The Media Memorability Prediction Task

Bharath Reddy Nagasetty-19211306
School of Computing,
Dublin City University,
Ireland
bharath.nagasetty2@mail.dcu.ie

ABSTRACT

In this paper, I present the Prediction of Media memorability task which was a part of MediaEval. The student was expected to design an automated system which predicts the memorability score for the videos which reflect the probability of a video being remembered. This approach uses three variables which are generated from the video clips along with NLP(Natural Language Processing) techniques and artificial neural networks which aid in predicting the memorability. The performance of the three features is compared and the best performance is chosen.

1 INTRODUCTION

In the current fast-paced times, the world has seen exponential growth in the generation of information from various sources in a day. A few materials stay in our memory while others fade away in due course of time. Identical pictures are often mistaken due to similarities or details that the human eye might miss. It is believed that media with faces, actions or events and subjects are more prone to be remembered in contrast with nature and landscapes which tend to be remembered quite less.

The ability to process this multimedia information has a potential benefit in developing recommendation systems according to the user's interest.Artificial Neural Networks have been useful to discriminate between cat images by viewing untagged videos.Memorability is a new field in computer vision which tells if an image/video is short term or long term memorable. NLP is a rapidly growing field with an expanding number of applications. Not much of research is done in the area of media classification and prediction.

Various types of features which were extracted from the original videos and were provided along with the development set. To grade the model's performance a spearman's rank coefficient is used. This paper proceeds in investigating a few among those features with prediction using neural network models along with preprocessing them to achieve a greater and efficient result.

2 RELATED WORK

In recent years work of paramount importance has been carried out in the field of the computer vision. Lot of research have been dedicated for identification of faces, landscapes and many more. However, only a few have been able to concentrate on the area of memorability. Although most of the research in this area is in progress, they are limited only to small scale[1].

To measure the memorability of images a set of employees from amazon were chosen for the visual memory game. The main goal of the experiment was to calculate the memorability scores of images which were presented. The participants were shown a sequence of images with a time interval of 1 second. The candidate's task was to press space bar whenever they saw a similar image in the sequence. The sequences in the game were broken into 120 images each such that each sequence gets the same number of images and then the memorability scores for each image were generated [2].

To automate this a model named MemNet was designed by fine-tuning of CNN and classified more than a thousand categories of images and scenes and achieved a correlation coefficient of 0.64. the model was applied on the overlapping between images to produce a memorability map and a non-photorealistic rendering was used to evaluate such memory maps[3].

In a competition to MemNet another model was developed which claimed to obtain better performance than the MemNet.The two main differences between the MemNet and this model were that it treated the prediction as a classification problem and combined the results that belonged to various classes. The base models included were SVR and KNN.[4]

3 APPROACH

To tackle the memorability task HMP, C3D and Captions were used. They were pre-extracted and provided with the problem. The main challenge was to prevent over fitting of the models to the data.

Four different models were developed for each of the features and comprised of two different models using captions.In the first approach a five layered ANN was used to train HMP. It consisted of 400 neurons in the input with a batch size of 60. In addition to that each layer had a dropout to prevent over fitting.The input and the hidden layers contained relu as the input and sigmoid as the output layer.

In the second, the array was formed by stacking the values on top of each other. The C3D model was formed by a five layer with alternating relu and sigmoid activation function

between the layers with adam optimizer.
The third model was made using captions by pre-processing the data and converting into vector counts and then passing them to input layer.The loss metric called mean squared error was used to measure accuracy.
The last model which was developed also used captions but in a different way. It uses embedding layer which converts the data into continuous variables in a smaller dimension. The ANN contained the embedding in the input layer followed by the dense layers.

3.1 Data Pre-processing

ANN models are trained with divergent pretrained features.The component captions contains recordings which was marked with a short description of what the video contains. The captions and the ground truth were merged to have a better view of the task. The captions feature which was in the form a text file consisted of many punctuations between the words. Preprocessing of captions included the removal of punctuations, cleaning of the text and removal of stop words. Tokenization was then applied to the text file and then the text was converted to number sequences using vectorization. Padding includes the building and fitting word index and converting the words as 2D array sequences.
In real-world, not all the sentences are of equal length, so to organise the data in equal length padding technique was used to fill the remaining array elements with a zero. When C3D and HMP features are concerned limited preprocessing is required. The array elements were stacked as mumpy array so as to make the development of model easier. The next step was to split the dev set to train and validation which was achieved my sklearn package.

3.2 Algorithms

Having assimilated the preprocessed data, the next step was to create a machine learning model to process the data and get the prediction. A forward feed neural network model has been used where the connection between the nodes does not form a cycle and the data is flown in a forward manner. Each feature uses a different density model. When HMP is considered a five-layered model with relu at the input and the sigmoid in the output layer.
Relu and sigmoid activation functions were used throughout this experiment.Relu was used to solve the vanishing gradient problem and sigmoid was used whenever the output was to be contained betwen[0,1].
Two types of optimizers were used which were rmsprop and Adam. RMSprop is used because there was a need to choose a different learning rate for each of the parameters in the model, whereas adam has given proven benefits while working with natural language processing. Along with the two optimizers, the MAE (Mean Absolute Error) and MSE(Mean Squared Error) were also used.

Table 1: Performance of Features

Runs	Feature	Short-term	Long-term
1	Captions	0.443	0.207
2	Captions with Embedding	0.412	0.204
3	C3d	0.275	0.124
4	HMP	0.325	0.145
5	Ensemble	0.457	0.221

3.3 Ensemble Model

After gathering prediction results from different features a simple ensemble model was designed. An ensemble model tries to gather all the models under a single roof and tries to increase its accuracy. There are various types of ensemble models but here the model which is taken into consideration is the simple average ensemble. The final predictions are taken as the average of all the member predictions.

4 RESULTS AND ANALYSIS

4.1 Tables

From the above table, we can observe the Captions gives a noticeable result. The Count vectorization model achieves a better result than the embedding model. By comparing the different models in captions, we can see that the CV model has an advantage only in the short-term memory and there is no significant difference between the long term results. In the visual features, the highest result achieved is in the HMP and we can conclude C3D was a poor choice with low memorability scores. The ensemble approach further enhanced the results where both short and long term scored showed a slight increase.

5 DISCUSSION AND OUTLOOK

From the above results, we can conclude that the short-term predictions are more trustworthy than the long term scores. We would have liked to add an extra feature mainly inceptionV3 and wanted to measure how it would have impacted the memorability scores. Another future improvement would be to ensemble all the models before prediction together and then train Adaboost classifier for the boosting of spearman's constant.
The main outcomes from this task would be that text features like captions gave a good score whereas the features based on the video. Another method which we wanted to try is Logistic regression which could be used for image features such as LBP which would be not that great with the results. A backpropagated neural network would also a good choice.

REFERENCES

[1] Quoc V Le. Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8595–8598. IEEE, 2013.

[2] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[3] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, 2014.

[4] H. Squalli-Houssaini, N. Q. K. Duong, M. Gwenaelle, and C. Demarty. Deep learning for predicting image memorability. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2371–2375, 2018.