Media Memorability Prediction Using Machine and Deep Learning

Akhileshkumar Mutuguppe Subbarao
School of Computing
Dublin City University, Ireland
akhileshkumar.mutuguppesubbarao2@mail.dcu.ie

***Abstract* — With increased access to the Internet, there is abundant video content on the web. Image/Video memorability has become important part of computer vison technologies. Thus, there is a requirement to research on method which help to organize these videos. Video memorability has used in fields like content summarization, content filtering, advertising. This paper uses ensembled way of weighted average to predict short-term and long-term memorability. The model is compared with spearman's correlation to see how model is representing the test set.**

***Keywords* — RNN , TFIDFvectorize, captions, ensemble model, SVR.**

I. INTRODUCTION

Growth of internet caused the researchers to look for new techniques that will help in formulating and fetching the digitalized data. The issue is critical as internet media like social media, browsers and endorsement systems are growing enormously day by day [1]. Video features can be useful to helps to compare different types of contents [2]

In this research, I have used features like Caption, aesthetic and C3D features to build the memorability model. Video features like C3D and aesthetic features are outperformed by caption features. Each model is measured against Spearman's Correlation. Final results show short-term memory scores are better than long-term memorability scores.

A weighted average Ensemble technique across the three model to get more accurate model are used rather than using individual model for short-term and long-term memorability. By predicting using RNN, SVR, Random Forest to determine the memorability scores, I have tried to fit different model to get more polished results.

key findings are:

(1) Short term memorability outperforms models for long term memorability.
(2) Models on video captions, Aesthetics are less accurate when compared with models based on Caption data.

I. RELATED WORK

In this paper, selection of features vital role for model predictions. Features such as captions, aesthetic, C3D, color histogram, HMP, HOG, Inception, LBP, ORB. Using all these features in the model does not provide accurate results as there is high correlation between feature would cause the problem of model overfitting. Captions, C3D models out performed other feature models [3]. Results obtained by Tanmayee Joshi [4] uses caption features by vectorizing the features using TFIDF features.

Authors in this research are able to distinguish between real world positive and negative words. I am using thee same terms with positive and negative coefficient values to determine the effects of terms in caption-based model.[5]

II. APPROACH

A. Features pre-processing

1. Captions

Processing of captions with the use of NLTK library. All the special characters, numbers are removed from caption data. Once the all the special characters are removed, captions are converted into lower case format to make data uniform across the data set. Once the cleaning is done captions are converted into numerical format using TFIDFVectorizer. TF-IDF uses n-gram approach to vectorize the captions. After TF-IDF the corpus will be tokenized with Tokenizer. Along with the caption tokenized values, the positive and negative coefficient word are concatenated based on the insights provided in previous research papers [5].

2. C3D features

C3D video features have been used in the model. It Coordinate 3D format which has biomechanical data. It has values of object, 3-dimensional model such as, curves points, bodies.

3. Aesthetic Features

Aesthetic features have been used in the model. Aesthetics which include (shape, color, texture, symmetry, and proportion) which are converted to numerical form in the dataset.

B. Selection of Models

Based on the previous research done, I have used below models for video memorability prediction.

1. Recurrent neural network (RNN)
I have used RNN model to determine the video memorability. In this model Dense layer with 200 nodes. Activations codes like **'softmax', 'relu'** are used to help forward propagation of neurons. Model uses **'Adamax'** as the optimizer, **'mean_absolute_error'** as loss function. It also uses **[cosine proximity], [accuracy],[mae]** metrics to increase the efficiency of the model.

2. Decision Tree Regression
3. Support Vector Regression (SVR)
4. Random Forest Regression

C. Ensemble

I have used weighted average technique to ensemble the different model to get optimum predictions. From the analysis, the captions features outperformed the visual and aesthetic features. Based on applying different model on captions. RNN performed best for captions followed by SVR and Random forest.

For both long term and shot term memorability, Output from the all the three models combined to reduce the noise and provide optimum values. Below is the flowchart of the ensemble model

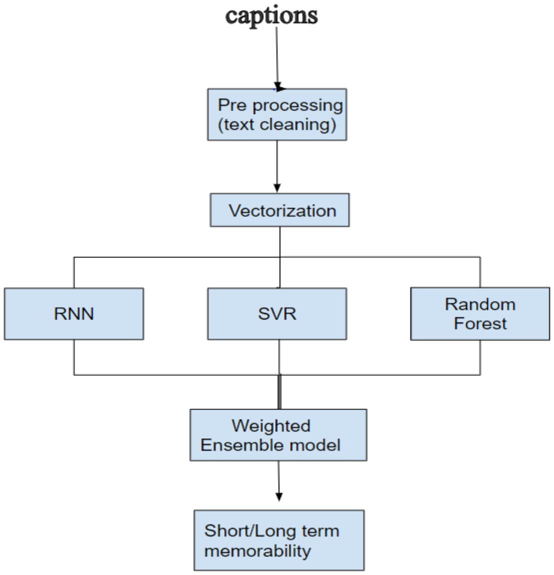

fig 1: Ensemble model flowchart

III. RESULTS AND FINDINGS

Tables 1 and 2 give an overall outcome of the research. Reading provided by different models are show in below tables Spearman's correlation are used to determine the model accuracy. Finally, best performing model in caption category are Ensembled to produce more accurate results.

(1) Short Term Ensemble: (0.60 * prediction_RNN) + (0.30*Prediction_SVR) + (0.10 * prediction RandomForest)

(2) Long Term Ensemble: (0.60 * prediction_RNN) + (0.30*Prediction_SVR) + (0.10 * prediction RandomForest)

Features	**RNN**	**SVR**	**Random Forest**
Captions	**0.442**	**0.427**	**0.419**
C3D	0.278	0.161	0.316
Aesthetic	-0.050	0.263	-0.006
Ensemble (caption)	**0.464**		

Table 1: Short-term memorability

Features	**RNN**	**SVR**	**Random Forest**
Captions	**0.174**	**0.179**	**0.171**
C3D	0.132	0.092	0.316
Aesthetic	0.002	-0.053	-0.027
Ensemble (caption)	**0.200**		

Table 2: Long-term memorability

IV. CONCLUSION AND FUTURE WORK

Research provide the insight on caption-based machine learning model and how it outperforms video and aesthetic feature. Short-term scores are high when compared to long term scores. Research shows the video features or aesthetic features cannot perform well in these models unless it is combined with any other features. It also describes the application of deep learning (neural network) features in predicting the video/image data. Finally, exploration shows Ensemble techniques like weighted average methods can help to combine different models to provide more noise free accurate hybrid models.

Future work includes increasing the model accuracy and shot/long term memorability scores. Exploration of effects of other features on the memorability scores. Finally, I aim to explore deep learning concepts like CNN which can enhance the predictive performance.

REFERENCES

[1] J. Han, C. Chen, L. Shao, X. Hu, J. Han, and T. Liu, "Learning computational models of video memorability from fmri brain imaging," IEEE transactions on cybernetics, vol. 45, no. 8, pp. 1692–1703, 2015.

[2] R. Cohendet, K. Yadati, N. Q. Duong, and C.-H. Demarty, "Annotating, understanding, and predicting long-term video memorability," in Proc. of the ICMR 2018 Workshop, Yokohama, Japan, June 11-14, 2018

[3] Cohendet, R., Demarty, C.H., Duong, N., Sjöberg, M., Ionescu, B. and Do, T.T., 2018. Mediaeval 2018: Predicting media memorability task.

[4] Aditya Khosla, Akhil S. Raju, Antonio Torralba, Aude Oliva, "Understanding and Predicting Image Memorability at a Large Scale".

[5] Rohit Gupta, Kush Motwani "Linear Models for Video Memorability Prediction"