Ensemble Models for Video Memorability Prediction Using Video and Image Features

Siddarth Shantinath Patil
Student No: 20210993
MCM Computing (Data Analytics)
Dublin City University, Dublin, Ireland
siddarth.patil2@mail.dcu.ie

ABSTRACT

According to Merriam Webster, "Memorability is defined as the state of being easy to remember or worth remembering." As of 2019, 720,000 hours of video is uploaded on YouTube alone, and 90% of people say that they discover new brands and products on YouTube. Moreover, due to the ongoing pandemic, the consumption of video content has increased exponentially. The MediaEval 2018 proposed a task of predicting the memorability of these videos which would help content creators to make more impactful content. [1] In this study, I have designed a model using provided Video and Image features to predict both the short-term and long-term memorability scores. Finally, the results showed that a weighted average ensemble of different models gives better results than a single model trained on a feature.

1 INTRODUCTION

For the task of predicting video memorability scores, I have made use of a dataset provided by MediaEval2018 which consists of a total of 8000 videos. These videos were divided into two groups, the first group is a Dev-Set which consists of 6000 videos and the second group is a Test-Set which contains the remaining 2000 videos. Video features like Convolution 3D (C3D) which is the output of the final classification layer of the C3D model and Histogram of Motion Patterns (HMP) are present in both Dev-Set and Test-Set. [1] Image features like Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP), Fc7layer from InceptionV3 (InceptionV3), Oriented FAST and Rotated BRIEF (ORB), Aesthetic feature, and Color Histogram are also present in both Dev-Set and Test-Set. [1] The Ground-Truth table consists of long-term and short-term Video Memorability (VM) scores for each video in the Dev-Set and I have created a model, trained on the features from Dev-Set to predict the long-term and short-term VM Score for the videos present in the Test-Set. The models were evaluated using Spearman's rank correlation coefficient score as a standard metric.

With the results of my analysis, I have identified the following conclusions:

- Video features (C3D & HMP) give better results when compared to image features in predicting both long-term and short-term VM scores.
- Ensemble models like Random Forest & Extreme Gradient Boosting (XGBoost) gives better results than Artificial Neural Networks (ANNs) and other simpler models.
- The weighted average ensembling technique provides a drastic improvement in the performance of the model.

Conference'17, July 2017, Washington, DC, USA

2 RELATED WORK

Work by Rohit Gupta and Kush Motwani [4] suggests that the use of linear models like LASSO regularized Logistic Regression, Linear Support Vector Regression & ElasticNet provides promising results. They trained these models on every available feature in the dataset and later performed weighted average ensembling on the best performing models to further improve the obtained results. Their study shows that Video features like C3D and HMP, and Image Feature LBP gave better results when compared to other features for predicting both short-term and long-term VM scores.

The usage of Support Vector Regressor (SVR) and k-Nearest Neighbor Regressor (KNR) was demonstrated in the work by Savii R.M., et. al, [7]. They trained the above-mentioned models on the available video features (C3D & HMP) and compared their results with a Neural Network (NN) trained on the same feature. The results of their study didnt give promising results.

Zhoo, T, et al, [8] presented their work where they trained SVR model on a dataset comprising of 590 short videos of length 1-8 seconds long. This dataset was trained on video feature C3D and Image Feature LBP. Both the features gave fairly good results when compared to their other models trained on the same features.

3 APPROACH

Based on the previous related work [4] [7] [8] [5] I decided to work with a total of five models which consists of four Machine Learning models and a Neural Network model. Machine Learning models used were SVR, Random Forest Regressor (RFR), Extreme Gradient Boosting (XGBoost), and KNR. For the neural network, I made use of Keras Sequential Neural Network (NN).

All the above models were trained using all the video features (C3D & HMP) and Image features which consisted of Aesthetic, LBP, InceptionV3, and ColorHistogram. I chose to omit ORB and HOG features based on their results in the previous work [4] [7] [8].

3.1 Data Pre-processing

It was observed during the loading process that features C3D, HMP, InceptionV3, LBP, and ColorHistogram were missing values for many of the videos. Therefore, I made two different sets of datasets. One where videos with missing values were dropped and the other where a list of zeros with appropriate length was added in place of missing values. I found out that the latter gave better results.

Further, video features (C3D & HMP) were used without any changes in them. For image features InceptionV3, LBP & ColorHistogram the features were extracted at three frames 0th, 56th, & 112th. I concatenated features of all three frames for each video to get the final feature list for the video. For Aesthetics feature similar approach was followed by concatenating its features on Mean and Median for every video. All these features were converted into NumPy arrays and were used to train the models. Finally, standardization of input was also performed on all the features to train an SVR model.

3.2 Model Training

A total of five models were trained on six features individually which resulted in thirty models. All thirty models were used to predict both short-term and long-term VM Scores. And for each model Spearman's rank correlation coefficient score was calculated to check for its results.

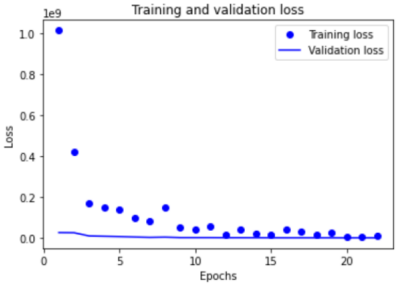

Figure 1. EarlyStopping for ColorHistogram

After the initial run, I found out that ensemble models RFR and XGBoost performed better than other models. I further tuned the models to improve their performance. For NN I added a second layer, added dropout, and made use of EarlyStopping to prevent it from over-fitting as shown in Figure 1. For SVR, I standardized the input to help the model understand the variable importance and perform better. The optimum *n_estimator* in case of RFR and XGBoost was found out by running the model on multiple values of *n_estimator* and selecting the best performing value for both short-term and long-term VM score. To find the optimum *n_estimator* for the KNR model, I plotted a graph with error rate vs *n_estimator* values ranging from 1 to 40 for every feature. Figure 2 shows the graph for the HMP feature.

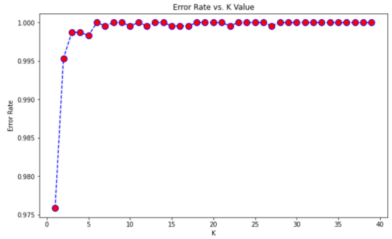

Figure 2. Finding optimum n_neighbors for HMP Feature

3.3 Creating Ensemble Model

HMP, C3D, and LBP features gave good results when trained on RFR and XGBoost. To further improve their results I decided to perform a weighted average ensemble on the best performing models for each of the features. As selecting appropriate weights could be a time-consuming task, I wrote a function that took two parameters as input. The first is a list of NumPy arrays of predictions obtained from different models, and the second argument is the NumPy array containing actual short-term and long-term VM scores. The function then tries every possible weight for each feature, passes it to another function which returns the Spearman's rank correlation coefficient score for it, the function then saves the weights, Spearman's score, feature related to weights, and finally it returns the best performing results.

For the short-term VM score, the weighted average ensemble technique helped in improving Spearman's rank correlation coefficient by 0.022 points, but in case of the long-term VM score, it proved to give bad results compared to the models trained on individual features.

4 RESULTS AND ANALYSIS

Table 1 below shows Spearman's rank correlation coefficient score for each model trained on all the features (after removing missing values) for short-term VM scores and long-term VM scores respectively.

Model/Feature	C3D		HMP		LBP		InceptionV3		Aesthetics		ColorHistogram	
	short-term	long-term	short-term	long-term	short-term	long-term	short-term	long-term	short-term	long-term	short-term	long-term
ANN	0.237	0.064	0.254	0.113	0.223	0.102	0.098	0.051	0.100	0.027	-0.03	-0.07
SVR	0.207	0.075	0.145	0.063	0.211	0.074	0.143	0.029	0.152	0.045	0.112	0.022
RFR	0.311	0.133	0.259	0.100	0.248	0.105	0.118	0.032	0.285	0.116	0.169	0.075
XGBoost	0.284	0.125	0.287	0.079	0.233	0.068	0.227	0.068	0.294	0.102	0.096	0.047
KNR	0.185	0.039	0.215	0.045	0.154	0.110	0.097	0.022	0.137	0.137	0.116	0.014

Table 1. Spearman's rank correlation coefficient score

The best performing models (RFR & XGBoost) from the above table were then trained on the whole dataset by replacing missing values with a list of zeros. The results of the same are shown below in Table 2. These models were then used for the weighted average ensemble.

Model/Feature	C3D		HMP		LBP		InceptionV3		Aesthetics		ColorHistogram	
	short-term	long-term	short-term	long-term	short-term	long-term	short-term	long-term	short-term	long-term	short-term	long-term
RFR	0.324	0.106	-	-	0.255	0.050	-	-	-	-	-	-
XGBoost	-	-	0.317	0.121	-	-	0.156	0.040	0.294	0.102	0.156	0.040

Table 2. Spearman's rank correlation coefficient score after replacing zeros.

The above results show that ensemble models perform better than other simpler models and Neural Network. Furthermore, applying a weighted average ensemble on the best performing ensemble improved the performance. The best performing ensemble model was:

- Short-term VM Score:
  *0.23*XGBoost on Aesthetics + 0.24*RFR on LBP + 0.26*XGBoost on HMP + 0.27*RFR on C3D*
  Spearman's coefficient - 0.346
- Long-term VM Score:
  XGBoost on HMP
  Spearman's coefficient - 0.121

5 CONCLUSION AND FUTURE WORK

The above study shows that video features perform better than Image features. I believe that to improve the results further we can try the following approach:

- Hyperparameter tuning of the best performing model might result in good scores.
- Instead of concatenating, taking means, median, or performing Principal Component Analysis (PCA) on Image features extracted on 3 different frames might help in improving the results.
- Generating new video or image-based features for more than 3 frames might help models to learn better and give better results. [3] [6] [2]

REFERENCES

[1] Romain Cohendet, Claire-Hélène Demarty, Ngoc Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. 2018. Mediaeval 2018: Predicting media memorability task. *arXiv preprint arXiv:1807.01052* (2018).

[2] Romain Cohendet, Claire-Hélène Demarty, and Ngoc QK Duong. 2018. Transfer Learning for Video Memorability Prediction.. In *MediaEval*.

[3] Romain Cohendet, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engilberge. 2019. VideoMem: Constructing, Analyzing, Predicting Short-term and Long-term Video Memorability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2531–2540.

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

[5] FINN GUSTAFSSON. Comparing Random Forest, XGBoost and Neural Networks With Hyperparameter Optimization by Nested Cross-Validation. (????).

[6] Jiaxin Lu, Mai Xu, Ren Yang, and Zulin Wang. 2020. Understanding and predicting the memorability of outdoor natural scenes. *IEEE Transactions on Image Processing* 29 (2020), 4927–4941.

[7] Ricardo Manhães Savii, Samuel Felipe dos Santos, and Jurandy Almeida. 2018. GIBIS at MediaEval 2018: Predicting Media Memorability Task.. In *MediaEval*.

[8] Tony Zhao, Irving Fang, Jeffrey Kim, and Gerald Friedland. 2021. Multimodal Ensemble Models for Predicting Video Memorability. *arXiv preprint arXiv:2102.01173* (2021).