Predicting Video Memorability using Ensemble Technique

Vinay Jagdish Jethmalani
Dublin City University
20210984
vinay.jethmalani2@mail.dcu.ie

***Abstract*—Human mind perceives and memorizes incidents and thoughts in a unique way of audio or visual. One such way to predict this ability is through analysis of video memorability scores. The goal of this task is to analyse and train the models on the features of short videos to produce short term and long term memorability scores. The problem is approached in two ways, firstly by using Convolutional Neural Networks (CNN) and then by using Stacking Ensemble technique. The Stacking Ensemble model outperformed the CNN model by providing better memorability scores.**

***Index Terms*—ensemble, neural networks, video memorability**

I. Introduction

In this paper, C3D is chosen as the prime candidate for memorability scores prediction. Apart from this, HMP and Inception V3 (Image Feature) are also grouped with C3D to record short and long term scores. Initially, selected feature values from the video is parsed and stored in a data frame which is then used for training and validation purpose.

In most of the researches, Deep Learning using Neural Networks has been inspected to be outperforming other traditional machine learning algorithms.Thus, CNN is initially selected for memorability scores prediction. Another model is also implemented by using Stacking Ensemble and Weighted Average Ensemble Technique. The results of both the models are then evaluated based on actual scores recorded in ground-truth file by calculating Spearman's Correlation Coefficients.

This paper is structured as follows: In Section II, Literature review on the topic of Video Memorability is discussed. In Section III, The approach and techniques implemented in this paper are explained. In Section IV, Results are analysed in detail and finally, in Section V, Conclusion and future work is summarized.

II. Related Work

Domain of human memorability has been assessed through multiple channels such as Deep Learning or Regression learning models. One such paper [1] has attempted to utilize Artificial Neural Network (ANN) and Natural Language Processing (NLP) technique to derive the semantic features from the video titles and model them for memorability. It also makes use of Support Vector Regression to compare the prediction results from Semantic and aesthetic features carved from the videos.

In another paper [2] modelling was implemented on the number of different features such as C3D, ColorHistogram, LBP, and HMP using LASSO regularized Logistic Regression, Linear Support Vector Regression, and Elastic Net and the resultant predictions were merged using Weighted Average Ensemble. The paper concluded that features such as C3D and HMP are more effective than others by calculating MSE, Pearson, and Spearman's Correlation Coefficient.

Since the size of the dataset for video memorability is large, dimensionality reduction techniques makes a difference in accuracy of predictions. [3] utilized Principal Component Analysis (PCA) to increase the prediction scores and built a deep learning model around C3D and other aesthetic features.

III. Approach

A. Preprocessing and Feature Selection

This work mainly focuses on the use of multiple video features, C3D and HMP, and image feature Inception V3. Features are grouped (as shown in figure 2) based on their individual prediction score to achieve higher efficiency. A data frame is constructed by importing and merging the feature values which is then used as model input in model implementation.

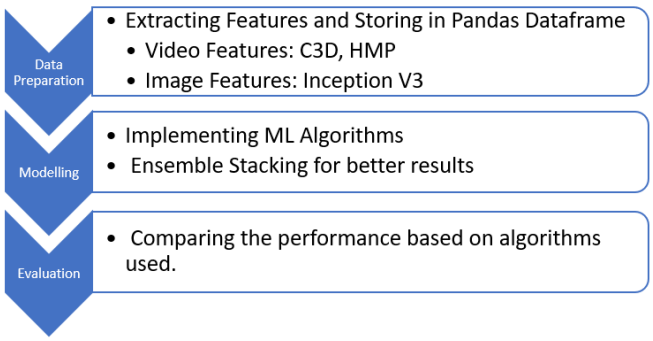

Fig. 1. Flow Diagram

B. Ensemble Implementation

Ensemble Technique is a machine learning technique which combines multiple weak learners to solve or predict the variable and provide better results as compared to the individual learners. Stacking is one such Ensemble technique in which

data set is trained through various models in first stage and then stacked for second layer learning.

The combinations of selected features (Figure 2) are trained individually on multiple Regression algorithms such as Random Forest Regression, Decision Tree, Bayesian Ridge, and Support Vector Regression by implementing the K-fold cross validation, with K value selected as 10. The resultant predictions from all the individual models is passed to the Linear Regression which acts as meta-regressor. Finally, Weighted Averaging method is implemented to assign appropriate weights based on results from Stacking ensemble.

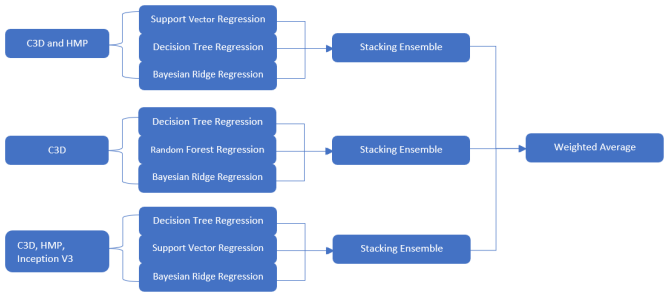

Fig. 2. Ensemble Stacked Model Structure

C. Neural Networks Implementation

A deep learning method of creating a network in layer-by-layer manner is implemented by using Sequential API from Keras library in Python. This model takes the input features as shown in Figure 3.

Feature	Short Term Score	Long Term Score
C3D	0.244	0.175
C3D and HMP	0.263	0.161
C3D, HMP & Inception V3	0.192	0.137

Fig. 3. CNN Prediction Scores

The Neural Network has total of 5 layers with 3 dense layers and 2 dropout layers which has drop-out of 0.5 to prevent over-fitting issue. The model uses the ReLU activation and the output layer utilizes Sigmoid activation for two outputs, long and short term memorability score of videos.

The model is compiled and Mean Square Error is used as loss function. Epoch count is kept at 20 and Spearman's Correlation Coefficient is used. (Scores are displayed in Figure 3)

IV. Results and Analysis

The output of all the implemented models is evaluated on the basis of the Spearman's Correlation Coefficients. Figure 3 and 4 provide the summary of memorability score from CNN model and Stacking Ensemble technique respectively. Following list of observations are deduced from the results:

1) Stacking Technique of Ensemble enhances the output of long-term video memorability considerably.
2) Ensemble methodology delivers better results when compared to Neural Networks (CNN).
3) Weighted average method considers the importance of each feature or feature combinations and assists in assigning appropriate weights to them.
4) Image feature Inception V3 individually predicts very poor memorability scores, but when combined with significant features such as C3D and HMP, the scores show a considerable improvement.

Feature	Technique Used	Short Term Score	Long Term Score
C3D	*Bayesian Ridge Regression*	0.290	0.173
	Decision Tree Regression	0.093	0.033
	Random Forest Regression	0.320	0.152
	Stacking Ensemble	**0.327**	**0.177**
C3D and HMP	*Decision Tree Regression*	0.124	0.049
	Support Vector Regression	0.286	0.152
	Bayesian Ridge Regression	0.289	0.176
	Stacking Ensemble	**0.293**	**0.175**
C3D, HMP and Inception V3	*Support Vector Regression*	0.246	0.178
	Bayesian Ridge Regression	0.276	0.193
	Decision Tree Regression	0.071	0.030
	Stacking Ensemble	**0.274**	**0.188**
Weighted Average of Stacking Ensemble		0.322	0.191

Fig. 4. Ensemble Prediction Scores

V. Conclusion and Future Work

The basic focus of this paper is to predict the memorability using the Neural Networks and Ensemble technique. The results of individual Decision Tree Regression model are below par making it a weak learner and other models such as Bayesian Ridge Regression, Random Forest Regression, and Support Vector Regression as strong learners.

From the results, it is completely evident that implementation of stacking ensemble significantly improves the outcome as compared to individual weak learner models and CNN.

In terms of future work, score related to emotions can be extracted from videos. Since, emotions plays an key role in memorability of videos or images.

References

[1] 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.

[2] Gupta, R. and Motwani, K., 2018. Linear Models for Video Memorability Prediction Using Visual and Semantic Features. In MediaEval.

[3] Constantin, M.G., Kang, C., Dinu, G., Dufaux, F., Valenzise, G. and Ionescu, B., 2019, October. Using Aesthetics and Action Recognition-based Networks for the Prediction of Media Memorability. In MediaEval 2019 Workshop.

[4] Analytics Vidhya. 2020. A Comprehensive Guide To Ensemble Learning.