Media Memorability prediction using Linear Models and Neural Network

Paritosh Gupta
Dublin City University, School of Computing, Dublin, Ireland
paritosh.gupta3@mail.dcu.ie

ABSTRACT:

The measure of ease to remember is defined as memorability. In today's world with the abundance of video content on the internet and social media, prediction of media memorability has the many possible application. In this paper, with the given visual, semantic and video caption features from MediaEval 2018 organizers for the video, used in models like Linear Regression, Neural Networks and Support Vector Regression to predict its memorability.

1. INTRODUCTION:

The primary task was to predict the Media Memorability score which will represent how memorable the video will be [1]. Video memorability depends on various features such as semantics, color features, saliency, etc. [2]. In this paper, the investigation was to the various given semantic and visual features to predict the media memorability for the video and to perform the depth analysis on the given features to develop the robust predictors for the media memorability. We have been provided with various visual features such as HMP, LBP and Color Histogram visual features and InceptionV3-Predictions & C3D-Predictions semantic features and video captions, where semantic or visual feature didn't contribute much towards score as that of video caption contribution [3]. I have worked with C3D-preds feature and captions for the videos to build the model. The models were evaluated using Spearman's rank correlation as the metric. Below are the key findings from this work-

a) Short-term memorability prediction score was higher as compared to the long-term memorability prediction.
b) Models based on C3D feature is outperformed by the model based on captions.
c) Model-based on both features combined, C3D & caption also outperformed by the model based on captions.

2. RELATED WORK:

Many researchers have taken a lot of interest in this field and looking at it as the potential application of Machine Learning in this area. In the recent work [1] [3] [4], deep learning based semantic feature representation (C3D-Preds), various levels of visual features and video captions were used for the memorability prediction. As per the findings, the captions are shown the best individual result as compared to the other features. Also, the researchers noticed that the CNN model, which is one of the state of the art technique, trained with high-level semantics features for image classification has shown best performance on numerous computer vision task [5].

3. APPROACH:

3.1 Model:

As most of the feature were the high dimension, overfitting and high variance were the potential concern for the model selection. I tried two simple linear models and one neural network model.

a) Linear Regression
b) Dense Neural Network
c) Support Vector Regression

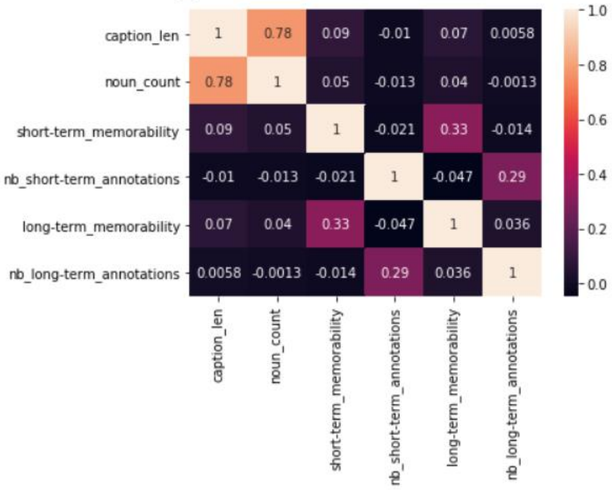
The caption was the highly prominent feature from the related work, so I tried to perform some Natural Language Processing operations. In addition to this, C3D-Pred feature was also extracted and applied to all these three models for short-term and long-term memorability prediction.

3.2 Features and Data Processing:

Visual feature such as C3D-Pred which is a generic feature for video and is compact, discriminative and efficient to compute. It is generated by training the 3D convolutional network on a video set with annotation. As this feature was already provided, I tried to load this extracted feature from each video file and further merged with the ground truth file.

Text feature like captions for the video was provided, which was pre-processed by removing stopwords and punctuation and converting them into lower case. Further, these cleaned captions were converted into Bag-of-Words using CountVectorizer and TfidfVectorizer from Scikit-Learn library. TfidfVectorizer showed better performance and chosen for the vectorization. Additionally, the length of the captions and count of noun words were also extracted as additional features.

Below is the correlation matrix for caption length, noun count, long-term and short-term annotations.



All these extracted features, visual and text were applied over all the three models. Regularization and finetuning were done by various parameters (such as C, alpha, epsilon, tolerance, regularizer, dropout) in the models.

4. RESULTS:

Below table shows the summary for the experimental results. Support Vector Regression model with caption feature performed well then other models in short-term prediction. Whereas for long-term prediction, Support Vector Regression with captions + C3D features performed well than others.

Table 1: Short-Term Memorability Prediction

Model	Features		Spearman
LR	Captions	CountVector+	0.291
		TfidfVector+	0.317
	Captions + C3D	TfidfVector+	0.306
DNN	Captions + C3D	TfidfVector+	0.085
		One hot encod	0.277
SVR	Captions	TfidfVector+	**0.407**
	C3D		0.235
	Captions + C3D	TfidfVector+	0.405

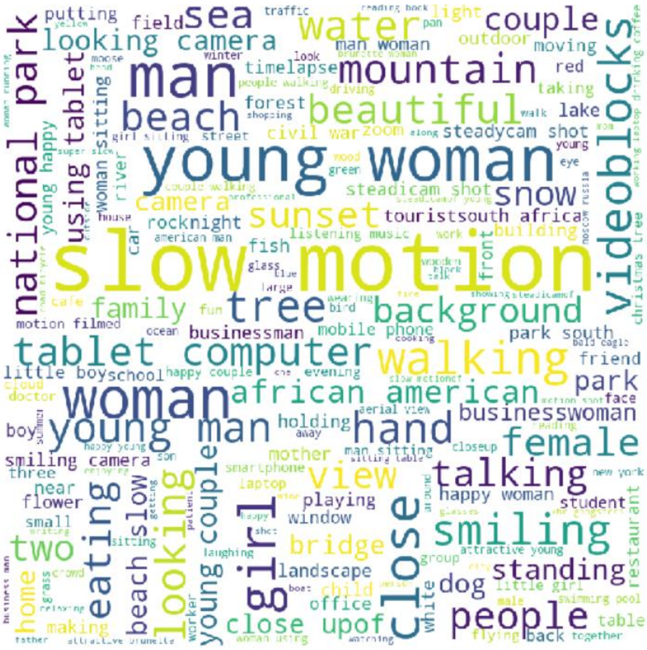
'+' : Caption length, Noun Count

Table 2: Long-Term Memorability Prediction

Model	Features		Spearman
LR	Captions	CountVector+	0.091
		TfidfVector+	0.087
	Captions + C3D	TfidfVector+	0.087
DNN	Captions + C3D	TfidfVector+	0.079
		One hot encod	0.115
SVR	Captions	TfidfVector+	**0.133**
	C3D		0.041
	Captions + C3D	TfidfVector+	**0.144**

'+' : Caption length, Noun Count

In my exploratory data analysis to interpret the caption feature and to view the most frequent words in it and I tried to plot the below word cloud which is representing bigger size for the word with the highest frequency.



5. CONCLUSION & FUTURE WORK:

This paper represents the simple regression models and Neural Networks on various visual, semantic and text feature to predict the media memorability. The text feature is shown good performance in predicting the video memorability but not that good. Due to high dimensionality the fetched features, models seem to perform not that well. If we would have more training dataset, then it would have been much better to apply some of the advanced state of the art techniques like CNN or LSTM which usually has higher chances of getting good accuracy.

REFERENCES

[1] C.-H. S.-T. RomainCohendet, "MediaEval2018: PredictingMediaMemorabilityTask.InProc. of the MediaEval 2018 Workshop," 2018.

[2] L.-V.-T. Duy-TueTran-Van, "PredictingMediaMemorability UsingDeepFeaturesandRecurrentNetwork," HoChiMinhCity, 2018.

[3] K. M. Rohit Gupta, "Linear Models for Video Memorability Prediction Using Visualand Semantic Features," Conduent Labs, India, 2018.

[4] D. S. a. H. S. Sumit Shekhar, "Show and Recall: Learning What Makes Videos Memorable," in *Computer Vision and Pattern Recognition. 2730–2739*, 2017.

[5] H. A. J. S. a. S. Ali Sharif Razavian, "CNN Features Off-the-Shelf: An Astounding Baselinefor Recognition. InThe IEEE

Conference on Computer Vision and PatternRecognition (CVPR) Workshops," 2014.

[6] S. S. S. B. N. P. Tanmayee Joshi, "Multimodal Approach to Predicting Media Memorability," TCS Research, Pune, India, 2018.