Prediction of Short-Term & Long-Term Video Memorability

Author

Amna Ashiq

Registration Number

00000203032

Supervisor

Dr. Syed Omer Gilani

DEPARTMENT OF ROBOTICS & ARTIFICIAL INTELLIGENCE
SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD

August 2021

Prediction of short-term & long-term Video Memorability

Author

Amna Ashiq

Regn Number

SMME-RIME17-203032

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Robotics & Intelligent Machine Engineering

Thesis Supervisor:

Dr. Syed Omer Gilani

Thesis Supervisor's Signature:

____

Department of Robotics & Artificial Intelligence
SCHOOL OF MECHANICAL & MANUFACTURING ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD
August, 2021

FORM TH-4

National University of Sciences & Technology

MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: **Amna Ashiq** with **Registration # 00000203032** Titled**: "Prediction of Short-Term and Long-Term Video Memorability**" be accepted in partial fulfillment of the requirements for the award of **MS Robotics & Intelligent Machine Engineering** degree.

Examination Committee Members

1. Name: Dr. Hasan Sajid Signature:_______________

2. Name: Dr. Yasar Ayaz Signature:_______________

3. Name: Dr. Asim Waris Signature:_______________

Supervisor's name: Dr. Syed Omer Gilani Signature:_______________

Date:___________________

______________________ ______________________

Head of Department Date

COUNTERSINGED

Date:__________________ ______________________

Principal

Thesis Acceptance Certificate

It is certified that the final copy of MS Thesis written by Amna Ashiq (Registration No. 00000203032), of Department of Robotics and Intelligent Machine Engineering (SMME) has been vetted by undersigned, found complete in all respects as per NUST statutes / regulations, is free from plagiarism, errors and mistakes and is accepted as a partial fulfillment for award of MS Degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in this dissertation.

Signature: ___________________

Name of Supervisor: Dr. Syed Omer Gilani

Date: __________________

Signature (HOD): __________________

Date: __________________

Signature(Principal): __________________

Date: __________________

Declaration

I certify that this research work titled "*Prediction of short-term and long-term video memorability*" is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources has been properly acknowledged/referred.

Signature of Student

Amna Ashiq

2017-NUST-MS-RIME-203032

Plagiarism Certificate (Turnitin Report)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

Amna Ashiq
203032

Dr. Syed Omer Gilani
(Supervisor)

Copyright Statement

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST School of Mechanical & Manufacturing Engineering (SMME). Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST School of Mechanical & Manufacturing Engineering, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the SMME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST School of Mechanical & Manufacturing Engineering, Islamabad.

Acknowledgements

I am thankful to my Creator Allah Subhana-Watala to have guided me throughout this work at every step and for the provision of knowledge, health and the strength to finish the research

I am profusely thankful to my beloved parents for their prayers, encouragement and moral support throughout my education.

I would also like to thank my supervisor Dr. Syed Omer Gilani, for his help throughout my thesis.

I would also like to pay special thanks to Dr. Hasan Sajid for his tremendous support and cooperation. I would also like to thank Dr. Yasar Ayyaz and Dr. Asim Waris for being on my thesis guidance and evaluation committee.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Dedicated to my exceptional parents and adored siblings whose tremendous support and cooperation led me to this wonderful accomplishment.

Abstract

With the explosion of user-generated video material on sites like Facebook and YouTube and e-learning platforms like Coursera, Udacity, and Udemy, new approaches for categorizing, annotating, and retrieving digital information are needed to make it more valuable in our daily lives. As a result, cognitive measure prediction, such as memorability, offers a wide range of possible uses. The attribute or state of being easy to recall is referred to as memorability. Memorability, like other important video cues like aesthetics or interestingness, can be used to assist people in choosing between otherwise similar videos. Therefore, a wide range of applications, such as Knowledge and training, content retrieval, summarization of content, story narration, online advertising, content promotion and screening, would benefit from systems able to rank videos according to their memory. In the proposed method, visual and semantic features have been used to train different Machine Learning and Deep learning models on the dataset of 10,000 soundless videos provided by the MediaEval Benchmarking Initiative to predict short and long term video memorability scores. According to the findings, short-term memorability is more predictable than long-term memorability since all models scored higher in short-term memorability than long-term memorability. The best results have been achieved with RNN using video captions and embedding.

Key Words: *aesthetic, interestingness, RNN, LSTM, TF-IDF Vectorizer, Word Embedding, Ensemble Learning*

Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

In recent years, media memorability has been intensively researched, and it has played a massive role in examining human perception and comprehension of the media content. The ability to retain and recall visuals content with incredible detail, often just after a single viewing, is a fundamental element of human cognition. Even though we are exposed to a large number of images and visual information every day, our memorability is capable of retaining a large number of objects with details from the content we have seen. However, not all information is kept and recalled in the same way; although some content can burn itself into our brain, others are swiftly forgotten.

Personal context and subjective consumption have an impact on the memorability of visual content. Our brain evolved to recall only the knowledge that is necessary for our survival. Memorable and forgettable content have distinct visual characteristics, making some information easier to recall than others. This shows that people naturally encode and discard the same information, regardless of their own experiences. Pictures of people, important actions and events, or central objects, for example, are more memorable than natural landscapes. Consistently forgotten content appears to lack objectivity and fine-grained representation in human memory.

The study of memorability from the computer vision perspective is a new area of research, driven on by the breakthrough of image memorability. The challenge is essential because media sites such as search engines, social networking sites, and recommendation systems continuously dealing with massive amounts of digital information. Memorability, like other important video cues such as aesthetics or interestingness, can be used to assist people in choosing between otherwise similar videos. As a result, systems that can rate movies based on their memorability will be useful in various applications, such as knowledge and training, information retrieval, content summarization, narration, online advertising, content distribution, and screening.

Despite its promise to be an active research topic in computer vision, video memorability prediction has a few significant challenges. One is that no precise definition of VM has been established in prior attempts to anticipate it, nor has a consistent and uniform mechanism for measuring it been defined. In contrast to image memorability, video memorability is a tricky problem due to the added complications of video. Videos also convey a wide range of visual notions to the consumer, making it difficult to determine the entire content's memorability.

Memorability (after ~2 days): 97%

Memorability (after ~2 days): 20%

In this research, we delve into the usage of various visual and semantic features to predict video memorability and conduct a thorough analysis of the features provided with the dataset. Our initial experiment was designed to identify the most important features, and it served as the foundation for our video memorability prediction engine. We examined all of the features included with the dataset, evaluated the performance of our deep learning network on these features, and demonstrated the significance of various features. We chose the video and semantic features for our second trial and eliminated the image features based on their performance. After that, we trained different deep learning models on a development dataset using video and semantic features. The best-performing systems from the development phase are then chosen, retrained on the entire devset using the best parameters, and then tested on the testset data. Finally, we created a model ensemble that outperforms models created with a single feature type. We have used Spearman's rank correlation (ρ) as the official evaluation metric for our models.

CHAPTER 2: LITERATURE REVIEW

2.1 What is Human Memory?

Our brains are bombarded with an enormous amount of information about ourselves and the environment around us from the instant we are born. The process of obtaining, preserving, keeping, and retrieving information is known as memory; it's our ability to recall and retrieve the observed information. Memory is an essential aspect of how we see the environment. Human Memory is being studied for thousands of years by scientists and philosophers, and it has now become one of the most popular areas in cognitive psychology.

From remembering significant events to finishing tasks and accomplishing objectives, human memory is a robust brain process that impacts many areas of our lives and how we view things. Humans have a plethora of memories that they keep for various amounts of time. Long-term memories last years, but short-term memories span seconds to hours. We also have an active memory, which allows us to repeat events to recall them for a limited time frame. We use our working memory whenever we repeat certain information to ourselves to remember it.

Another approach to classifying memories is by the memory's subject and whether or not we are constantly mindful of it. Declarative memory, often referred to as explicit memory, refers to memories that you have consciously. Some of these memories are factual statements or "general knowledge," such as Korea's capital (Seoul) or the number of cards in a conventional deck of playing cards (52). Others are made up of memories from your history, such as your graduation date.

Non-declarative memory, often known as implicit memory, accumulates unintentionally. Procedural memories are used by your body to remember the skill sets you've learned. Do you ride a bike or play an instrument? At work, those are your procedural memories. Non-declarative memories can also alter your body's automatic responses, such as drooling when you see your favourite meal or stiffening up when you see something you're afraid of.

Declarative memories are generally easier to form than non-declarative memories. Knowing a country's capital is easier than learning to play the violin. Non-declarative memories, on the other hand, tend to linger around longer. Bicycle riding is something you're not likely to forget once you master it.

2.2 How memories are formed, stored, and recalled

Since the 1940s, scientists have assumed that memories are stored in cell assemblies, groupings of neurons or nerve cells. Those interlinked cells ignite as a group in reaction to a particular stimulus, such as the glimpse of a mother's face or the aroma of freshly made bread. More neurons firing together means more intercellular connections. When a prospective stimulus triggers the cells, the entire assembly is more likely to fire; we experience the nerves' collective activity that transcribes memory.

Memory consolidation is a process that allows a short-term memory to become a long-term memory by strengthening short-term memory for long-term storage. Several techniques are thought to be involved in consolidation. One such process is in which individual nerves alter themselves to develop and communicate with their surrounding nerves differently in a process known as long-term potentiation. This remodeling changes the connections between nerves in the long term, which stabilizes the memory.

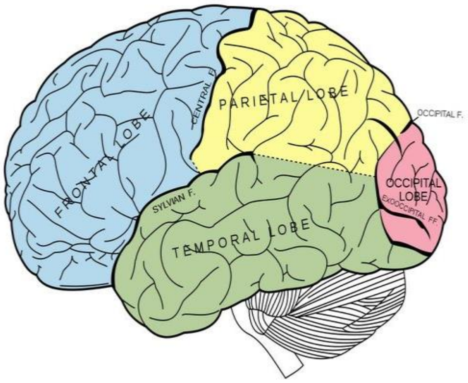


Figure 1: Human Brain

2.3 Memory Types

In the brain, memory is divided into three types: long-term memories, sensory memories, and short-term memories.

2.3.1 Sensory Memories

Originated from sensory organs, sensory memories are the ones that or usually kept for tiny periods. They are generally held for less than 500 milliseconds. Visual information is referred to as iconic memory. The raw data that the brain receives from the eye is known as visual memories (through the optic nerve). We store and process sensory memories without even acknowledging them – that is, without making a deliberate attempt to do so. Attentive processing is the term used to describe how this information is handled. It's a form of image processing that concentrates on a small number of picture elements, such as hues, patterns, tilt, bending, sharpness, and so on, rather than attempting to decipher the full image received.

2.3.2 Short-Term Memories

Short-term memory stores the information we are currently subconscious of or thinking about, which is also known as rational awareness in Freudian psychology. When attention is being paid to sensory memories, they create information in short-term memory. Even though most short-term memories fade quickly, paying attention to them allows them to move to the next level of memory: long-term memory. Active memory tends to store data for only 20 to 30 seconds most of the data.

2.3.3 Long-Term Memories

The memories that are transmitted to our short-term memory are usually lost fast. This is most likely a positive thing. We might easily get overloaded with information if we didn't remember the massive amounts of data we see on a daily basis, and digesting it in a meaningful way would become impossible.

The technique of keeping information indefinitely is known as long-term memory. Long-term memory is also described as subconscious and intuitive memory in Freudian psychology. Although much of this information is concealed from our conscious awareness, it may be recalled and preserved in working memory for later use. Some of this stuff is easy to remember, while others are considerably more challenging.

2.4 Image Memorability

Imagine a world where humans couldn't retain visual memories. To identify individuals, locations, and objects, the great majority of the population depends largely on visual cues. People have an exciting tendency to recall and forget the same visuals, despite having had distinct personal experiences [1, 12]. The capacity of methodologies to evaluate the memorability of a particular video is the subject of this research study. A good memorability prediction might be beneficial in various applications, including enhancing educational materials, saving for the items that we tend to forget, creating memorable advertisements, and even displaying visuals in a more digestible format for the general public.

In computer vision, the memorability of multimedia such as photos and recordings has lately been a hot topic. Information extraction and recommendation engines need a lot of power to keep up with constantly expanding data in today's fast-paced environment. Hence, in such media systems, the capacity to grasp the material plays a crucial role in helping them improve processing. When it comes to comprehending information, many elements such as memorability, visual saliency, Image aesthetics [1, 2], social popularity, engagement, and relatability [8, 9] may play a significant role. Visual memory has been extensively researched in psychology for decades, even though image memorability has just lately been explored in computer vision. Researchers have shown that people may recall thousands of photographs they've only seen once, even after being exposed to numerous more comparable images [10, 11]. There are a variety of elements that might influence how well an image is recalled, including intrusive visual appearance and the context of the user [8]. Although the user's context is likely to impact memorability, there is a broad agreement (consensus) across groups of individuals on how to judge a given image [12] Further research is needed to determine how memorable a given image is to the average observer.

There are inherent and extrinsic features that make an image memorable [9, 13, 14] in the field of computer vision. There is no link between memorability and color according to basic picture characteristics derived from pixel statistics or object statistics. Also, beauty and interestingness, which are subjective ideas, are not associated with memorability. Object and context meanings, on the other hand, have a strong link to memorability. Furthermore, tests in [9,16] revealed that human consistency was enough when annotating visual memory, proving the possibility of predicting image memorability. A key observation is that both the type of scene and objects in the image are highly related to its memorability.

Several deep learning methods for predicting memorability have been developed over time using the large-scale dataset LaMem and MemNet. MemNet [5] from MIT have made substantial progress in predicting image memorability (IM). Using deep learning, visual attention, and recurrent neural networks, Fajtl et. al. [3] developed a technique in 2018 that predicted memorability with a near-human consistency level on this dataset. Authors' Copyright is owned by the owner/author in [8]. (s).
There's little doubt that deep learning has surpassed human consistency with = 0.72 at MediaEval'19, 29-31 October 2019, Sophia Antipolis, France.

2.5 Video Memorability

The research in image memorability has recently been expanded to videos [3, 7, 15], following the rapid growth of the research area of image memorability [2, 5, 10, 11]. The need for innovative ways to help store and retrieve digital information, making it more valuable in our everyday lives, is a major motivator for VM prediction. Because social media platforms such as Facebook, YouTube and Twitter, search engines such as Google search, and recommendation systems are continuously dealing with the explosion of content daily, this research problem has become essential. Like other video features such as aesthetics and interestingness memorability can also play an important role in choosing from available content. It can play a huge role in learning, video summarization, information retrieval and video filtering.

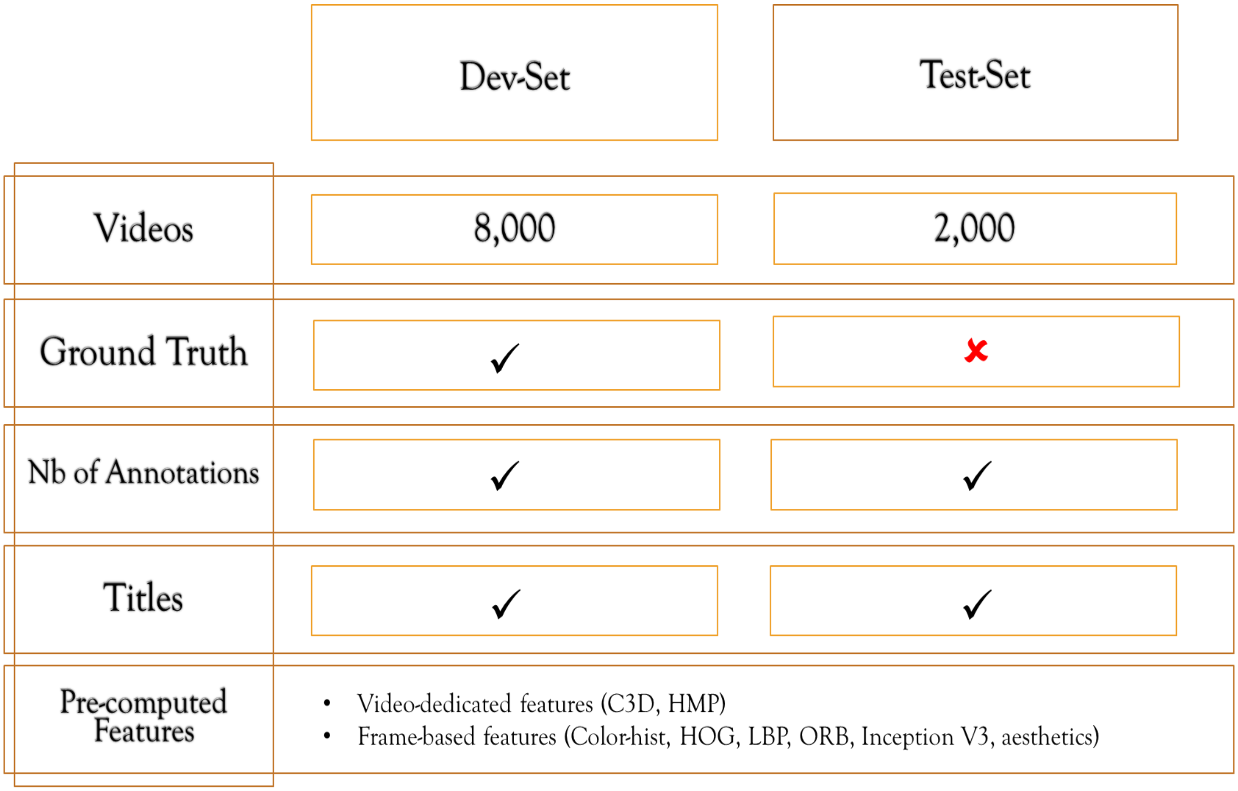
In recent years several techniques have been used to perform video memorability. The researchers have been agreed that video semantics from this research it has been observed that short term memory is more predictable than long term memorability.

CHAPTER 3: DATASET

The dataset consists of 10,000 soundless videos that can be used and redistributed under the MediaEval license for research purpose. Ten thousand videos were split into 8,000 developments and 2,000 test set. They were taken from the raw footage of professional content creator's. Each video is 7s long and feature a variety of scenarios.

The dataset includes the number of annotations and ground truth for each video sample for both long term and short term task. Each video also has its own caption. These captions are often interpreted as a collection of tags (textual metadata) that can help determine video memorability. The dataset also includes its own set of pre-extracted features. It includes some frame-based features which were retrieved by evaluating the first, middle, and end frames of each video.

Frame-based features include the Local Binary Patterns (LBP) [12], Histogram of Oriented Gradients (HoG) [8], ORB features and Color Histogram in HSV space. The fc7 layer from InceptionV3 was also extracted. Color, texture and object-based descriptors are represented by Aesthetic Visual Features (AVF), which aggregate the mean and median values taken from every 10 frames in a clip. And pre-extracted video-level features that examine the video as a whole and naturally describe the movement in these videos and also known as motion and temporal features this includes HMP Histogram of motion patterns and C3D the output of the final classification layer of the convolutional neural network.



	Dev-Set	Test-Set
Videos	8,000	2,000
Ground Truth	✓	✗
Nb of Annotations	✓	✓
Titles	✓	✓
Pre-computed Features	• Video-dedicated features (C3D, HMP) • Frame-based features (Color-hist, HOG, LBP, ORB, Inception V3, aesthetics)	

Table 1: Dataset Description

3.1 GROUND TRUTH AND ANNOTATION PROTOCOL

Annotations for short-term and long-term memorability were produced using performance tests. A collection of target samples (clips repeated after a specific amount of time) and filler samples were shown to the participants in these tests. In the short-term phase, participants were shown 40 target clips that reappeared in the testing phase and 140 filler clips that were only shown once, for a total of 180 videos in the short-term phase. They were next shown 40 videos from the prior filler collection and another 120 new ones, for a total of 160 videos in the long-term phase. A random interval ranging from 45 to 100 videos is used to determine how many times the repeat will occur. Participants were required to press the space bar whenever they believed that a video clip was repeated. For short-term memorability scores, the number of annotations is higher than for long-term memorability scores due to the difficulties of collecting data after a long wait through crowdsourcing. In the short-term and long-term recognition tasks, each video got an average of

38 and 13 annotations, respectively. The dataset provider gave the number of annotations for both tasks for each video clip in the development.

3.2 Memorability scores calculation

For both short-term and long-term memory performance, the dataset supplier provided each video an initial memorability score, which was defined as the percentage of correct detections by participants. The percentage scores are shown as floats between 0 and 1.

Correcting/normalizing the short-term raw scores involves applying a linear transformation that considers the memory retention length. A linear adjustment has been made since it has been discovered that memorability declines linearly with increasing retention period. Nevertheless, note that the applied adjustment has little influence on the memorability score, both absolute and relative terms. There was no adjustment applied to long-term scores, however. Gathered scores have shown no correlation between retention length and long-term memorability. According to the test procedure, the second measurement was taken 24 to 72 hours following the first. There's no reason to believe that the memory performance will suffer any more degradation from the retention period after such a long retention period.

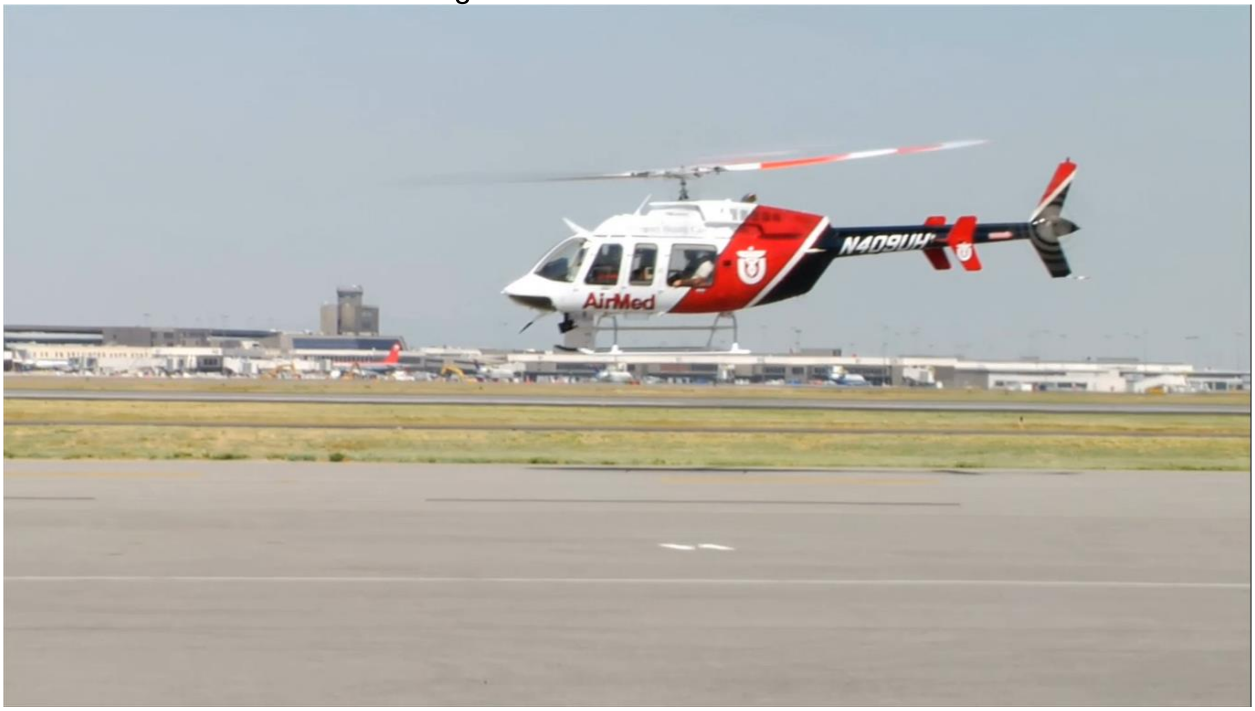
3.3 Features:

The dataset comes with a set of pre-computed features, which includes semantic and visual features. Each feature is organized in a different folder, one per feature.

3.3.1 Semantic Features

Each video comes with its caption. These captions are often interpreted as a collection of tags (textual metadata) that can help determine video memorability.

Figure 2: Dataset Illustration



Caption: medical-helicopter-hovers-at-airport

Caption: couple-relaxing-on-picnic-crane-shot

Caption: woman-eating-healthy-lunch-in-the-restaurant

Caption: student-hanging-out-outside-of-school-pan

Caption: steadcamof-two-healthy-men-peddling-free-wheeling-down-hill-with-cycling-road-bicycle-at-sunset

3.3.2 Visual Features

The visual features are inspired by different properties of images such as saliency and aesthetics. We assumed that memorability of the video is affected by the properties of images comprising the video. It is composed of both video-based and image-based features.

Video specialized features

Two video-specific features have been provided with the dataset C3D and HMP.

C3D features: C3D represents the output of the final classification layer of the C3D model. It comes in the format of a text file one per video, which consists of a list of numbers on one line(dimension=101).

HMP: It represents the histogram of motion patterns for each video consisting of a single list of pairs of numbers with the format: bin: number (dimension = 6075) on one line.

Image-Based features

These image based features were extracted on three key-frames (first (0), middle (56) and last (112)) for each video. There are three files per video, with names video nb-0.txt, video nb-56.txt, video nb-112.txt.

HoG: Histograms of oriented gradients are calculated on 32x32 windows on a greyscale image. It comes in the form of a text file where each file consists of a single list of numbers on one line (dimension = depends on the image size)

LBP: Local Binary Patterns represent local texture information. Its values are calculated for of 8x15 pixel's patches. It comes in the form of a text file where each file consists of a single list of numbers on one line (dimension = depends on the image size).

InceptionV3: Inception V3 is the output of the fc7layer of the InceptionV3 deep network. It comes in the form of a text file, where each file consists of a single list of pairs of numbers with format imagenet_class: activation (max dimension = 1,000).

Color Histogram: classic color histogram (three channels) comes in the form of a text file where each file consists of 3 lists (Red, Green, Blue in that order) of 255 pairs with format bin: number, e.g., 254:1008. One list per line.

CHAPTER 4: METHODOLOGY

4.1 Features

4.1.1 Convolution 3D (C3D) -- Generic features generated for video analysis.

C3D is generated by using a huge annotated video dataset to train a deep 3D convolutional network. Objects, activities, settings, and other commonly recurring categories in videos are all included in the dataset.

4.1.2 Histogram of Motion Patterns (HMP)

For each video, a histogram of motion patterns is generated. As the motion continues, HMP provides a static image template that aids in comprehending the motion location and route. The temporal motion data is condensed into a single image template, where intensity is a function of motion recency, with brighter values indicating more recent motion. In computer vision, a color histogram is a representation of the distribution of colors in an image. In digital images, a color histogram depicts the number of pixels with colors in each of a specified set of color ranges. The human mind is greatly influenced by color and its distribution. In the HSV space, this is computed using 64 bins in each color space for 3 key-frames for each video.

4.1.3 Saliency

Aspects of visual information that grab human attention are known as saliency. Computing visual saliency [19, 46, 21] and its applications to recognition problems have been extensively studied [42]. Memorability may be predicted using saliency. All three key frames (beginning, middle, and end frames) of each video have been saliency-mapped.

4.1.4 InceptionV3

For image processing and object detection, Inception v3 uses a convolutional neural network. Object detection deep network InceptionV3, trained on the ImageNet dataset, has reached its final class activation.

4.1.4 HOG

A feature descriptor called HOG counts the number of instances when gradient orientation occurs in a certain area inside an imaging frame of reference. In the HOG descriptor, an object's structure or form is emphasized. HOG can offer the edge direction in addition to the feature description. In this case, the gradient and orientation of the edges are extracted. Aside from that, these angles are computed in 'localized' sections of the image space. For each of these areas, the gradients and orientations are computed and HOG would create a Histogram. An oriented Gradient Histogram is a histogram generated by combining gradients and pixel values. Each key frame is divided into 32x32 windows to compute HOG descriptors (Histograms of Oriented Gradients), with each feature having 256 main components.

4.1.5 Local Binary Pattern (LBP)

The Local Binary Pattern (LBP) is a basic yet effective texture operator that identifies pixels in an image by thresholding the pixels' immediate surroundings and treating the output as a binary integer. Local Binary Patterns for patches of 8x15 pixels are given in the dataset.

4.1.6 Captions

Objects and activities in the videos were used as the basis for creating the textual information in the dataset. Captions may be a convenient way to express video content, making them helpful for predictions.

4.2 TF-IDF Vectorizer

Inverse Document Frequency (TF-IDF) is an acronym for Term Frequency Inverse Document Frequency. That's a very common technique used to convert text into a meaningful numerical representation for machine learning algorithms. To calculate the relevance of each word in the text, it is common to assign a weight to each word in the document. In the fields of Information Retrieval and Text Mining, this method is frequently used. For instance, "This House is quite lovely." We can easily grasp the statement since we are familiar with the meaning of the words and the structure of the sentence. If the machine cannot grasp this statement, what hope do we have? The computer can only interpret data in numerical form. To make the machine more intelligent, we vectorize all of the texts. When we vectorize the documents, we can then do a variety of activities such as identifying the relevant documents or performing ranking, grouping and other similar operations. Similar results may be obtained via a google search. A document is a web page, and a query is the search phrase that you use to find a document. Every document on Google has a standard representation. With a query, Google finds the relevance of all papers, ranks them in order of relevancy, and gives you the top k documents. All of this is done using vectorized query and document data. As you can see, Google's algorithms are quite smart and tuned.

4.3 Glove Embedding’s

GloVe (global vectors for word representation) is a Stanford-developed unsupervised learning method of creating word embedding’s from a corpus by aggregating a global word-word co-occurrence matrix. In vector space, the resultant embedding’s reveal intriguing linear substructures of the word. It is based on word-context matrices and matrix factorization methods. In a big corpus, you create a large matrix of co-occurrence information and count each "word" (rows) and how frequently we encounter this word in some "context" (columns). For each term, we look for context terms within a window size before and after the term. We also assign less weight to statements that are further away. Since it is combinatorial in size, the number of "contexts" is enormous. After that, we factorize this matrix to get a lower-dimensional matrix with each row corresponding to a vector representation of each word. This is often accomplished by reducing a "reconstruction loss."

This loss seeks out lower-dimensional representations that can account for the majority of the variation in high-dimensional data. We can use pre-trained word vectors that can be used quickly.

4.4 SVR

In the field of machine learning, SVMs, or Support Vector Machines, are one of the most popular and extensively used algorithms. Regression of SVM is also known as Support Vector Regression or SVR.

To predict discrete values, Support Vector Regression is a supervised learning technique. Similar to SVMs, Support Vector Regression is based on the same idea. There is a fundamental principle behind the use of SVR, which is to identify the best fitting line. The best fit line in SVR is the hyperplane with the most points. Other regression models aim for a minimum error, whereas SVR aims for the best line that fits the data within a threshold value. Hyperplane and boundary line distance is the threshold value. With more than 10000 samples, SVR has a fit time complexity that is more than quadratic with the sample number. In the case of big datasets, the Linear SVR or the SGD Regressor is employed as the regression model. But it just examines the linear kernel. Due to the cost function, the Support Vector Regression model only uses a small fraction of the training data to build a model.

4.5 CNN

CNN) is a form of ANN that is specially intended to analyze pixel input and is used in image recognition and processing.

CNN's are image processing, AI systems that apply deep learning to do both generative and descriptive tasks. They frequently use machine vision, which includes image and video recognition, as well as NLP and recommender systems.

When it comes to neural networks, they are a collection of hardware and/or software systems that are modelled after how neurons work in the brain. For example, traditional neural networks aren't appropriate for image processing since they require images to be supplied in smaller parts with lower resolutions. According to CNN, their neurons resemble those of the human frontal lobe, which is responsible for processing visual inputs in humans and other animals, then the rest of the brain.

This type of system is similar to a multilayer perceptron but has been optimized for speed. These include a convolutional layer that incorporates several convolutional layers as well as pooling and fully connected layers, as well as a normalization layer. A system that is significantly more effective and simpler to train due to the removal of constraints and increase in efficiency for image processing has been developed.

Using the right filters, a ConvNet may capture the spatial and temporal relationships in a picture. In part, this is because the number of parameters involved is reduced, and the weights are reusable. So, the network may be taught to better grasp the complexity of the image.

4.6 RNN

RNN are the most advanced algorithm for sequential data and are adopted by Apple's Siri as well as Google's voice-search service. As the first algorithm with internal memory, it is well suited to machine learning challenges that include sequential data, because it retains its inputs. It's one of the algorithms that's been behind the scenes of deep learning's spectacular successes in recent years.

Recurrent neural networks, like many other deep learning techniques, are rather ancient. They were first developed in the 1980s, but it wasn't until recently that their entire potential became apparent. A rise in processing power, huge quantities of data, and the discovery of long-short-term memory (LSTM) in the 1990s have all contributed to bringing neural networks to the forefront.

Their internal memory allows RNNs to recall critical information about the input they received, allowing them to predict the future with great precision. As a result, they're the ideal method for sequential data, such as time-series and speech. Compared to other algorithms, recurrent neural networks can build a far deeper knowledge of a sequence and its environment.

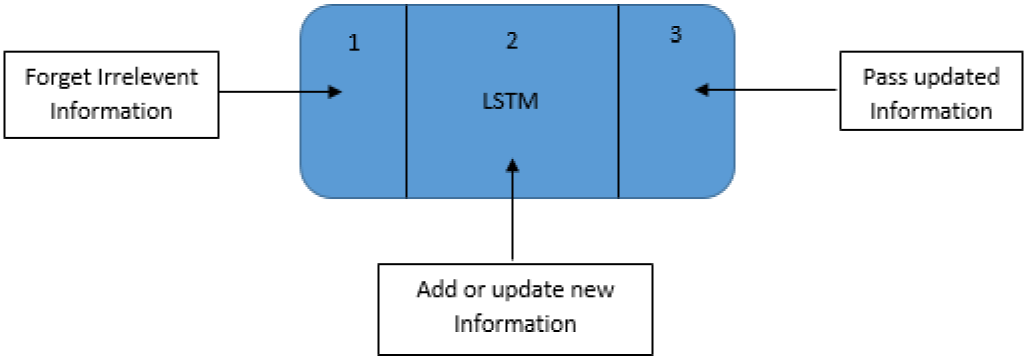
In addition to its benefits, the RNN also has certain drawbacks, such as its inability to handle very lengthy sequences if employing tanh or relu as an activation function and gradient vanishing and exploding issues.

4.7 LSTM

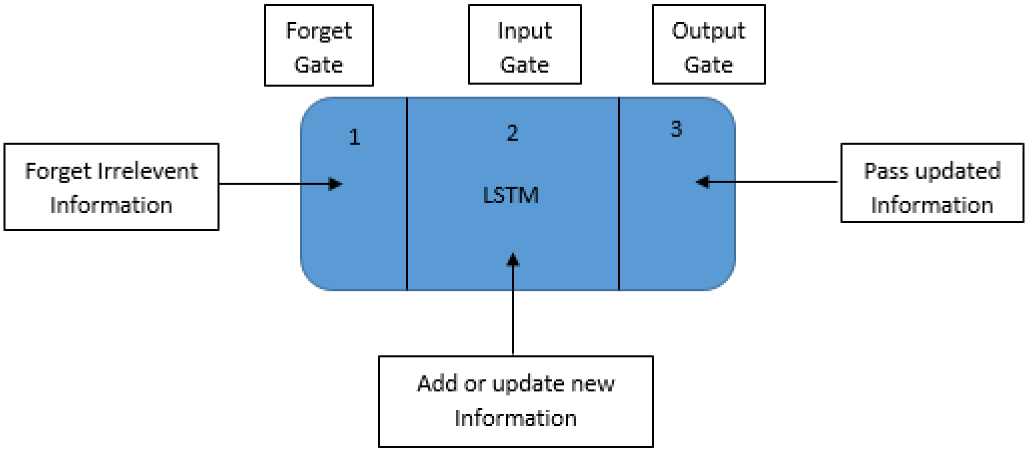
An enhanced RNN, the Long Short Term Memory Network allows information to remain. In addition, it can deal with the vanishing gradient issue that RNN has to deal with. When given time delays of undetermined duration, LSTM is ideal for classifying, processing, and predicting time series. Backpropagation is used to train the model. Think about it: when viewing a movie, you recall the previous scene or while reading a book, you recall the previous chapter's events. RNNs function in a similar way, in that they remember information from the past and utilize it to process the present input. Due to the vanishing gradient, RNNs are unable to recall long-term dependencies. To minimize long-term dependence concerns, LSTMs were specifically built.

4.7.1 LSTM Architecture

High-level, LSTM is quite similar to an RNN cell in terms of functionality. The LSTM network's internal workings are shown here. Every one of these pieces serves a different job, as you can see in the figure below.



Firstly, it decides if prior timestamp information should be remembered or if it is useless and may be ignored. Cell strives to learn new information from input in the second stage. Finally, in the third section, the cell propagates the modified data from the current timestamp to the next timestamp. Gates are the three components of an LSTM cell. The Forget gate is the first portion, the Input gate is the second, and the Output gate is the third.



Similar to an RNN, an LSTM contains a hidden state where H(t-1) is the previous timestamp's hidden state and Ht is the current timestamp's hidden state. C(t-1) and C(t) indicate the cell state of the LSTM for the previous and current timestamp, respectively. Cell state is referred to as Long term memory, whereas the concealed state is called Short term memory. Please see the image below. All timestamps and data are stored in this cell's state.

4.8 Ensemble Learning

In ensemble learning, numerous models are combined to produce better outcomes in machine learning. When compared to a single model, this strategy offers better prediction performance. Bagging to reduce variance, boosting to reduce bias, or stacking to improve predictions, ensemble approaches combine various machine learning techniques into one predictive model.

Bagging

Bootstrap aggregation is what Bagging stands for. Multiple estimates can be averaged together to minimize the variation of a prediction. Example: Using a random selection of subsets of data, we may train M distinct trees and compute the ensemble.

$$f(x) = 1/M \sum_{m=1}^{M} f_m(x)$$

To generate the data subsets for training the base learners, bagging employs bootstrap sampling. Using voting for classification and averaging for regression, bagging aggregates the outputs of base-level learners.

4.8.1 Boosting

An algorithmic technique called boosting converts poor learners into strong ones. According to boosting, weighted data is fitted using a series of weak learners' models that are only marginally better than random guessing. In this round, we assign more weight to cases that were misclassified in the first round.

A weighted majority vote (classification) or weighted sum (regression) is then used to integrate the predictions to generate the final forecast. As opposed to bagging, base learners are trained sequentially on a weighted version of the data in boosting methods.

4.8.2 Stacking

By using a meta-classifier and/or meta-regressor, stacking is an ensemble learning approach that combines several classifications and regression models. Basic models are trained using a complete training set, and the meta-model is learned using the outputs of basic models as features. As a result, stacking ensembles are typically diverse.

4.9 Correlation Score

The correlation coefficient is a statistical metric used to determine how two random variables are related. For example, you can look at the correlation between two random variables to see how they relate to each other. The strength of a link may be determined by looking at the correlation coefficient. The correlation coefficient has a range of -1 to +1. A positive correlation exists when two variables increase or decrease in lockstep, while a negative correlation exists when one variable increases while the other decreases. If a change in one variable does not influence the other, the two variables have a zero correlation.

4.10 Covariance:

The covariance of two random variables is a measure of how much they differ by their respective means. When the scale is changed, it affects it. The covariance coefficient has a value between -∞ and +∞.

4.11 Pearson Correlation Coefficient (PCC):

Pearson correlation in statistics is a measure of how closely two random variables are related to one another between the range +1 to -1 coefficient's value. The standard deviation of each random variable is used to normalize Pearson correlation. In Pearson, both variables need to be normally distributed.

$$PCC(X,Y) = \frac{COV(X,Y)}{SD_x * SD_y}$$

Here X and Y are two random variables. COV is covariance and SD is the standard deviation.

4.12 Spearman Rank Correlation Coefficient (SRCC)

Spearman's correlation is one of the most prominent methods for evaluating the correlation between variables. For both continuous and discrete ordinal variables, it is a suitable measure to use. The Spearman's correlation score will always be somewhere between -1 (perfect negative) and 1 (perfect positive).

Some of Pearson's correlation score shortcomings are addressed by Spearman’s correlation score. As a result, it does not make any assumptions about how data will be distributed. To assess the degree of connection between two variables, SRCC assigns a rank to each random variable and computes PCC from it.

Given two random variables X and Y, find the sum of their squares. For each random variable, assign a rating of 1 to the lowest value. To calculate the SRCC, first, apply the Pearson correlation coefficient to Rank(X) and Rank(Y). When using monotonically rising or decreasing functions, SRCC has a range of -1 to +1.

4.13 Proposed Method

The dataset is composed of 10,1000 videos, 8000 for development set and 2000 for test set. Only development set had the labels. So I divided the development set into 7,000 training set and 1000 validation set. Then I used the validation set choose best performing features and hyper-parameters and then evaluated the performance of my models.

Experiment 1:

In our first experiment We leveraged our held-out validation set to choose best performing features.

Features Used:

I have used Different visual & semantic features to predict the memorability score:

Semantic Features

- Captions

Video Features

- C3D
- HMP

Experiment 2:

In our second experiment we have used validation dataset to choose hyper-parameters and evaluated the performance of our models.

First, we developed traditional Machine Learning and highly regularized linear models:

i) Support Vector Regression.

Second, I have implemented Deep Learning techniques such as:

i) CNN
ii) RNN
iii) LSTM

Data Preprocessing:

After loading the dataset for all 3 features (Captions, C3D, HMP), video features such as C3D and HMP were used directly while an extensive work was done on semantic features.

Caption dataset has been cleaned by stripping special characters, converting all the words to lowercase, and removing the stop words. Then a bag of words was created from cleaned data and those bag of words were used to extract features using Tf-idf Vectorizer and Glove Embedding.

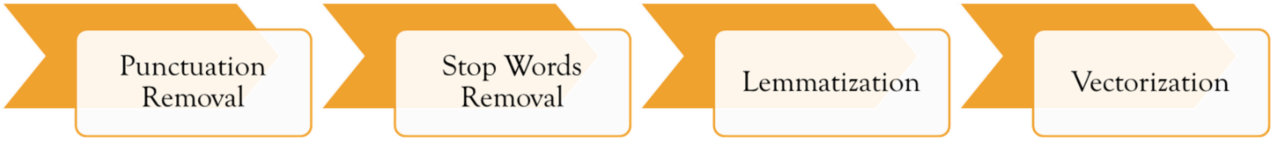


Figure 3: Caption Preprocessing

SVR:

For the implementation of SVR I have used scikit-learn library with regularization parameter as 100 and RBF kernel.

CNN:

I have design the CNN model using an embedding layer and 1D convolutional layer. Convolutional layer has 128 filters with window size of 5. Then another layer of 10 neurons has been added to reduce the dimensions of the network. To reduce over fitting the data, we add dropout layer and each layers uses ridge regularization to reduce dimensions with lower weights.

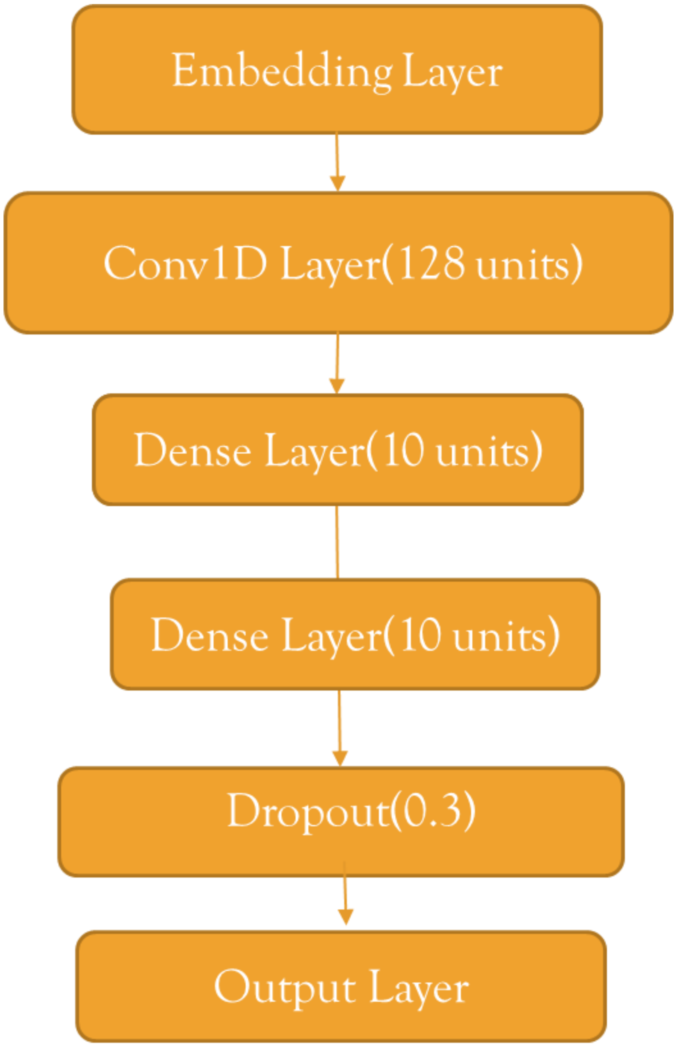


Figure 4: CNN Architecture

LSTM:

For LSTM, I have design the model with embedding layer as input layer followed by a hidden LSTM layer with 150 units which is followed by another Dense layer of 30 units. For output layer, I have used a Dense layer with two neurons each for short and long term score. To reduce overfitting, the data I have used dropout layer and as a final optimization algorithm I have used Adamx Optimizer.

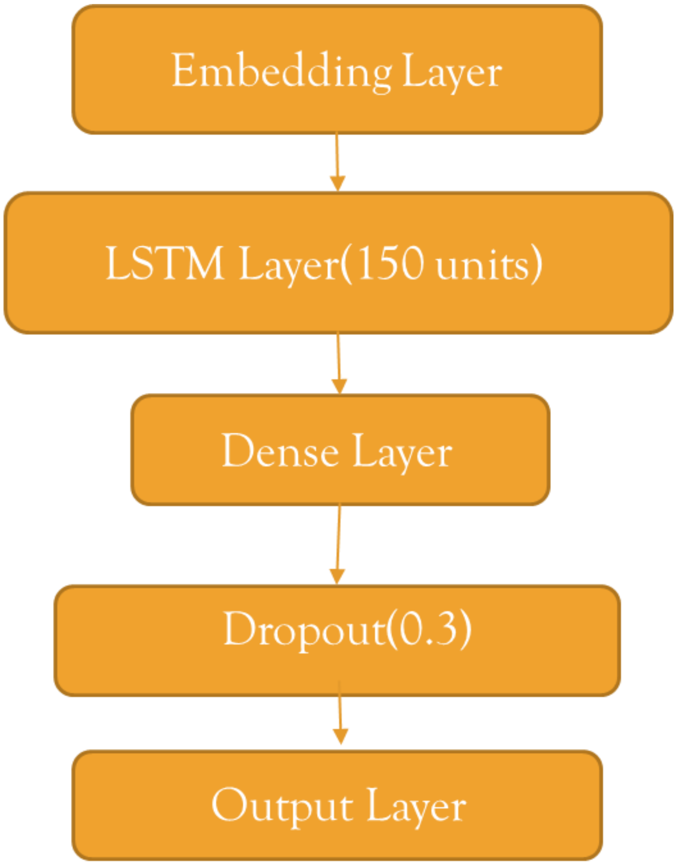
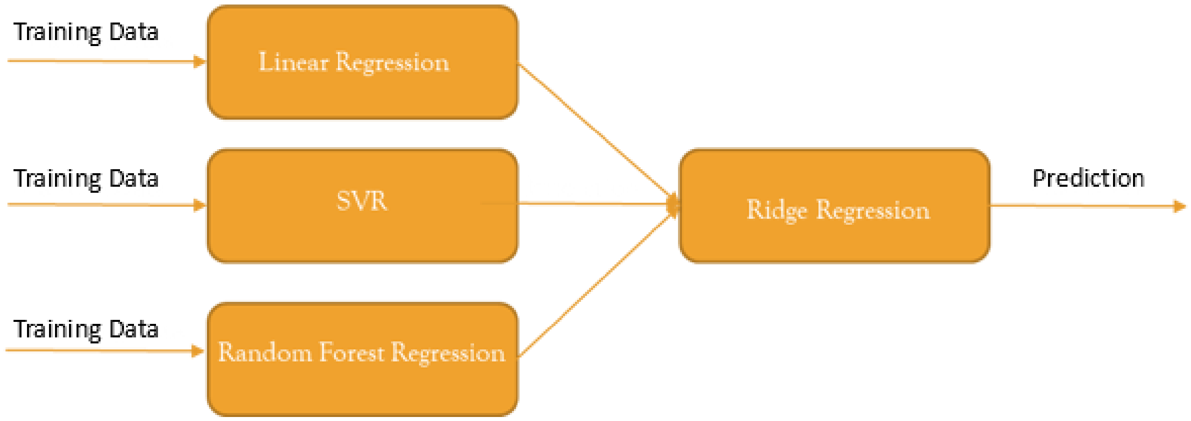


Figure 5: LSTM Architecture

4.14 Ensemble of all these Models

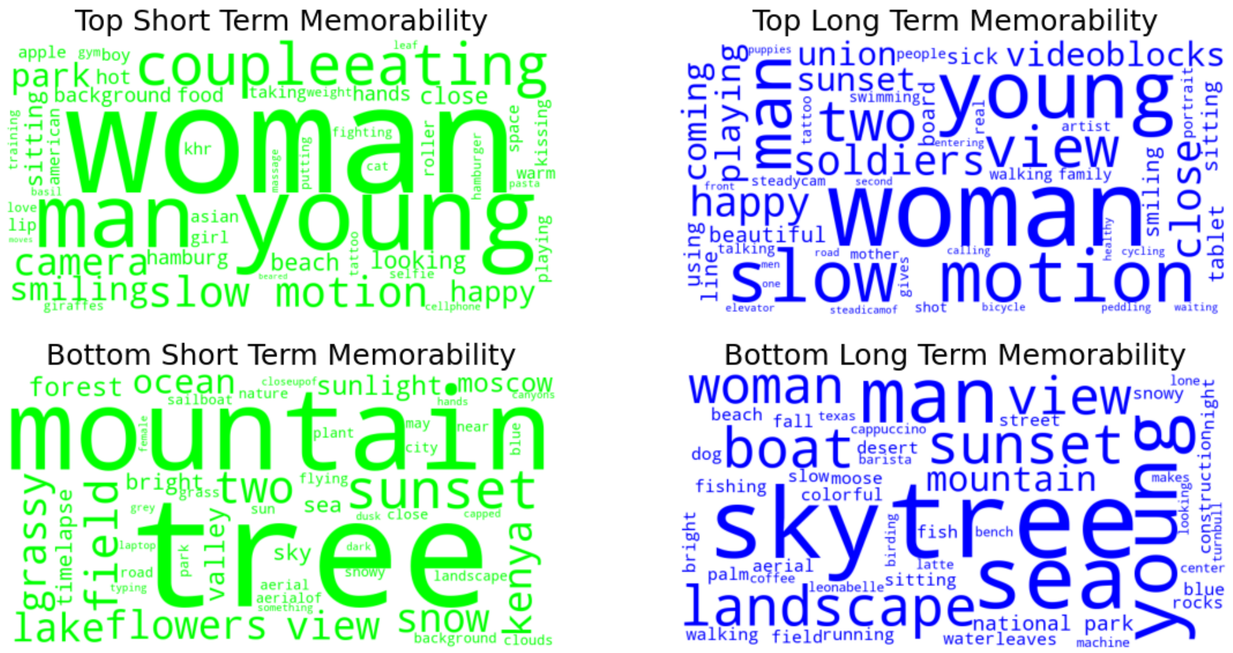
I have used stacking regressor with the following initial estimators with the final estimator as Ridge Regression:

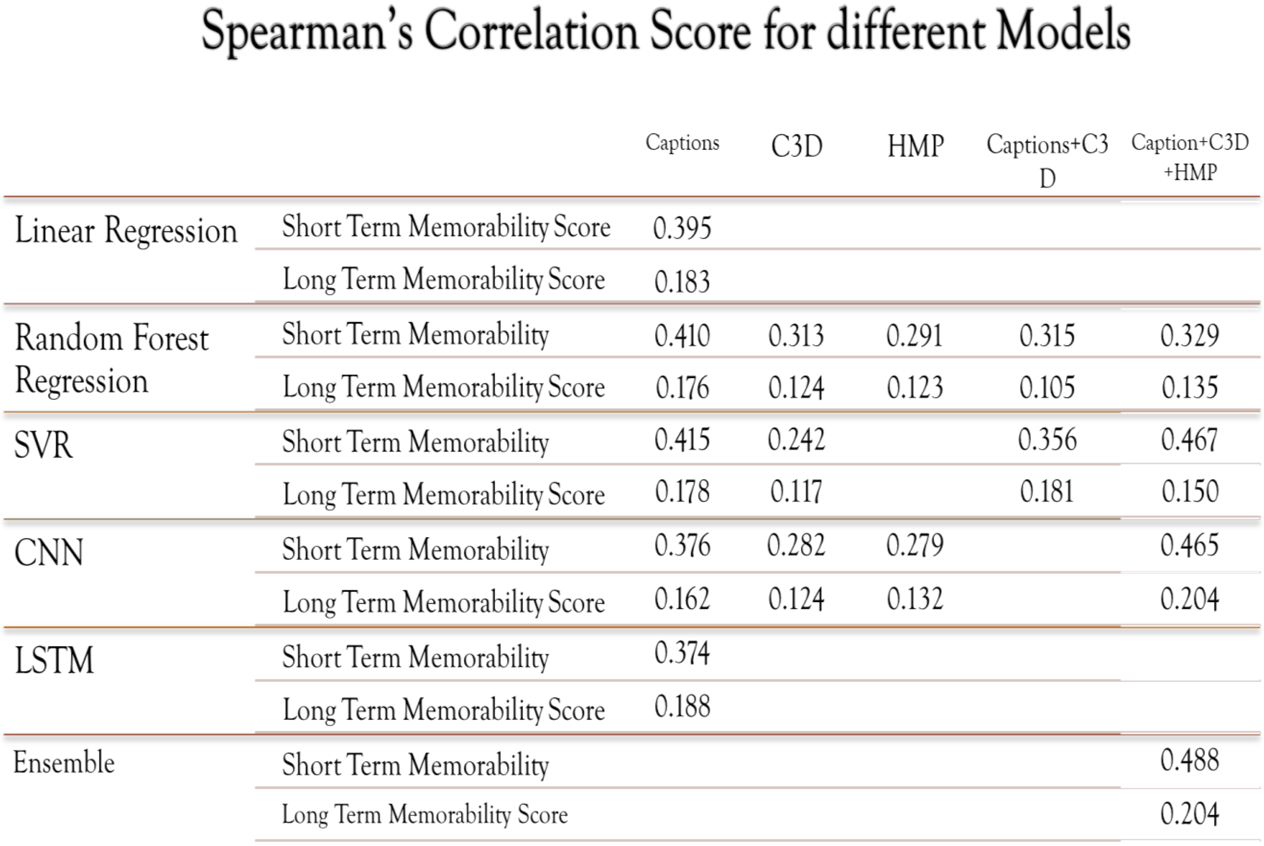
- Linear Regression
- SVR
- Random Forest Regressor



CHAPTER 4: RESULTS AND ANALYSIS

In experiment 1, the model scored badly on image-level features, with a poor Spearman's rank correlation on the validation set. This demonstrates that low-level image features alone are insufficient to comprehend media memorability, implying that videos are far more than a collection of frames. I employed textual information and video-based visual features in the second experiment. The research revealed that subtitles produced greater outcomes than any other video feature available. Combining both visual and textual elements yields better results. As a final predictor, the ensemble model for predicting media memorability was adopted.





Spearman's Correlation Score for different Models

		Captions	C3D	HMP	Captions+C3D	Caption+C3D +HMP
Linear Regression	Short Term Memorability Score	0.395				
	Long Term Memorability Score	0.183				
Random Forest Regression	Short Term Memorability	0.410	0.313	0.291	0.315	0.329
	Long Term Memorability Score	0.176	0.124	0.123	0.105	0.135
SVR	Short Term Memorability	0.415	0.242		0.356	0.467
	Long Term Memorability Score	0.178	0.117		0.181	0.150
CNN	Short Term Memorability	0.376	0.282	0.279		0.465
	Long Term Memorability Score	0.162	0.124	0.132		0.204
LSTM	Short Term Memorability	0.374				
	Long Term Memorability Score	0.188				
Ensemble	Short Term Memorability					0.488
	Long Term Memorability Score					0.204

Table 2: Results

Compared to all other features examined, visual semantic features generated from picture captioning give higher prediction results. Since the cameras record visual information throughout the scene, which is known to play a significant part in human memory, this is not surprising. A recent study on Image memorability found that image captioning-based features were more accurate for predicting IM than traditional CNN features. As they encapsulate the visual spatiotemporal information of the dataset, C3D features proved to be fairly successful in predicting VM on the dataset.

CHAPTER 5: CONCLUSION AND FUTURE WORK

In this regard, we have the following discoveries and contributions to make:

- Methods such as embedding's and recurrent networks may get extremely high results for captions. Deep Learning CNN models might outperform models trained using captions and other visual features in terms of short-term memorability.

- Although we were unable to show it in this research, we believe that trained CNN models would perform better as feature extractors when fine-tuned with enough training data and iterations.

- It's possible to overcome memory restrictions by utilizing prediction-based model assembly rather than training lengthy feature vectors.

- To get optimal results, it is important to combine multiple types of models, such as emotion models accompanied by captioned images or high-level representations from CNNs or visually computed features.

REFERENCES

[1] Romain Cohendet, Karthik Yadati, Ngoc Q. K. Duong, and ClaireHélène Demarty. 2018. Annotating, Understanding, and Predicting Long-term Video Memorability. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (ICMR '18)*. ACM, New York, NY, USA, 178–186. https://doi.org/10.1145/3206025.3206056 Arndt, N., 1993, " Blade Row Interaction in a Multistage Low Pressure Turbine", ASME Journal of turbomachinery, Vol. 115, pp. 370-376.

[2] M. G. Constantin, B. Ionescu, C-H. Demarty, N. Q. K. Duong, X. Alameda-Pineda, and M. Sjöberg. 2019. The Predicting Media Memorability Task at MediaEval 2019. In *Proc. of the MediaEval 2019 Workshop*. Sophia Antipolis, France.

[3] R. M. Savii, S. F. dos Santos, and J. Almeida. 2018. GIBIS at MediaEval 2018: Predicting Media Memorability Task. In Proc. of the MediaEval 2018 Workshop. Sophia Antipolis, France.

[4] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. 2016. Deep learning for image memorability prediction: The emotional bias. In Proceedings of the 24th ACM international conference on Multimedia. ACM, 491–495.

[5] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. 2008. Visual long-term memory has a massive storage capacity for object details. Proceedings of the National Academy of Sciences 105, 38 (2008), 14325–14329.

[6] Mihai Gabriel Constantin, Miriam Redi, Gloria Zen, and Bogdan Ionescu. 2019. Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates. ACM Computing Surveys (CSUR) 52, 2 (2019), 25.

[7] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. Amnet: Memorability estimation with attention. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 6363–6372.

[8] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. 2014. Learning computational models of video memorability from fMRI brain imaging. IEEE transactions on cybernetics 45, 8 (2014), 1692–1703.

[9] Ronald A Rensink, J Kevin O'Regan, and James J Clark. 1997. To see or not to see: The need for attention to perceive changes in scenes. *Psychological science* 8, 5 (1997), 368–373.

[10] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier AlamedaPineda, Elisa Ricci, and Nicu Sebe. 2017. How to Make an Image More Memorable? A Deep Style Transfer Approach. In *ACM International Conference on Multimedia Retrieval*.

[11] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[12] Duy-Tue Tran-Van, Le-Vu Tran, and Minh-Triet Tran. 2018. Predicting Media Memorability Using Deep Features and Recurrent Network.. In *MediaEval*

[13] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. 2017. Show and Recall: Learning What Makes Videos Memorable. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2730–2739.

Prediction of Short-Term and Long-Term Video Memorability

ORIGINALITY REPORT

8%	3%	4%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	towardsdatascience.com Internet Source	1%
2	Submitted to National College of Ireland Student Paper	1%
3	Romain Cohendet, Claire-Helene Demarty, Ngoc Duong, Martin Engilberge. "VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability", 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019 Publication	1%
4	R Dhanya, Irene Rose Paul, Sai Sindhu Akula, Madhumathi Sivakumar, Jyothisha J Nair. "F-test feature selection in Stacking ensemble model for breast cancer prediction", Procedia Computer Science, 2020 Publication	1%
5	Submitted to Camarines Sur Polytechnic Colleges Student Paper	1%
	www.ijert.org	

6	Internet Source	<1 %
7	Submitted to California State University, Sacramento Student Paper	<1 %
8	Submitted to Colorado Technical University Online Student Paper	<1 %
9	Submitted to cdsl Student Paper	<1 %
10	"Long-term memory", Salem Press Encyclopedia of Health, 2013 Publication	<1 %
11	Submitted to Manchester Metropolitan University Student Paper	<1 %
12	Submitted to University of Greenwich Student Paper	<1 %
13	Submitted to University of Wales Institute, Cardiff Student Paper	<1 %
14	Kai Shu, Deepak Mahudeswaran, Huan Liu. "FakeNewsTracker: a tool for fake news collection, detection, and visualization", Computational and Mathematical Organization Theory, 2018 Publication	<1 %

15	stock.adobe.com Internet Source	<1 %
16	"Identification of Default Payments of Credit Card Clients using Boosting Techniques", International Journal of Recent Technology and Engineering, 2020 Publication	<1 %
17	Akihiro Matsufuji, Haruka Sekino, Eri Sato-Shimokawara, Toru Yamaguchi. "A Calculation Method of the Similarity Between Trained Model and New Sample by using Gaussian Distribution", 2020 International Joint Conference on Neural Networks (IJCNN), 2020 Publication	<1 %
18	Akihiro Matsufuji, Eri Sato-Shimokawara, Toru Yamaguchi, Lieu-Hen Chen. "Adaptive Multi Model Architecture by Using Similarity Between Trained User and New User", 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), 2019 Publication	<1 %
19	Huaichun Fu, Shouwei Gao, Yan Peng, Nan Zhao. "Prediction of fishing vessel operation mode based on Stacking model fusion", Journal of Physics: Conference Series, 2021 Publication	<1 %

20	Submitted to iGroup Student Paper	<1 %
21	Submitted to La Trobe University Student Paper	<1 %
22	section.iaesonline.com Internet Source	<1 %
23	Sandholm, William. "Vital Statistics", Oxford University Press Publication	<1 %
24	Qi Lin, Sami R. Yousif, Marvin M. Chun, Brian J. Scholl. "Visual memorability in the absence of semantic content", Cognition, 2021 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		