMP which have been the input to our models. Convolution 3D (C3D) feature, a generic spatio-temporal feature obtained by training a deep 3-dimensional convolutional network on a large, annotated video dataset comprising objects, scenes, actions, and other frequently occurring concepts. C3D has three main advantages. First, it is generic: achieving state of the art results on object recognition, scene classification, and action similarity labelling in videos. Second, it is compact: obtaining better accuracies than best hand-crafted features and best deep image features with a lower-dimensional feature descriptor. Third, it is efficient to compute: 91 times faster than hand-crafted features, and two orders of magnitude faster than current deep-learning based video classification methods. [1]. It consists of 101 dimensions. HMP on the other hand is helpful to recognize the movements in a video by identifying the patterns of motion. It consists of 6075 dimensions.

The classical machine learning algorithms Random Forest and Support Vector Regression (SVR) were used to process the input features and predict memorability. SVR acknowledges the presence of non-linearity in the data and provides a proficient prediction model [6] whereas Random forest is the combination of multiple individual decision trees to act as an ensemble. [7]. The motivation behind using Random Forest is that the combination of outputs of mutually exclusive nodes will outperform any individual models which then vote for the said the predicted output.

IV. Results And Evaluation

Spearman’s Correlation Coefficient is used to evaluate the different models. The value of Spearman’s coefficient lies between -1 and 1. The Below table depicts results after fitting and prediction.



S no.	Feature	ML Model	Spearman’s Coefficient	
			Short Term Memorability	Long Term Memorability
1	C3D	Random Forest	0.275	0.104
2	C3D	Support Vector Regression	0.206	0.117
3	HMP	Random Forest	0.233	0.103
4	HMP	Support Vector Regression	0.190	0.084
5	C3D+HMP	Random Forest	**0.330**	**0.130**
6	C3D+HMP	Support Vector Regression	0.206	0.117

Fig 1: Spearman’s Coefficient for Evaluation

We find that C3D and HMP when used with Random Forest provides the best results of all the combinations. Hence, we use this model to predict the memorability scores.

V. Conclusion and Future Work

In conclusion, using more than one feature is more useful in prediction. Using only HMP or C3D did not yield acceptable results. A combination of HMP and C3D resulted in the most optimum results with the classical machine learning model, Random Forest.

This paves a path for further studying deep learning and convolutional neural networks and comparing their results with classical machine learning algorithms. Using video annotation to extract more features or using other features available like Aesthetics, Inception, HOG, Color Histogram and to create an ensemble of these methods is also interesting and would be undertaken next.

VI. References

[1] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, “C3D: Generic Features for Video Analysis”, 2014.

[2] J. Han, C. Chen, L. Shao, X. Hu, J. Han, and T. Liu., “Learning computational models of video memorability from mri brain imaging”, IEEE transactions on Cybernetics, 45(8):1692–1703, 2015.

[3] A. Khosla, A. S. Raju, A. Torralba and A. Oliva, "Understanding and Predicting Image Memorability at a Large Scale," 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[4] R. Gupta and K. Motwani, “Linear Models for Video Memorability Prediction Using Visual and Semantic Features.” , 2018.

[5] D. Azcona, E. Moreu, F. Hu, T. E. Ward, A. F. Smeaton, “Predicting Media Memorability Using Ensemble Models”, 2019.

[6] “Benchmarking Initiative for Multimedia Evaluation”, MediaEval, 2019.

[7] A. Sethi, “Support Vector Regression Tutorial for Machine Learning”, www.analyticsvidhya.com, 2020.

[8] Prem, “Which is better – Random Forest vs Support Vector Machine vs Neural Network”, www.iunera.com, 2021.