Quest for Insight: Predicting Memorability based on Frequency of N-Grams

Muhammad Mustafa Ali Usmani[1], Sumaiyah Zahid[1] and Muhammad Atif Tahir[1]

[1]National University of Computer and Emerging Sciences (FAST-NUCES), Karachi, Pakistan

Abstract

With every passing minute the video content being generated is multiplying exponentially. To filter and analyze this content in order to keep the relevant videos, we take the help of memorability scores. In this paper we discuss a new approach to calculate the memorability of the video clips using the captions that are provided as descriptions. This quest by FAST-NUCES team is based on basic text normalization techniques, and memorability is predicted using both unigrams and bigrams. Results show that bigrams give a better accuracy as compared to unigrams as they add more context and meaning to the model. It was also inferred that the frequency and memorability of the words were inversely correlated.

1. Introduction

The world has progressed in such a way that we see media all around us all the time. This media can come in the form of text written in newspapers, images on billboards, video advertisements on television, or music on the radio, to name a few. We continually perceive a lot of the media through different senses and the brain decides if we want to remember it or get rid of it. Since there is a huge amount of content generated continuously that is floating around the brain needs to filter what is important enough to remember. Memorability is a measure of how likely any media is to be retained by the brain.

Memorability can be affected by a lot of factors, the most important of which are the features present in the media. Other things that can affect memorability can be the association of the media with the person consuming it, or its cultural or religious importance to a particular group of people.

In the MediaEval challenge [1] the goal is to predict video memorability. As a side quest we have used the text captions associated with videos to predict their influence on the memorability of a video. The prediction of memorability will enable a system to identify if the video is relevant and can be used for different purposes such as education, summarizing or storytelling.

2. Motivation

Explainable artificial intelligence aims to decipher and decode the black box that is machine learning. Using this concept as the motivation, after building the models for memorability prediction, we move towards finding the factors that contribute most to the memorability of a video.

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online

*Corresponding author.

†These authors contributed equally.

mustafaausmani@gmail.com (M. M. A. Usmani); sumaiyah@nu.edu.pk (S. Zahid); atif.tahir@nu.edu.pk (M. A. Tahir)

Video captions are analyzed to see what words independently contribute to the most memorability. Furthermore, the captions are used to calculate the memorability score of the videos. A comparison of off-the-shelf models is also presented as a foundation for more complex models.

3. Approach

3.1. Data and Preprocessing

The Memento10k [2] dataset provided by the organizers was used to conduct the experiments. This dataset consists of 10,000 three second videos with their memorability scores. Each video clip has 5 human generated textual captions associated with it.

To make the captions processable by the model and reduce randomness, we normalize it. The captions are converted to lowercase then the stopping words and punctuations are removed. The words in the captions are then lemmatized to reduce them to their base words. The captions are broken down into pairs of words called bigrams [3] or are tokenized into singular words called unigrams.

3.2. Data Mining

After preprocessing, the frequency of each pair of words in the captions is computed and stored in a dictionary. This pair will be referred to as a bigram. Each time the bigram appears in a caption, the associated memorability is added to the total memorability score of the bigram. At the end the total score of each bigram is divided by its frequency, this gives the average memorability score of the bigram. The importance of the pair is given by

$$Importance = 1 - (objectoccurrences/totalobjects)$$

To get the final memorability score we multiply the average memorability of the bigram by its importance. By multiplying the average memorability score by the importance we ensure that the percentage of the bigram in the whole dataset is considered while computing the final score. As a way of comparing the results obtained on bigrams and to see what value a pair of words brings, we also apply the above outlined procedure on unigrams, those are single words.

3.3. Exploratory Data Analysis

The dataset and n-grams was explored to get a summary of its main characteristics and see how each statistic affects the memorability. The dataset was analysed and explored while keeping in mind the memorability scores that were provided in the dataset.

3.4. Predicting Memorability using a caption

The caption whose memorability we want to predict, known as the query caption, is preprocessed in the same way as the data. Bigrams are created from this caption. For each bigram, its Jaccard similarity [4] with every other bigram in the dataset is computed.

Jaccard similarity can have values ranging from 0 to 1. Where 0 indicates no similarity and 1 indicates the most similarity. After the similarities with each bigram in the dataset are obtained, the most similar bigram is chosen and its memorability is added to the total memorability of the query caption.

In the same way the most similar memorability for each bigram in the query caption is obtained by using jaccard similarity. All the obtained memorabilities are added and divided by

Table 1
Summary of bigrams

Bigram Top Frequency	Bigram Top Score	Bigram Bottom Score
'group', 'people'	'girl', 'lip'	'race', 'tricycle'
'two', 'people'	'lip', 'pierce'	'mountain', 'peak'
'little', 'girl'	'mumble', 'something'	'game', 'soccer'

Table 2
Correlation coefficient and error on unigrams and bigrams

n-gram	Spearman	Pearson	RMSE
uni-garm	0.293	0.262	0.018
bi-gram	0.45	0.43	0.02

the total number of bigrams in the query caption to give the predicted memorability score of the query caption.

4. Results and Analysis

As a result of the exploratory data analysis, some statistics were summarized. It was seen that in amongst all colors white appeared 2394 times throughout the dataset. Pink color had the most memorability and blue had the least memorability. The frequency of men and women is 5097 and 2625 respectively, with women having a higher memorability score. Car had the most occurrences in the dataset. Amongst the vehicles van had the highest memorability while helicopters had the lowest memorability.

From the analysis of the dataset it was seen that the bigram ['group', 'people'] had the highest frequency in the dataset, and ['girl', 'lip'] was the most memorable bigram from the dataset. Table 1 lists some other bigrams that had high frequencies and memorability scores. Conversely it also lists bigrams that have the worst memorability scores.

Analysis of the unigram showed that man and people were the single most occurring words in the dataset while sweaty, hotdog and artillery were some of the most memorable words in the dataset.

It was observed that memorability score and frequency of the word were inversely proportional, i.e. rare words and objects are more memorable. The correlation of frequency and memorability score came out to be -0.30.

The prediction model using captions was tested on both unigrams and bigrams; a significant decrease in error was observed when bigrams were used. Here the memorability scores that came with the original dataset are used as ground truth. This is used to calculate the error and accuracy with respect to the predicted memorability score of the caption. It was perceived that the better accuracy when using bigrams was because they add more meaning and context as compared to unigrams. (Table 2)

For further analysis as a separate quest, some off the shelf machine learning models were applied on the features that were provided by the organizers. The models were trained, tested and validated on the same dataset, but video features instead of captions. The results were then compared and captured. AlexNet, VGG and ResNet gave the best memorability scores compared to features extracted from any other model. On the other hand it was seen that AdaBoost[5] gave the best performance on the features, and Linear Regression the worst. The

Table 3
Spearman correlation coefficient of different models using different features on validation set

Features	Linear Regression	Ridge Regression	KNN	Random Forest	XGBoost	AdaBoost
ResNet	0.327	0.342	0.515	0.506	0.501	0.521
VGG	0.424	0.432	0.472	0.481	0.538	0.537
DenseNet	0.42	0.43	0.48	0.499	0.504	0.503
AlexNet	0.098	0.103	0.511	0.531	0.537	0.55
EfficientNet	0.337	0.339	0.404	0.481	0.488	0.488

best performance given by AdaBoost is owed to the ensemble techniques used in the model.

5. Discussion and Outlook

It was seen that frequency is inversely correlated to memorability score which means that rare words and objects are more memorable. In the dataset it was observed that some words that occurred only once had a memorability score of 1.0.

Bigrams were better than unigrams at predicting the memorability scores, as they provided more meaning. Bigrams also performed better than the regression techniques. A potential problem with prediction using text can be that long sentences will lose their context and meaning.

From the results we can conclude that AdaBoost gives the best results on features from all models, and features from AlexNet give the best result as compared to other features. AdaBoost gives the best results beacuse it used ensemble techniques to reduce overfitting and increase the accuracy.

6. Acknowledgement

This work was supported in part by the Higher Education Commission (HEC) Pakistan, and in part by the Ministry of Planning Development and Reforms under the National Center in Big Data and Cloud Computing.

References

[1] L. Sweeney, M. G. Constantin, C.-H. Demarty, C. Fosco, A. García Seco de Herrera, S. Halder, G. Healy, B. Ionescu, A. Matran-Fernandez, A. F. Smeaton, M. Sultana, Overview of the MediaEval 2022 predicting video memorability task, in: MediaEval Multimedia Benchmark Workshop Working Notes, 2023.

[2] A. Newman, C. Fosco, V. Casser, A. Lee, B. A. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, CoRR abs/2009.02568 (2020). URL: https://arxiv.org/abs/2009.02568. arXiv:2009.02568.

[3] C.-M. Tan, Y. fang Wang, C. Lee, The use of bigrams to enhance text categorization, Inf. Process. Manag. 38 (2002) 529–546.

[4] S. Niwattanakul, J. Singthongchai, E. Naenudorn, S. Wanapu, Using of jaccard coefficient for keywords similarity, 2013.

[5] T. Chengsheng, L. Huacheng, X. Bing, Adaboost typical algorithm and its application research, MATEC Web of Conferences 139 (2017) 00222. doi:10.1051/matecconf/201713900222.