Predicting long-term and short-term Video Memorability using Semantic and Video features.

Raj Singh,
Department of Computing,
Dublin City University, Dublin, Ireland,
raj.singh5@mail.dcu.ie
Student ID: 19210213

Abstract – We are performing the task of predicting the long term and short-term memorability of a video using Semantic feature and video features of a video like captions, HMP, C3D. Memorability is a factor that shows us how important or influential an item or thing was to leave a mark on a person's mind so that the person can recall the image or gather some information about the item with ease. Nowadays, video contents generated are increasing exponentially and are used for entertainment, advertisement, social awareness, etc. So by predicting memorability scores of a video can have a lot of applications. By this, we can know what type of videos a higher impact on the viewer, this can be used in marketing, advertisement, entertainment industry, and by different businesses leading to a higher impact on the customers thereby increasing sales. We are using Random Forest, SVR, and RNN on Semantic and Video features of a video for predicting the memorability scores. Finally, we have also combined some models to give better spearman's score.

Keywords—Random Forest, SVR, RNN, C3D, HMP, Captions, Video Memorability, Ensemble Model

I. INTRODUCTION

As the number of videos generated every day is increasing at a rapid speed there is the need of studying these videos to know what type of videos make a higher impact on the end-user and stays in their memory. Memorability can be used to help make choices by comparing videos that can have applications in advertising, education, entertainment, content recommendations, and filtering. So, many industries can be benefited from a model capable of judging videos upon their memorability. We are provided with a training set of 6000 videos with different pre-computed features. Each video also has a ground truth associated with it. After the model training, we are further provided with a 2000 test set videos for memorability prediction. In the research, I have analyzed 3 features and their combinations: a) Captions b) HMP c) C3D d) Captions + C3D e) Captions + C3D + HMP. We have trained our model with all these features and their combinations using SVR, Random Forest, and RNN and evaluated them using the Spearmen correlation coefficient. Then the combination of the best model and the best feature is used to predict the memorability of video. As per our observation of the spearman scores, the Random forest model has yielded the best results for captions. Also, we have combined Random forest with Captions, Random Forrest with C3D, Random Forest with Captions + C3D and Random forest with Captions + C3D + HMP to create an ensemble model and yield a better spearman's co-efficient than others. In the rest of the paper, there will be a detailed explanation about the models and the features used for prediction and highlights of the results obtained.

II. RELATED WORK

Different models were used on different features like C3D, Histogram, captions, HMP, Inception. The main finding from previous papers is that captions provide the best memorability scores among all the features. [3] Many new machine learning techniques have been used to study different features. [3] Further, Gupta and Motvani have used Logistic Regression, Support Vector Regression, and ElasticNet on different video features. As per their results, the short-term memorability scores obtained by them were 0.5 and 0.26 for the long term using the ResNet model.[3]
Also, a multi-layer perception model has been used on C3D, HMP and more, and lastly, they have used an ensemble model for prediction for more efficient prediction.
Also, we have seen that there is a correlation between long-term and short-term memorability scores, it's clear that all the models score higher in the short-term memorability score if compared to long-term memorability scores.[6] Loong-term score range between 0.2-0.8 and short term score is distributed between 0.4-1. Furthermore, an investigation found out that there was a higher memorability score for outdoor scenes rather than human-made scenes.[6] Han, Liu and Fan, 2018 motivated the use of Inception V3 feature on a pre-trained model using transfer learning to have good memorability outcomes in the area of image classification from small datasets.[7]

III. APPROACH

This section describes how memorability scores were predicted using different models and features.

3.1 Models and Features Used:

A) Random Forest with Captions
B) SVR with Captions
C) RNN with Captions
D) Random forest with C3D
E) SVR with C3D
F) Random Forest with Captions+C3D
G) SVR with Captions+ C3D
H) Random Forest with Captions+C3D+HMP
I) SVR with Captions+C3D+HMP
J) Combination of all Random-Forest implementation.

3.2 Features Description

- **Captions**: Captions are a one-line description of videos.

- **C3D**: C3D is a new generic feature for Videos. We get C3D by training a deep 3D convolution network on videos. It is a single list of numbers on one line.(Dimension = 101)

- **HMP**: It is a single list of pairs of numbers with the format: bin: number (dimension = 6075) on one line.

3.3 Data Cleaning and Flow

First of all, We have written the function to load Captions, C3D, HMP and calculate the Spearman's score. Then we have loaded the dataset for all the 3 features and the ground truth file with the long-term and short-term memorability scores. To avoid dimension issues when trying to predict the test scores we will be implementing CounterVectorizer. Count Vectorizer is used to tokenize a text file and establish a vocabulary of known words and also encode the file using it.[5]. It created a bag of words for each video caption.
Now we cleaned our caption dataset by removing punctuations, converting all the words to lowercase, and removing the stop words. After this, we have combined the different features as mentioned above.

	video	caption
0	video3.webm	blonde-woman-is-massaged-tilt-down
1	video4.webm	roulette-table-spinning-with-ball-in-closeup-shot
2	video6.webm	khr-gangsters
3	video8.webm	medical-helicopter-hovers-at-airport
4	video10.webm	couple-relaxing-on-picnic-crane-shot

<Captions before cleaning>

	video	caption
0	video3.webm	blonde woman massaged tilt
1	video4.webm	roulette table spinning ball closeup shot
2	video6.webm	khr gangsters
3	video8.webm	medical helicopter hovers airport
4	video10.webm	couple relaxing picnic crane shot

<Captions after cleaning>

3.4 Training the Model

Now we have trained our model with Random Forest and Support Vector Regression for all the above features. We have split our data into a training set and a Test set, here we are using the test set for validation and testing. This split in the data was used to calculate the Spearman's correlation coefficient score for short-term and long-term memorability scores. Once we get the scores for all the combinations of models and features, we select the best model and the feature and train the entire 6000 development set and use the results to predict for the test set.

IV. Model Description

We have used Random Forest, Support Vector Regression, and RNN models on different features to predict the memorability of videos. Out of these, the Random forest Model yielded the best results when combined with the Captions feature.

A) Support Vector Regression

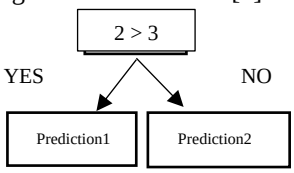
These are similar to simple regression. In simple regression, we try to minimize the error rate and in SVR we try to fit the error with a certain threshold. [8] SVR provides us with a lot of flexibility for our model as we can decide how much error rate is acceptable.

B) Recurrent Neural network

RNN is a class of artificial neural networks. It uses previous inputs and outputs within the calculation. For example, We can predict future words much better if remember the previous letters.

C) Random Forest Model

With libraries like Scikit-Learn, it has become very easy to use many machine learning algorithms in Python, it is important to know how the Random Forest algorithm works.
Random forest is made from many decision trees and hence it is a building block of Random Forest. Decision trees are nothing but a series of Yes/No questions asked about data which leads us towards the predicted class. It makes classifications as humans do, we can ask a series of questions about the training data and conclude.[9]



When we use the random forest algorithm to solve a problem we use the MSE to how our data goes from one node to another.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2$$

- N is the number of data points
- fi is the value returned by the model
- yi is the actual value for data point i

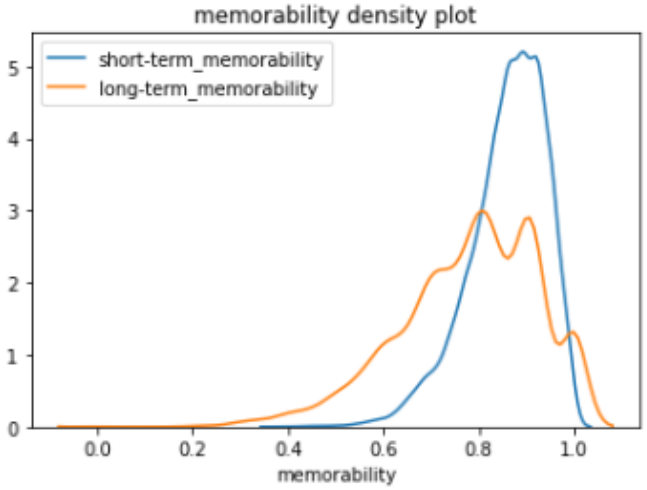
The formula calculates how far every node is from the value that is predicted, this helps to decide on which branch to select for a better result in your forest

D) Ensemble Model

Finally, we have used the all predictions done by the Random Forest model on Captions, C3D, Captions + C3D and Captions + C3D + HMP and combined them to calculate the spearman's co-relation co-efficient. This model performed the best if compared to all other models and had the spearman's score of 0.435 for short-term and 0.201 for long term memorability.

V. Results and Analysis

5.1 Memorability Score Distribution for the development set.



< Short-term and Long-term Memorability score distribution >

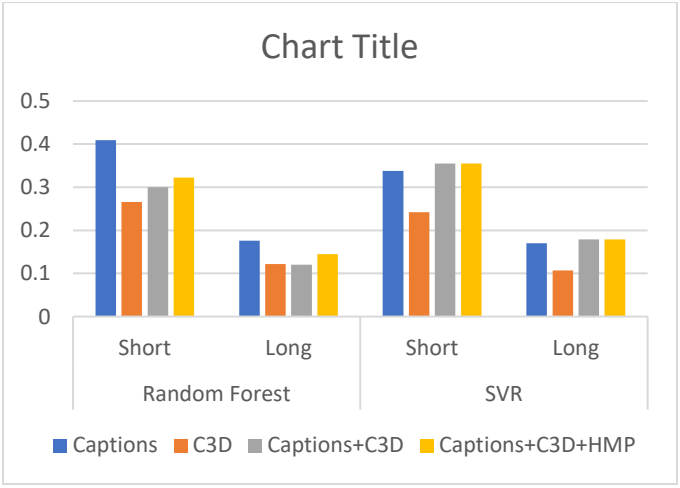
We can see that the memorability score for short-term is highly concentrated in a single area, whereas for long term score is more deviated.

5.2 Model Outcome

After applying all the models to the features, we get their spearman's correlation coefficient as follows.

	Random Forest		SVR		RNN	
Features:-	Short Term	Long Term	Short Term	Long Term	Short Term	Long Term
Captions	0.409	0.176	0.338	0.17	0.168	0.075
C3D	0.266	0.122	0.242	0.107		
Captions+ C3D	0.3	0.12	0.355	0.179		
Captions+C3D+HMP	0.322	0.145	0.355	0.179		

<Chart 1 for Spearman values for different combinations>



<Chart 2 for Spearman values for different combinations>

The best model that gives us the best performance is the random forest model with captions. The Short-term memorability score for Random forest with captions is 0.409 and the long-term memorability score is 0.179. Also, it is very clear that the long-term memorability score is very less for all the combinations if compared to its short-term memorability scores. This variance might be because the annotations i.e. number of people who appeared for the experiment is far less in the long-term if compared to the short-term. This might have happened because far fewer people might have appeared for the long-term test for memorability. Also, after combining all the Random Forest features results we created an ensemble model which gave Spearman's coefficient of 0.435. We also tried SVR on Captions to give a spearman's coefficient of 0.338. As for our final prediction we have used Random Forest with Captions for predicting the long term and short term memorability scores.

Furthermore, for the spearman's score's we previously divide our dev-set into a test set and train set. Now, we will train the model with a full 6000 dev-dataset and predict the scores for the 2000 testing set. After training, we will load the test ground truth file with video id and annotations and no long-term, short-term scores and the test-set for caption which we will use to predict the long-term and short-term scores and save it in the test ground-truth file with the same video-id as in test-set for captions. Finally, we will export the predicted results.

The long term and the short-term memorability scores are distributed in the following manner in the development set.

	short-term	long-term
count	6000	6000
mean	0.860243	0.778942
std	0.080655	0.144692
min	0.388	0
25%	0.811	0.7
50%	0.867	0.8
75%	0.923	0.9
max	0.989	1

<Development Set Short-term and Long-term scores Description >

	short-term	long-term
count	2000	2000
mean	0.848028	0.751163
std	0.034564	0.067863
min	0.687093	0.397318
25%	0.830368	0.721224
50%	0.849698	0.757218
75%	0.871004	0.792872
max	0.953749	0.94947

<Predicted Short-term and Long-term scores Description >

We can see that the mean for the short-term and the long-term scores for the development set and the mean for the long-term and short-term scores for the predictions are very close.

5.3 Analysis of Prediction

	Caption	Short-term	Long-term
video7494.webm	green jeep struggling drive huge rocks	0.854635	0.7127
video7495.webm	hiking woman tourist walking forward mountains...	0.898883	0.7815
video7496.webm	close african american doctors hands using sph...	0.841901	0.80563
video7497.webm	slow motion man using treadmill gym regular ph...	0.915282	0.82699
video7498.webm	slow motion photographer national park	0.866134	0.71173
video7499.webm	group mixed race american patriotic peoples am...	0.829725	0.74051
video7500.webm	business people train draw diagrams board lear...	0.829797	0.71951
video7502.webm	father daughters smiling	0.849537	0.72965
video7503.webm	mechanic using rotary polisher paintwork black...	0.778917	0.79046
video7504.webm	young couple conversation hotel corridor	0.866541	0.67518

<Predicted Output>

The top 5 highest score obtained for short-term memorability score is 0.9537 for caption “winter road traffic snowy highway mountains” and 0.9365 for caption “Potter teaches craft child” and 0.9341 for caption “mountain climber repelling cliff face” and 0.92908 for caption “stunning hillside sweep” and 0.9288 for caption “vibrant sunset field”

The top 3 lower scores obtained for the short term memorability is 0.68709 for caption “women eating healthy breakfast cereals milk” and 0.70994 for “lone moose munching grass” and 0.72745 for “playing videogames gamepad”

The top 5 highest scores obtained for long-term memorability scores is 0.94948 for “potter teaches craft child” and 0.92004 for caption “Asian man fastens pearl necklace Asian womans” and 0.91752 for caption “Stunning hillside sweep” and 0.91482 for caption “Cutting fruit Vietnam street” and 0.91071 for caption “videoblocks soldiers getting injured shooting”

The top 3 lowest scores for long term memorability are 0.39732 for caption “homosexual couple gay people young lesbian woman” and 0.48227 for caption “waterbucks walking around waterpool kruger” and 0.499794 for caption “dealer shuffling splitting deck”

We have observed that the memorability scores for both long-term and short-term are higher for outdoor views like mountains, sunset, rocks, hills, sunset, snow, street, etc and the scores are lower for indoor and task related to human beings like breakfast, munching grass, videogames, shuffling cards, gay couple. This same behavior is observed in the memorability scores of development set for the captions as well.

CONCLUSION AND FUTURE WORK

We have used 3 different models on 3 different features and their combinations and calculated spearman’s correlation coefficient for validating the results. The Random forest yielded the best results with the captions feature of the video after the ensemble model. Also, spearman’s coefficient clearly showed that captions outperformed other features. However, the spearman’s coefficient was very low for long-term memorability scores if compared to short-term scores for all the models used, the main reason for it can be a smaller number of annotations or fewer test subjects as observed.

We also observed that the memorability scores were high for outdoor variables like the sky, sea, sun, snow, traffic, etc if compared to indoor items like videogames, shuffling cars, eating, etc, for both captions in the development set and the captions in the test set for which we predicted memorability scores. This observation can be further studied and can be used in advertising, education, entertainment, etc.

Still, many different models can be used and tried on a different combination of features for obtaining better results. Many new features can also be evolved involving sounds, pixels, and colors which can make the problem and the solution more interesting. An ensemble model is also proposed in the paper which combines the results of different models for a much efficient prediction. Different ensemble models can also be implemented for higher accuracy in prediction.

REFERENCES

[1] 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.

[2] https://github.com/arunabellgutteramesh/VideoMemorabilityPredictionUsingML/blob/master/Report/Aruna_BellgutteRamesh_18210858_CA684.pdf, Video Memorability Prediction Using Machine Learning.

[3] https://github.com/harshalchaudhari35/MediaEval-Media-Memorability/blob/master/docs/report.pdf

[4] https://github.com/simonina/ML-Predicting-Video-Memorability/blob/master/Alex_Simonin_18212250_CA684.pdf

[5] https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/

[6] https://github.com/NavaneethanRajasekaran/predicting-the-short-term-and-long-term-Human-memorability-of-videos-using-NLP/blob/master/ML%20Report.pdf

[7] https://github.com/justprophet/Memorability-of-videos/blob/master/Paper-MachineLearningAssingment-Akash-18210613.pdf

[8] https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff

[9] https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76