Video Memorability Prediction using Captions and Video Features

Rebekah Jennifer Manimaran (19210518)

Dublin City University

rebekah.manimaran2@mail.dcu.ie

ABSTRACT

Video Memorability is a media analytics prediction task where the probability of how much a human can remember a set of videos is analyzed and scores are determined. In this paper, captions of the video files and C3D features are used in machine learning algorithms to design a short-term and long-term memorability score. The performance of the model is compared with each other and the best-fit algorithm is identified for video captions and C3D.

KEYWORDS

Memorability, Machine learning, Random forest, K-Nearest neighbour

1. INTRODUCTION

Video Memorability is a challenge of MediaEval Predicting Media Memorability Task which predicts the memorability percentage of a video to the viewers. The is a public dataset that consists of 8000 soundless short clips divided into 6000 for development or training set and 2000 for the testing set. The ground truth of memorability also consists of a number of annotations that were used to calculate the short term and long-term scores. The titles of each video are termed as the video captions and there were several pre-computed features such as C3D, HMP to proceed with model building. This paper investigates C3D, Pre-extracted visual features and video captions to develop a model on various machine learning algorithms. And, finally combine both the features to obtain a single best-outperformed model. The Machine learning model is evaluated using a metrics called as spearman's rank correlation coefficient. This metric assesses the relationship between the predicted short term and long-term memorability scores with the actual ground truth value of the short-term and long-term scores.

RELATED WORK

In the work done by Romain et al [1], they have designed computation models for predicting the intrinsic features of the video clips from the visual content. The study involved the collection of 10000 soundless videos of 7 seconds [1] that are extracted from raw footage. The annotations for each video are collected with the human measure for the performance of short term and long-term score through recognition test [1].

The experiment done follows machine learning approaches [1] where 5 different models are developed for this memorability prediction of the standard regression problem. The first two models do Image memorability prediction using Memnet and pre-trained VGG16. They have been finetuned on ImageNet and LeMem dataset which is combined using MLP as a classifier [1]. The third model uses Image caption features to builder an encoder using CNN and LSTM [1] for retrieving the textual metadata embeddings of image title files. The fourth model used is ResNet18 and Resnet101 finetuned using CNN and data augmentation techniques. The final model developed is a 2-layer MLP semantic embedding model where it gives a better performance score of 0.565 (short-term) and 0.275 (long-term) than the other models using spearman's correlation metric.

In a study conducted by [2] for the MediaEval 2019 challenge, they have constructed an ensemble Machine learning model and Deep learning models with high regularisation. They have used support vector regression and Bayesian ridge regression models for traditional learning approaches that ensembles with video captions and pre-computed features. The challenge has introduced a set of features [3] such as video features (C3D and HMP), Image frame features (LBP, InceptionV3, Color histogram, aesthetics)

and Textual features (Image Caption). They have used transfer learning methods to extract their relevant features and fine-tune the neural network (CNN) based on Resnet architecture [2] for developing ensemble models on memorability score prediction. The score achieved on ResNet is 0.50 (short-term) and 0.20 (long-term).

APPROACH

The approach used in this paper is to extract relevant contextual meaning using Textual information and video-based information using NLP techniques and machine learning algorithms.

a. Machine Learning with Textual feature

Textual features used in this model building process are titles of the Images called Captions. To make the textual captions understandable to machines, it must convert into sequences of array numbers or matrices. The approach is to count the number of words in the caption file and use logarithms to assign weights for the sequence of array numbers. To perform this pre-processing of Captions files are implemented using Stop words removal, Punctuation marks removal and Lemmatizing the words using WordNet to its root form. Later, cleaned textual files are fed into the TFIDF vectorizer where the collection of raw titles is converted to matrixes of TFIDF features. An argument is passed into the vectorizer that calculates the inverse document frequency. The result of IDF allows the most frequently occurring words to have low weights and rare words to assign high weights. Also, another argument that does the smoothening for the caption files is passed where every feature word is normalized and standardized for the unknown test datasets. Thus, better contextual information is retrieved for scoring the short-term and long-term.

Using the above features and vocabulary as the caption input files, Linear regression, Decision Tree Regression, K-Nearest Neighbors regression and Random Forest regression are used to train and predict the short-term and long-term memorability scores using Spearman's Correlation Coefficient. The results of the model were KNN regression showed a better score of **0.417** for short-term and Random Forest of **0.168** for long-term.

b. Machine Learning with Video-based feature

In Video-based features, C3D is a spatiotemporal visual feature that is from a 3-Dimensional Neural network model. The final classification layer of NN is the output feature of a single list of 101 dimensions numbers.

In the C3D video features, different machine learning algorithms are used to predict the probability of remembering the video. Linear regression is used with a normalization technique for training and prediction on the test set. Decision Tree regression with Friedman mean square error method is used for training and prediction of memorability. K-Nearest Neighbors regression technique is used where the weights are assigned to the distance points. This makes the closer relevant feature point to give more importance than the weights which are further away. The results of the designed model show that Random forest regression for both short-term and long-term have better scores of **0.327** and **0.151**.

c. Machine Learning Combining Textual and Video Feature

The textual caption features and video based C3D features are combined to perform effectively. After careful analysis, K-Nearest Neighbor regression is used for the training of the 6000 video files and testing using the 2000 video files. The training set is divided into train and validation to get the Spearman rank correlation value. To assign the number of neighbors in KNN regression, visualization is done by giving the range from value 20 to 30 where it shows that n_neighbors = 26 is the model that does not overfit with training and testing accuracy. The short-term memorability score is **0.381** and the long-term memorability score is **0.169**.

2. RESULTS AND ANALYSIS

Model	Linear Reg	DTree Reg	KNN Reg	RForest Reg
Caption + tfidf + Lemma-tizer	0.142	0.300	**0.417**	0.415
C3D feature	0.278	0.117	0.220	**0.327**
Caption +C3D	-	-	**0.381**	-

Table 1: Spearman's Correlation for short-term memorability

Model	Linear Reg	DTree Reg	KNN Reg	RForest Reg
Caption + tfidf + Lemma-tizer	0.080	0.110	0.158	**0.168**
C3D feature	0.119	0.055	0.063	**0.151**
Caption +C3D			**0.169**	

Table 2: Spearman's Correlation of long-term memorability

From the above Table 1 and Table 2, the different approaches used to predict the short-term score and long-term scores are shown to analyze the efficiency of each method. The results show that Textual features like captions perform better than video specialized features. Also, long-term memorability does not provide the same results as short-term memorability. This could be because long-term memorability depends on the human brain of an individual's memory.

3. CONCLUSION AND FUTURE WORK

The approach was on the Caption and C3D features, therefore further analysis and investigations can be done for the other pre-computed features such as HMP and ColorHistogram. Thus, Future work can also be improved on the short-term and long-term memorability scores using Neural network models and Pre-trained architectures of CNN.

REFERENCES

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

[2] Azcona, David, Enric Moreu, Feiyan Hu, Tomás Ward, and Alan F. Smeaton. "Predicting media memorability using ensemble models." CEUR Workshop Proceedings, 2019.

[3] Mihai Gabriel Constantin, BogdanIonescu, Claire-Hélène Demarty, Ngoc Q.K. Duong, Xavier Alameda-Pineda, and Mats Sjöberg. 2019. The Predicting Media Memorability Task at Mediaeval2019. (2019).