Media Evalualtion: Predicting Media Memorability

Shivani Bhat
shivani.bhat4@mail.dcu.ie
Machine Learning Assignment-CA684
Dublin City University
20211408

***Abstract*—Media memorability is defined as a task of predicting a value which indicates how memorable a video is by the user.Memorability can be regarded as a useful metric of video importance to help make a choice between competing videos. It requires participants to automatically predict memorability scores for videos that reflect the probability a video will be remembered.**

***Index Terms*—memorability, Spearman's correlation, correlation, Random Forest, Linear Regression, Neural Network**

I. Introduction

Image memorability has initially been defined as the probability for an image to be recognized a few minutes after a single view, when presented amidst a stream of images. [3]. Knowledge of the memorability of a video has potential in advertisement and content recommendation applications. Although highly subjective, it has been shown that media memorability is measurable and predictable [7]. The objective of this task is to investigate the video memorability scores and predict short and long term score as part of the MediaEval Predicting Media Memorability challenge.

A set of features are provided which describe the image. These features can be broadly classified into Video Features(C3D,HMP) and Image features(InceptionV3,HOG,ORB,LBP). Among these features, this approach works on detecting the best suitable models with selected features that predict a good memorability score. The models employed to predict memorability are Random Forest, Linear Regression, SVR and Neural Networks on the C3D features, HMP Features and InceptionV3 .A lot of previous works has shown good results by using captions as their features. In this study, we train different models to predict the score on features other than caption. The models are evaluated using Spearman Rank Correlation Coefficient which will be discussed later.

The paper is organised into the following sections. Section II gives a brief review of the previous work conducted in this area,Section III outlines the features incorporated , the data set construction and feature extraction to be fed as an input to the models where then in Section IV, the Machine Learning approaches on the image and video features are described with a comparative study and result analysis. The outcomes are then concluded in Section V with some future scope.

II. RELATED WORK

In recent times, a lot of interesting work is conducted on video memorability. A lot of work has been conducted to determine video memorability score and how various low and high level visual features,image features and captions are investigated [6].Multiple low level and high level visual features and some deep learning based action recognition representation (C3DPreds) along with image and video captions are used for memorability predictions [3], [4]. Analysis of the responses of the high-level CNN layers shows which objects and regions are positively, and negatively, correlated with memorability, allowing us to create memorability maps for each image and provide a concrete method to perform image memorability manipulation [5].Annotations also were a good feature to predict memorability [2]. Although, CNNs are widely used [1] for image classification tasks, a lot of work is conducted using RNNs to predict the scores in one of the case.

III. MEMORABILITY DATA SET CONSTRUCTION

A. Task Description

The objective of this task is to predict the memorability score for video samples provided.The ground truth data contains scores for both "short-term" and "long-term" memorability, created via memory performance tests that were conducted. The short term score is measured a few minutes after the memorization process whereas long term scores are the ones that's were measured 24-72 hours after the process.

B. Features and Data pre-processing

The proposed data set consists of 8000 video features, split into 6,000 videos for the development set (dev-set) and 2,000 for the testing set (test set). Participants must train their systems on the dev-set and submit runs containing memorability scores for the test set. There are ground truth values provided by annotators for each video in the dev-set and test set. In this task, as mentioned, we have both video and image features. On modelling the data of image features, the evaluation metric did not perform well on them. On the other hand, better results were shown for the video features. The features selected are read from the C3D and HMP files and stored as a vector data frame ,C3D features contain 101 values for each video while HMP feature has 6075 values for each video.But, a substantial amount of HMP features had values lesser than 6075. To handle this, zeros were appended but this leads to the issue or sparse matrix. Also, the output of the fc7 layer of InceptionV3 was used to model the predictions [8]. The data frame is then merged with the ground truth values.

The features are then trained against the ground truth to predict the memorability scores for the test set.

IV. MEMORABILITY PREDICTIONS: MODELS AND ANALYSIS

The analysis is conducted using C3D features, HMP and InceptionV3 on a few models. Since this looks like a regression concept, the initial model that was used to train the data is **Linear regression**.Subsequently, **Random Forest**, defined as a supervised learning algorithm that uses ensemble learning method for regression which operated by merging a lot of decision trees where the output is the mean of all predictions from all trees.The third model used is **Support Vector Regressor** which uses the concept of support vector machine but outputs a real number. A sequential **neural network** is built with an alternate of RelU and dropout layers with 500 neurons each. A final dense layer with a sigmoid activation function provided the output.

Feature	Model	Short Term Memorability Score	Long Score Memorability Score
C3D	a) Random Forest N-estimators=100	0.302	0.102
	b) Linear Regression	0.277	0.103
	c)Neural Networks	0.227	0.070
HMP	a) Random Forest N-estimators=100	0.296	0.130
	b) Decision tree	0.053	0.012
	c)Linear Regression	0.317	0.129
	d)Neural Networks	0.317	0.129
Inception V3	a) Random Forest N-estimators=100	0.095	-0.037
C3D+HMP	a) Random Forest N-estimators=100	0.330	0.139
	b) Support Vector Regressor	0.206	0.117
	c)Linear Regression	0.011	0.012
	d)Decision Trees	0.114	0.025

Figure 1. Models with Spearman correlation coefficient

The evaluation metric used to compare the models is Spearman's correlation coefficient. This metric allows comparison between algorithms along with normalising the output of different systems by taking into account monotonic relationships between ground truth and system output. Though primarily a prediction task, the use of Spearman's rank as the official metric will allow for the evaluation of the systems based on the ranking of different video samples from the test set.

As in Figure 1, Random Forests and SVR have shown consistent results while training C3D and HMP features but also when the two are combined with an ensemble with the two models, have shown even better predictions. Best results were obtained with a Random Forest with estimators=100.

In this exploration, C3D merged with HMP performed the best with Random Forest and SVR and this was finally used to train the 6000 dev-set and then test the 2000 test-set to predict the final memorability scores for short and long term. As observed, when it comes to visual features, Inception V3 performed poorly while predicting the score while video features like C3D and HMP gave significantly better results.

V. CONCLUSION

The study of video memorability keeps evolving and it is envisioned that many applications can be developed from this domain. memorability is a study that keeps on evolving. New visual materials for learning and education could benefit from the memorability maps approach, which reinforces forgettable aspects of a picture while retaining memorable ones. The exploration showed that Video features showed better results than Image Features. C3D and HMP when merged with an ensemble provided higher results when modelled with Random Forest and SVR. While, predictions were made, there is more scope to make predictions with an ensemble of image features(weak learners). Another aspect could be introduced where additional features can be extracted from the video to predict memorability.

REFERENCES

[1] Dan Claudiu Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *Twenty-second international joint conference on artificial intelligence*, 2011.

[2] Romain Cohendet, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engilberge. Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2531–2540, 2019.

[3] Romain Cohendet, Karthik Yadati, Ngoc QK Duong, and Claire-Hélène Demarty. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 178–186, 2018.

[4] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1469–1482, 2013.

[5] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2390–2398, 2015.

[6] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2730–2739, 2017.

[7] Wensheng Sun and Xu Zhang. Video memorability prediction with recurrent neural networks and video titles at the 2018 mediaeval predicting media memorability task. In *MediaEval*, 2018.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.