Predicting Video Memorability Using Captions and C3D Features for MediaEval 2018

Alexander Simonin

Dublin City University, Ireland
Alex.simonin2@mail.dcu.ie

ABSTRACT

This paper outlines an approach taken to compute the memorability of short video clips for the 2018 MediaEval Predicting Media Memorability Task. The approach taken herein is based on video captions and C3D aesthetic features associated with videos. These inputs are merged and fed into a Gradient Boosting Regressor model to predict both short-term and long-term memorability.

Key Words

Media Memorability, C3D, Captions, Deep Learning, Gradient Boosting, Decision Trees

1 INTRODUCTION

This paper is concerned with predicting the memorability of short video clips based on aesthetic and descriptive features, for the MediaEval Benchmarking Initiative for Multimedia Evaluation. The application of this work is broad and if properly applied has the potential to be very valuable in fields such as the digital media marketing.

The tangible goal of this project is to automatically predict through deep learning techniques the memorability of a video. Participants in this project are provided with a dataset of 6,000 short video clips. These clips are soundless and have numerous pre-computed visual features attached to them. Each of the videos also has a ground truth associated with it. Once you have trained a model, a further 2,000 test set of videos is provided to test your memorability prediction model against [1].

The approach outlined in this paper is my final product and focuses on taking pre-processed captions and combining these with 3D convulsion (C3D) pre-computed features [2]. This is then fed into the Gradient Boosting Regressor deep learning model [3] to predict both short-term memorability and long-term memorability. The rest of this paper will describe in depth how this achieved and highlight the results produced. This paper should be read in conjunction with my python code.

2 RELATED WORK

I used Eoin Brophy's example solution as a base for extracting and pre-processing my captions as well as calculating my Spearman Coefficient [4]. After researching a number of the previous year's submission papers, the main one I ended up leveraging is – 'GIBIS at MediaEval 2018: Predicting Media Memorability Task' [5]. This paper use C3D among other features to predict video memorability. I used this paper to get ideas as to how to apply C3D features effectively and what model factors to consider.

3 APPROACH

3.1 Motivation for Approach

The main motivation for my combination of features was a result of doing research on past teams that performed well and achieved high Spearman and Pearson coefficient scores. The highest spearman scores achieved tended to only use 2 or 3 feature sets. This meant teams could focus on maximising the value output from each of these feature sets and apply a model that was tailored to fit with custom parameters.

Two teams who participated in the previous year's competition rated each of the pre-computer features with captions scoring highest for both [1]. C3D features was also rated as one of the best features by both teams. C3D is discriminative, compact, and efficient to compute [2]. For this reason I came to the conclusion that a combination of both C3D and captions would yield a strong solution to the challenge. I then decided I would start off with the Keras Sequential modal as used in Eoin Brophy's approach solution [4]. Using this model I achieved quite low results as can be seen in Table 1 below.

I then did a host of research on finding a model that would suit the input and features I was using. This research brought me upon decision tree models and specifically the Gradient Boosting Regressor model.

3.2 Explaining Approach

The final approach I have taken, as mentioned, uses both the videos descriptive captions and a pre-computed C3D features set. The deep learning model I implement is an ensemble decision tree regressor model. A Google Colab notebook is used to run the python code.

3.2.1 Video captions are loaded into a data frame from Google Drive using the Pandas library. A host of pre-processing is then carried out on the captions. This includes counting the occurrences of each of the words and splitting up each caption using python tokenization. Each word is then mapped to an index and the captions are stored in an integer sequence of length 50. The sequences are padded with zeros where necessary. This is stored in a data frame.

3.2.2 The C3D feature set is pulled in similarly to captions. The data frame is again pre-processed to split out each feature value into a separate column. This resulted in 101 feature columns. These columns represent different types of scene, object and action classifications.

3.2.3 Once both of these data frames are created and pre-processed, they are merged using Pandas join, which is then fed into the Gradient Boosting Regressors model as parameter X. This model is an ensemble decision tree regressor model constructed in SciKit Learn. One big advantage of this model is it is very flexible which is important when passing multiple features for training. The model is trained on 12 layers and has 650 n_estimators which equates to 650 decision trees. The models learning rate parameter is set to 0.01. These parameters are true for both short-term and short-term memorability alike. The ground truth scores for each video are fed into the model as Parameter Y.

The model is first trained on 80% (4,800 videos) of the development set and validated against 20% (1,200 videos). This split was used to calculate short-term and long-term spearman's coefficient scores, as seen in table 1 within the results section.

Once this network was developed I train the model once more using the entire 6,000 development set. I then make a prediction against the test set of 2,000 videos. This test set was pre-processed the exact same as the 6,000 development set with the same model parameters being fed in. For further low level detail on my approach, read through my code and corresponding comments.

4 RESULTS AND ANALYSIS

4.1 Results

The evaluation metrics for my results are calculated using Spearman's rank correlation. See table 1. Here I show the spearman scores I achieved and how these scores fare against the sequential model I used initially, per Eoin's example solution [4].

Table 1: Short-term and Long-term Spearman Coefficient Scores by Model

Model (C3D & Captions)	Short-term	Long-term
Gradient Boosting Regressor Model	0.341	0.12
Sequential Model	0.039	0.007

4.3 Analysis

As you can see from the Spearman results the Gradient Boosting Model produces much better scores. This is because the model fares much between when multiple features are fed in.

In all cases the short term memorability of videos is more predictable than long term, since all model predictions score higher in the short term memorability of videos. This can be seen clearly by examining Figures 1 and 2 below. These graphs plot the trained true memorability scores versus the predicted memorability scores for both long term and short term memorability.

Figure 1: Short-term – Predictions vs. Trained.

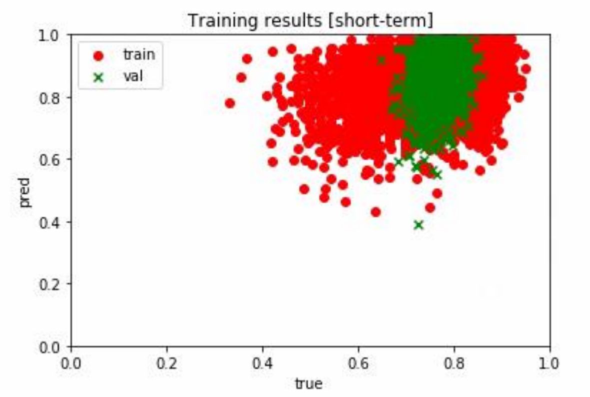
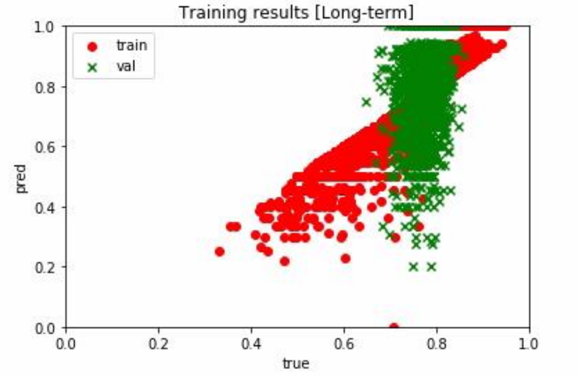


Figure 2: Long-term – Prediction vs. Trained



5 Discussion and Outlook

In conclusion I am quite proud of the results I achieved, however I do see scope for improvement. In hindsight I would have front loaded slightly more as a very large proportion of this project was taken up by trying to learn Python and Machine Learning techniques, having no prior experience in either. I also think a much better prediction could be achieved if a larger data set was used for training. I believe the introduction of the HMP Visual feature could have improved my results through my research of other team's submissions last year.

REFERENCES

[1] Cohendet et al., R.(2018). MediaEval_Problem_Slides. In: *MediaEval 2018*. MediaEval.

[2] Du Tran, M. (2014). *C3D: Generic Features for Video Analysis - Facebook Research*. [online] Facebook Research. Available at: https://research.fb.com/c3d-generic-features-for-video-analysis/ [Accessed 25 Apr. 2019].

[3] Open Data Group. (2018). *Gradient Boosting Regressor*. [online] Available at: https://opendatagroup.github.io/Knowledge%20Center/Tutorials/Gradient%20Boosting%20Regressor/[Accessed 25 Apr. 2019].

[4] Brophy, E. (2018). *Google Colaboratory*. [online] Colab.research.google.com. Available at: https://colab.research.google.com/drive/1X7l5MGrDZa2IdMCOxwgILCD5CzHKE8NF?authuser=1[Accessed 25 Apr. 2019].

[5] Savii et al., R. (2018). [online] Ceur-ws.org. Available at: http://ceur-ws.org/Vol 2283/MediaEval_18_paper_40.pdf [Accessed 25 Apr. 2019].