Predicting Video Memorability Using Captions and Image Features

Akash Bhargava- 18210613
MS in Computing
Machine Learning(CA684)
Dublin City University
Dublin, Ireland
akash.bhargava2@mail.dcu.ie

Abstract—**Memorability is defined as the quality or state of being easy to remember or worth remembering. With billions of hours of user generated video content on online platforms, prediction of a cognitive measure like memorability can have many potential applications. In this paper, the use of three variables in predicting Video Memorability was investigated. With the use of various intrinsic features, we were able to determine the best option that outperforms other models to predict video memorability.**

I. Introduction

In this work we investigate the use of both captions and video features which helps us to predict video memorability. We conducted an extensive analysis on the three features and the ensemble to determine the most robust method amongst them. Among the features provided for the model, we used captions, HMP (Hierarchical Matching Pursuit) and Inception V3 features in a deep learning model. Based on the results we can conclude the following.

1. Short-term predictions were more accurate than the long-term predictions.
2. Due to the limit in the number of observations, captions proved to be a better approach against the video features.
3. As the sample size increases video features prove to be more robust in terms of performance.
4. It was observed that models trained on high level CNN trained on captions or image classification (visual features) give a comparatively better result than traditional models.

II. Literature Review

In recent years, the work on video memorability has been a subject of great interest. The works of (Cohendet, no date), (Isola *et al.*, 2013) and (Shekhar and Singal, no date) tells about the use multiple low level and high level visual features and some deep learning based action recognition representation (C3D-Preds), and image and video captions for memorability prediction. The important conclusions from these papers are that models using captions give the best individual results and the features when passed through deep learning models significantly improve those results. Adding to that, it was found out by the researchers that high level semantic features learned by CNNs trained for image classification give state of the art performance on a variety of computed vision tasks. (Han, Liu and Fan, 2018) emphasized the use of Inception V3 features on a pre-trained model using transfer learning to have given astonishing results in areas of image classification for smaller datasets.

III. Approach

A. Models

To tackle the problem of video memorability we used captions, HMP and Inception V3 features. All these features were provided as a part of problem statement. Due to high dimensionality of the features, high variance and over-fitting were the major concerns of the task.

- We used captions that are the one sentence text descriptions of the videos for which we had to predict memorability. As per the research conducted in previous papers, the use of captions was important in smaller datasets.
- We used HMP features (Bo and Fox, no date) as it was expected to give better results in terms of image features.
- We used Inception V3 features (Szegedy *et al.*, 2014) as well to compare against HMP features.

B. Data Pre-Processing

- For pre-processing captions, we used Bag of Words feature of Skikit-Learn. We used two methodologies namely Count Vectorizer and TF-IDF Vectorizer to create bag of words. The idea behind Count Vectorizer was to give weights to the words based on the memorability score provided. For this, we used word unigrams and bigrams and removed English stopwords. The TF-IDF algorithm assigns weights to the words based on its occurrence in the text. In this case, we selected English dictionary as a parameter. In both cases we used tokenization using NLTK library. We also tried Porter and Lancaster stemming to generate more accuracy while creating bag of words and found out that Lancaster stemming gives better results in this scenario.
- HMP features were provided in text files as a part of the competition. We used a method to extract those features in an array and converted the resultant array to Numpy Array. Similarly, for Inception V3 features, we used 56nd frame data.

The assumption behind this was that most relevant information would be present in the 56th frame. The data was extracted and stored in a Numpy array.

C. Modelling

In place of traditional models, we used dense networks. The idea behind this was to create a deep learning regression model that will predict the memorability score using all the latent features we extracted in the previous section (He and Sun, no date). See fig 1.

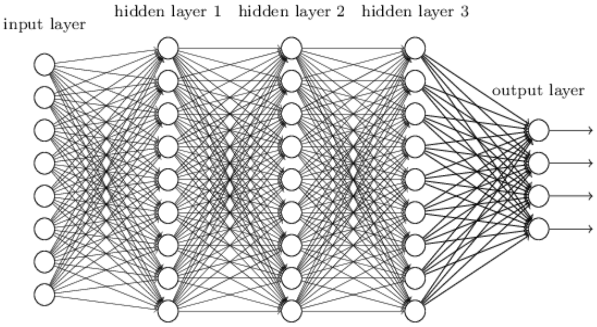

Fig 1

To compile 'adam' has been used as the optimizer and 'mean_squared_error' has been used as the regression algorithm. We have added validation split and early stopping monitor to address the issue of overfitting of the training data.

D. Creating an Ensemble Model

After generating results from separate features, we created two ensemble models. An ensemble model tries to contain the features of all existing models thus giving a bit more in terms of accuracy (Sollich and Krogh, no date). We focused on two ensembles i.e. simple average ensemble and weighted average ensemble.

In simple average ensemble, the ensemble prediction is calculated as the average of member prediction. A limitation to this model is that each model has equal contribution in the final result irrespective of its performance.

A weighted ensemble is an extension of a simple average ensemble where the contribution of each member to the final prediction is weighted by the performance of the model. In this case, the custom weights are all positive and the sum of all weights equals to one which allows the weights to indicate the percentage of trust or expected performance from each model.

IV. Results

Fig 2 and 3 figures gives an overall summary of the experimental results.

- Two modes for Captions namely CV (Count Vectorizer) and TF-IDF
- Average Ensemble takes equal weights for Captions.
- Weighted Ensemble is 0.50*Captions (CV) + 0.25*Caption (TF-IDF) + 0.25*HMP

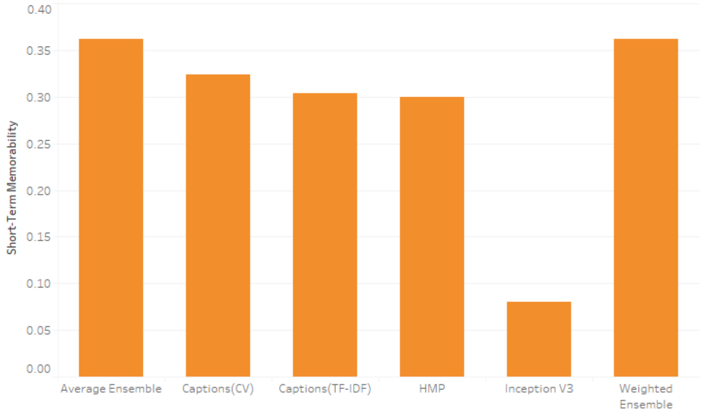

FIG 2. SHORT TERM MEMORABILITY SCORES

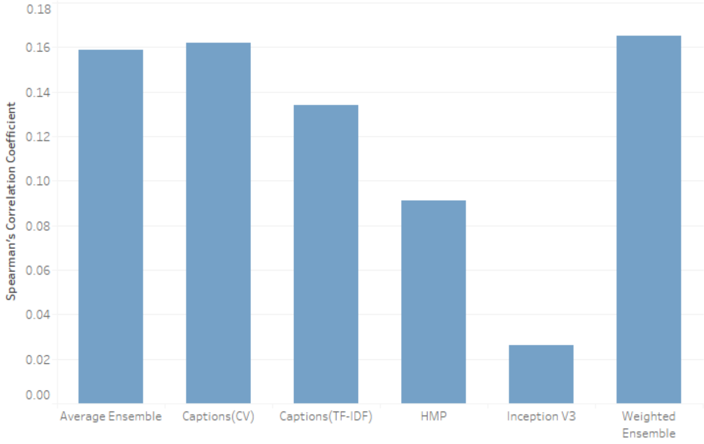

FIG 3. LONG TERM MEMORABILITY SCORES

V. Analysis and Discussion

From the above-mentioned results, it can be said that Short-Term memorability predictions are more credible and accurate than Long-Term predictions. Also, models with captions (both CV and TF-IDF) achieve better results in terms of individual models.

In terms of Visual Features, Inception V3 seems to be a poor choice with the low memorability scores in both departments where as HMP gives significantly better results.

Ensemble approaches seem to enhance the memorability scores. From the two, weighted ensemble helps in significantly increasing short-term memorability score.

For further improvements we would have liked to extract Resnet features to measure the impact on memorability scores. We wanted to try Logistic Regression for image features and use methods such as Z-Score Normalization and Support Vector Machines for Captions.

In Conclusion, memorability is a study that keeps on evolving. With the addition of some more features such as sound makes this problem even more complex and interesting as it brings processing of signals into the picture. It has a lot of applications in the market and will keep attracting the interest of data scientists in the years to come.

References

Bo, L. and Fox, D. (no date) ‘Hierarchical Matching Pursuit for Image Classification : Architecture and Fast Algorithms’, pp. 1–9.

Cohendet, R. (no date) ‘Annotating , Understanding , and Predicting Long-term Video Memorability’.

Han, D., Liu, Q. and Fan, W. (2018) ‘A new image classification method using CNN transfer learning and web data augmentation’, *Expert Systems with Applications*. Elsevier Ltd, 95, pp. 43–56. doi: 10.1016/j.eswa.2017.11.028.

He, K. and Sun, J. (no date) ‘Deep Residual Learning for Image Recognition’, pp. 1–9.

Isola, P. *et al.* (2013) ‘What makes a photograph memorable ?’, pp. 1–14.

Shekhar, S. and Singal, D. (no date) ‘Show and Recall : Learning What Makes Videos Memorable’.

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

Szegedy, C. *et al.* (2014) ‘Rethinking the Inception Architecture for Computer Vision’.