Predicting Media Memorability Using Extracted Features.

Rohit Nair
School of Computing
Dublin City University
Dublin, Ireland
email@email.com

ABSTRACT

We see a lot of images and videos on multiple platforms, but we do not remember all of them and not all aspects of them. The goal of the project is to use these extracted features to predict the memorability of each video for short and long term. Videos consist of multiple image frames that are shown in succession which result in us seeing a continuous video. The features associated with these videos are available as a part of the Media Memorability Task at MediaEval. The ground truth of each video being memorable in the long and short term is available. The extracted features are being evaluated individually and then these features are being merged to generate the final output.

KEYWORDS

• XGBoost • Random Forest • Extra Trees • Media Memorability • Extra Trees

1 INTRODUCTION

Images and videos contain a lot of information that are extracted as features. It can be analysed to gain insights from the data. With the amount of data available online, the information at our disposal is endless.

Memorability is defined as the quality or state of being easy to remember (Gupta & Motwani, 2018). Certain videos tend to be remembered while others are not. The tendency of any media to be remembered is known as its memorability or media memorability. This project explores different models and features to predict the short and long-term memorability of the videos in the data set provided. The data set being used consists of features that have already been extracted and the source videos files are also available. Captions, a text feature, cannot be used to predict the memorability of a video. The models are trained on HMP and Colour Histogram; which are visual features, and C3D prediction semantic features. The features were first trained and tested individually and then merged for the final prediction, the results of this have been included in the result section of the paper.

2 RELATED WORK

There are papers and projects that have come up with their own solution for this problem statement. In their work (Gupta & Motwani, 2018) combine features and create an ensemble model using weights for each ensemble. They have compared Linear Support Vector Regression, LASSO and ElasticNet. For LBP, HMP and ColorHistogram, Lasso Regression worked the best while ElasticNet gave better results for the rest of the features.

In a different paper, the authors (Goswami, Bhardwaj, Singh and Vatsa, 2014) have used memorable frames for face recognition. The proposed methodology uses deep learning architectures in unison with memorability scores of face images to pick out frames that are relevant and while ignoring the rest.

Khosla, Raju, Torralba and Oliva have used Convoluted Neural Networks to fine tune and predict memorability (Khosla, Raju, Torralba and Oliva, 2015). They have also created an annotated data set that have 60000 images from different sources. Their model managed to achieve a raked correlation of 0.64 which is close to human observers.

Isola, Xiao, Torralba and Oliva, in their research, explore different factors that affect the memorability of images (Isola, Xiao, Torralba and Oliva, 2011). They consider factors like repetition, object types, placement, background etc.

In a paper submitted in MediaEval 2018, the authors have extracted features from the videos, fen them into Inception-v3 convolution network to extract 2048 dimensional features (Tran-Van et al., 2018). This output becomes the input for the recurrent neural network.

3 APPROACH

3.1 Data Set, Features and Pre-Processing

We have been provided with a data set of features for 8000 videos which are split into training set and testing set, 6000 and 2000 records, respectively. The ground truth has been provided for the training data set while the ground truth for the testing data set must be generated by the model.

There are a total of 9 features extracted, using the videos made available by MediaEval. Out of these 9, captions were not to be used. The features used for this project are C3D, HMP and Colour Histogram.

For HMP and C3D, we read each file to create a vector for it. We add these vectors to a matrix to create a data frame such that the video file name becomes the name of the row.

For colour histogram, each file has RGB values ranging from 0-225. There are 3 such files for each video because they are 3 frames for each video. We fetch the RGB values from each file to create a vector. The 3 files for each video, is split using the name and then merged to create a complete record for that video. This

process is repeated for all videos and then a final data frame is created by merging these vectors.

3.2 Models

Each feature was used to train different models individually. The data set was split in an 80:20 ratio, 80 for training and 20 for validation. Using the ground truth provided, we train the models and test them with the validation data. The algorithms used for testing are Support Vector Regressor, Extreme Gradient Boost, KNN, Random Forest, Adaptive Boost, Extra Trees Regressor and Gradient Boost. Each feature was used to test the models separately first.

Then, based on the value of Spearman's correlation coefficient, we choose the models that will be used in this step. The features are merged to create a data frame of combined features which are again split into 80:20 ration, trained and tested using the ground truth table. The result of the predictions are in Table 4.

Finally, a data frame is created by merging the 3 selected features from the entire dev-set provided (all 6000 records), we train the model without splitting it. We use this model to populate values in the ground truth table for test data and create a CSV file for the same.

4 RESULTS AND EVALUATION

The values for spearman's correlation coefficient for each algorithm with the validation data set, from highest to lowest, when the features were not combined are as follows:

Table 1: Spearman's Correlation Coefficient for C3D

C3D		
	Short	**Long**
Random Forest	0.304	0.134
Gradient Boost	0.274	0.137
XGBoost	0.274	0.136
Extra Trees	0.274	0.115
SVR	0.262	0.086
Adaptive Boosting	0.252	0.106
KNN	0.117	0.073

Table 2: Spearman's Correlation Coefficient for Colour Histogram

Colour Histogram		
	Short	**Long**
Extra Trees	0.257	0.062
Random Forest	0.274	0.137
XGBoost	0.274	0.136
Gradient Boost	0.274	0.115
SVR	0.117	0.092
Adaptive Boost	0.154	0.042
KNN	0.148	0.093

.

Table 3: Spearman's Correlation Coefficient for HMP

HMP		
	Short	**Long**
Random Forest	0.280	0.138
Extra Trees	0.263	0.110
XGBoost	0.261	0.113
Gradient Boost	0.250	0.118
SVR	0.246	0.097
Adaptive Boost	0.194	0.041
KNN	0.118	0.090

After training the features individually, the features were merged, models were trained on these merged features and validate using the validation data set. The data set's size increases when you combine the features, so the top 5 algorithms were chosen to be tested after combining the features. The result of that is as follows:

Table 4: Spearman's Correlation Coefficient after combining features

Merged Features		
	Short	**Long**
Random Forest	0.372	0.180
Extra Trees	0.372	0.148
XGBoost	0.349	0.143
Gradient Boost	0.348	0.15
SVR	0.311	0.146

From Table 4, we can say that Random forest and Extra trees are almost similar when it comes to the short-term memorability scores. For long term memorability, random forest outperforms extra trees by a small margin. But during the multiple iterations, Random Forest took a long time to train compared to any other model.

4 CONCLUSION AND FUTURE WORK

Out of all the algorithms used, random forest and Extra Trees worked the best. Compared to random forest, the training time required for extra trees regressor was far less. That is why, to create the final output Extra Tree Regressor was used. The key findings from this project are as follows:

- Random Forest performs best overall but tends to take a lot of time for training.
- Random Forest and Extra Tree Regressor perform almost the same for short term memorability.
- KNN, consistently performed bad for the three selected features.
- Combining features helped in increasing the prediction score by a small margin.

An attempt was made to combine these models, but the resultant model performed worse than the individual models itself. For future work, these models can be combined and the performance can be improved with weighted averages.

REFERENCES

[1] Gupta, R., & Motwani, K. (2018). Linear Models for Video Memorability Prediction Using Visual and Semantic Features. *MediaEval*.

[2] Goswami, G., Bhardwaj, R., Singh, R., & Vatsa, M. (2014). MDLFace: Memorability augmented deep learning for video face recognition. *IEEE International Joint Conference On Biometrics*. doi: 10.1109/btas.2014.6996299

[3] Khosla, A., Raju, A., Torralba, A., & Oliva, A. (2015). Understanding and Predicting Image Memorability at a Large Scale. *2015 IEEE International Conference On Computer Vision (ICCV)*. doi: 10.1109/iccv.2015.275

[4] Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? *CVPR 2011*. doi: 10.1109/cvpr.2011.5995721

[5] Tran-Van, D., Tran, L., & Tran, M. (2018). Predicting Media Memorability Using Deep Features and Recurrent Network. *MediaEval, 2283*. http://ceur-ws.org/Vol-2283/MediaEval_18_paper_24.pdf.

[6] Chaudhry, R., Kilaru, M., & Shekhar, S. (2018). Show and Recall @ MediaEval 2018 ViMemNet: Predicting Video Memorability. *MediaEval, 2283*. http://ceur-ws.org/Vol-2283/MediaEval_18_paper_15.pdf.