Predicting Media Memorability Scores Using Semantic Features

Joseph Cherian
Dublin City University
joseph.cherian2@mail.dcu.ie

***Abstract*—**In this paper, the Predicting Media Memorability task is presented, which is running for the second year at the MediaEval 2019 Benchmarking Initiative for Multimedia Evaluation [3]. Memorability can be defined as the state or quality of being easy to remember or worth remembering [4]. Prediction of memorability has many potential applications as a large amount of video content is streamed online by users every day. The short-term memorability and long-term memorability scores are predicted using the captions and ground truth dataset features.

Keywords—captions, video memorability, bag of words, Count Vectorizer, random forest

1 INTRODUCTION

Prediction of the media memorability task mainly deals with the identification of the probability of video remembrance. Automatically predicting memorability scores for videos is the task requirement for the participants [1]. The ground thruth dataset was provided with the short term and long term (after 24-72 hours) memorability scores along with their respective number of annotations which refer to the probability of being remembered after two different durations of memory retention. Video features like C3D features and HMP (Histogram of Motion Patterns) along with the image features like HoG (Histogram of Oriented Gradients),LBP(Local Binary Pattern),InceptionV3,ORB (Oriented FAST and Rotated BRIEF) and Color Histogram extracted on three key-frames (first (0), middle (56) and last (112)) on each video were provided. Description of the video in a short sentence(caption) which is a semantic feature was also provided.

From the paper [4] I have identified that the caption feature is more effective in calculating the spearman scores for the short-term and long-term memorability compared to the other features provided.

2 RELATED WORK

Many features and their combinations using various models are used to predict the media memorability. The papers [4][6] experimentally shows that using captions with various models like TF-IDF(Vectorization) give a better media memorability score when compared with C3D, HMP and other pre-extracted visual features. In the paper [4] captions are used with a bag-of-word approach using TF-IDF and with Embeddings and Neural Networks. The bag of words approach has been used with captions using Count Vectorizer instead of TF-IDF in this paper. Captions using Random Forest and Support Vector Regression models have also been implemented.

3 APPROACH

8,000 short soundless videos which are shared under a license that allows their use and redistribution in the context of MediaEval 2018 are present in the dataset. These 8,000 videos were split into 6,000 videos for the development set and 2,000 videos for the test set [1]. The captions feature is used along with three models to predict the short-term and long-term memorability scores.

3.1 *Models*

High variance and over-fitting are a potential concern in this task as most of the features provided are very high dimensional and the number of videos is of the same order of magnitude as the dimensionality of the features. Traditional Machine Learning and highly regularised linear models are developed namely [4]:
(1) Random Forest with Captions
(2) Support Vector Regression with Captions
(3) Vectorizing Data using the Bag-Of-Words approach

These models are used on the caption feature on the Dev-set to determine the best performing model for predicting the short-term and long-term memorability scores. The model is then applied on 2000 videos for

the caption feature in the test set to get the predicted short-term and long-term memorability values.

3.2 *Data Pre-Processing*

The captions and ground truth values are merged into a single data frame. Bag of Words feature of Skikit-Learn is used for pre-processing captions. Sentence connecting words like ‘and’, ‘the’, ’a’ etc are called Stopwords. Natural Language Processing techniques are used to pre-process captions to remove Stopwords. The punctuation used in the captions is discarded and replaced with space by importing the **string** module. The words in the caption text are converted into lower case. Tokenization is used to split the entire caption text for the videos into individual words which are used as the input for parsing. The text to vector conversions are done using Count Vectorizer.

4 RESULTS AND ANALYSIS

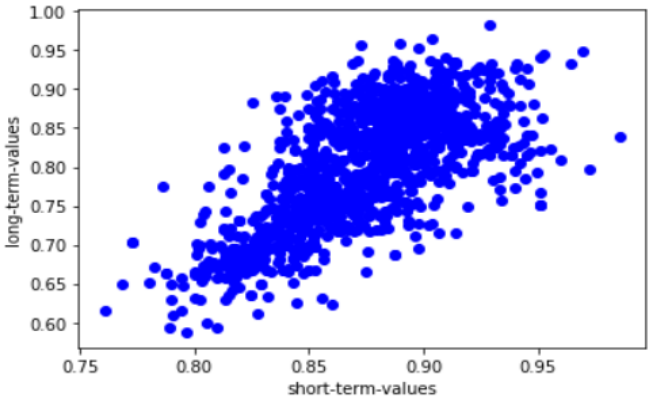
Calculation of the results are done using a non-parametric measure for rank correlation which is Spearman’s rank correlation coefficient. The best results for Short-Term Memorability and Long- Term Memorability are obtained by using random forest regressor on the captions feature as shown below.

Short-Term Memorability Spearman’s Scores

Features	Random Forest	SVR	Vectorizing Data: Bag-Of-Words
Captions	0.404	0.338	0.401

Long-Term Memorability Spearman’s Scores

Features	Random Forest	SVR	Vectorizing Data: Bag-Of-Words
Captions	0.186	0.170	0.173



The Short-term memorability are plotted against the long-term memorability values after testing the model using the Vectorizing Data: Bag-Of-Words approach.

5 DISCUSSION AND OUTLOOK

My findings and contributions in this area are the following:

1.Increasing the random state for train_test_split function in Skikit-Learn library increases the spearman scores for all the three models used.

2.Increasing the random state for Random Forest Regressor in Skikit-Learn library decreases the spearman scores for both short-term and long-term memorability.

3. Decreasing the maximum features(max_features) parameter in the CountVectorizer function in Skikit-Learn library increases the short-term and long-term memorability spearman score for SVR with captions. Example: The spearman scores for the SVR short term and long term memorability change from 0.338 and 0.170 to 0.416 and 0.196 when the maximum features in the CountVectorizer function is changed from 3112 to 3000.

6 CONCLUSIONS AND FUTURE WORK

The short-term and long-term memorability scores are not entirely dependent on the model used on the caption feature as varying the random state for the various functions like Random Forest Regressor and train_test_split function can alter the scores. Sound as a feature makes the prediction of media memorability more interesting as processing of signals comes into the picture. In the future, I will try using a combination of features using an ensemble model to measure the media memorability.

REFERENCES

[1] C.-H. D. N. Q. K. D. M. S. B. I. Romain Cohendet, "MediaEval 2018: Predicting Media Memorability," 2018.

[2] H. Squalli-Houssaini, N. Q. K. Duong, M. Gwenaelle and C.-H. Demarty, "Deep Learning for Predicting Image Memorability," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[3] M. I. B. D. C. D. N. A.-P. X. a. S. M. Constantin, "Predicting Media Memorability Task at MediaEval 2019," in *Proc. of MediaEval 2019 Workshop*, Sophia Antipolis, France, 2019.

[4] E. M. F. H. T. E. W. A. F. S. David Azcona, "Predicting Media Memorability Using Ensemble Models," in *CEUR Workshop Proceedings*, 2019.

[5] S. S. S. B. N. P. Tanmayee Joshi, "Multimodal Approach to Predicting Media Memorability," in *MediaEval*, 2018.

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

[7] L.-V. T. M.-T. T. Duy-Tue Tran-Van, "Predicting Media Memorability Using Deep Features and Recurrent Network.," in *MediaEval*, 2018.

[8] D. S. H. S. M. K. A. S. Sumit Shekhar, "Show and Recall: Learning What Makes Videos Memorable," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.