Predicting Video Memorability Using Machine Learning Algorithms

Mitul Verma
19210961
MCM Computing (Cloud Computing)
Dublin City University
Dublin, Ireland
mitul.verma3@mail.dcu.ie

Abstract—Every day a large number of videos explode the internet. This explosion surges a need to create research methods for video analysis. Such a method must take human cognition into account. One of the most important features of human cognition is memorability. Memorability is defined as the ability to recall or remember visual content after watching it. Creating such a model helps us to understand the underlying functioning of the human brain while processing an image or videos. Such revelations help us to improve model performance and extending future possibilities in feature extraction of image and neural networks. In this work, I have tried to highlight important features for video analysis by predicting short-term and long-term memorability scores working with multiple features, videos, and semantic features, individually as well as in combination.

Keywords—HMP Features, C3D Features, Semantic Features (TF-IDF and Count Vectorizer).

I. Introduction

For my investigation, I have used various features for video analysis. Such features include HMP Feature, C3D Features, HOG Features and Semantics which include Captions. Each feature helps me to predict the short-term and long-term memorability score of videos up to some extent. The model accuracy is checked using spearman's correlation coefficient. The features with high spearman's correlation coefficient are selected and used in combination for further analysis to improve model performance. The dataset is split into a train and test set to build the Train model and then to test it. The model with the best accuracy score is used to predict memorability.

Important discoveries:

- Some features do not contribute anything in predicting memorability such as annotations present in ground truth CSV. Removing such features does not affect model accuracy.
- The Prediction for short-term memorability score is better as compared to the long-term memorability score for any given model.
- C3D feature performed well as compare to other video features.
- While Captions performed best out of all the features available.
- Semantic features consistently performed well when used with non-parametric models rather than parameter ones.

The Rest of the investigation is further divided into the following sections. Section II consists of the literature review of tremendous work done by people. Section III walks people through the path I followed providing details on ML Models, Data Preprocessing and Feature extraction. Section IV consists of results. Section V consists of a conclusion that puts light on future possibilities. Section VI consists of a brief list of references.

II. Literature Review

In [3], People from MIT use the LaMem dataset. They used hybrid-CNN and tuned it accordingly and achieved a co-relation rank of 0.64. In [2], A team working in Conduent labs, in India, uses HMP features, C3D and color histogram. They also use features extracted from images of frames of videos using CNN. They also add some weight to the captions and use SVR, LASSO and Elastic Net. But to improve their accuracy they built an ensemble of models using some of their best models. They used a simple weighted average technique to blend their previously obtained outputs. In [4], A CNN Model is used with 2-layer, each layer consists of a batch normalization, 2D convolution and max pool operation. Finally, they all are vectorized into a fully connected linear model. In [5], The features are divided into two categories, low level and high level. SVR is used for low features. While, CNN_LSTM and BI-LSTM are applied on high level features.

III. Approach

This section consists of an approach I used to produce good results.

A. *Models*

I chose one Parametric Model (Linear Regression) and Two Non-Parametric Model (Decision Tree and Random Forest Model) followed by Ensemble of Random Forest and neural networks:

1) *Linear Regression Model*
2) *KNearestNeighbours*
3) *Decision Tree Regression Model*
4) *Random Forest Regression Model*
5) *Neural Networks*

B. *Features and Data Pre-Processing*

Video features which include HMP and C3D were used to predict both short-term and long-term memorability scores individually. All three features performed poorly when used alone. Out of all the three C3D features performed well and hence selected for further analysis. **HMP** and **C3D** were individually used to predict memorability.

Semantic feature – When Captions were used, they performed very well as compared to Video Features. **Captions** gave better results compared to video features. For Captions, we used two very important features of Bag of Words. The first one is TF-IDF. The second method used was Wordcount. Using Corpus of stopword, downloaded using the NLTK library, we removed stopword from the captions. Then each caption, corresponds to a video, transformed into a TF-IDF score and supplied to the models in the form of vector of equal length. The same methodology was used for CountVectorizer. Where instead of the TF-IDF score, word count was used. Both the features were individually used to train the model and both short-term and long-term memorability was calculated. Using spearman's correlation.

the accuracy of both models was calculated. TF-IDF performed better than CountVectorizer. TF-IDF was selected for further analysis.

Video Features with Semantic features – In this Model, a combination of Video and Caption features were used. The C3D feature which performed well then rest of the video features is combined with TF-IDF which outperformed CountVectorizer. The new model containing both the features performed better than the C3D feature alone but failed to perform better then captions alone. Captions alone outperformed every other feature.

One of the most famous English idioms says that "Words are more powerful than guns". So, we try the same thing here. Some words contain extra weights than other words [2]. i.e. some words have a more positive impact, while some may have a strong negative impact than then rest of the words. So, during the calculation of the TF-IDF score, we add some weights to few words, which may have more impact than other words. We then used this dataset to train our model and later test it. Even for this model, we used the same set of machine learning algorithms. First, we use the linear regression model, followed by a decision tree and then the ensemble of Random Forest. For each model, the results were improved significantly. The best model was an ensemble of random forest. In the end, I used neural networks on my best model. The long-term memorability score finally improved.

After working with different model using different features, I analyzed that best features for video memorability is captions with TF-IDF Score + weights. The best model to use with such a feature is an ensemble of random forest for short-term scores and neural networks for the long-term. Hence, I used these two models for my final computation and the final results are saved in Predicted_Score.csv.

IV. Result

The Table below consists of results generated using different Models and Features. The Short-Term and Long-Term Memorability score is Spearmen's coefficient.

Feature	Model		Short-Term Memorability Score	Long-Term Memorability Score
Video feature: HMP features	a) Linear Regression		0.020	0.006
	b) KNN Regressor(n=67)		0.075	0.036
	c) SV Regressor		0.066	0.041
	d) Decision Tree		0.05	0.017
	e) Random Forest Regression	n_estimators =10	0.068	0.03
		n_estimators =100	0.285	0.095
C3D features	a) Linear Regression		0.27	0.112
	b) KNN Regressor(n=55)		0.208	0.065
	c) SV Regressor		0.195	0.032
	d) Decision Tree		0.123	0.050
	e) Random Forest Regression	n_estimators =10	0.317	0.146
		n_estimators =100	0.312	0.128
Semantic feature:	a) Linear Regression		0.130	0.034
	b) Decision Tree		0.242	0.060
	c)SV Regressor		0.392	0.166
	d) Random	n_estimators =10	0.382	0.122
Captions with TF_IDF Score	Forest Regression	n_estimators =100	0.371	0.133
Captions with wordcounts.	a) Linear Regression		0.127	0.058
	b) Decision Tree		0.245	0.068
	c) Random Forest Regression	n_estimators =10	0.348	0.134
		n_estimators =100	0.337	0.138
Semantic feature: Captions with weights (TF-IDF + Weight)	a) Linear Regression		0.125	0.127
	b) Neural Networks		0.403	0.243
	c) SV Regressor		0.419	0.160
	d) Decision Tree		0.287	0.142
	e) Random Forest Regression	n_estimators =10	0.445	0.200
		n_estimators =100	0.447	0.203
Video + Semantic: (C3D + Caption with TF_IDF + weights)	a). Support Vector Regressor		0.382	0.205
	b). KNN Regressor		0.327	0.167
	c). Random Forest	n_estimator =100	0.370	0.181
Video + Semantic feature: (C3D + Caption with TF-IDF score)	Random Forest Regression	n_estimators =100	0.312	0.158

Table 1. Results

V. Conclusion And Future Work

After working with a different set of features and using both parametric and non-parametric sets of machine learning algorithm I concluded that captions are a more important feature than others.

A simple addition of weights to only 16 term changes the result significantly. I think more emphasis should be given to this area of research. Not only this the combination of feature although did not perform better than the caption, but it did perform better than the rest of the model. Therefore, I also think that the right combination of feature and right model might provide as much need breakthrough in this area and we might achieve the desired the results.

VI. References

[1] Multimediaeval.org. 2020. Media Memorability. [online] Available at: http://www.multimediaeval.org/mediaeval2019/memorability

[2] K. M. Rohit Gupta, "Linear Models for Video Memorability Prediction Using Visual and Semantic Features," MediaEval-2018.

[3] People.csail.mit.edu. 2020. [online] Available at: https://people.csail.mit.edu/khosla/papers/iccv2015_khosla.pdf

[4] Chaudhry, R. and Kilaru, M., 2020. [online] Ceur-ws.org. Available at: http://ceur-ws.org/Vol-2283/MediaEval_18_paper_15.pdf

[5] Joshi, T. and Pedanekar, N., 2020. [online] Ceur-ws.org. Available at: http://ceur-ws.org/Vol-2283/MediaEval_18_paper_41.pdf

[6] Medium. 2020. How to Calculate TF-IDF (Term Frequency–Inverse Document Frequency) From The Beatles Biography In…. [online] Available at: https://iyzico.engineering/how-to-calculate-tf-idf-term-frequency-inverse-document-frequency-from-the-beatles-biography-in-c4c3cd968296

[7] Medium. 2020. First Neural Network for Beginners Explained (With Code). [online] Available at: https://towardsdatascience.com/first-neural-network-for-beginners-explained-with-code-4cfd37e06eaf

[8] Cohendet, R., 2018. Medieval 2018: Predicting Media Memorability.

ARC Center of Excellence for Robotic Vision, University of Adelaide, Australia, [online] Available at https://arxiv.org/pdf/1807.01052.pdf.