Predicting Media Memorability using various Machine Learning models and features.

Bidish Basu – 19210427

MSc. in Computing

Dublin City University

bidish.basu2@mail.dcu.ie

***Abstract -* Memorability of videos plays an important role in today’s world where millions of videos are generated on a daily basis. The MediaEval media memorability[1] task aims at predicting the memorability of videos using various machine learning models. The machine learning models are trained on various features and the results are then tabulated. The best combination of the model and the feature is then selected to predict the short-term and long-term memorability scores.**

*Keywords - Media Memorability, Multilayer Perceptron, Decision Tree Regressor, Bagging Regressor, Linear Regression, Captions, C3D, Aesthetic Features, Spearman’s Correlation Coefficient, Hashing Vectorizer, NLTK.*

I . INTRODUCTION

The task of video memorability involves how successfully can a video be remembered, both in short and long term intervals. To aid the task, the ground truth values of multiple features such as Captions, C3d,HMP, Inception V3 etc. were provided. The ground truth contains the values of both the short-term and the long-term memorability scores of 6000 videos. For the prediction, 2000 more videos along with their features were provided. The evaluation metric used is the Spearman’s Correlation Coefficient.

The approach that is used to predict the long and the short term memorability scores focuses mainly on using multiple features i.e., Aesthetic Features, Captions and C3D and training them using multiple machine learning models such as Multilayer Perceptron, Bagging Regressor with the base estimator as Decision Tree Regressor and Linear Regression models and finding the best combination of the model and the features which is then finally used to predict the long term and the short term memorability scores of the videos.

It was observed that the combination of Multilayer Perceptron along with Captions gave the best long-term Spearman’s Correlation Coefficient values and a combination of Bagging Regressor with Captions gave the best short-term Spearman’s Correlation Coefficient values. The long-term predictions are comparable for both the models but there is a large difference in the short-term scores. Therefore, the combination of Bagging Regressor with the base estimator as Decision Tree Regressor along with Captions as the feature was then used to predict the scores. The short-term scores predicted were better than the long-term scores and due to the limited amount of observations, the captions proved to be better than the other features like the Aesthetic features and the C3D which were also tested.

II. LITERATURE REVIEW

There has been a substantial amount of work done in the field of video memorability. In the paper titled “Linear Models for Video Memorability Prediction Using Visual and Semantic Features”[2], Rohit Gupta and Kush Motwani, investigate different semantic and visual features to predict the memorability of the videos. As part of their analysis, they were able to uncover many intrinsic semantic features that affect the memorability of the videos. For their approach, they used LASSO(L1) regularized Logistic Regression, Linear Support Vector Regression and ElasticNet to train and predict the scores.

Another approach as proposed in “VideoMem : Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability”[3] uses various deep neural network-based models for the prediction of long-term and short-term memorability.

III APPROACH

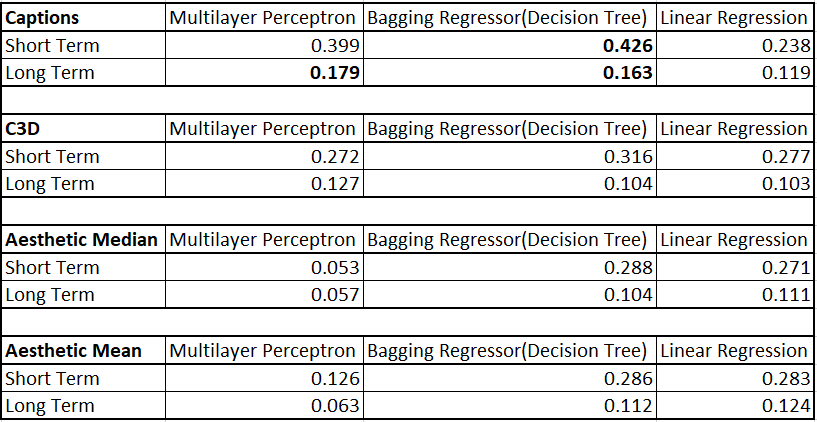
*A. Features*

In order to find the long-term and short-term memorability scores, multiple features were tested to begin with. The features used were, Aesthetic Features, Captions and C3D.

*B. Data Preprocessing*

Before the models were trained, the data was preprocessed to get an optimal result. The captions were preprocessed using the NLTK libraries. Firstly, all the characters were converted to lowercase and the punctuations were replaced with spaces. Next, the captions were tokenized into individual words at the same time, the non-alphabetic characters were filtered out. Once tokenization was performed, all the related words were then stemmed as this helped with increasing the accuracy of the final predictions. Next, the common English words that do not have any likely benefits in Natural Language Processing were removed. The words were then finally rejoined into a single line.[4] The captions were then converted into a collection of token occurrences using Hashing Vectorizer. A main advantage of the Hashing Vectorizer is that it features low memory scalability when working with datasets which in turn means that there isn’t a necessity to store a vocabulary dictionary in the memory, this in turn results in improving accuracy.

Table 1 : Spearman’s Rank Score



For performing analysis on C3D, Aesthetic features, aggregated both by mean and median, the features of each video is first loaded onto a data frame and then, the feature values are then split into individual elements which are then used for the analysis.

*C Machine Learning Models*

The machine learning models used for training and the analysis were, Multilayer Perceptron, Bagging Regressor with the base estimator as Decision Tree Regressor and Linear Regression. For all the features, to train the various models, the training dataset is split into 80% for training and 20% for testing and validation.

The first model that was tested was the Multilayer Perceptron. In order to prevent over fitting, Ridge Regularization has been used. Furthermore, a dropout layer has been added to reduce the overfitting of the models. To compile the model, “Adadelta” is used as the optimizer and loss is measured using “mse”. An early stopping monitor is also defined in order to address the overfitting and under fitting issue, as too many epochs can lead to overfitting the training dataset and too less can lead to an underfitted model.[5]

The features were then tested with Bagging Regressor with the base estimator as Decision Tree Regressor. A Bagging Regressor, takes random subsets of the original dataset and fits the base regressor, in this case a Decision Tree Regressor on each of these random subsets and then aggregates each of the predictions to form the final prediction[6].In order to further reduce the overfitting in the Decision Tree Regressor, a random splitter was used. This is because, it selects a set of features and splits it randomly, and adding to that, there is no additional overhead to compute the best split.[7]

Finally, Linear Regression was performed on the features. Linear Regression is a simple model which was used as the features are of higher dimensions, this in turn means that there are higher chances of multicollinearity. This means that there can be more chances of overfitting.

IV RESULTS AND ANALYSIS

The results were calculated using the Spearman’s rank Correlation Coefficient. From the table below, we see that the combination of Multilayer Perceptron along with Captions gave the best long-term Spearman’s Correlation Coefficient values and a combination of Bagging Regressor with Captions gave the best short-term Spearman’s Correlation Coefficient values.

Along with this, it is also needs to be noted that the short-term memorability scores are far better than the long-term memorability scores across all models and the features. It must be noted that other features namely, C3D and Aesthetic Features(both measured in mean and median) do not perform as well as the captions across all the three models.

For Captions, long-term predictions are comparable for both Multilayer Perceptron and Bagging Regressor but there is a large difference in the short-term scores. Therefore, for the final predictions, the combination of Bagging Regressor with the base estimator as Decision Tree Regressor along with Captions was then used to predict the scores. The model is trained on the complete dev-set that has 6000 instances and predictions are done on the test-set that has 2000 instances.

V CONCLUSION

I believe that there is a scope of improvement, approaches can be used to combine the results from multiple models to improve the final score. In addition to this I believe that having a bigger dataset could have helped improve the performance of the models selected as well. In future works, I would combine multiple features together and feed that to the models to generate better results.

REFERENCES

1. Cohendet, Romain, Claire-Hélène Demarty, Ngoc Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. "Mediaeval 2018: Predicting media memorability task." arXiv preprint arXiv:1807.01052 (2018).
2. Gupta, R. and Motwani, K., 2018. Linear Models for Video Memorability Prediction Using Visual and Semantic Features. In MediaEval.
3. Cohendet, Romain, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engilberge. "VideoMem: Constructing, Analyzing, Predicting Short-term and Long-term Video Memorability." In Proceedings of the IEEE International Conference on Computer Vision, pp. 2531-2540. 2019.
4. Allen, Michael. “101: Pre-Processing Data: Tokenization, Stemming, and Removal of Stop Words,” December 14, 2018. Available at: <https://pythonhealthcare.org/2018/12/14/101-pre-processing-data-tokenization-stemming-and-removal-of-stop-words/>.
5. Brownlee, Jason. “How to Develop Multilayer Perceptron Models for Time Series Forecasting,” November 9, 2018. Available at: <https://machinelearningmastery.com/how-to-develop-multilayer-perceptron-models-for-time-series-forecasting/>.
6. “Bagging Regressor,” Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingRegressor.html>.
7. Mithrakumar, Mukesh. “How to Tune a Decision Tree?,” November 11, 2019. Available at: <https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680>.