Predicting Video Memoribility using Machine Learning

Jai Warde
MSc. in Computing
Dublin City Univerity
Dublin, Ireland
Student ID: 20210699
jai.warde3@mail.dcu.ie

***Abstract*— This paper outlines the approach taken for MediaEval 2018 predicting media memorability task. The approach is based on merging only the video features like C3D, HMP & LBP. The merge sequence is then provided as input to Random Forest and SVR models to predict the short-term and long-term memorability scores respectively.**

Keywords—Video memorability, short-term, long-term, C3D, HMP, LBP, Random Forest, XGBoost, SVR

I. INTRODUCTION

The role of predicting memorability is primarily concerned with determining the likelihood of a video being remembered. The ground truth gathered by the organizers was presented in this mission. Short-term and long-term memorability ratings, as well as the number of annotations, were included in this ground reality. Image features such as Color Histogram, ORB, InceptionV3, LBP, and HOG that were extracted at the beginning, middle, and end of the video as a frame were provided, as were video features such as C3D, HMP. Along with these, a semantic feature that described a video in a short sentence (caption) was also provided. In my work, I explored video features C3D, HMP, and LBP extensively to build a stable predictive model.

I have used the above-mentioned video features and identified that merging C3D and HMP perform best for the long-term spearman score. On the other hand, merging C3D, HMP, and LBP performs best for the short-term spearman score. The approach outlined that Random forest outperforms for short-term and SVR outperforms for long-term scores.

II. RELATED WORK

R. Gupta et al. (2018) **[1]**, the winners of the MediaEval 2018 competition, demonstrated a model that combined semantic and visual features to predict memorability ratings. Video features such as C3D and HMP outperformed image features such as Color Histogram, InceptionV3-Preds, and LBP, according to the researchers. In addition, they built their best model using the semantic features captions. An ensemble of Caption Predictor and Resnet Predictor served as their final predictor. Therefore, I explore more on pre-computed features C3D, HMP and LBP as restricted to use semantic feature captions.

'GIBIS at MediaEval 2018: Predicting Media Memorability Task' **[3]** is the key one I ended up exploring after reviewing a number of previous year's submission articles. C3D is one of the features used in this study to predict video memorability. This paper helped me come up with ideas about how to use C3D features effectively and what model variables to consider.

III. APPROACH

A. Motivation of Approach

The approach to ensemble the feature set was a result of previous work carried out in various research papers of MediaEval to get a high spearman coefficient score. The key feature used in the research papers is the captions semantic feature to explain the video semantics. As part of this work, I was restricted to use the semantic feature i.e., caption. Another outperforming pre-computed feature in C3D as it was rated best feature by previous research papers.

I conclude using the C3D feature along with other pre-computed video features to yield a high spearman score for this challenge. I even tried Eoin Brophy's solution **[2]** of the Keras Sequential model. Without the use of captions semantic feature, the above approach didn't perform well and leads to a very low spearman score for short-term and long-term.

Using the different feature set combination tested, the short-term spearman score was the highest in Random forest for C3D, HMP, and LBP features. The long-term spearman score was the highest in Support Vector Regressor for C3D and HMP features.

B. Explaining Approach

The final approach I have taken uses the combination of pre-computed feature set for both the short-term and long-term memorability scores. I implemented random forest over C3D, HMP, and LBP which outperforms other techniques in the short-term score. And, I implemented SVR over C3D and HMP which outperforms other techniques in the long-term score. The Collab notebook is used as an IDE for python code.

I loaded the pre-computed features i.e., C3D, HMP, and LBP into a python data frame. These features are then pre-processed by extracting feature values in

separate columns rather than a list of NumPy arrays. The C3D feature for each video has **101** values, the HMP feature for each video has **6075** values, the LBP feature has 3 frame files (0, 56, 112) for each video, and each file has **122** values i.e., resulting in 366 value for each video. The HMP feature has a very high dimension of 6075, so I used **PCA (Principal Component Analysis)** to reduce the dimension by preserving **99.99%** of the data variance. The feature set is then merged using NumPy concatenate and different machine learning techniques are applied.

The various techniques including Eoin's solution are applied and compared for the short-term and long-term video memorability spearman score. The Random Forest with n_estimators = 100 performs best in the short-term score. I tried different hyper-parameter tuning to come up with the best value of estimators. The SVR is performing the best for long-term memorability score.

The model is first trained on **4800** videos and validated against the **1200** videos. Once, this training is done I train the best performing models mentioned above over the entire **6000** videos and predicted the short-term and long-term memorability of **2000** test videos. For further details on the approach, please refer to the collab notebook.

IV. Results and Analysis

The evaluation metric used in this paper is the **Spearman correlations score**. The short-term and long-term memorability spearman score is computed for various models. The look-up table with scores of various feature sets and models tested is represented in below tabular figures. The various models implemented on the feature space are **K Nearest Neighbor** (KNN), **Random Forest**, **Neural Network** (NN), **Support Vector Regressor** (SVR), and **Extreme Gradient Boosting** (XGBoost) algorithms.

The combination of **C3D, HMP,** and **LBP** outperforms others in short-term memorability spearman score as shown in the figure. The **Random Forest** models with **estimators = 100** are performing with a short-term spearman score of **0.348 (34.8%)**. The short-term score is usually higher than the long-term score irrespective of the feature set combination and models.

Short-Term Memorability Spearman Score					
Feature set	Models				
	KNN	RF	NN	SVR	XGBoost
C3D	0.251	0.314	0.266	0.242	0.295
HMP	0.237	0.247	0.09	0.199	0.268
LBP	0.23	0.318	0.212	-0.088	0.253
C3D & HMP	0.2555	0.303	0.265	0.259	0.313
C3D & LBP	0.274	0.333	0.285	0.228	0.324
C3D, HMP & LBP	0.277	**0.348**	0.279	0.242	0.337

The combination of **C3D and HMP** outperforms others in the long-term memorability spearman score as shown in the below figure. The **SVR (Support Vector Regressor)** model is performing with a long-term spearman score of **0.137 (13.7%)**.

Long-Term Memorability Spearman Score					
Feature set	Models				
	KNN	RF	NN	SVR	XGBoost
C3D	0.097	0.117	0.121	0.107	0.071
HMP	0.075	0.065	0.048	0.087	0.073
LBP	0.055	0.057	0.068	-0.04	0.051
C3D & HMP	0.1	0.123	0.118	**0.137**	0.077
C3D & LBP	0.104	0.099	0.112	0.094	0.095
C3D, HMP & LBP	0.101	0.095	0.105	0.099	0.087

The exploration and applying of machine learning/ deep learning models over the different feature sets show that the combination of certain video features provides better spearman scores for both short-term and long-term memorability. The merge input set of C3D, HMP, and LBP perform better as not considering captions in this paper. However, Random Forest and SVR outperform by providing better spearman scores for short-term and long-term memorability.

V. Conclusion

The exploration and applying of machine learning/ deep learning models over the different feature sets show that the combination of certain video features provides better spearman scores for both short-term and long-term memorability. The merge input set of C3D, HMP, and LBP perform better as not considering captions in this paper. However, Random Forest and SVR outperform by providing better spearman scores for short-term and long-term memorability.

In the future, I would incorporate the caption feature to enhance the spearman score of both short-term and long-term memorability. The earlier work is done which showcases best on the leaderboard puts high emphasis on the caption feature.

References

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

[2] Brophy, E. *Google Collaboratory*. [online] Colab.research.google.com. Available at: https://colab.research.google.com/drive/1X7l5MGrDZa2IdMCOxwgILCD5CzHKE8NF?authuser=1, 2018.

[3] Savii et al., R. [online] Ceur-ws.org. Available at: http://ceur-ws.org/Vol 2283/MediaEval_18_paper_40.pdf, 2018.