Video Memorability Prediction Using Machine Learning Models

Anshika Sharma
19210993
MCM Computing (Cloud Computing)
Dublin City University
Dublin, Ireland
anshika.sharma27@mail.dcu.ie

Abstract— Memorability is defined as the state of being easy to remember or worth remembering [1]. With the advent of technology, the rate at which videos are created is increasing exponentially. Through access to social media, there are billions of videos available on these platforms that catch one's attention on a daily basis. Each video has its own impact on an individual. But the question is if people remember such videos and if they do, is there any connection between such videos. The MediaEval proposed a task to predict the memorability scores of given videos. [2] Some visual and semantic features were provided by the organizers for these videos. In this project, I have used the provided features to predict the short-term and long-term memorability scores.

Keywords— video memorability, captions, Random Forest.

1. INTRODUCTION

In this study, I chose captions to predict the memorability as it is implied from the previous work that when machine learning models were applied using captions as features, better results were generated. Instead of messing up the model with other features like CD3, Color Histogram, HMP, HOG, ORB, etc, I decided to stick with captions only. I explored two algorithms, Random Forest and Support Vector Machine. Eventually, I took random Forest to evaluate the short-term and long-term scores.

Spearman's correlation coefficient was selected as a measure of evaluation. The features were processed through two vectorizers; Count Vectorizer and TfIdf Vectorizer, before sending them to the models.

2. LITERATURE REVIEW

R. Gupta et al. (2018) [3], the winner of Mediaeval 2018 competition demonstrated a model that considered visual and semantic features. predicted the memorability scores using semantic and Visual features combined. According to them C3D and HMP exceeded image features like Color Histogram, InceptionV3-Preds, and LBP. In their model, they used captions as the feature for prediction. An ensemble of the Caption Predictor and Resnet Predictor was their final model. The major takeaway from their work was that they classified the captions with positive and negative coefficients. They concluded that nature-related words had a negative impact as when compared to human-related factors, which had a positive impact. Henceforth, I've learned the effect of captions on the prediction.

3. APPROACH

This section describes how I approached to the prediction.

3.1 Features and Data Pre-Processing

I've only used one feature in this investigation. Semantic feature – Captions gave better results compared to video features. According to previous work, semantic feature caption gave better results. Therefore, extensive work is carried out only captions. The cleaning of the caption was performed by eliminating the special characters and stop words. A bag of words was created through cleaned captions This bag of words was run with Count Vectorizer and TfIdf Vectorizer, separately, to obtain features. These features were sent as independent variables to my Machine Learning (ML) model. TfIdf is a statistical measure used to evaluate the importance of a word in a document in a corpus [4]. So, the TfIdf Vectorizer calculates the TfIDf value that is; Term Frequency Inverse Document Frequency of each word in the given corpus and forms a feature. Similarly, the bag of words was also run with Count Vectorizer. Count Vectorizer

calculates the frequency of occurrence of a word in the given corpus. However, TfIdf Vectorizer outperformed Count Vectorizer.

Mathematically, TF-IDF is calculated as follows:

$$\text{Tf-idf Score} = tf * \log\_e(N/df) \qquad (1)$$

tf = term frequency of a word in the sentence
df = number of documents containing that particular word
N = total number of documents/lines

The captions are initially sent to the Vectorizer and then to the model.

3.2 Models

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

1) Random Forest Regressor Model

2) Support Vector Machine Regressor Model

Both the models were run on Count Vectorizer as well as TfIdf Vectorizer. The TfIdf generated better results in both the Random Forest and Support Vector Machine.

After running 4 models, Random Forest Regression Model performed the best amongst all the others. The final results are stored in **Anshika_Sharma_19210993_Results .csv**

4. RESULT AND ANALYSIS

The results were calculated using Spearman's rank correlation coefficient which is a non-parametric measure for rank correlation. The best results were obtained with captions used with weights and TFIDF vectorizer, also using Random Forest regressor results were consistent with random validation datasets. The results are tabulated as below (Table 1).

Model with Captions as a Feature	**Count Vectorizer**		**TFIDF Vectorizer**	
	Short -term	Long-term	Short -term	Long-term
Random Forest	0.411	0.172	0.433	0.157
Support Vector Machine	0.340	0.164	0.364	0.169

Table (1): Spearman's correlation.

5. CONCLUSION AND FUTURE WORK

The exploration showed better results for the captions than any other semantic or visual features. A more in-depth analysis of various other combinations of features might develop better results. Therefore, I think there's much scope for research in this field. More work can be done on finding impact coefficiency for each term in the corpus and give weights accordingly.

Refernces

[1] "Merriam Webster," [Online]. Available: https://www.merriamwebster.com/dictionary/ memorability. [Accessed 2019].

[2] "Multimediaeval," [Online]. Available: http://www.multimediaeval.org/mediaeval2018/memorability/. [Accessed 2019].

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

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