Predicting Video Memorability

Amit Kumar - 19210716
Dublin City University, Ireland
amit.kumar8@mail.dcu.ie

ABSTRACT

Media significance has increased over the time. This paper is going to predict the memorability of a video for a watcher. Watchers need to instinctively forecast the video memorability outcome. The present paper deals with short term and long-term video memorability of videos and tries to find the best results. Moreover, this paper finds out memorability scores for the merger of these different features. Spearman's rank correlation coefficient is used to evaluate the model. The present paper utilizes features like Captions and HMP (Hierarchical Matching Pursuit) for training the model using deep learning model

1 INTRODUCTION

Now that we have a huge demand for new methods to arrange and compute digital material in real life, it inspires us to forecast video memorability [1]. Nowadays, media platforms tackle with big amount of data every day. As a result of this, the models which can sort the videos based on memorability could help a lot of applications [1]. Let's see some of the barriers for predicting the video memorability. First of all, video memorability is not properly explained by earlier works as compared to image memorability. The other problem is that dataset is not sufficient enough to to make the models [1]. Based on some studies, we can say that long-term memory can store huge amount of features for an item knowing that we deal with a lot of images and photos everyday [2]. Studies have also thrown light on the statement that people have huge regularity to hold on to details considering the fact that individual circumstances and individual utilization [2]. I came to know that CNN trained models are better than other models when it comes to outcome of a prediction. Moreover, I came to know that predictivity of short-term memorability of a video is more easily achievable than the long-term memorability.

2 RELATED WORK

I went through some research papers based on video memorability prediction and the dataset is provided by a competition named Media Eval 2018. Captions and HMP were among the main features for a good prediction. The main outcomes according to (Cohendet) and (Shekhar and Singal) informs about deep learning methods with features like Captions, HMP, C3D, Color Histogram, Inception V3 to predict memorability with deep learning method. Moreover, when CNN Trained models are used for the features the results are improved. Long-term memorability annotations are considered to be better performance measures for long-term memory performance [1]. Contestants are required to perform two subtasks which are Long-term memorability Subtask and short-term memorability Subtask. Contestants are provided with different video dataset like Dev-set, Test-set with annotations of the memorability.

3 APPROACH

3.1 Feature Extraction and data Pre-processing

Now that I went through research papers, I decided to use features like Captions and HMP for training the model. All these features are already provided to us but in this situation if we are not removing noise from our data there will be an issue of overfitting. Captions are basically small explanation of the video.

HMP (Hierarchical Matching Pursuit) provides better results when it comes to image features. Two Techniques utilized for making bag of words are TF-IDF Vectorizer and Count Vectorizer which are used to Pre-process the the captions. TF-IDF Vectorizer is used for converting text data into digit form for the model to understand the data clearly. Lancaster and Porter stemming is used to get better results. Since we already have the HMP features we will use an array to pull out the feature. The final computed array is stored as a NumPy array.

3.2 Model

Dense networks are used for building regression model based on deep learning. Mean Squared error is the algorithm for regression and Adam is used as an optimizer. Split and stopping monitor are utilized in this situation to deal with the problem of overfitting.

Now that we have received our results, ensemble models which considers all the features of the present models is taken into account for better results [3]. Simple and weighted ensembles are used in the present work.

Simple ensemble considers the average of the predictions of different parts but the outcome of every model is similar even if the results are better. A weighted ensemble considers the weights of results of every model. Predicting video memorability is a regression task so it was good to go with the neural network approach. The root mean square error value is predicted for both Long-term and Short-term video memorability.

4 Results and Analysis

In this situation, we have considered 20% test set and 80% train dataset to carry on the analysis with our deep learning model. The ensemble approach utilizes text and image features and improves the efficiency. Neural networks improve the performance in an even better way and when it comes to Short-term memorability results provided by neural networks are way too better. The results are basically the differentiation of the Spearman's correlation coefficient scores. Here, we will see the Spearman's correlation coefficient

scores for different features like caption and HMP (Hierarchical Matching Pursuit) for both Long-term and short term memorability predictions. The following tables namely table 1, table2 show the Long-term and short-term memorability scores.

Table 1. Short Term Memorability Scores

Model	**Spearman**
Captions	0.320
HMP	0.285
Ensembled model	**0.386**

Table 2. Long Term Memorability Scores

Model	**Spearman**
Captions	0.176
HMP	0.121
Ensembled model	**0.181**

1 CONCLUSIONS

Now that we know all the results, we can say that predictivity of short-term memorability of a video is more easily achievable and accurate than the long-term memorability. HMP which is an image feature provides superior. Captions and HMP Feature with neural network provided the best results. Improvements are needed in the field of feature engineering for several features for video memorability predictions. Memorability feature has a great importance in the field of data science and it is in high demand in the present market. There can be some more combinations of the features in the coming time which can lead to an efficient and better performance and these techniques can provide a higher video memorability prediction.

2 Acknowledgments

I would like to thank the research paper authors from which I gained a lot of information about this task. I got an opportunity to learn about neural network and ensemble model. I also gained information about the Count vectorizer and TF-IDF which are used to create bag of words which is very helpful in Pre-processing. These techniques definitely lead to a better result.

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] Shekhar, Sumit & Singal, Dhruv & Singh, Harvineet & Kedia, Manav & Shetty, Akhil. (2017). Show and Recall: Learning What Makes Videos Memorable.

[3] Sollich, P. and Krogh, A. (no date) 'Learning with ensembles : How over-fitting can be useful', pp. 4–10

[4] He, K. and Sun, J. (no date) 'Deep Residual Learning for Image Recognition', pp. 1–9

[5] [1]S. Lathuiliere, P. Mesejo, X. Alameda-Pineda and R. Horaud, "A Comprehensive Analysis of Deep Regression", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1, 2020. Available: 10.1109/tpami.2019.2910523 [Accessed 29 April 2020].