Video Memorability Prediction Using Machine Learning

Aruna Bellgutte Ramesh
18210858
MCM Computing (Cloud Computing)
Dublin City University
Dublin, Ireland
aruna.bellgutteramesh2@mail.dcu.ie

Abstract— Memorability is defined as the state of being easy to remember or worth remembering [1]. With the advent in technology, video content generated every unit of time is increasing exponentially. Some of these videos have more impact than the others resulting in higher memorability scores for such videos. Predicting such memorability sure does have some applications. For example, we could use the contents of the videos with high memorability scores for target marketing, creating greater impacts on people leading to greater sales. In this study, I have designed a model to predict both short-term and long-term memorability scores of videos. The model uses both video as well as semantic features to predict memorability scores.

Keywords—C3D, HMP, Semantic, TfIdfVectorizer, CountVectorizer, Captions with Weights

I. Introduction

In this study, I investigated the usage of various video and semantic features for predicting the memorability scores of video clips. During this research, I analyzed different visual features such as C3D, HMP and Color Histogram to design a model for prediction. Initially, I trained the models with all these available features individually. I also trained my model with the semantic feature, captions, that described the video in a sentence. Later, I took a combination of the best performing video feature, C3D, and captions to train my model. The models were evaluated using Spearmann's correlation score as a standard measure.

My key analysis and findings are as follows:

- Certain features in the ground truth dataset like the number of annotations did not make any contributions to prediction.
- For any given model, short-term memorability scores were predicted more accurately than the long-term memorability scores.
- Amongst video features C3D outperformed all other video features.
- Captions despite being the only semantic feature available outperformed all other provided features.
- Using simple non-linear regression models with controlled parameters performed well consistently with semantic features.

The rest of the paper is organized into following sections: Section II is a literature review of the previous related work, Section III is brief explanation about the approach I took - with details on the Machine Learning (ML) models, data preprocessing and feature extraction, Section IV shows the results, Section V discusses the future work and conclusion and Section VI provides the references.

II. Literature Review

In [2], three simple, linear and regularized models were chosen – L1 Regularized Logistic Regression, Linear Support Vector Regression, ElasticNet. These selected models were run for each of the provided features. Video features like HMP and C3D were used directly, whereas frame-level features like ColorHistogram and LBP were concatenated using frames. The semantic features were processed under CountVectorizer by removing the stopwords, using unigrams and bigrams. 1st, 56th and 112th frames of the video are averaged and normalized from the penultimate layer of ResNet50 and DenseNet121. This results in feature vector for DenseNet and ResNet respectively. But in order to improve their accuracy they built an ensemble of models using some of their best models. They used a simple weighted average technique to blend their previously obtained outputs.

III. Approach

The following section describes how I approached to the solution.

A. Models

I chose simple linear regression models for the prediction. I ran the provided features over three simple models:

1) *Linear Regression Model*
2) *Decision Tree Regression Model*
3) *Random Forest Regression Model*

B. Features and Data Pre-Processing

Video features like **HMP** and **C3D** were individually used to predict memorability. The HMP features alone were simply read into frames and sent as independent variable with short-term and long-term memorability scores as dependent variables joined on video names. The same was followed for C3D features as well. Both yielded poor results. But of the two, C3D gave better results.

Semantic feature – **Captions** gave better results compared to video features. Hence, extensive work was done on semantic feature. The captions were cleaned by removing special characters, converting captions into small case and removing stop words. The cleaned words were used to create a bag of words. This bag of words was run with **TfIdfVectorizer** to obtain features. These features were sent as independent variables to my Machine Learning (ML) model. TfIdf is a statistical measure used to evaluate how important a word is to a document in a collection or corpus [4]. So, the TfIdfVectorizer calculates the TfIDf value of each word in the given corpus and forms a feature.

Similarly, the bag of words was also run with **CountVectorizer**. But, TfIdfVectorizer outperformed CountVectorizer. CountVectorizer calculates the frequency of occurrence of a word in the given corpus.

Video and Semantic features – here, **C3D and captions** - were combined and sent as independent variables to ML model. Captions were run through with TfIdfVectorizer before sending to the ML model. C3D feature was taken as is. Even though this performed better than C3D feature alone, it failed to perform better than captions alone.

In [2], it has been said that few terms have more positive impact on memorability than others. These terms are also given in their paper with the coefficiency of their effect. Contrary to the popular belief, terms pertaining to nature had a negative effect i.e., less memorability score and terms pertaining to people or indoor actions had a positive effect i.e., more memorability score. Using this concept, I gave certain extra weights to the terms with **positive coefficiency** (terms were given in the paper [2]). These terms were searched in captions, if found the weight for the caption was cumulatively increased. This model of mine performed the **best** with **Random Forest Regression Model** and n_estimators=100.

In my exploration I learnt that the model with weighted captions worked best. Hence, I used the same model for my final computation. Therefore, my ML model is on semantic feature (captions), with TfIdfVectorized features, along with weighted captions. The final results are stored in **testSetResult.csv**.

IV. Result

The results are tabulated as below (Table 1). The scores are calculated using Spearmann's correlation.

Feature	**Model**		**Short Term Memorability Score**	**Long Term Memorability Score**
Video feature: HMP features	a) Linear Regression Model		0.066	0.045
	b) Decision Tree Model		0.093	0.005
	c) Random Forest Regression Model	n_estimators =10	0.168	0.067
		n_estimators =100	0.285	0.095
Video feature: C3D features	a) Linear Regression Model		0.266	0.090
	b) Decision Tree Model		0.081	-0.007
	c) Random Forest Regression Model	n_estimators =10	0.161	0.080
		n_estimators =100	0.278	0.104
Semantic feature: Captions Using **TfIdfVectorizer**	a) Linear Regression Model		0.191	0.060
	b) Decision Tree Model		0.227	0.104
	c) Random Forest Regression Model	n_estimators =10	0.351	0.158
		n_estimators =100	0.396	0.182
Semantic feature: Captions Using **CountVectorizer**	a) Linear Regression Model		0.080	-0.010
	b) Decision Tree Model		0.257	0.049
	c) Random Forest Regression Model	n_estimators =10	0.362	0.102
		n_estimators =100	0.394	0.114
Semantic feature: Captions with weights Using weighted scores for **positive** words	a) Linear Regression Model		0.136	-0.006
	b) Decision Tree Model		0.280	0.097
	c) Random Forest Regression Model	n_estimators =10	0.370	0.152
		n_estimators =100	0.415	0.178
Video & Semantic feature: C3D & Captions	Random Forest Regression Model	n_estimators =100	0.369	0.158

Table 1. Results

V. Conclusion And Future Work

The exploration showed that captions provided better results than any other provided video features. More in depth exploration on captions can give even better results.

A simple weight for 16 terms provided better results in prediction. Hence, I think there's much scope for research in this field. More work can be done on finding impact coefficiency for each term in corpus and give weights accordingly.

VI. References

[1] "Dictionary by Merriam-Webster: America's most-trusted online dictionary," [Online]. Available: https://www.merriam-webster.com/.

[2] K. M. Rohit Gupta, "Linear Models for Video Memorability Prediction Using Visual and Semantic Features," MediaEval-2018.

[3] "Dictionary by Merriam-Webster: America's most-trusted online dictionary," https://www.merriam-webster.com/.

[4] "Tf-idf :: A Single-Page Tutorial - Information Retrieval and Text Mining," [Online]. Available: http://www.tfidf.com/.

[5] "Machine Learning A-Z: Download Practice Datasets - SuperDataScience Pages - Big Data | Analytics Careers | Mentors | Success," [Online]. Available: https://www.superdatascience.com/pages/machine-learning/. [Accessed 2019].