Video Memorability Prediction with Recurrent Neural Networks and Video Titles at the 2018 MediaEval Predicting Media Memorability Task

Wensheng Sun
Michigan Technological University, Houghton, USA
wsun3@mtu.edu

Xu Zhang
Saginaw Valley State University, Saginaw, USA
xzhang21@svsu.edu

ABSTRACT

This paper describes the approach developed to predict the short-term and long-term video memorability at the 2018 MediaEval Predicting Media Memorability Task [1]. This approach utilizes the scene semantics derived from the titles of the videos using natural language processing (NLP) techniques and a recurrent neural network (RNN). Compared to using video-based features, this approach has a low computational cost for feature extraction. The performance of the semantic-based methods are compared with those of the aesthetic feature-based methods using support vector regression (ϵ-SVR) and artificial neural network (ANN) models, and the possibility of predicting the highly subjective media memorability with simple features is explored.

1 INTRODUCTION

Knowledge of the memorability of a video has potential in advertisement and content recommendation applications. Although highly subjective, it has been shown that media memorability is measurable and predictable. As with most other machine learning problems, finding the most relevant features and the right model is the key to the successful prediction of the media memorability. In [2], the authors investigate possible features that are correlated with image memorability. It is shown that simple image features such as color and number of objects show negligible correlation with image memorability, whereas semantics are significantly correlated with the memorability.

Even though images are reportedly different from videos in many aspects [3], the similarity and connection between images and videos motivate this work to explore the possible connection between semantics of a video and its memorability at the 2018 MediaEval Predicting Media Memorability Task [1]. This hypothesis is confirmed in [4], where the authors show that visual semantic features provide best prediction among other audio and visual features. Different from [4], an RNN is used to extract the semantics from the video titles and to predict the video memorabilities in this work.

Compared to video-based features, the extraction of video semantics from its title requires relatively low feature extraction cost. Moreover, the authors in [5] demonstrate a strong connection between aesthetic features and image interestingness. Thus in this work, models to predict video memorability using precomputed aesthetic features [6] provided by the organizer are also developed and compared with the semantic-based models in performance.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

2 APPROACH

2.1 Semantic-based Models

Table 1: official test results: Spearman's rank correlation

Run	Method	Short-term	Long-term
1	SVR+AF(Median)	0.315299	0.083562
2	SVR+AF(Mean)	0.347227	0.091239
3	ANN+AF(Mean)	0.121194	0.057660
4	**RNN+Captions**	**0.356349**	**0.213220**
5	SVR+Captions	0.230784	0.111450

The main model corresponding to run 4 is a three-layer neural network with a recurrent layer; the structure of the model is depicted in Fig. 1. After importing the titles, punctuation and white-space is removed. The texts are then tokenized to integer sequences with length equal to 20. Longer titles are truncated and short titles are padded with zeros. After the preprocessing, 80% of the training dataset is randomly chosen to train the model, and the remaining 20% is used for model evaluation.

The tokenized titles are fed to an embedding layer with the output dimension equal to 15. The embedding matrix is initialized following uniform distribution. No embedding regularizer is used. The semantics are extracted by adding a fully connected recurrent layer with 10 units after the embedding layer. The activation function for the recurrent layer is hyperbolic tangent. The layer uses a bias vector, which is initialized as zeros. Initializer for the kernel weight matrix used for the linear transformation of the inputs is chosen as "glorot uniform". Initializer for the recurrent kernel weight matrix used for the linear transformation of the recurrent state is set as "orthogonal". A 10-node fully connected dense layer follows using rectangular linear unit (ReLU) activation function. The kernel regularization function used is $l_1 - l_2$ regularization with $\lambda_1 = 0.001$ and $\lambda_2 = 0.004$. The initialization scheme is the same as that of the RNN layer. The last layer is a 2-node dense layer predicting the short-term and long-term memorability simultaneously, where a linear activation function is used. This model is trained using RMSprop optimizer against the mean absolute error (MAE). The model is trained 10 epochs with batch size equal to 20.

Similar to the model in [4], the semantics are combined with a support vector regression (ϵ-SVR) model to generate run 5, whose structure is also shown in Fig. 1. After the preprocessing stage, the dimensionality of the tokenized titles is reduced to explain 90% of the variance through principle component analysis (PCA). The output is then fed into an ϵ-SVR model. The penalty parameter C of

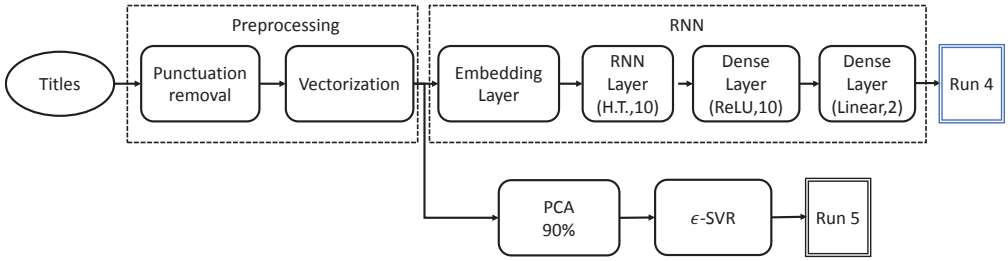

Figure 1: Semantic-based models: the recurrent neural network model and ϵ-SVR model correspond to run 4 and 5, respectively.

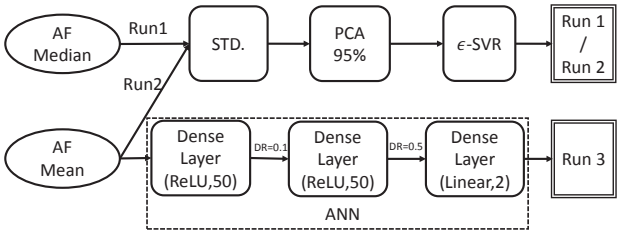

Figure 2: Aesthetic feature-based models: ϵ-SVR models with median and mean aesthetic features correspond to run 1 and 2, respectively; ANN with mean aesthetic features generates run 3.

the error is set to be 0.1. The ϵ, which defines a tube within which no penalty is associated, is equal to 0.01. Radial basis functions are used as the kernel function. The above hyper parameters are obtained through a grid search cross-validation using the Spearman's rank correlation as the scoring matrix.

2.2 Aesthetic Feature-based Models

Details of the models using precomputed aesthetic features [6] are described in this section. As shown in Fig. 2, run 1 and run 2 are generated by ϵ-SVR models using aesthetic visual features aggregated at video level by median and mean methods, respectively. In both runs, the input features are standardized first, and a PCA module is applied to reduce the dimensionality of the data to count 95% of the data variance. Radial basis function is chosen in both runs. The grid search cross-validated best parameters for the eSVR model are $C = 0.01$ and $\epsilon = 0.1$.

The evaluation results show that the mean aesthetic features are more relevant to the video memorability. Thus run 3 is generated using ANN and mean aesthetic features as illustrated in Fig. 2. The ANN model consists of three dense layers, the first two layers are fully connected dense layers with 50 nodes, where ReLU activation function is used, and l_2 regularization is applied. The regularization penalty constant is set to 0.001. Dropout rates for the first two layers are equal to 0.1 and 0.5, respectively. The output layer has two nodes and uses linear activation functions. Mean square error (MSE) is used as the loss function during the training process, where the validation data is randomly chosen from the training data within each epoch. 20 epochs are trained in total with the batch size equal to 32.

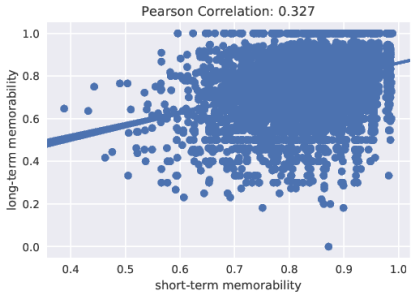

Figure 3: Correlation between two types of memorabilities

3 RESULTS AND ANALYSIS

From the returned evaluation results in Table. 1, the following conclusions can be observed: 1) The model using RNN and semantics is the best among all the five models. It confirms that the semantics of the videos are more relevant to both short and long-term memorability than aesthetic features. Especially for long-term memorability, the semantic based models outperform the aesthetic feature-based models unanimously. 2) Without the recurrent layer, the performance decreases. Thus it can be inferred that interaction between objects in a video has more impact on the video's long-term and short-term memorability than knowing only the objects. 3) Even though there is certain correlation between short and long-term memorability as depicted in Fig. 3, results have shown that short-term memorability is more predictable than long-term ones since all models score higher in short-term than long-term memorability. As illustrated in Fig. 3, long-term scores range from 0.2 to 1 and exhibit higher variance than the short-term scores, which distribute from 0.4 to 1. Thus, one possible reason is that the long-term memorability is more subjective and depends more on individual's memory.

It is observed that the SVR models using median and mean aesthetic features have close performance as run 4 in terms of short-term memorability prediction. However, the long-term performance is far worse than run 4. Further investigations are needed to clarify this. Performance of run 3 is worse than that of run 2, even though both of them use mean aesthetic features. Possible reasons are over-fitting and the missing standardization procedure in run 4. In the future, ensemble methods are expected to further enhance the prediction accuracy.

REFERENCES

[1] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. MediaEval 2018: Predicting Media Memorability Task. In *Proc. of the MediaEval 2018 Workshop*, Vol. abs/1807.01052. 29-31 October, 2018, Sophia Antipolis, France, 2018.

[2] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence* 36, 7 (2014), 1469–1482.

[3] S. Shekhar, D. Singal, H. Singh, M. Kedia, and A. Shetty. 2017. Show and Recall: Learning What Makes Videos Memorable. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2730–2739. https://doi.org/10.1109/ICCVW.2017.321

[4] Romain Cohendet, Karthik Yadati, Ngoc Q.K. Duong, and Claire-Hélène Demarty. 2018. Annotating, understanding, and predicting long-term video memorability. In *Proc. of the ICMR 2018 Workshop, Yokohama, Japan, June 11-14.*

[5] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. 2013. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1633–1640.

[6] Andreas F Haas, Marine Guibert, Anja Foerschner, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A Sandin, Jennifer E Smith, Mark JA Vermeij, and others. 2015. Can we measure beauty? Computational evaluation of coral reef aesthetics. *PeerJ* 3 (2015), pp.1390.