Video memorability prediction

Rashmiranjan Das
19210554
MCM Computing(data Analytics)
Dublin City University
Rashmiranjan.das2@mail.dcu.ie

Abstract- We watch multiple movies and advertisements in a day. What makes a video stand out and be retained in our memory. In this paper, we will be addressing that question and predict the memorability scores of videos presented under "MediaEval: Predicting Media Memorability challenge"[2]. The challenge composed of 8000 soundless short videos associated with two scores of memorability i.e. their probability to be remembered after two different durations, a few minutes after the memorization of the videos, and then 24 to 72 hours later. After exploratory data analysis, I test multiple deep neural networks for video predictions. My best model was an ensemble of the independent models to give a short-term score of 0.443 and a long-term score of 0.233.

Key Terms: C3D, HMP, Tokenizer, Captions, Sequential neural network, TF-IDF and Ensemble.

I. Introduction

With increasing technology and growing streaming services, our video consumption has increased exponentially. With everyone quarantined due to coronavirus, Netflix has 160 million active subscribers. Even after consuming large content, we remember scenes that we watched many years back. Humans have an exponential ability to remember fine details but as we are unable to remember all incidents. The question that arises is, what makes a video memorable. In this paper, we will be addressing that question and create a model to predict the short term and long-term score for sample videos provided. This study will definitely benefit the content creators, artists, digital marketing team and many more to analyze the outcome and focus on key insights to make their content more memorable.

One of the papers for Media Eval 2018 competition [3] demonstrated a model that predicted the memorability score using the semantic and visual features combined. They have identified that the video features C3D and HMP have outperformed the image feature such as ColorHistogram, inception V3-Preds and LBP.

Using Deep learning we propose a solution to the memorability problem by estimating the VM score[4].

II. Data

A video consists of various features i.e. semantic features like captions and video features. Set of precomputed features using CNN (Convolutional neural network) like C3D, HMP, LBP, InceptionV3 etc[1].

The model was created based on sequential network with dense layers. The selected model was run on each of the features. The Semantic feature was preprocessed and tokenize to convert into a vector for model input. Video features like C3D and HMP were directly used with no preprocessing. In order to improve the accuracy, an ensemble model was built on top of all the distributed models. The models were assigned with weights and averaged to fetch the best predictions.

III. Approach

A. Models selected

The model selected for all the features is a sequential neural network with two dense layers and 'selu' as the activation function. It was followed by a dropout layer of 0.5 intensity to make sure that the data doesn't overfit.

B. Feature engineering

The semantic feature was preprocessed to eliminate unnecessary characters like punctuations. Special characters and stop words. A counter subclass of the collection is imported to store a key-value pair. Using class tokenizer, vectorized the text corpus by turning into a sequence of integers. Fitted the corpus as a matrix with binary as the mode of other modes (i.e. 'count', 'tfidf' and 'freq').

Video features like HMP and C3D are used straight away as input to the sequential model. HMP has 6075 attributes while C3D has 101 columns as input. Both the model yielded a poorer score compared to captions while C3D had the best individual short-term score of the rest.

Performed grid search to obtain the best dropout parameter, activation function, Learning rate, network weight, batch size and number of epochs.

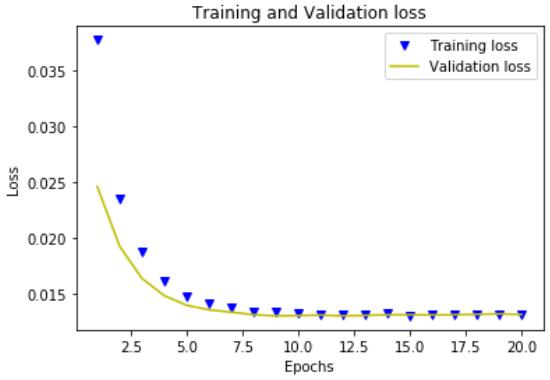

Fig 1

All three features (caption, C3D and HMP) were combined to form a model that gave a long-term best of 0.421 and 0.231.

Decided to ensemble all four models with equal weights as all the features are independent of each other which allows the model to focus on the respective features.

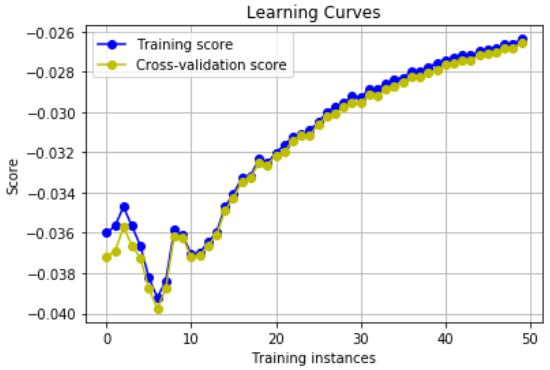

Fig 2

The following table depicts the score obtained by each model and its gradual growth after multiple iterations.

Features	***Short term score***	***Long term score***
HMP	*0.282*	*0.152*
C3D	*0.357*	*0.15*
Captions	*0.403*	*0.198*
Ensemble 1	*0.432*	*0.222*
Combined model	*0.421*	*0.231*
Ensemble 2	*0.443*	*0.233*

Fig 3

IV. Conclusion

The key analysis for the paper is:

1. Caption is the only semantic feature available and gives the best memorability score
2. Short term feature is more accurately predicted than long term score
3. Few of the video features don't contribute at all towards the memorability score
4. After caption, HMP and C3D provide a better score than other features

The importance of semantic feature and new words is important to generate weights, which will lead to improving the VM score. However, an ensemble model helps to factor in all the features for improved results.

I'm proud of the result that I was able to achieve in the given time. I do feel there is scope for improvement. As future work, I would like to use more features and search for the optimum their best respective score.

References

[1] "arXiv Fulltext PDF." Accessed: Apr. 29, 2020. [Online]. Available: https://arxiv.org/pdf/1807.01052.pdf.

[2] "arXiv.org Snapshot." Accessed: Apr. 29, 2020. [Online]. Available: https://arxiv.org/abs/1807.01052.

[3] "Cohendet_VideoMem_Constructing_Analyzing_ICCV_2019_supplemental.pdf." Accessed: Apr. 29, 2020. [Online]. Available: http://openaccess.thecvf.com/content_ICCV_2019/supplemental/Cohendet_VideoMem_Constructing_Analyzing_ICCV_2019_supplemental.pdf.

[4] "Smeaton et al. - Dublin's Participation in the Predicting Media Mem.pdf." Accessed: Apr. 29, 2020. [Online]. Available: http://ceur-ws.org/Vol-2283/MediaEval_18_paper_14.pdf.