Predicting Video Memorability Using Multi-Output Regressor

Nivedita Mahato
Department of Computing
Dublin City University
Dublin, Ireland
nivedita.mahato2@mail.dcu.ie
19210398

***Abstract*—With the advent of technology, the video content produced by each unit of time is rising exponentially. Some Videos have a different impact on people and stay in their memory. If we ask someone to explain an image or picture, he will mention some important details of it, which means that he only holds the key points of the image that allow him to recall. In this paper are trying to develop a model which will predict memorability of various video based on video features and Semantic features using various machine learning algorithms.**

***Index Terms*—C3D,Aesthetic,CounterVectorizer,Captions**

I. INTRODUCTION

Social media is rapidly expanding its horizon with different social media applications. The media data, especially the data in the form of video, has increased.The media memorability prediction is useful to analyzed which video can be watched more or remembered by the audience.As humans can remember specific video based on some particular features of a video or by the Caption or title of the video. In our research, we are training our models with the video and semantic features to check which features are more helpful in the media memorability score compared to other features. In this paper, various video features such as C3D and Aesthetic features are used for evaluating the media memorability score of our videos. The semantic features (Captions) are also used in our study. Semantics is the description of the videos. The models were evaluated using Spearman's correlation score as a standard measure.

II. LITERATURE REVIEW

A lot of interest has been taken on the video memorability and work has been done to improve various machine learning model prediction, and recent works [1] also used various high level visual features such as deep learning based action recognition representations (C3D-Preds), and image features of the videos and captions as semantic features for memorability prediction [2]. The key findings on the memorability of these papers are that the models which used caption have the best outcomes. In addition, researchers have found that high-level semantic features learned by CNNs trained in image classification achieve state-of-the-art success on a variety of computer vision tasks [4]. With the changing world the tools that provide powerful parallel machines (GPUs) along with a large amount of training data, Convolutional neural networks (ConvNets) have come back to provide new advances in visual recognition. ConvNets has also been applied to the issue of human pose estimation in both pictures and videos [3].In this paper we will focus on C3D feature for our analysis.The Authors have proposed a novel approach which uses both the features of audio-visual and fMRI.Model was built which learned on fMRI-features which mirrors the brain activity of memorizing videos. In second step they were able to predict VM without fMRI-scans. The extensive feature engineering was performed on several computed features before building the final predictor. The analysis included C3D deep learning features, color features, semantic features obtained from video captions, dense trajectories and saliency features. Caption feature turned out to be best parameter for predicting VM [6].

III. DATA DESCRIPTION

The dataset contains two parts development-set (dev-set) and test-set making a total data of 8000 videos. Of these 8000 videos, 6000 videos are used as a development set and rest 2000 videos for the test set. Raw footage by professionals was used as the source of extraction of this content. The development set contains the memorability scores for each video with the short and long-term memorability score, while the test set data with only video name. Machine learning algorithms are used to find the memorability score for the test set of our dataset. The dataset consists of CSV file which contains all the original title of the videos. These titles are most of the time seen as Bag of Words (textual data) which is also used for training the data for memorability of the videos. It also contains pre-extracted Video features like C3D features and HMP (Histogram of Motion Patterns) and Image features such as HoG (Histogram of Oriented Gradients), LBP (Local Binary Pattern), InceptionV3, ORB (Oriented FAST and Rotated BRIEF) and Color Histogram.

IV. APPROCH

The following section describes how I approached to the solution.

A. Features Selection and Data Pre-Processing

For our analysis, we have used Video features like Aesthetic Features and C3D and Semantic feature (Caption). The Semantic Features were cleaned by eliminating special characters, turning captions into small cases, and deleting stop words. The collected terms were used to make a bag of words. This bag of words was run with CountVectorizer to get features. These features are then sent to Machine Learning models as independent variables. The four models are trained with each of the features, and the results are recorded.

B. Models

The most useful feature of regression analysis is that it helps in crunching the numbers to help make better decisions for the company now and in the future. The model selection has proven to be challenging and fascinating. The approach is to shorten list models based on the understanding of the model and choosing models based on previous research performed by others. Different regression models are used for our study in this paper. Semantic and Video features are used over three simple models:

1) Extra Trees Regressor
This class implements a meta estimator that applies several randomized decision trees to different dataset sub-samples and uses averaging to boost predictive precision and control over-fitting of data.

2) Multi-Output Regression Model
The Multi-Output Regression is used to fit one regressor per target. It is an easy technique to expand regressors that do not usually support multi-target regression.

3) K-Nearest Neighbors Regression Model
K-nearest neighbors is a straightforward algorithm that stores all possible cases and predicts the result based on a similarity measure (e.g. distance functions).

4) Gradient Boosting Regression Model
Gradient Boosting is used to build smaller weaker trees which will be then combined together to generalize them. This Machine Learning technique is used as for regression problems. Each of the above models is explored with Captions,

Aesthetic and C3D features of Dev-set. The results of these models are tabulated below. The scores are calculated using Spearmann's correlation.

V. RESULT

The result shows that the model with Captions works best with Multi-Output Regressor Model. Therefore,Multi-Output Regressor Model with Semantic feature (captions) containing CountVectorized features is used for final computation.

TABLE I
SHORT TERM MEMORABILITY USING SPEARMAN CORRELATION

Machine learning Models Short-Term Memorability	**Model Features**		
	Captions	***C3D***	***Aesthetics***
Extra Trees Regressor	0.349	0.254	0.254
Multi-Output Regressor	**0.418**	0.330	0.330
K-Nearest Neighbors Regressor	0.314	0.258	0.258
Gradient Boosting Regressor	0.400	0.313	0.310

TABLE II
LONG TERM MEMORABILITY USING SPEARMAN CORRELATION

Machine learning Models Long-Term Memorability	**Model Features**		
	Captions	***C3D***	***Aesthetics***
Extra Trees Regressor	0.115	0.079	0.079
Multi-Output Regressor	**0.177**	0.114	0.114
K-Nearest Neighbors Regressor	0.127	0.112	0.112
Gradient Boosting Regressor	0.130	0.162	0.111

VI. CONCLUSION AND FUTURE WORK

In this paper, experiment with different regression models has been performed on different features. The result of each model indicates that semantic features were the excellent predictors of memorability scores in both long term and short term, while C3D and Aesthetic features performed poorly. The Multi-Output Regressor Model outperformed all the other regression models. The future work would be for improving both short term and long-term memorability with the focus on different features of videos such as Inception, Histogram of Motion Pictures along with ensemble regression models.

REFERENCES

[1] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino.2018. AMNet: Memorability Estimation with Attention.In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 6363–6372

[2] 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.

[3] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia,and Akhil Shetty. 2017. Show and Recall: Learning What Makes Videos Memorable. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2730–2739.

[4] Cohendet, R., Demarty, C.H., Duong, N., Sjöberg, M., Ionescu, B. and Do, T.T., 2018. Mediaeval 2018: Predicting media memorability task. arXiv preprint arXiv:1807.01052.

[5] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva,and Patrick Le Callet. 2016. Deep Learning for Image Memorability Prediction: the Emotional Bias. In Proc. ACM Int. Conf. on Multimedia (ACMM). 491–495.

[6] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han,and Tianming Liu. 2015. Learning computational models of video memorability from fMRI brain imaging. IEEE Transactions on Cybernetics 45, 8 (2015), 1692–1703.