Comparison of Simple Regression Models for Video Memorability Prediction Using Semantic and Video Features

Archana Kalapgar
19210184
MCM Computing
Dublin City University, Ireland
archana.kalapgar2@mail.dcu.ie

ABSTRACT

Memorability is defined as the condition of being easy to remember or worth remembering. It has been said that few words have a more positive impact on memorability than others. This concept is useful in determining sentiments which has many applications. To perform this, participants need to involuntarily predict video memorability scores which represent the likelihood of remembering a video. Some video has a greater effect than others, significantly resulting in higher ratings for memorability. In this paper, several simple regression models are compared with simple deep learning regression models which use both videos as well as semantic features to predict memorability scores.

KEYWORDS

C3D, HMP, Caption with Weights,TFIDFVectoizer, CountVectorizer, Visual, Semantic

1 INTRODUCTION

Explored and extracted visual features such as C3D, HMP. Trained the model with individual features using simple regression models such as Linear Regression, Decision Tree, Random Forest, Ridge Regression, XGBoost, Bayesian Ridge Regression, Support Vector Regression and simple deep learning neural network regression model. Also extracted semantic feature that described the video in a sentence. The models were evaluated using Spearman's correlation score as a standard measure. A combination of the best performing video and the semantic feature was chosen to train the final model for prediction.

Based on the outcomes, the following are the analysed conclusions:

(1) Short-term predictions are more accurate than long term predictions.
(2) Visual Feature based on HMP do not work well and are outperformed by those based on C3D.
(3) Captions despite being the only semantic feature available outperformed on models trained on visual features.
(4) Analysis of models trained on video captions allows us to identify semantic factors affecting video memorability.
(5) Linear Bayesian Ridge Regression outperformed in the captions features using count vectorizer.
(6) Combination of C3D and captions yields best results than simply C3D features.
(7) It was observed that models trained on high-level CNN trained on captions or image classification (visual features) give a comparatively better result than traditional models.

, ,

2 RELATED WORK

Work on video memorability has recently begun to generate a lot of curiosity, and recent works [2], [6], investigate the use of neural networks in various low level and high-level visual features, image and video captions for video memorability prediction. In [2], Simple linear and regularized models were selected – L1 Regularized Logistic Regression, Linear Support Vector Regression, ElasticNet. These selected models were run for Video features like HMP and C3D. The semantic features were processed under Count Vectorizer by removing the stop words. The key findings of all these papers on memorability are that model using captions produce the best individual outcome.

3 APPROACH

3.1 Models

Given that most of the features provided are very high dimensional and the number of videos is of the same order of magnitude as the dimensionality of the features, overfitting is a major potential challenge in this task.

Feature Selection: Used semantic features such as captions that are one-sentence text descriptions of the videos to predict memorability. As video features were expected to give better results, used HMP against C3D feature.

Model Selection: Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable. Selected simple regression models for such forecasts as mentioned below:

- Linear Regression [4][5]
- Decision Tree[5]
- Random Forest[5]
- Ridge Regression
- XGBoost
- Linear Support Vector Regression[4][3]
- Linear Bayesian Ridge Regression
- Simple Neural Network [1]

3.1.1 Traditional Regression Models: All simple regression model was applied for training and validation of individual features such as C3D, HMP, Captions with CountVectorizer(CV), Captions with TFIDFVectorizer(TFIDF), Captions with weights(Weights). TFIDF gave extraordinary results out of other sematic features so tested combination of the semantic feature with video features such as C3D+Captions(TFIDF) which outperformed all the other models. Values for various model hyper-parameters controlling the extent of regularization (such as alpha, compute score and n_estimators) were picked by using grid search over the dev set.

3.1.2 Deep Learning Regression Models: Applied Deep Learning Regression model for HMP features and Combination of captions with HMP such as HMP and Captions with CountVectorizer(CV), HMP and Captions with TFIDFVectorizer(TFIDF). Added validation split, early stopping monitor to tackle the problem of overfitting, 'adam' has been used as the optimizer and 'mean_squared_error' has been used as the regression algorithm. Applied all the above-mentioned learning model for training and validation as a single as well as a combination of the selected features [1]. The most paramount model was selected further for testing purpose.

3.2 Features and data pre-processing

3.2.1 VIDEO FEATURES:. Extracted features of HMP and C3D. Selected HMP feature as an independent variable with the ground truth of Dev-set. Selected values of short-term memorability and long-term memorability from dev-set of ground-truth. The HMP features alone were simply read into frames and sent as the independent variable with ground-truth as dependent variables joined on video. The same was also performed on C3D features. Both resulted in poor results. But C3D performed well compared to HMP.

(1) Method 1: Analysed C3D and HMP feature with traditional regression models. C3D feature outperformed HMP using Bayesian Ridge Regression Model.
(2) Method 2: Analysed HMP on Deep Learning Regression model to get improved results. Used two dense layers with activation function 'relu' and two drop-out layers. Added an output layer to get results. Compiled the model with accuracy metrics.

3.2.2 SEMANTIC FEATURES:. Extensive research has been done on semantic features proving it to perform better than other video features. Cleaning and standardization of text, making it noise-free and ready for analysis is performed using stop words on the commonly occurring words in the corpus of caption feature. Captions feature was processed using three different methods to increase the correctness.

(1) Method 1: TFIDF is a weighted model commonly used for information retrieval problems. Converted the cleaned corpus of words into vector models based on the occurrence of words using TfidfVectorizer.
(2) Method 2: The bag of words was also run with Count vectors as features. Count based features might seem trivial but show a great impact on learning models. Count Vector is a matrix notation of the corpus which represents the frequency count of a term in a corpus.
(3) Method 3: In [4], it has been said that few terms have a more positive impact on memorability than others. Based on this concept, Certain words are given high weights as they are very useful in determining sentiments. For example, the word pregnant, couple and baby have high weights. This can be used for predicting video memorability using semantic features. This new caption feature with weights is used for training and validation. Applied extra weights to the words with positive coefficiency [2]. These terms were searched in captions if found the weight for the caption was cumulatively increased.

Out of all three methods, captions with TFIDFVectorizer gave extraordinary results. This model performed best with Bayesian Ridge Regression Model with true compute the score.

3.2.3 COMBINATION OF VISUAL AND SEMANTIC FEATURES:

.

(1) Method 1: C3D and Captions with TfidfVectorizer were merged to form a single feature against ground truth values. It not only performed better than the C3D feature alone, but it also gave outstanding prediction values. In my exploration, the model with TFIDF captions worked best. Used the same model for my final computation in testing purpose. Therefore, my final model is on a semantic feature (captions), with TfIdfVectorized features. The results are stored in Final_Result.csv.
(2) Method 2: HMP featured merged with TFIDF Captions to form a single feature. Similarly, concatenated HMP with CountVectorizer Captions and used against ground truth values. Analysed both features on Deep Learning Regression. Used two dense layers with activation function 'relu' and two drop-out layers. Added an output layer and compiled the model with accuracy metrics. Out of both the models, HMP with TFIDF Captions gave better results but it was not as precise as C3D with TFIDF captions.

4 CONCLUSION AND FUTURE WORK

Like in the past work of video memorability, my analysis showed that the combination of TFIDF captions with C3D provided superior results than any other provided video features. A more in-depth exploration of captions can give even better results. In contrast to past work, however, after examining the work of simple layers of CNNs and observe that simple trained models can work well too. Hence, I think there's much scope for research in this field. More work can be done on finding impact coefficiency for each term in corpus and give weights accordingly.

4.1 Tables

Tables 1 and 2 give an overall summary of our experimental results on short-term and long-term memorability scores on single visual and semantic features. Results for the best model for each feature are presented. For HMP, C3D and Captions, Bayesian Ridge is the best model. Table 3 give an overall summary of our experimental results on the combination of features. For both combination of C3D with TFIDF as well as only caption with TFIDF which alone gave extraordinary results.

Model	*HMP*	*C3D*	*CV*	*TFIDF*	*Weight*
Linear Regression	0.071	0.290	0.114	0.059	0.095
Decision Tree	0.082	0.093	0.238	0.268	0.269
Random Forest	0.291	0.331	0.392	0.383	0.382
Ridge Regression	0.246	0.296	0.392	0.392	0.392
XGBoost	0.303	0.278	0.350	0.338	0.335
SVR	0.274	0.201	0.369	0.359	0.369
Bayesian Ridge	0.261	0.293	0.397	***0.404***	0.242

Table 1: Short-Term Memorability of single of Features

Model	*HMP*	*C3D*	*CV*	*TFIDF*	*Weight*
Linear Regression	0.004	0.139	0.038	0.029	0.095
Decision Tree	0.020	0.057	0.104	0.111	0.084
Random Forest	0.142	0.129	0.203	0.191	0.184
Ridge Regression	0.137	0.142	0.203	0.205	0.205
XGBoost	0.112	0.095	0.164	0.183	0.188
SVR	0.093	0.091	0.204	0.203	0.242
Bayesian Ridge	0.116	0.128	0.244	***0.236***	0.242

Table 2: Long-Term Memorability of single of Features

Feature	*Model*	*Short-Term*	*Long-Term*
TFIDF+C3D	Bayesian Ridge	***0.410***	***0.229***
TFIDF+HMP	CNN	0.314	0.130
CV+HMP	CNN	0.353	0.169

Table 3: Short-Term and Long-term Memorability of Combination of Features

REFERENCES

[1] Akash Bhargava. Predicting Video Memorability Using Captions and Image Features. 18210613 (????).

[2] Romain Cohendet, Karthik Yadati, Ngoc Q.K. Duong, and Claire Hélène Demarty. 2018. Annotating, Understanding, and Predicting Long-term Video Memorability. *ICMR 2018 - Proceedings of the 2018 ACM International Conference on Multimedia Retrieval* (2018), 178–186. https://doi.org/10.1145/3206025.3206056

[3] R. E. Fan, P. H. Chen, and C. J. Lin. 2005. Working set selection using the second order information for training {SVM}. *Journal of Machine Learning Research* 6 (2005), 1889–1918. https://doi.org/10.1007/s00249-014-0965-x

[4] Rohit Gupta and Kush Motwani. 2018. Linear models for video memorability prediction using visual and semantic features. *CEUR Workshop Proceedings* 2283 (2018), 2–4.

[5] Aruna Bellgutte Ramesh. Video Memorability Prediction Using Machine Learning. Ml (????).

[6] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and recall: Learning what makes videos memorable. *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017* 2018-January (2017), 2730–2739. https://doi.org/10.1109/ICCVW.2017.321 arXiv:1707.05357