Media Memorability Predictions using C3D and HMP features

Prasad Arvind Govardhankar
MSc in Computing
Dublin City University
20210305
prasad.govardhankar2@dcu.mail.ie

***Abstract*—This paper discusses an approach focused on C3D and HMP features for the task of Predicting Media Memorability, which is part of the MediaEval 2018 Multimedia Assessment Bench-marking Initiative. The system is programmed to predict video memorability scores automatically. A video containing action is more interesting than a video containing natural scenery or landscape. The C3D and HMP attributes are used on multiple machine learning models to solve the problem.**

***Index Terms*—C3D, HMP, Exrta decision tree, Gradient Booster, ML Regressor, Random forest, Liner Regression, XGB**

I. Introduction

The need for new strategies to better organize and retrieve digital content in order to make it more accessible in our everyday lives is a big impetus for predicting video memorability (VM) [1].The issue is becoming more urgent as media outlets such as social networks, search engines, and recommendation systems struggle with ever-increasing amounts of content data.Our brain evolved to be capable of remembering only the information required for our survival, development, and happiness [2]. This explains why we as humans have a strong propensity to memorize/forget the same images, resulting in a high human consistency in image memorability (IM) and video memorability (VM).The data set, annotation protocol, pre computed features, and ground truth data are all specified in the task overview document. By measuring a memorability score for each video, participants in the Predicting Media Memorability Task will build mechanisms that can predict how memorable a video is.

II. Related Work

Participants in the Predicting Media Memorability Task would create systems capable of predicting how memorable a video is by calculating a memorability score for and video. An comprehensive dataset of videos with memorability annotations will be presented to participants. Since the ground truth was gathered through recognition assessments, it represents objective memory efficiency measures.The dataset includes both ”short-term” and ”long-term” memorability annotations, in comparison to previous work on image memorability prediction, which assessed memorability a few minutes after memorization [3] [4].The deep aesthetic model, which was based on ResNet 101’s architecture, was fine-tuned. For the action recognition networks, they used features derived from the I3D and TSN networks, as well as C3D features provided by the task organizers, to try to increase these features [5].They then performed some late fusion experiments to improve the efficiency of these individual runs even further.

The task organizer trained models on HMP, LBP, and ColorHistogram Visual Features, as well as InceptionV3, C3D semantic features, based on the features provided by the task organizer [6]. Models based on InceptionV3 Preds, LBP, and ColorHistogram don’t suit well and are outperformed by C3D Preds and HMP based models, according to the findings. Models trained on all of the visual features mentioned above outperform models trained on the video captions. A sample of 8 frames from a single video is extracted in the paper [7], The frames are then fed into a pre-trained InceptionV3 convolution network, which extracts 2048-dimension features. When the video frames are withdrawn, the second stage of a recurrent neural network with one LSTM layer sequentially accepts them as inputs.The output of the last dense layer for the last sequence’s input, i.e. the video’s final frame, corresponds to the memorability value for the input of the last series. When compared to models trained solely on pre-computed features, the model performs admirably.

III. Approach

The memorability dataset intotal includes 8,000 videos, in which 2,000 videos are test collection, and 6,000-videos are Development set. The dataset set was split into two parts: training (4,800 videos) and validation (1200 videos). To select hyper-parameters and evaluate the output of all models, I used our held-out validation dataset.

My approach is to develop individual models for C3D and HMP features. Also, by combining them. Firstly, I developed Multiple traditional Machine Learning models for C3D, listed below

- Extra Decision Tree
- Gradient Booster
- MLP Regressor
- Random Forest
- Liner Regression
- XGB

and evaluated the scores. I found out Gradient boost regressor outperformed every other model and then executed the same models for HMP features. After evaluation for it, Gradient boost regressor was the top-performing model. Afterward, I executed those models by combining both the features and evaluated the results, here also Gradient boost regressor outperformed other models, hence I choose the Gradient boost regressor to evaluate results on the test dataset by using the same approach which is combining C3D and HMP Test dataset.

IV. Evaluation

The Spearman's rank correlation between expected memorability scores and ground-truth memorability scores computed across all test videos will be the official assessment metric. Since the task is still a prediction task, the official metric will only be used to rate the different videos. The decision to use the Spearman's rank correlation as the official measure reflects a desire to normalize the performance of the various systems and make comparisons easier.

V. Results

Scores for C3D Devset

Models	Short Term	Long Term
Extra DT	0.205	0.057
ML Regressor	0.287	0.103
Gradient boost	0.319	0.162
Random Forest	0.291	0.128
Linear Regression	0.272	0.112
XGB	0.314	0.092

Scores for HMP Devset

Models	Short Term	Long Term
Extra DT	0.206	0.046
ML Regressor	0.257	0.111
Gradient boost	0.310	0.113
Random Forest	0.315	0.128
Linear Regression	0.009	0.073
XGB	0.292	0.092

Scores combining both Features

Models	Short Term	Long Term
Extra DT	0.160	0.066
ML Regressor	0.199	0.082
Gradient boost	0.299	0.117
Random Forest	0.275	0.126
Linear Regression	0.010	0.002
XGB	0.273	0.101

VI. Conclusion

I introduced and compared several machine learning models in this paper, focusing on C3D and HMP. Not only one of them, but both of them have their own predictive power in each form of prediction, resulting in the best outcome. By integrating both features, tried to achieve more accurate predictions. This reinforces the concept of short-term prediction and long-term memorability in the MediaEval task.As a result, C3D and HMP can be useful in predicting video memorability.

References

[1] R. Cohendet et al., "MediaEval 2018: Predicting Media Memorability Task," arXiv:1807.01052 [cs], Jul. 2018, Accessed: Apr. 29, 2020. [Online]. Available: http://arxiv.org/abs/1807.01052.

[2] M. G. Constantin, B. Ionescu, C.-H. Demarty, N. Q. K. Duong, X. Alameda-Pineda, and M. Sjöberg, "The Predicting Media Memorability Task at MediaEval 2019," p. 4.

[3] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable?. In Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, 145–152.

[4] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In Proc. IEEE Int. Conf. on Computer Vision (ICCV). 2390–2398.

[5] M. G. Constantin, C. Kang, G. Dinu, F. Dufaux, G. Valenzise, and B. Ionescu, "Using Aesthetics and Action Recognition-based Networks for the Prediction of Media Memorability," p. 4.

[6] R. Gupta and K. Motwani, "Linear Models for Video Memorability Prediction Using Visual and Semantic Features," p. 3.

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