Video Memorability Prediction Modelling

Qinyan Chen  
 School of Computing  
 Dublin City University   
Dublin Ireland  
Qinyan.chen7@mail.dcu.ie

ABSTRACT

Predicting Media Memorability score is a competition proposed by MediaEval 2018, which required participants to predict the “short-term” and “long-term” memorability score of videos. The memorability score is defined as the probability of remembering a video. The features had been provided including semantic and visual features, some of them are pre-computed [1]. In this paper, I introduced the methods I used to preprocessing the features and the model I used to predict the memorability score. The results were evaluated by the Spearman Correlation Coefficient. Finally, I chose the model with highest Spearman score as to predict the memorability score.

KEYWORDS

Video Memorability, Captions, Count Vectorizer, TF-IDF, Random Forest

1 INTRODUCTION

Video memorability can be used to measure the importance of a video, which can help people improve the quality of the video and use it in various fields [2]. When solving machine learning problems, it is very important to find the most relevant features and suitable models. Therefore, during this project, I investigated several features which had been provided. Video captions is a semantic feature which can be used to explore the impact of semantic factors on memory. In addition, I used spatiotemporal visual feature C3D [3], which extracted from the output of the final layer of the C3D model, and HMP [4], and InceptionV3 features for modeling, and conducted a lot of exploration and experiments to obtain the best performing predictor. These are my findings:

1. Caption feature preprocessed by CountVectorizer trained with random forest model got the highest Spearman score
2. C3D performs better in linear models than in other models
3. Inception V3 seems to be the most under-performing feature, and its score is much lower than the modeling results of any other feature
4. Among the video-dedicated features, the experimental results of C3D are better than HMP
5. All models have better prediction effects on short-term memory than on long-term memory

2 RELATED WORK

The work [5] used RNN with semantic feature got the best performance, they also used SVR on caption feature, but the result of this model is worse than the RNN model. In addition, they used ANN and SVR models for aesthetic feature respectively, and the experimental results of these models are not as good as the RNN model using Caption features mentioned earlier. Comparing to the winner of this competition [6], the experimental results of the model built using captions feature is better than other models they use. Therefore, I have explored this feature the most.

3 APPROACH

3.1 Feature Selection and Data Pre-Processing

I used semantic feature (Captions) firstly because as mentioned before it performed best in previous works. I removed the English stop words then I used two methods—CountVectorizer and TfidfVectorizer to extract semantic features and created the bag of words [7]. In CountVectorizer, I used a tokenizer to convert the text to a vector and serialize it. In addition, I used C3D, HMP, InceptionV3 features to explore the visual features. Some of these features have a high dimensionality, for example, HMP has 6075 values for each video. Hence when using these features, pay attention to whether they are over-fitting.

3.2 Model Selections

In my project, I used 3 machine learning algorithms for each feature, and for caption feature, I used two more algorithm to find a better prediction model.

1. Linear Regression (LR)

Linear Regression is used to be the base mode for all features, and C3D has the highest score for Spearman in this model, but it does not perform well for other features

1. Artificial Neural Network (ANN)

In ANN model, I used RMSProp as optimizer, and for caption features that are preprocessed in different ways (CountVectorizer and TfidfVectorizer), I trained them with ANN respectively.

1. Random Forest (RF)

Random Forest is an ensemble learning method which has good performance in solving regression and classification questions [8]. This model had been used as the final model to predict the memorability in our test-set.

Furthermore, when modelling using captions, I tried to add pooling layer to ANN, and I performed 1D Convolutional Neural Network on Caption feature. It is not as good as other models in this experiment.

4 RESULTS

The following two tables show the Spearman correlation coefficient scores obtained by using different parameters for each model. Overall, for features, Caption features the best performance, followed by C3D, and in comparison, InceptionV3 performs the worst. For models, in the short-term memorability table, the random forest performed best, followed by the ANN. While in the long-term memorability table, the ANN got the highest score for all features, followed by random forest.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Short-Term Memorability Spearman Correlation Score | | | | |
| Model  Feature | Linear Regression | Artificial Neural Network | Random Forest | Convolutional Neural Network |
| Captions (CountVectorizer) | 0.239 | 0.302 | **0.383** | 0.205 |
| Captions (TfidfVectorizer) | 0.115 | 0.266 | **0.357** |  |
| C3D | 0.296 | 0.283 | 0.296 |  |
| HMP | 0.068 | 0.165 | 0.243 |  |
| Inception V3 | 0.084 | 0.118 | -0.076 |  |

Table 1: The Results of Short-Term Memorability

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Long-Term Memorability Spearman Correlation Score | | | | |
| Model  Feature | Linear Regression | Artificial Neural Network | Random Forest | Convolutional Neural Network |
| Captions (CountVectorizer) | 0.081 | **0.169** | **0.154** | **0.168** |
| Captions (TfidfVectorizer) | 0.054 | 0.156 | 0.131 |  |
| C3D | 0.149 | **0.175** | 0.112 |  |
| HMP | -0.020 | 0.096 | 0.061 |  |
| Inception V3 | 0.021 | 0.085 | -0.057 |  |

Table 2: The Results of Long-Term Memorability

Based on the above results, I chose the Caption feature preprocessed with CountVectorizer and used random forest as the final model.

ANALYSIS AND DISCUSSION

From the experimental results, we can see that Caption and C3D are two features that can get higher scores compared to others, and Random Forest and ANN are two models with better results. For C3D, using ANN can get better results than random forest. In my future work, I should try to merge the Caption and C3D models for modeling. At the same time, I should try to further optimize the parameters of the ANN and random forest models. Furthermore, convolutional neural networks are used for modeling on other features.

CONCLUSION

In conclusion, my results show that caption feature that can get the best results, and the results obtained by using caption to predict are better than other features. It is worth noting that the different preprocessing methods for caption features will also have an impact on the prediction results. In my experiments, using CountVectorizer is better than using TfidfVectorizer. No matter which model and feature we use, our short-term memorability prediction is always better than long-term memorability prediction. It is worth noting that according to the data in Ground truth dataset, we found that the number of people participating in the long-term memory test is much smaller than the number of people participating in the short-term memory. The prediction result of ANN for long-term memorability is better than that of random forest, and for short-term memorability prediction, the effect of random forest is far better than that of ANN. In the end, we chose the caption feature preprocessed by CountVectorizer and the random forest as our predictor.

REFERENCES

1. 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.*
2. Cohendet, R., Yadati, K., Duong, N.Q. and Demarty, C.H., 2018, June. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval* (pp. 178-186).
3. Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*(pp. 4489-4497).
4. Almeida, J., Leite, N.J. and Torres, R.D.S., 2011, September. Comparison of video sequences with histograms of motion patterns. In *2011 18th IEEE International Conference on Image Processing* (pp. 3673-3676). IEEE.
5. Sun, W. and Zhang, X., 2018. Video Memorability Prediction with Recurrent Neural Networks and Video Titles at the 2018 MediaEval Predicting Media Memorability Task. In *MediaEval*.
6. Gupta, R. and Motwani, K., 2018. Linear Models for Video Memorability Prediction Using Visual and Semantic Features. *In MediaEval*.
7. Eshan, S.C. and Hasan, M.S., 2017, December. An application of machine learning to detect abusive bengali text. In *2017 20th International Conference of Computer and Information Technology (ICCIT) (pp. 1-6). IEEE*.
8. Segal, M.R., 2004. Machine learning benchmarks and random forest regression.

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00