Random Forest Model for Media Memorability Prediction Using Visual and Semantic Features

Lulu Zha
Dublin City University, Ireland
lulu.zha2@mail.dcu.ie

ABSTRACT

Video and image memorability has become an important research topic in the computer vision field. Video memorability can be used in education, search, advertising, and other fields. This article uses video- and image-based features, as well as text-based features, to predict long-term(1 to 3 days later) and short-term(minutes later) memorability scores for video. In this paper, multiple features are used to model media memorability for the "Predicting Media memorability" task at MediaEval 2018. The Spearman level correlation is used to evaluate the model to see that the model proposed in this paper is a good representation of the test data.

1 INTRODUCTION

With the rapid development of the field of media memorability predictability, various social platforms such as YouTube and TikTok are processing more and more content data[6]. Video memorability(VM) is useful for solving this problem[6][8]. It ranks videos based on their memorability, enabling recommendations and filtering. In this paper, I developed a video memorability prediction model by studying various features of images and videos to predict video memorability. Among the features provided[1], I trained models on C3D, Captions, HMP and Color Histogram. Besides, I combined the four features provided and trained the model again. Use Spearman's ranking correlation as an indicator evaluation model. The main findings and contributions are as follows:

1. The short-term memorability model is much better than the long-term memorability model.
2. In a single feature, models based on Captions or HMP do not perform well. However, models based on C3D provide much better performance.
3. In multiple features, models based on two or three features are better than individual features.

2 RELATED WORK

With the explosion of video content on the Internet, it is necessary to study video analysis methods that take into account human cognition[6]. Memorability is an important aspect of cognition and is inherent in visual content[8]. Although image memorability has been studied by many people, the problem of video memorability modeling has not yet been solved well. Recent work[2][4] has examined the use of different levels of visual features, predictions based on semantic features, and the use of motion recognition representations based on deep learning. This work shows that models that use captions score the most in a single feature and how to use multiple features to improve these results. Also, the researchers found that advanced visual semantics helped better predict the memorability of video[3].

3 APPROACH

3.1 Models

Since images and videos often have large dimensions, they pose a huge challenge to analytical tasks[5]. To effectively solve the problem of overfitting[7], we prefer simple, faster training models: Random Forest(RF) Model. For each set of features, I tried a Random Forest Model. Besides, I use simple calculation methods to calculate short- and long-term memorability scores to select the fit video attributes and models.

3.2 Features and Data Pre-Processing

Video-level features, such as C3D and HMP, are used as it is, while frame-level features such as ColorHistogram are connected across frames, especially with color parameters combining frame0, 56, and 112. When processing Captions, I replaced punctuations with space and converted words to lower case. I used the tokenizer to vectorize a text corpus, by turning each text into either a sequence of integers. After pre-processing single features, I concatenated these features into one multi-feature by row level.

4 RESULTS AND ANALYSIS

Tables 1 and 2 provide an overview of my experimental results. The results of a random forest model for each feature are presented. C3D gets the best score for a single feature. For multiple features, C3D plus HMP plus Captions is the best combination.

Table 1: RF Memorability Short-term Score

Features	Short-term
Captions	0.258
HMP	0.294
C3D	0.314
C3D+Captions	0.321
C3D+HMP+ColorHistogram	0.319
C3D+Captions+HMP	0.328
C3D+Captions+HMP+ColorHistogram	0.315

Table 2:RF Memorability Long-term Score

Features	Long-term
Captions	0.130
HMP	0.097
C3D	0.117
C3D+Captions	0.111
C3D+HMP+ColorHistogram	0.148
C3D+Captions+HMP	0.130
C3D+Captions+HMP+ColorHistogram	0.146

For video features, I use the C3D and HMP features. The C3D feature is a single list of numbers on a line with a dimension of 101. The HMP feature is a single list of pairs of numbers with the format: bin: number (dimension is 6075) on one line. For image features, I use the ColorHistogram feature. The ColorHistogram feature is a list of 3 pairs of 255 pairs (red, green, blue order) in the bin: number, for example, 254:1008. A list per row. Also, the Captions feature has a size of 50. All features include 6000 elements that match 6000 videos. As a result, for the multiple features(C3D+Captions+HMP), the dimension is $d = 101 + 50 + 6075 = 6225$.

To verify the effectiveness of the combined features, I used the development set to compare the performance of Random Forest in the original and combined features. I use 20 percent of the data as the test set. The Spearman coefficient in Table 1 and 2 show that the performance of the single video and image features is lower than that of combined features in the short- and long-term. The score of C3D is better than other individual features. Although the score of the combination feature is better in the short- and long-term, the result is far from satisfactory. The cause may be that a combination of high-dimensional features may contain redundant or noisy information.

5 CONCLUSIONS

Based on memorability prediction, this paper presents a combination of image, video and semantic features for training in Random Forest models. The result is better than the average of MediaEval 2018 Predicting Media Memorability Task[1]. It verified the validity of my method. However, unlike other papers, the score for a single Captions feature is not the highest, which may have a lot to do with the Random Forest model I chose. In the future, I will further improve the score of the Random Forest model and its best parameters. Besides, I'm going to try different deep learning models to compare the pros and cons of each model horizontally.

ACKNOWLEDGMENTS

In this paper, I would like to thank Dr. Tomas Ward(Dublin City University) for giving me this chance to learn Machine Learning and Ta. Eoin(Dublin City University) for helping me to solve different difficult problems.

REFERENCES

[1] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q.K. Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. 2018. MediaEval 2018: Predicting Media Memorability Task. In Working Notes Proceedings of the MediaEval 2018 Workshop.

[2] Romain Cohendet, Karthik Yadati, Ngoc Q.K. Duong, and Claire Hélène Demarty. 2018. Annotating, Understanding, and Predicting Long-term Video Memorability. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. ACM, 178–186.

[3] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[4] Sumit Shekhar, Dhruv Singal, and Harvineet Singh. 2017. Show and Recall: Learning What Makes Videos Memorable. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2730–2739.

[5] Gupta, Rohit, and Kush Motwani. 2018. Linear Models for Video Memorability Prediction Using Visual and Semantic Features. In MediaEval.

[6] Joshi, T., Sivaprasad, S., Bhat, S., and Pedanekar, N. 2018. Multimodal Approach to Predicting Media Memorability. In MediaEval.

[7] Tran-Van, D. T., Tran, L. V., and Tran, M. T. 2018. Predicting Media Memorability Using Deep Features and Recurrent Network. In MediaEval.

[8] Peng, H., Li, K., Li, B., Ling, H., Xiong, W., and Hu, W. 2015. Predicting image memorability by multi-view adaptive regression. In Proceedings of the 23rd ACM international conference on Multimedia. 1147-1150.