MediaEval 2019: Predicting Media Memorability

Olabode Cole

Department of Computer Science
MCM1: Data Analysis
Dublin City University
Dublin, Ireland
olabide.cole3@mail.dcu.ie

Abstract— **This report is about the mediaeval memorability task, which seems to be the part of a medieval benchmarking initiative for evaluation. Participating members are meant to implement a system utilizing machine learning algorithms that would naturally anticipate memorability scores for videos in terms of features which emulates the probability of the videos being remembered. In relations to existing work within image memory probability, whereas image memorability was calculated right after being memorized, however, the dataset is composed of short and long-term memorability annotations which are utilized to calculate are efficient the system is.**

Keywords— Memorability, Machine learning, Annotations.

I. INTRODUCTION

Videos visual content is widely spread, in terms of entertainment or message being passed across. Statistics in 2018 shows that's Facebook averages over 8 billion of videos being viewed daily, with squared videos getting up to 275% more views more shares, 523% more comments, and 349% more reactions than the average. However, squared videos with embedded sounds would draw users attention and videos longer than 90 would often do a lot better, therefore, improving the probability of being remembered compared to soundless short clip videos which would be discussed further within this paper. Knowing all this information it becomes very interesting to see if there's a way of implementing a system that would take short soundless videos into accounts in terms of analyzing their cognitive and phycological factors.

A crucial aspect of human cognition is memory or the capability to recall visual content after reviewing it. Previous studies have proven that training specific features can be utilized to predict image memorability efficient enough that it could be compared to humans. Even though image memorability has been a fascinating aspect in recent years, modeling a system for video memorability have not been greatly explored as there are several aspects that need to be considered such as frame rate and the duration of the videos, etc.

Visual content can be used to transfer a message and the way certain user perceive does content are different based on their preference, therefore makes it difficult to determine the memorability of the overall content.

II. TASK DESCRIBED

The main purpose of this task is to implement a system that would be able to predict video memorability using different features and computing a memorability score for each video. The dataset that would be utilized consists of memorability annotation which are classified into two categories, long and short memorability score along with pre-extracted state of art visual features which means all the user has to do is to implement a system and train the pre-extracted features. Ground truth was compiled using a recognition test along with memorability result obtain through objective measurement. Users are enforced to alternate computational models competent enough to interpret memorability of a video through its visual contents.

However, users are also given the option to utilize the descriptive title of each video as a feature which in our case is called captions. The models implemented will be assessed using evaluation metrics used in ranking task such as Spearman rank correlation. A crucial aspect of the task is understanding the patterns within the dataset and improve the probability of an algorithm being able to identify such patterns

III. PREPROCESS

The caption feature was the only feature utilizes within this project as it was easy to understand and preprocess compared with the rest of the features for someone who doesn't have great programming experience. This is because the caption features consist of just one text file and within that text file there are two columns, one for each video and the other for the captions related to those videos, whereas for other features there are multiple text file within the features and I found it quite difficult to understand how I would preprocess those features in order to train them and achieve a good result. I also thought about using a bag of words analysis to preprocess the caption feature to see if it would produce better accuracy score but I ran out of time as I spent most of my time trying to understand how the algorithm I utilized works and to improve their predictions.

IV. APPROACH

Semantic-based Approach

Runs	Methods	Short Term	Long Term
1	RNN + Captions	**0.412**	**0.199**
2	NN + Captions	0.305	0.162
3	KNN + Captions	0.338	0.090
4	CNN + Captions	0.162	0.089

Table 1.0. Prediction Results

The main model that predicted the best result is a three-layer Recurrent neural network, the structure of this algorithm is later shown in figure 1.1.

After importing the video's captions text file along with the ground truth. The captions text was tokenized into a series of integer sequence with even length. After preprocessing 80% of the dataset were chosen at random for training while the remaining 20% was utilized for testing the models.

The preprocessed titles were transferred into the embedded layer with an input dimension of 5191, output dimension equal to 20 and input length of 50, along with the matrix being initialized by the uniform distribution. Dropout at the rate of 0.5 was utilized as the embedding regularization, the semantics was obtained by enumerating a fully connected recurrent layer with 10 units. A 10 node fully connected dense layer utilizing a rectified linear Relu activation function and for the linear transformation initializer for the kernel weight matrix was set as Orthogonal. The kernel regularization function implemented is l1−l2 regularization with λ1 = 0.001 and λ2 = 0.004, however, the initialization scheme is the exact same as the Recurrent neural network layer.

The last layer within the model happens to be a 2 node layer used for predicting the short and the long term memorability concurrently along with a linear activation function. The model is trained using ADAM optimizer contrary to the mean squared error. The model is prepared with 5 epochs with the validation data being the X test and Y test.
The convolution neural network process was quite similar to the Recurrent neural network in terms of preprocessing, however, the dimensions of the token titles weren't reduced as the principal component analysis was set to 0 and the output was imported into the CNN.
The convolution neural network needs to be aware of the input shape to expect, therefore, the first layer in a sequential model need to receive the input shape which was set to (4800,50,1200). The convolution 2d layer was set as 64 unit, along with a 3 by 3 kernel size followed by an activation layer known as rectified linear. In the case of max pooling, the spatial neighborhood was defined and the larget features within the rectified window were chosen. CNN layer is 2d but however the dense layer requires 1D, therefore, the layers within the models were needed to be flattened.

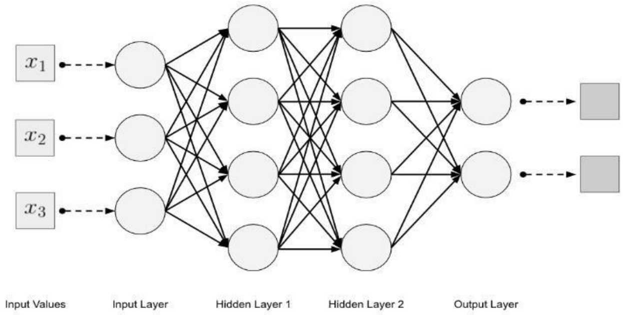

Figure 1.0. Recurrent neural network Architecture.

V. EVALUATION & RESULT

Based on the result shown inTable 1 the following conclusion can be observed that Recurrent neural network gave the best result among the rest of the features and algorithm utilized, it also tells us that the semantics of the video captions seems to be a lot more memorable compared to other features. Especially for long term memorability the video captions seems to surpass the other features, however, apart from the recurrent neural network, the performance decreases.

The activities occurring within each video has a huge impact on how memorable the video is and I believe the captions is capable of grasping the user's attention by summing up what's happening within the videos into short texts, therefore, would make it a lot easier to remember and better to predict.
Even though there are few interrelations between the short and long memory, the results have shown that the short term memory seems to be a lot more predictable than long term memory, however, I believe long term and short term memory are often subjective and depending on the individual.

The combination of the convolutional neural network along with captions didn’t predict a good score in terms of short or long term performance. I am not really sure why because the model seems to correct in terms of the parameters and it couldn’t have been the training methods as a higher epoch accuracy also gave a poor score while a lower epochs accuracy gave a fair score. When I plot the epoch along with the higher accuracy the graphs show that the neural network is not overfitting and seems to be accurate so I am not sure about the negative score. Meanwhile, the lower epoch accurate visualization graph which can be seen within the python file with a score of 0.162 - 0.089 shows that the data was overfitting/underfitting. Therefore I believe the Convolutional neural network is not as efficient when it comes to predicting memorability score.

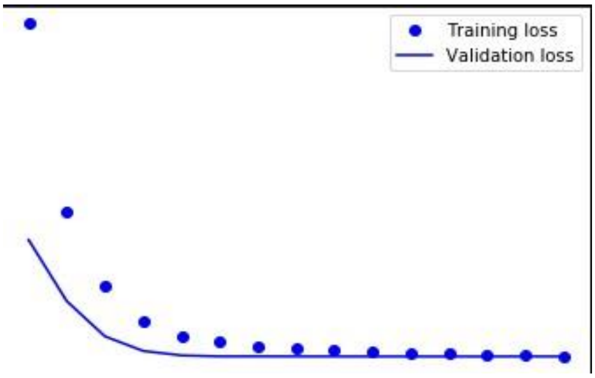

Figure 1.1. CNN Training & Validation loss Visualized

The prediction with KNN was quite decent as it was the second-best score from all the algorithms utilized. However, I was not able to display the result visually because KNN doesn’t seem to support the attribute history giving this error ‘ KNeighborsRegressor' object has no attribute 'history’. It is probably because the function history utilizes training loss and validation loss which is within the epochs but the version I implemented which gave me the best score does not include those attributes.

VI. CONCLUSION

Implementing this project was a huge learning curve for me as I learned a lot about how different algorithm works and what they are and are not good for. Looking at similar project online that have been implemented in that past the dataset mainly utilized were well-labeled csv table format which kind of makes the task a lot easier as there’s barely any preprocessing that needed to be done. However, the dataset we were given was a lot more challenging for example each video having their own CSV file within a specific feature, but from working with the caption feature I learned about how words/sentences could be preprocessed in term of tokenizing or BagOfWords methods. Therefore, I really enjoyed working on this project because the new additional skill and knowledge I have obtained can be utilized while working on my practicum which has a lot to do with Machine and Deep Learning algorithms.

VII. REFERENCE

(Romain Cohendet*, 2018)

(WenshengSun, 2018)

(Sumit Shekhar, 2017)

(Romain Cohendet, 2018)