Predicting Video Memorability Scores Using Machine Learning Models

Swathikiran Srungavarapu
18210866
School of Computing
Dublin City University
Dublin 9, Ireland
swathi.srungavarapu2@mail.dcu.ie

Abstract— **"Memorability is defined as the state of being easy to remember or worth remembering" [1]. With access to social media, billions of videos are available on such platforms that grabs one's attention on daily basis, but the question is, whether these videos are remembered by people and if so, there is there any pattern between such memorized videos. The MediaEval 2018 proposed a task of predicting the memorability scores of videos. [2] The organizers provided various visual and semantic features 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, NLP text-processing, C3D features

I. INTRODUCTION

Predicting memorability task is mainly focused on identifying the probability of a video being remembered. In this task, the ground truth which was collected by the organizers was provided. This ground truth had short-term and long-term memorability scores along with their respective number of annotations. Video features like C3D, HMP were provided, similarly, Image features such as ColorHistogram, ORB, InceptionV3, LBP and HOG that were extracted at the beginning, middle and end of the video as a frame were provided. Along with these, semantic feature that described a video in a short sentence (caption) was also provided. In my work, I explored on video and semantic features extensively to build a stable predictive model.

I have identified that captions contribute more effectively than any other given feature. Therefore, I have used captions as a mandatory feature in building the model. When Video features were explored, I have identified that C3D features were giving better results than HMP features.

This paper is structured as follows. In Section II, I will be reviewing the existing work on MediaEval 2018 competition and Natural Language Processing methodologies. In Section III, I will be describing approaches used for extracting the various features. In Section IV Machine Learning algorithms that were used will be discussed, finally, I will be summarizing my analysis with results in Section V and future steps in Section VI.

II. LITERATURE REVIEW

The winners of the MediaEval 2018 competition R. Gupta et al. (2018) [3] demonstrated a model that predicted the memorability scores using semantic and Visual features combined. They have identified that video features C3D and HMP have outperformed the image features such as ColorHistogram, InceptionV3-Preds and LBP. Also, they have used the semantic features captions in building their best model. Their final predictor was an ensemble of Caption Predictor and Resnet Predictor. Major take away from their research is that they have identified the words in captions with positive and negative coefficients. They have concluded that words related to nature had negative coefficients and words related to humans had positive coefficients, Therefore, I explore more on captions and the impact of certain words in the captions.

III. FEATURE EXTRACTION

A. Semantic Features

Feature extraction of captions was done using NLTK libraries and defined methods. Initially the captions were extracted from the file and then were processed to remove all the regex leaving out only meaningful English words. Along with this, the stop words were removed and stemming of words was done for normalization. Later the words were vectorized using their occurrence counts, TF-IDF scores. Typically, TF-IDF stands for term frequency-inverse document frequency that reflects how important a word is to a document in a collection or corpus [4]. Mathematically, TF-IDF is calculated as follows:

Tf-idf Score = tf * log_e(N/df)
tf = term frequency of a word in the sentence
df = number of documents containing that particular word
N = total number of documents/lines

Also, I have provided extra weights for certain words that had positive coefficients according to R Gupta at el (2018). Detailed analysis on these weights and their impact is discussed in section VI.

B. Video Features

Considering that videos in reality create an impact during their course but not due to a particular picture in frame. Therefore, considering any of the image features that were extracted from start and end of the video did not seem to be appropriate features for predicting the memorability. So, I used only C3D and HMP features that are directly extracted features from the video. C3D feature file consists of 101 float values where as HMP had 6074 float values describe a certain video.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

IV. Models

I preferred simple Linear and Non-Linear regression models for this task, as major features are of high dimensionality, where multicollinearity exists and finding the variables with high P-values is hard task. Hence, models can be overfitting. Therefore, using simple models was adequate and I was able to verify the overfitting problem with techniques like K-Fold cross validation and tuning parameters of the models. Also, I have built an Artificial Neural Networks with standard parameters [5].

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- Keras Sequential Artificial Neural Network

For each model, I have used the above models on the captions, C3D and HMP features independently and also combining few features together.

V. Results

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.

Short- Term Memorability Spearman's Scores				
Features	**Linear**	**Decision Tree**	**Random Forest**	**ANN**
Captions	0.093	0.237	**0.404**	0.320
C3D	0.272	0.142	0.260	0.227
HMP	0.019	0.066	0.142	0.083
Captions + C3D	0.094	0.130	0.354	0.355

Fig. 1. Results for Short-term predictions.

Long- Term Memorability Spearman's Scores				
Features	**Linear**	**Decision Tree**	**Random Forest**	**ANN**
Captions	0.054	0.132	**0.176**	0.176
C3D	0.076	0.066	0.094	0.130
HMP	0.018	0.061	0.067	0.077
Captions + C3D	0.028	0.064	0.166	**0.194**

Fig. 2. Results for long-term predictions.

Final Model was built with Captions along with word weights using Random Forest Regressor varying the size of trees in the Random Forest. The accuracy scores of with varying Trees are plotted as below:

VI. Ananlysis and Discussion

I have explored various visual, semantic features and methodologies for extracting these features in this project. The major emphasis was on extracting as many features I could using captions. During this study, I explored that using TF-IDF Vectorizer gave better results compared to count vectorizer. This is due to the importance given by TFIDF vectorizer to words that are rarely appeared, Therefore, words that appeared regularly in all captions will be weighted less and only rare words are given higher weights. This ensures that the memorability scores are dependent on one word rather than whole sentence that is generalized. Therefore, when such rare words are seen in test sets the model recognizes their importance and predicts better scores.

Along with this, I gave weights to certain words to improve the model. These words were chosen from R Gupta et al. (2018) [3] paper, authors described words with higher positive coefficients and certain words with negative coefficients that influenced their model. I gave higher weights to the words with positive coefficients in order of their positivity and similarly lesser weights to the words in order of their negativity. This boosted the accuracy of the model.

Once, TFIDF with weights features were providing better results on Random Forest Regressor consistently, I have varied the range of trees from 10 to 1000 until I found the threshold limit. The same can be seen in the Fig. 3.

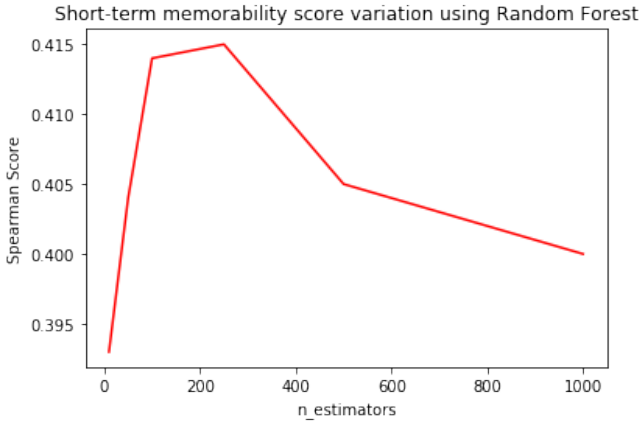

Fig. 3. Graph showing the variation in accuracy with Random Forest Model and around 250 trees as the threshold.

Using C3D features, there was not much to explore in creating new features, I have tried to remove the first and last 15 values in the dataset assuming those values belong to the start and end of the video, but that resulted with low accuracy results.

When C3D features and Captions were combined, I have processed these features into a Keras sequential Artificial Neural network with 2 dense layers and 10 epochs, the model performed well on training data, but not with validation data. This shows that the model was overfitting. Therefore, I discarded that model for my prediction though it predicted better results for long- term memorability.

VII. Conclusion *and Future work*

I conclude this paper, with an opinion that captions provide better results when used as features to predict the memorability scores. Also, understanding the importance of rare words and their associations is useful to extract new features such as weights. These new features will contribute in building better prediction models.

However, Combining ResNet50 and captions fetched better results to other researchers, Resnet is a pre-trained model which was built to perform well on image features. As future work, I would like to combine captions and image features together to build a predictor using Resnet50. Also, focusing on increasing the scores for long-term memorability is another aspect that can be anlysed.

References

[1] "Merriam Webster," [Online]. Available: https://www.merriam-webster.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] "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Tf%E2%80%93idf. [Accessed 2019].

[5] "SuperDataScience," [Online]. Available: https://www.superdatascience.com/pages/machine-learning/. [Accessed 2019].