Regression Models for Video Memorability Prediction Using Visual and Text Features

Rohit Shrinivas Toshniwal
Dublin City University
Dublin,Ireland
rohit.toshniwal2@mail.dcu.ie
Student Id: 19211138

ABSTRACT

Memorability of a video or an image has been a significant research topic in the area of computer vision. Visual Memorability may be seen extensively in areas such as curriculum, marketing, sponsorship, and many more. This paper aims to forecast the long-term and short-term memorability of a video utilizing various visual features as well as text-based features. This article utilizes the Random forest model approach in a mixed way utilizing the weighted average. We will be predicting the probability that the video will be remembered based on a given set of pre-computed features. The model is tested using a spearman rank correlation to see how well the test data can be interpreted in our model.

1. INTRODUCTION

Throughout our fast-paced environment, communications channels such as social networking networks and apps, television advertisement companies, knowledge collection and suggestion systems will cope with details relating to communications on an unprecedented level day after day. This allows the need to comprehend this information to be of the utmost importance for these distribution structures to allow them to maximize their delivery (Squalli-Houssaini, H et al, 2018).

Given its potential to be an important field of research in the computer vision world, Visual prediction suffers from a number of hurdles, there is no unmistakable rule or thumb rule for determining video memorability. Any system that can chip away at a dataset will most likely be unable to run on another dataset, rendering it a profoundly erratic field (Cohendet, R et al, 2018).

In this paper, I have analysed different video features and text features to predict video memorability, detailed research is performed to select the right features, and an ensemble method is used to construct a robust model. Among the features offered by Media Eval, I used captions, HMP features and C3D for the training my model. Models are analysed using the Spearman rank correlation as a measure. The findings suggest that it is somewhat harder to estimate long-term memorability than short-term memorability.

2. RELATED WORK

The dataset in which I developed my model was part of the Media Eval2018 Predicting Media Memorability competition. Several organizations around the globe have engaged in this initiative. I have referred a variety of articles from the teams that have worked in this challenge. Rohit Gupta and Kush Motwani [2] from Conduent Labs, India Three models were used in their study, namely: LASSO (L1) Regularized Logistic Regression [3], Linear Support Vector Regression [4] and ElasticNet (L1 and L2 Regularized Linear Regression) [5]. For-range of specifications, these versions were evaluated and the better picked. In their research, HMP and C3D features are used as they are, whereas LBP and Color Histogram are concatenated through sets.

In their study, Alan F. Smeaton et al. [6] carried out six techniques which were focused on various features. In their first step, the pre-computed features given by the organizers were combined into a single vector and the neural network was introduced. In the second method, three keyframes of videos were used as inputs and at each epoch, one frame was randomly chosen as a type of data increase. Image and video saliency models that produce regions that draw the most human interest have been used in their third method. The fourth method concerned a real human being, where they analyzed a person's reactions by displaying the middle frame of each film, and their EEG (Electroencephalography) signals were captured utilizing a neural methodology. They used the visual aesthetics classifier on the picture attributes in their fifth solution. Eventually, in the last step, the effects of all the strategies described above were combined and a linear model was educated. The findings of their study have shown that the Image Saliency Method has provided great output and the Neuronal Method has generated poor outcomes.

3. DATASET

In this analysis, a production dataset of 6000 rows was given, where each row correlates to a video along with short-term memorability scores, long-term memorability scores, and the amount of annotations representing the number of people interested in the project. 80% of the dev-set was divided to train our model, and the rest was used for testing. Recently, a study dataset feature has been published, enabling us to predict both short-term and long-term memorability ratings.

4. APPROACH

4.1 Support Vector Regression

The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real

Dublin City University, Rohit Toshniwal, Ireland

number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

4.2 Bayesian linear regression

The Bayesian linear regression model produces straight relapse by using probability disseminations instead of point gauges (Koehrsen, 2019). The accompanying technique is utilized for the Bayesian straight relapse model with the response extricated from the normal dispersion.

$$y \sim N(\beta^T X, \sigma^2 I)$$

The exhibition, y, is produced by a typical (Gaussian) conveyance characterized by mean and variance. The incentive for direct relapse is the transposal of the weight variable aggravated by the indicator work. The difference is the opposite of the standard deviation (increased by the Identity Matrix since it is a multi-dimensional capacity of the model) (Koehrsen, 2019). The objective of the Bayesian Linear Regression isn't to find the "most highest" estimation of the model parameters, however rather to assess the posterior distribution of the model parameters (Koehrsen, 2019).

The posterior likelihood of the model parameters depends on the inputs and outputs of the training:

$$P(\beta|y, X) = \frac{P(y|\beta, X) * P(\beta|X)}{P(y|X)}$$

Here, P(β, X) is the back likelihood appropriation of the model parameters subject to data sources and yields. It is comparable to the probability of the outcomes, P(y, X), duplicated by the earlier probability of the parameters and isolated by the constant of normalization (Koehrsen, 2019).

4.3 Random Forest Regression

A Random Forest is an ensemble methodology capable of providing both regression and classification activities utilizing several decision trees and a method named Bootstrap Aggregation, widely known as bagging. Bagging, in the Random Forest Process, requires the preparation of each decision tree on a separate data sample where sampling is performed with substitution.

4.4 Models

On the basis of study, I chose to create a layout focused on captions, C3D and HMP features. All of these features have a high dimensionality, which makes it necessary for us to use a regularization technique to eliminate any noise or distortion in our data that may contribute to a data overlay. Using Bayesian Ridge Regressor, Support Vector Regressor and Random Forest Regressor, this allows one to implement regularized models. Memorability attributes, in a simplistic way, are the proportion of viewers who recall the video from the amount of characters seen in the picture. In a sense, it determines an individual's capacity to remember the video. I built these three models on following features:

1. Captions
2. C3D
3. HMP
4. Combination of Captions and C3D
5. Combination of Captions, C3D and HMP

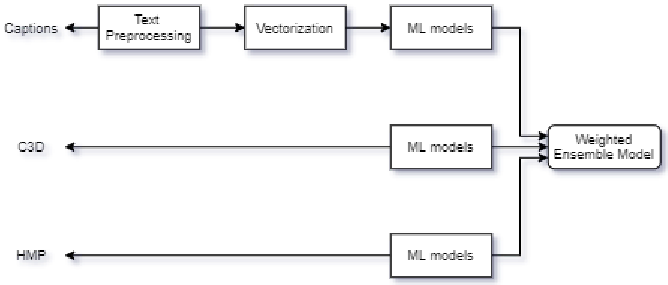

Figure 1. Model Flow Chart

5. RESULTS AND ANALYSIS

We have used the following features: Captions, C3D and HMP for this work. Random Forest (RF), Support Vector Regressor (SVR) and Bayesian Ridge Regressor(BRR) Machine Learning models from Scikit-learn [9] were used to predict the memorability scores. In the first approach, only Captions were used in which each caption was converted into Bag of words using CountVectorizer from Scikitlearn. In the second approach only C3D features were used. In the third approach Captions and C3D features were flattened into one single vector similarly Captions, C3D and HMP features were also combined and used. Spearman's rank correlation coefficient [2]. The following table 1 and table 2 shows the accuracy scores obtained in each approach with Random Forest SVR and BRR models. The optimal weights for our model can be defined using the following formula:

1. Short-term= 0.6*captions+0.1*HMP+0.3*C3D
2. Long-term = 0.5*captions+0.2*HMP+0.3*C3D

	Random Forest		**BRR**		**SVR**	
Features	**Short-Term**	**Long-Term**	**Short-Term**	**Long-Term**	**Short-Term**	**Long-Term**
Captions	**0.404**	0.171	0.338	0.17	0.355	0.171
C3D	0.266	0.122	0.286	0.126	0.242	0.107
Captions_C3D	0.321	0.104	0.373	0.228	0.355	**0.179**
Captions_C3D_HMP	0.342	0.127	0.373	0.228	0.355	**0.179**

Table 1. Spearman's coefficient Results

Model	**Short-Term**	**Long-Term**
Ensembled Model	**0.405**	**0.181**

Table 2. Ensembled Model results

6. CONCLUSIONS

The findings of my models demonstrate that my random forest-based model with captions provides excellent efficiency relative to other features when used independently, but by integrating the results of such models in a weighted average way we may obtain much stronger output than before. In a way, we can assume that because the quality of the video can be explained by means of the captions, we can get a reasonable approximation of the memorability of the video. Features such as C3D and HMP help us to utilize visual functionality where no information is available to estimate memorability and can be helpful in the unsupervised calculation of real-world visual memorability.

Future Work, more analysis will be undertaken to use much more advanced methods, such as the Bayesian Model Neural Networks, to boost performance.

REFERENCES

[1] Squalli-Houssaini, H., Duong, N.Q., Gwenaëlle, M. and Demarty, C.H., 2018, April. Deep learning for predicting image memorability. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2371-2375). IEEE.

[2] K. M. Rohit Gupta, "Linear Models for Video Memorability Prediction Using Visual and Semantic Features," in MediaEval, EURECOM, Sophia Antipolis, France, 2018. K. Elissa, "Title of paper if known," unpublished.

[3] Cohendet, R., Demarty, C.H. and Duong, N.Q., Transfer learning for video memorability prediction.

[4] P.-H. C. C.-J. L. Rong-En Fan, "Working Set Selection Using Second Order Information for Training Support Vector Machines," Journal of Machine Learning Research 6 (2005) 1889–1918, p. 30, 2005.

[5] M. 2018, "MediaEval-2018," [Online]. Available: http://www.multimediaeval.org/mediaeval2018/. [Accessed 24 04 2019].

[6] Koehrsen, W. (2019). Introduction to Bayesian Linear Regression. [online] Towards Data Science. Available at: https://towardsdatascience.com/introduction-to-bayesianlinear-regression-e66e60791ea7 [Accessed 24 Apr. 2019].

[7] O. C. P. D. C. G. 1. G. H. F. H. K. M. E. M. T. W. Alan F. Smeaton, "Dublin's Participation in the Predicting Media Memorability Task at MediaEval 2018," in MediaEval 2018, EURECOM, Sophia Antipolis, France, 2018.

[8] Gupta, R. and Motwani, K., Linear Models for Video Memorability Prediction Using Visual and Semantic Features.

[9] "Scikit-learn: Machine Learning in Python Documentation," [Online]. Available: https://scikit-learn.org/stable/. [Accessed 24 04 2019].