CA684 Machine Learning Predicting Media Memorability Project

Akihisa Nishikawa

Security and Forensics

18210393

akihisa.nishikawa2@mail.dcu.ie

ABSTRACT

Importance of the media has significantly increased over the decade. Any activity through the internet, it is almost impossible to avoid digital contents like videos, and images. At the same time, the impression those media have become competitive. Predicting memorability is inevitable obstacles in many industries. This paper tackles the problem as the part of CA684 machine learning project and it presents three methods to examine the graphical and semantic features. In addition, analyzed how the combination of those features affects video memorability

1. INTRODUCTION

Variety of the digital contents are widely spread and bloated over the internet. The zillions of contents we come across are mainly from advertisement or social network. How memorable the media are is highly depending on the visual contents. In my work, I focused on the factors that are likely to be related to media memorability in real life. My research concentrated on the statistics rather than psychological factors or theoretical relation between features and memory. In order to predict the memorability correctly differentiation between short-term and long-term memory is necessary. Since long-term memory is not recalled as often as short-term memory, long-term memories are less likely to be remembered in details than short-term memory. [1] The human memory problem has been studied enough to show that individual context like hobbies or experience matter for long-term memory [2]. It indicates that predicting memorability of media in long-term memory from video features only is relatively harder since those features are not taking personal context in account.

1. RESEARCH

In order to identify the relation between memorability and pre-extracted features that are given in MediaEval 2018, first I looked into related work in human memory and memorability predicting with machine learning.

Many studies have proved that media memorability is highly related to the visual information. Especially motions. It is found that an image with salient motions tend to be considered more memorable. Also, depth of the image is remarkably important for memorability. [3] So those studies indicate that the flow of the consecutive frames is more important rather than the one frame out of video. Moreover, converting videos into 3D representation will create a type of feature and examining it helps to predict video memorability precisely.

With regards to the memorability of images or video, existence of objects, impression of objects, duration of appearance does matter. [4] has been able to reliably evaluate the memorability level of novel pictures or faces. Nevertheless, the prediction of memorability of images using learning algorithm are still limited. Especially when multiple objects are in image or videos, features of those objects should be taken into account as well. For instance, when a person is in the video the facial expression can influence the memorability. Each emotional expression affects differently to memorability score [5]

To express the features of objects that are in the videos, semantic features are the suitable in this case since they contain ambiguity of features like emotions also, they can be used without processing each object that are in each video. So here I consider captions as features of objects which appear in video.

1. APPROACH

For the algorithm part, I implemented Neural Network, Linear Regression, and Random Forest. Comparison of those models are main purpose of using multiple models. I keep all features input as single feature or simple concatenation between features in order to make it comparable.

* 1. Features Extraction

I used three different features and certain combination of those features. The selection of the features is based on the research. Mainly focused on the motion and flow of the video as the [3] [4] have shown in their work. Those features that contain the information of continuous frame, or motion of objects rather than single frame. To avoid overfitting, all frame should be considered or taken into account for some case the appearance of objects is too sudden and short. In that case, it is hard to predict the memorability even though the video was frightening. Here are the features combinations that I used.

* **Captions**: Textual feature describing the video. Captions are manually created. As I mentioned in last section, here I consider captions as features of objects in the videos.
* **C3D**: C3D is obtained by training a deep 3D convolutional network. It contains various concepts including objects, actions, and, scenes.
* **HMP**: Histogram of motion patterns. They encode the global temporal dynamics of an actions
  1. Features Combination

For each feature, all algorithms are applied to examine Spearman’s score. In this section I present the combination of feature.

* **Captions + HMP**: the importance of motion and what object is causing the movement are proven by [3] [5]
* **C3D + HMP**: C3D is also considered as description of objects, frequency of occurring. Those features of objects are combined with Motion histogram pattern
* **All three combined**: Motion pattern histogram is combined with other two features. Objects features with motion is considered.

1. MODEL DESCRIPTION

In this section I am going to describe the algorithm and parameters that I pick. There are three algorithm I implemented Neural Network, Linear Regression, Random Forest (Regression). Since memorability evaluation is not classification problem, it is relatively harder for linear regression to predict memorability. However, I focused on comparison rather than getting high correlation score.

For splitting dataset, I used ratio; training:test = 80:20 For the neural network, I implemented it using keras. there are three layers and every layer is set to use Scaled Exponential Linear Unit (Selu) as an activation. Other activation function like softplus or softsign decrease the validation accuracy and spearman’s score significantly. Dimensionality of output shape for layers is 30, 25, 2. If they are set lower accuracy start going down. I intentionally reduced this parameter as going down the layers in order to assume memorability precisely. For same reason, dropout is set lower between first layer and second layer. As the data go down to layers dropout rate increase. Most importantly, Root Mean Square Prop algorithm is used for the optimizer. Only learning rate is changed and set to 0.0009 everything else is default. Mainly those parameters are choses by trial and error. For linear regression model is set quite simple only regularize is set to true. For the random forest each maximum depth of decision tree is set to 10.

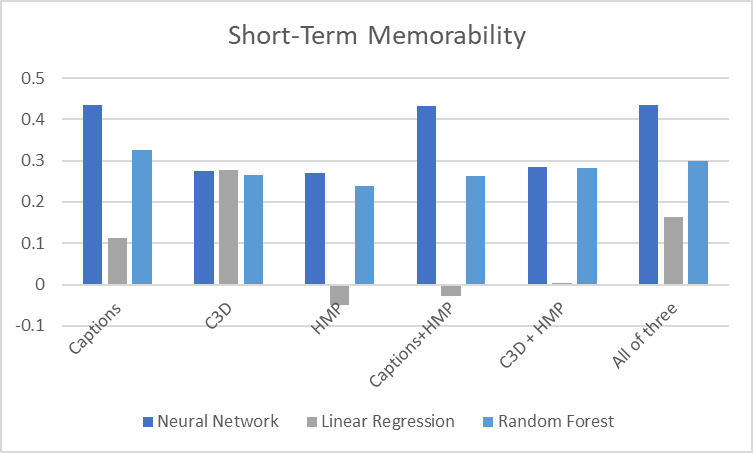
1. Analysis of the Results

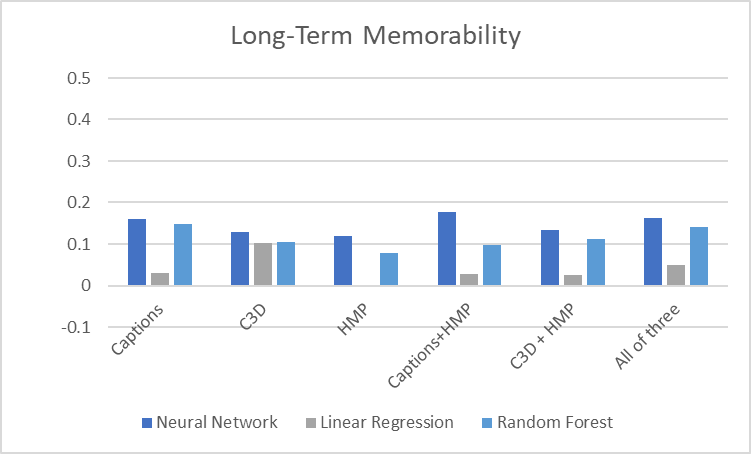
The results of my work is described in Table 1 and Chart 1 which are the comparison of Spearman’s score across the algorithm and input.





**Table 1 Spearman’s score**





**Chart 1 Comparison**

1. CONCLUSION

Overall short-term memorability scores are quite high in neural network. The highest is ensemble of three features with Neural Network. Only C3D represented almost no difference amongst the algorithms. For, the short-term memorability, every caption involvement increases the spearman’s score. As for the long-term memorability, neural network with captions and HMP remarked the best score, however most of the features are hard to be used for differentiating short and long term considering what [3] proved. The entire dataset is 6000 videos and the all features are pre-extracted in this project. Scaling up feature engineering in terms of variety of features is definitely necessary to predict the memorability in human cognition level.

# References

|  |  |
| --- | --- |
| [1] | D. Norris, "Short-Term Memory and Long-Term Memory are Still Different," 2017 May 22. |
| [2] | Markopoulos, Gerry, "Distinctiveness and memory," New York, 2007. |
| [3] | Sathisha Basavaraju, Paritosh Mittal, Arijit Sur, "Image Memorability: The Role of Depth and Motion," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 06 September 2018. |
| [4] | Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, Bernard Ghanem, "What Makes an Object Memorable?," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 13 Dec 2015. |
| [5] | Aditya Khosla, Akhil S. Raju, Antonio Torralba, Aude Oliva, "Understanding and Predicting Image Memorability at a Large Scale". |