Predicting Video Memorability Using Machine Learning

Kirthy Francis
19210588
MCM Computing (Data Analytics)
Dublin City University
Dublin, Ireland
kirthy.francis@mail.dcu.ie

ABSTRACT

A torrent of videos has been appearing on the internet in the current era of widening Internet access. It is becoming increasingly important to research the human cognitive factors influencing the consumption of these images, so that they can be coordinated and curated effectively. Video Memorability is one such significant cognitive element, which is the ability to remember content of a video after viewing it. Predicting video memorability has a lot of applications. For example, we might use the video content with high memorability scores for target marketing, creating greater impacts on people that contribute to higher sales. This study uses semantic as well as video features to predict long-term and short-term memorability.

KEYWORDS

Media Memorability, C3D, Captions, HMP, Deep Learning, Gradient Boosting, Decision Trees, Linear Regression, Random Forest

1 INTRODUCTION

Through this study, I explored the use of various visual and semantic features in predicting video memorability, and then perform comprehensive analysis of the features chosen to establish a reliable predictor of video memorability. While designing the model for prediction, I analyzed visual features like C3D spatio-temporal visual features that are obtained by extracting the output of the final classification layer of the C3D model, a 3-dimensional convolutional network proposed for generic video analysis, and HMP, the histogram of motion patterns for each video [3] and semantic feature - captions of each video. Initially, I trained the models individually with each of these features and then I tried different combinations. Later, to train my model I took a combination of the best performing video elements - C3D,HMP and captions. The models were assessed using a standard measure of Spearmann's correlation value.Within the total of 8,000 videos that were annotated, 6,000 of them were used as dev-set for training and testing different models, and the remaining 2,000 videos were reserved for the test-set.

My key observations are as follows

- The short-term memorability scores were predicted more precisely, for any given model, than the long-term memorability scores
- Higher number of estimators in Random Forest and Gradient Boosting Regressor give better results, even though it creates high computation overhead.
- Random Forest model constantly gives good accuracy, but Gradient Boosting gave the best accuracy using Captions, C3D and HMP features.

The remainder of this paper is divided into the following sections: Section II is a literature review that discusses the previous relevant work, Section III is a brief overview of the feature extraction, data preprocessing and the Machine Learning (ML) models used. Section IV illustrates the results and Section V addresses the future work and conclusion.

2 RELATED WORK

Three simple linear models-L1 Regularized Logistic Regression, Linear Support Vector Regression, ElasticNet-were selected in [4] which used video features, such as HMP and C3D directly, while frame-level features such as ColorHistogram and LBP were concatenated using frames. To boost the precision, they built an ensemble of their best models. Meanwhile, [5] uses C3D, among other features and explains the model factors that need to be considered to apply C3D features effectively. Also, I used the example solution from Eoin Brophy as a basis for extracting and pre-processing my captions and calculating my Spearman Coefficient [2]

3 APPROACH

While analysing the work of past teams that achieved good accuracy, I discovered that a combination of the features achieved higher Spearman scores. Also, the highest scoring teams used only two to three features and still performed well. It was also observed that C3D features were rated as one of the best features. Taking all these factors into consideration, I concluded that I would try out combinations of Captions, C3D, and HMP to find the best model.

After further research on which model would suit both the features and the data, I came upon decision trees and decided to try different decision tree based models.

3.1 Models

I started off with simple linear regression models and then worked with boosting regression models. Totally I ran the features on five models:

(1) Linear Regression Model
(2) Decision Tree Regression Model
(3) Random Forest Regression Model
(4) XGBoost Regression Model
(5) Gradient Boosting Model

After running each of these models individually and in combinations of the features, I decided to use the Gradient Boosting Regressor with a combination of all the features.[1]

3.2 Features and Data Pre-Processing

The final model used both semantic and video features. The deep learning model that I implemented is an ensemble decision tree. The python code was implemented on a Jupyter Notebook.

The **semantic feature** – **captions** consisted of a one-sentence description of each video. The video captions were loaded into a data frame using the Pandas library. A lot of text processing needed to be done on the captions before they were run on the models. The captions were cleaned by removing special characters, converting captions into lower case. The occurrences of each word were counted and the captions were split up using python tokenization. Then each word was mapped to an index and the captions were stored an integer sequence of length 50. If the length of the sequence was less than 50, it was padded with zeroes. These sequences were stored in a dataframe.

The collection of the **video features** – **C3D** and **HMP** features are loaded into dataframes similar to the way captions were. The data frame is again pre-processed to divide each characteristic value into a separate column. That has resulted in 101 feature columns for C3D and 6075 feature columns for HMP. These columns reflect different types of classifications for a scenario, object, and action for C3D and histogram features for HMP.

When all of these data frames have been generated and pre-processed, they are merged using Pandas reduce function on video names into another dataframe. First, the individual feature data frames are passed to each of the models after splitting into training and validation sets. The models are first trained on the development set 80 percent (4,800 videos) and validated against 20 percent (1,200 videos). This split was used to measure the Spearman coefficient scores of the short-term and long-term spearman and arrive at the best model. Then a combination of captions and C3D was run on the models. And finally a combination of all three features – C3D, HMP, and captions- was run. The Random Forest Regression Model run with n_estimators=100. Gradient Boosting Regressors is an ensemble decision tree model that is very flexible. The model had n_estimators set to 650 and 12 layers which take a lot of computing power. The extreme Gradient Boosting method is a specific implementation that uses more accurate approximations to find the best tree model. This model also had n_estimators set to 100.

Once all the above models were applied to the features, **Gradient Boosting Regressors** were found to give the best accuracy with a combination of all the features. The model was then trained using the entire 6,000 development set and a prediction was done against the test set of 2,000 videos. Preprocessing for the test set was performed in the same manner as the development set.

4 RESULTS AND ANALYSIS

The evaluation metrics for my results are calculated using Spearman's rank correlation. The results are tabulated in Table 1.

Feature	Model	Short-term	Long-Term
Captions	Linear Regression	-0.003	-0.042
	Decision Tree	0.143	0.058
	Random Forest	0.254	0.083
	XGBoost	0.246	0.132
	Gradient Boosting	0.250	0.075
C3D	Linear Regression	0.277	0.103
	Decision Tree	0.124	0.033
	Random Forest	0.332	0.113
	XGBoost	0.151	0.034
	Gradient Boosting	0.320	0.097
HMP	Linear Regression	0.036	-0.004
	Decision Tree	0.109	0.042
	Random Forest	0.309	0.114
	XGBoost	0.202	0.076
	Gradient Boosting	0.279	0.123
Captions + C3D	Linear Regression	0.273	0.081
	Decision Tree	0.086	0.011
	Random Forest	0.303	0.116
	XGBoost	0.154	0.029
	Gradient Boosting	0.323	0.104
Captions + C3D + HMP	Linear Regression	0.019	0.001
	Decision Tree	0.085	0.044
	Random Forest	0.341	0.132
	XGBoost	0.170	0.075
	Gradient Boosting	0.350	0.142

Table 1: Short-term and Long-term Spearman Coefficient Scores by Model

5 CONCLUSION AND FUTURE WORK

From the table it can be observed that the Gradient Boosting Model produces much better scores. This is because the model fares much better when multiple features are fed in. In all cases the short term memorability of videos is more predictable than long term, since all model predictions score higher in the short term memorability of videos. I think a larger dataset would give better predictions.Also for further research, there is room for the exploration of captions as some of the past teams have achieved high accuracy with the captions features, by adding weights to words, etc.

REFERENCES

[1] 2019. Gradient Boosting Regressor. (2019). https://opendatagroup.github.io/KnowledgeCenter/Tutorials/GradientBoostingRegressor

[2] Eoin Brophy. 2019. Google Colaboratory. (2019). https://colab.research.google.com/drive/1X7l5MGrDZa2IdMCOxwgILCD5CzHKE8NF?authuser=1

[3] Romain Cohendet, Claire-Hélène Demarty, Ngoc Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. 2018. Mediaeval 2018: Predicting media memorability task. *arXiv preprint arXiv:1807.01052* (2018).

[4] Rohit Gupta and Kush Motwani. Linear Models for Video Memorability Prediction Using Visual and Semantic Features.

[5] Ricardo Manhães Savii, Samuel Felipe dos Santos, and Jurandy Almeida. 2018. GIBIS at MediaEval 2018: Predicting Media Memorability Task.. In *MediaEval*.