Predicting Video Memorability Scores Using Machine Learning Models

SAHITHI RAYIDI
20210916
School of Computing
Dublin City University
sahithi.rayidi2@mail.dcu.ie

ABSTRACT

"Memorability is characterized as the condition that is worth recollecting or simple to recall ". In this digital world, there are billions of recordings accessible that get one's consideration on an everyday premise. If these recordings are recalled by individuals then, is there any kind of pattern among these recordings because of which the individuals are able to remember them? This paper explains how the machine learning models considering spearman's score as an evaluation metric are used in analysing the pre-extracted features like HMP, C3D in predicting the short and long term memorability of the video recordings.

Keywords - Memorability, Machine learning, Spearman's scores

I. Introduction

In recent years, there have been many advancements in computing devices that have a greater positive impact on applications that can increase memorability. The focus of this work is predicting the probability of an individual remembering a video/recording using ground truth values which are provided. These ground truth values are composed of long & short-term memorability scores of each video and its respective annotations, in a .CSV file. To train our models, pre-extracted video, image and semantic features of these recordings,such as- C3D, HMP, ColorHistogram, ORB, Inception V3, LBP, HOG and captions are provided. The image features are extracted by pulling 3 frames from the beginning(0th frame),the middle(56th frame), and the end((112th frame) of every recording [Herrera 2020]. This paper explores and analyzes the video features (C3D, HMP) vastly to develop a predictive model on the probability of memorising the given recordings.

This paper is aligned as follows. Section II. introduces related work on MediaEval 2018 competition and also on existing video memorability approaches. In Section III, discuss the video features extraction and how they are explored. Section IV gives a description of the machine learning models used in this system. Section V presents the results and their analysis. Concluding remarks and future possibilities are outlined in Section VI.

II. Related Work

Author R.Gupta et al. [2018] builds a model that has achieved the best memorability scores in the MediaEval Competition 2018 task. The model was developed using semantic(captions) and a combination of video features. In their paper, the authors emphasise that video features(C3D & HMP) outperform image features. They train their best model with a combination of semantic features and video features.It is an ensemble of caption and Resnet predictors. The drawback of their work is that they have calculated the positive and negative coefficients to identify the words in the captions sentences and concluded that words associated with nature have negative coefficients and words associated with humans have positive coefficients.

III. Feature Extraction and Exploration

The model presented in this paper is trained on video features(C3D & HMP) which were extracted directly from the recordings in text files with 101 float values in C3D and 6074 float values in HMP, firstly these text files are appended into a list and then converted into dataframe to fit them into models by splitting 80 to 20 ratio as train and test. Video features are considered in this approach because, while taking real life into account, we know that videos create an impact during their course, however we cannot pinpoint the specific image in the video that causes this impact. All the image features extracted in this dataset are obtained by pulling frames from the beginning, middle, and the end of each video. However, these frames are not necessarily representational therefore, the proposed model was not trained with them

C3D is a deep 3 dimensional Convolutional Neural Networks with homogeneous architecture consisting of 3x3x3 convolutional Kernels and with 2x2x2 pooling at each layer[Almeida 2011]. This architecture helps in extracting the generic features of the videos and their compact representation in an efficient manner.

HMP is defined as a process of recognizing patterns of motions which are extracted from the video stream and their occurrence histogram is demonstrated to be a powerful feature for describing the content in video[Binu 2018]. HMP features given in our data has 6074 float values which describe certain vedio.

IV. Models

Many features in this task have high multicollinearity, making it difficult to predict the high P-values, increasing the risk of overfit models. To alleviate this problem, The model underwent parameter tuning and was trained with K-fold validation. I chose to work with effective yet computationally light models such as-

- Decision Tree Regressor
- Linear Regression
- KNN Regressor
- Random Forest Regressor

Each of the models listed above is trained on C3D and HMP(video features) individually.

V. Results and Analysis

Spearman's score is considered as a non-parametric measure for rank correlation and is used to calculate the results in this task. Models trained on C3D and HMP features, get the best results with the Random Forest Regressor algorithm for both short-term and long-term scores. Random Forest regressor results were found to be consistent with various training and validation splits.

Short Term Memorability Spearman's Scores				
Features	**Decision Tree**	**Linear Regression**	**KNN**	**Random Forest**
C3D	0.087	0.305	0.303	**0.338**
HMP	0.085	0.158	0.14	0.223

Table 1 : Results for short-term prediction

Long Term Memorability Spearman's Scores				
Features	**Decision Tree**	**Linear Regression**	**KNN**	**Random Forest**
C3D	0.021	0.150	0.133	**0.159**
HMP	0.061	0.096	0.000	0.063

Table 2 : Results for long-term prediction.

After exploring and experimenting with several visual features and their corresponding characteristics I chose C3D & HMP to develop this predictor model. While being goode features they do not allow us to engineer any new features.

HMP features consistently underperform, when compared to C3D. I attribute this to the time series based extraction of the HMP features which can better describe the temporal variation of video sequence than the spatial arrangement. In contrast, the C3D features are generated using generic features and compact representations.

When C3D Spearman's scores are compared in Table.2 there is no much variation between both the Linear Regression and KNN model but a slight difference is observed when these both are compared with the Random Forest model which attained the highest value for long-term scores. Short-term scores from table.1 incur a slight variation between Linear regression and Random Forest model where the Random forest outperformed all other 3 models with the highest score.

The Final Predictor Model was built using Random Forest Regressor with C3D varying the size of trees in the Random Forest. The predicted short and long-term scores are generated and stored in predictions.csv file.

VI. Conclusion and Future Work

I conclude this paper, by stating that C3D provides better results when used as features to predict the memorability scores. However, Combining C3D with other features such as captions, HMP fetched better results for other researchers, captions are semantic features which can be used to extract new features as well. As a part of further work, I would like to develop a model of C3D and image features together to build a predictor using ResNet50. Also, concentrating on increasing the long-term memorability scores is another aspect that can be analysed.

REFERENCES

[1] Herrera, A. G. S. D., Rukiye Savran Kiziltepe, Jon Chamberlain, M. Constantin, C. Demarty, F. Doctor, B. Ionescu and A. Smeaton. 2020. Overview of MediaEval 2020 Predicting Media Memorability Task: What Makes a Video Memorable? ArXiv abs/2012.15650..

[2] R. Gupta, 2018. Linear Models for Video Memorability Prediction using Visual and Semantic Features. MediaEval.

[3] Almeida, Jurandy & Leite, Neucimar & Torres, Ricardo. 2011. Comparison of video sequences with histograms of motion patterns. Proceedings - International Conference on Image Processing, ICIP. 3673-3676. 10.1109/ICIP.2011.6116516.

[4] Binu M. Nair. 2018. Deep Dive into Convolutional 3D features for action and activity recognition (C3D). https://medium.com.