Ensemble Approach to Predict the Media Memorability for MediaEval 2018

Gautam Shanbhag[1]
[1]Dublin City University, Ireland
gautam.shanbhag2@mail.dcu.ie

ABSTRACT

Predicting Media Memorability Task focuses on the problem of predicting how memorable a video will be [1]. This paper describes the approach developed to predict the short-term and long-term video memorability for the 2018 MediaEval Benchmarking Initiative for Multimedia Evaluation. Predictions are scored between 0-1 where closer to 1 refers to videos being more memorable. The approach taken for this research is simple and understandable as it is based on finalising models, by comparing 3 algorithms shortlisted based on prior domain knowledge (text, aesthetic feature, video feature) subsequently selecting the one that suits our dataset, finally ensemble model is used to retrieve scores by using a weighted average approach.

1 INTRODUCTION

Video memorability (VM) is an important field of research in computer vision. The huge amount of media data in various fields like education, social networking, advertisement and public relations have given rise to the problem of video memorability. The main issue with this is that there are no standards or benchmark which can be used to evaluate VM, another problem is the lack of enough data and tools that can be used to perform the same [3].

In this paper, I have started my research by going through previous work in this domain to identify features which provide the best results possible with the use of various visual, textual and semantic features to predict video memorability. Among the features provided, I have trained my model by shortlisting captions as the prominent feature subsequently, I proceeded with training my model on aesthetic feature which is derived by taking the median and C3D features which as per experiments and studies have provided the promising outcome amongst the other visual features provided.

An Ensemble model is created by using weighted average technique across the 3 features to extract a better prediction than individual probabilities of short term and long term memorability.
By getting predictions using Linear Regression to determine the base mark predictions, I have tried out other different models to find a better model.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

My key findings are:

(1) Models for short term memorability perform much better than models for long term memorability.
(2) Common Model does not fit Short term predictions & Long term predictions
(3) Models based on the video captions provided outperform models trained on other visual features.

2 RELATED WORK

Feature Selection plays an important role for model outcome. Given a bag of features to use such as captions, aesthetic (mean & median), visual features such as C3D, color histogram, HMP, HOG, Inception, LBP, ORB and combining all those did not actually make sense due to high correlation between them as well a high dimensionality as it would introduce the problem of overfitting. Models based on captions, C3D out performed other feature selections as seen by Rohit et.al. [2]. Results compared by Tanmayee Joshi et.al [4] have derived text features by using TfIdf vectoriser by 100 dimensional limit. Ensemble approach resulted in better results for Smeaton et.al [5] where the team used linear model to ensemble on subset of the training data to identify which were the most important predictors.

3 APPROACH

In this section, I will discuss the task of predicting Video Memorability. Feature engineering played a major role for my outcome predictions. Section 3.1 focuses on feature engineering related to captions. Section 3.2 visual and aesthetic feature selection and talks more about the model selection.

Section 3.1 Caption Based Feature Engineering

Caption provided the best result for short term and long term memorability predictions. Captions can be a compact source of representing the video content, and thus is useful for predictions.
ML models expects caption to be vectorized before feeding as words cannot be processed in raw form. Captions had to be cleaned and extracted to form a corpus to pass it to a model.
I started with removing special characters present in caption followed by converting the captions to lower case. Later captions were tokenized from the review to extract individual word. I used Porter Stemmer on words extracted from tokenizing. And lastly before vectorising converted it back to a single sentence to be passed to vectorizer.

TFIDF Vectoriser has been used with max feature count set to 1500 and followed by n-gram approach (specifically uni-gram and bi-gram).

caption	corpus
blonde-woman-is-massaged-tilt-down	blond woman massag tilt
roulette-table-spinning-with-ball-in-closeup-shot	roulett tabl spin ball closeup shot
khr-gangsters	khr gangster
medical-helicopter-hovers-at-airport	medic helicopt hover airport
couple-relaxing-on-picnic-crane-shot	coupl relax picnic crane shot

Fig 1. Corpus extracted after Cleaning on Captions

Section 3.2 Model Selection

Model selection proved to be challenging and interesting. My approach to short list models was based on understanding of the model working and selection of models based on prior work done by others. Keeping the readings of Logistic Model as my base model for comparison, for Captions I went ahead with comparing Linear Regression, SVR & Bayesian Ridge.

Similarly Aesthetic, Linear Regression, Decision Tree Regressor, Bayesian Ridge. C3D was tested on Linear Regression, Lasso Linear Regression by normalizing the dataset and setting cross validation to 5 & Random Forest by using estimator count of 1500 and max feature set to square root of count.

Section 3.3 Ensemble

Ensemble is basically a weighted average technique to fetch optimum predictions by weighting features which contribute towards model. Adding prominent weights to intermediate model output which contribute significantly towards better prediction increases the overall predication percentage of the model.

From my analysis, I figured out that for short term memorability captions dominated the model. Out of the models selected Bayesian Ridge outperformed the others for Captions and Aesthetic and Random Forest worked best with C3D

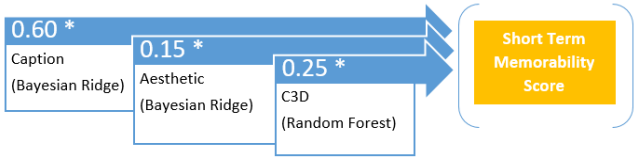

Fig 2. Ensemble Approach for Short Term Memorability

For Long Term memorability, Captions and C3D gave similar performance as compared to Aesthetic which performed less accurately thus I have added weights equally to both followed by Aesthetic to be 1/5 of the entire model. As per my research Bayesian dominated in Captions, followed by Random Forest for C3D similar to short term. One thing to note was Linear Regression worked better for long term predictions when compared against short term. Thus ensemble of these 3 for long term yielded better results for my research.

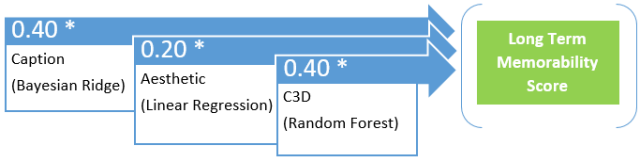

Fig 3. Ensemble Approach for Long Term Memorability

4 RESULTS AND ANALYSIS

Tables 1 and 2 give an overall synopsis of my exploratory outcomes. Individual reading are provided taken from K-Fold approach to evaluate training and test split to capture overall accuracy of model. I have used Spearman correlation matrix to determine the overall model accuracy. Finally Ensemble of top performing models in each feature was short listed and combined for final accuracy of model.

(1) Short Term Ensemble: (0.60 * Caption score) + (0.15 * Aesthetic Score) + (0.25 * C3D score)

(2) Long Term Ensemble: (0.40 * Caption score) + (0.20 * Aesthetic Score) + (0.40 * C3D score)

SHORT TERM MEMORABILITY		
Feature	**Model**	**Spearman**
Caption	Linear Regression	0.356
	SVR	0.303
	Bayesian Ridge	0.417
Aesthetic	Linear Regression	0.237
	Decision Tree Regressor	0.150
	Bayesian Ridge	0.248
C3D	Linear Regression	0.251
	Lasso Linear Regression	0.259
	Random Forest	0.318
Ensemble	**Bayesian Ridge**	**0.439**
	Bayesian Ridge	
	Random Forest	

Fig 4. Short Term Memorability Using Spearman Correlation

LONG TERM MEMORABILITY		
Feature	**Model**	**Spearman**
Caption	Linear Regression	0.130
	SVR	0.112
	Bayesian Ridge	0.154
Aesthetic	Linear Regression	0.088
	Decision Tree Regressor	0.028
	Bayesian Ridge	0.062
C3D	Linear Regression	0.074
	Lasso Linear Regression	0.066
	Random Forest	0.113
Ensemble	**Bayesian Ridge**	**0.174**
	Linear Regression	
	Random Forest	

Fig 5. Long Term Memorability Using Spearman Correlation

5 CONCLUSIONS

This paper presents the ensemble model for predicting media memorability. I have used textual features such as captions alongside visual features and aesthetic features of images taken from frames. The results show that better results are obtained by combining visual, textual and aesthetic features. On the other hand, short term and long term scores need to be computed separately as common model does not hold true for both. In future, I plan the explore effects of other visual and aesthetic features by trying out different combinations. Secondly, I aim to explore more sophisticated methods to improve prediction performance.

ACKNOWLEDGMENTS

This work is supported by Prof. Tomas Ward as part of Data Analytics curriculum where the competition helped me learn, understand and tackle such problems with ease. Supporting and enhancing my learning abilities was a great take away from this project.

REFERENCES

[1] http://www.multimediaeval.org/mediaeval2018/memorability/
[2] Gupta, R. and Motwani, K., Linear Models for Video Memorability Prediction Using Visual and Semantic Features.
[3] 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.
[4] P. Bahl, R. Chancre, and J. Dungeon. 2004. SSCH: Slotted Seeded Channel Hopping for Capacity Improvement in IEEE 802.11 Ad-Hoc Wireless Networks. In Proceeding of the 10th International Conference on Mobile Computing and Networking (MobiCom'04). ACM, New York, NY, 112–117.
[5] Smeaton, A.F., Corrigan, O., Dockree, P., Gurrin, C., Healy, G., Hu, F., McGuinness, K., Mohedano, E. and Ward, T.E., 2018. Dublin's participation in the predicting media memorability task at MediaEval 2018.