An Exploration of Ensemble Methods to Predict Video Memorability using Captions

Souradip Goswami (19210273)
MSc. In Computing
Dublin City University
Dublin, Ireland
Souradip.goswami2@mail.dcu.ie

Abstract

Memorability is defined as the quality or state of being easy to remember or worth remembering. With the exponential growth of media content worldwide, there is an urgent need for research on methods for video analysis which takes human cognition into account and keeping this in mind the media memorability task was created. This task uses different image and video features to predict the short-term and long-term memorability scores of video clips. In this paper we use the Caption feature to build two ensemble learning methods namely Stacking and Simple Average to predict the long-term and short-term memorability scores.

Keywords

Captions, Vectorizer, Random forest (RF) regressor, Multi-layer Perceptron (MLP), Recurrent Neural Network (RNN), Simple Average Ensemble, Support Vector Regressor (SVR), Gradient Boosting Regressor (GBR), Stacking Ensemble.

1. Introduction

In this paper we investigate the video memorability scores using caption as the feature which performs the best among all others as revealed by the previous studies. We have used ensemble methods in a quest to build models that can predict the long-term and short-term video memorability scores with good accuracy. The two Ensemble methods that have been used are Stacking and Simple Average.

Based on the results it was observed that the short-term memorability scores outperformed the long-term scores. It was also seen that the Simple Average method using RF Regressor, MLP and RNN outperforms all the others.

2. Related Work

Over the past decade, a large amount of studies has been carried out to predict the video memorability scores of the short video clips. The works of Cohendet [1], Isola et al [2], Shekhar and Singhal [3] suggests that use of captions give the best individual results compared to the other features.
The works of Harshal Chaudhuri [4] presents an approach where he has compared three Neural Network models namely MLP, RNN and Convolutional Neural Network (CNN) using captions as features.
Khaleel [5] worked on captions using SVR and RF and I have extended that using the Stacking Ensemble method with SVR and RF regressor as the base model and GBR as meta model.

3. Data Pre-processing

The Caption feature was used as the predictor for predicting the memorability scores. The Caption dataset consists of a short caption explaining what each and every video is about. The dataset was preprocessed before using the captions for training the models. The captions had stop words along with punctuations which were removed from the dataset. The count vectorizer was used to vectorize the captions and store them in an array and finally generate a captions bag. For the neural network models we have used sequence encoding and one-hot encoding to the captions before fitting the model. The encoding converts the captions to binary formats.

4. Approach

Two ensemble models namely Stacking Ensemble and Simple Average has been used to predict the scores. The building mechanisms of the models have been discussed below.

4.1 Simple Average Ensemble:

The Simple Average is one of the simplest forms of ensemble learning. It gives equal importance to all the models that are being ensembled. The simple average ensemble is a combination of the RF Regressor, MLP and RNN.

4.1.1 Random Forest Regressor:

The RF Regressor used has 200 estimators which signifies the number of decision trees to be used in the algorithm. The random forest builds the decision trees on various sub samples of the dataset and features and uses averaging to improve accuracy. The sub samples are drawn with replacement i.e. boot strap methods.

4.1.2 Multi-layer Perceptron Model:

We have used a 3-layer MLP Model with two hidden layers consisting of 20 neurons each and output layer of 2 neurons each to predict the long-term and short-term scores. The dropout function is added to the MLP model to reduce overfitting and improve generalization error. A Scaled Exponential Linear Unit (SELU) activation function has been used for the two initial layers which provides self-normalization. Sigmoid activation is used as the output layer which works well in cases where the predictions range between 0 and 1. We went with 'adamax' as the activation function and Mean squared error (MSE) as the loss function. We have used 30 epochs where one epoch is when an entire dataset is passed through the neural network once.

4.1.3 Recurrent Neural Network Model:

We have used Long Short Term Memory (LSTM) for the RNN Model. LSTM has feedback connections and works very well for sequences of data like video. The RNN is designed with 150 hidden layers followed by another layer of 30 neurons. For output we went for two neurons each for predicting the long-term and short-term scores. We have also used dropout to avoid over fitting. SELU has been used as the activation function for the initial layers and sigmoid activation for the output layer. We have used adamax as the optimizer function and mean squared error as the loss function.
All these models are trained separately and validated with 1200 video samples to generate the prediction score. A combination of the above models are used for Simple Average.

4.2 Stacking Ensemble:

Stacked Ensemble is a supervised ensembled machine learning algorithm that finds the optimal combination from a collection of prediction algorithms. It involves base learners and a meta learner. The base learners are trained using the training data and fed to the meta learner which generates the final prediction. We have used SVR and RF Regressor as the base learner and fed that to the GBR which is the meta learner.

4.2.1 Support Vector Regressor:

Support Vector Regressor is a non-parametric technique and it relies on the kernel functions to create the hyperplanes in feature space. The kernel can be linear, radial basis function or rbf, poly, sigmoid and others. In our model we have used rbf which works better than the linear model.

4.2.2 Random Forest Regressor:

The RF Regressor is similar to the one used for Simple average ensemble method.

4.2.3 Gradient Boosting Regressor:

The GBR builds an additive model in a forward stage wise fashion and in each stage a regression tree is fitted on the gradient of the loss function. The main aim is to reduce the loss function by adding weak learners using a gradient descent. The number of estimators used are 200 which indicates the number of boosting stages to perform. The learning rate is kept quite small which ensures a good performance.

4.2.4 StackingCV Regressor:

The StackingCV Regressor extends the stacking algorithm using out-of-fold predictions to prepare the input data for level 2 regressor. In the standard stacking procedure, the first level regressors are fit to the same training set that is used to prepare the inputs for the second level regressor and this may lead to overfitting. The StackingCV Regressor, however, uses the concept of out-of-fold predictions to avoid this overfitting. We have used the parameter 'cv' which is the cross-validation generator that guides the splitting strategy. The value of the cv signifies the number of folds in K-fold. We have used the StackingCV Regressor to perform the stacking ensemble using SVR and RF regressor as the base model and GBR as the meta model. The model is being trained with 80 percent of the dev set i.e. 4800 video samples out of 6000.

We will be looking into the results to see how each models fare in terms of the scores and whether the ensemble models outperform the individual models in prediction accuracy.

5. Results and Analysis

The individual models as well the Ensemble models have been validated with 1200 samples (20 percent of the dev set). Let us look into the validation results and analyse how each model predict the short-term and long-term score.

The Spearman's Coefficient calculation method is used to calculate the memorability scores. Below we show the scores of each individual model as well as the ensemble models.

Random Forest Regressor:

Model Name	Feature	Short-term Score	Long-term Score
Random Forest	Caption	0.414	0.179

Multi-layer Perceptron Model:

Model Name	Feature	Short-term Score	Long-term Score
MLP Model	Caption	0.401	0.211

Recurrent Neural Network Model:

Model Name	Feature	Short-term Score	Long-term Score
RNN	Caption	0.369	0.189

Support Vector Regressor:

Model Name	Feature	Short-term Score	Long-term Score
SVR	Caption	0.419	0.179

Simple Average Ensemble:

Model Name	Feature	Short-term Score	Long-term Score
Simple Average	Caption	0.45	0.213

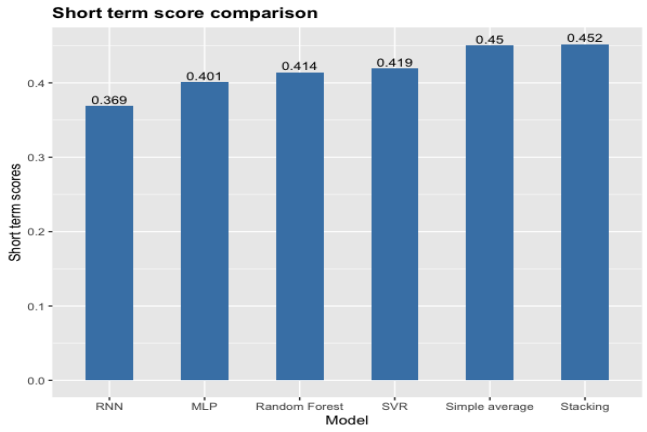
Stacking Ensemble Model:

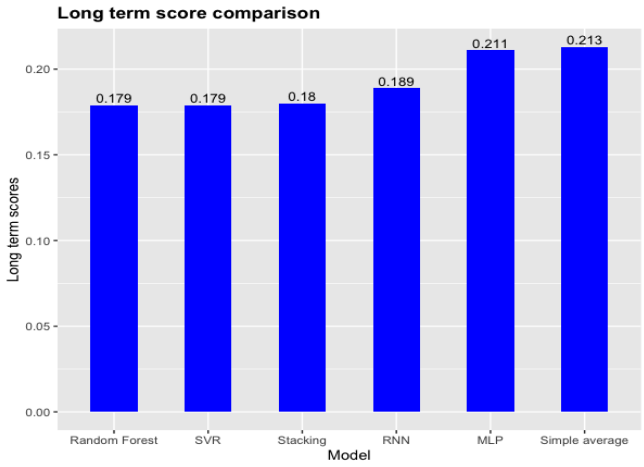
Model Name	Feature	Short-term Score	Long-term Score
Stacking	Caption	0.452	0.18

From the results it is quite evident that the Simple Average model performs better than the individual models as well as the Stacking Ensemble. The results also suggest that short-term scores are better than the long-term ones for all the models.

The Stacking method can be improved by incorporating a number of base models and tuning the hyper-parameters properly. Although based on the fact that we have used a single feature and very few base models, Stacking gave quite good results and it might be a way to move forward.

The comparison between the model scores have been plotted below:





The graph for short-term score is suggestive of the fact that both the ensemble models perform better than the base models. However, for the long-term scores Simple Average performs the best. So, we have used the Simple Average model to generate scores for the test dataset. The average short-term and long-term scores have been recorded below:

Short-term Average	Long-term Average
0.85630913	0.76549109

6. Conclusion and Future Work

In this paper we have tried to dig into ensemble models like Simple Average and Stacking. The Simple Average performs the best among all the models. However, the Stacking Ensemble can be improved highly by incorporating a few more models and adding some more features. As a part of future work, the Stacking Ensemble can be a very good approach as it provides good scores in spite of the fact that we have used limited base models and just a single feature. Moreover, other ensemble learnings like bagging and boosting can also be applied to check whether the performance improves.

7. References

[1] Cohendet, R., Yadati, K., Duong, N. Q. K., and Demarty, C. (2018). Annotating, understanding, and predicting long-term videomemorability. In Aizawa, K., Lew, M. S., and Satoh, S., editors,Proceed-ings of the 2018 ACM on International Conference on Multimedia Retrieval,ICMR 2018, Yokohama, Japan, June 11-14, 2018, pages 178–186. ACM.

[2] Isola, P., Xiao, J., Parikh, D., Torralba, A., and Oliva, A.(2014). What makes a photograph memorable? IEEE Trans. Pattern Anal.Mach. Intell., 36(7):1469–1482

[3] Shekhar, S., Singal, D., Singh, H., Kedia, M., and Shetty,A. (2017). Show and recall: Learning what makes videos memorable. In2017 IEEE International Conference on Computer Vision Workshops, ICCVWorkshops 2017, Venice, Italy, October 22-29, 2017, pages 2730–2739. IEEEComputer Society

[4] https://github.com/harshalchaudhari35/MediaEval-Media-Memorability

[5] https://github.com/Khalees2/Video-Memorability-Predection-using-Machine-Learning