Major Project-II Report on

# Design Of Framework For Recommendations Of Videos Based On Previous Uploads Of Users

Submitted in partial fulfillment of the requirements for the degree of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

by

**Sarthak Jain (191IT145)**

**Niranjan Mahabaleshwar Hegde (191IT235)**

*under the guidance of*

**Mr. Dinesh Naik**


DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE - 575025

February 2023

# DECLARATION

I/We hereby *declare* that the Major Project-II Work Report entitled ***"Design Of Framework For Recommendations Of Videos Based On Previous Uploads Of Users"*** , which is being submitted to the **National Institute of Technology Karnataka, Surathkal**, for the award of the Degree of Bachelor of Technology in Information Technology, is a *bonafide report of the work carried out by us.* The material contained in this Major Project-II Report has not been submitted to any University or Institution for the award of any degree.

*Name of the Student (Registration Number) with Signature*

(1) Sarthak Jain (191IT145)

(2) Niranjan Mahabaleshwar Hegde(191IT235)

Department of Information Technology
Place : NITK, Surathkal
Date : 20/02/2023

# CERTIFICATE

This is to *certify* that the Major Project Work Report entitled ***"Design Of Framework For Recommendations Of Videos Based On Previous Uploads Of Users"*** submitted by

*Name of the Student (Registration Number)*

(1) Sarthak Jain (191IT145)

(2) Niranjan Mahabaleshwar Hegde (191IT235)

as the record of the work carried out by them, is *accepted as the B.Tech. Major Project-II work report submission* in partial fulfillment of the requirement for the award of degree of Bachelor of Technology in Information Technology in the Department of Information Technology, NITK Surathkal.

**(Mr.Dinesh Naik)**
Assistant Professor
Department of Information Technology
NITK Surathkal

# ABSTRACT

Video recommendation has become one of the essential aspects and challenges in the ever blooming field of technology. With enormous amounts data being uploaded every second, It has become essential to map the content created to the right user in order to make the platform engaging for the users. Matching uploaded videos with already presented videos in the database is not feasible. We propose a framework that requires lesser computational requirements by creating a system of recommendations that presents the user with relevant content by using the created descriptions as input. For this, we use the genetic algorithms to select important frames rather than taking the frames at equal distances. The pre processing steps are applied on these frames where three types of features are selected i.e. Global features, motion features and local features. Text preprocessing is done on the corpus which helps in speeding up the computation. The main framework consists of a Gated recurrent units model which is combined with an attention mechanism. GRU is computationally lightweight compared to LSTM and thus helps in improving results. The description obtained as output is fed as input to the recommender. The recommender is a combination of 3 models which are TF-IDF based,sentence transformer based and page rank base. All three scores arrays are normalised and normalized values are added. Final rank list is obtained by adding all three scores. Based on the scores top 10 video suggestions are shown to the users.

***Keywords***— Video Recommendation, GRU, Genetic Algorithm, Attention Mechanism

# CONTENTS

# List of Figures

# List of Tables

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

Increased availability of the internet has resulted in the transfer of enormous amount of data over the internet. This data can be of any form for example videos, text, voice etc. Advancements in technology has made it possible to provide users with the content they wish to see or the content they might be interested in. One such task is a recommendation of videos. It has become extremely important to map the content created to the right user in order to make the platform engaging for the users. Additionally, it makes content more visible and ensures that viewers are shown only relevant and interesting stuff. Users' overall satisfaction with the site can be raised by a well-designed video recommendation system that helps them choose content they're likely to appreciate. As a result, the platform may see an increase in user engagement, longer viewing sessions, and more regular use. A video recommendation system can also contribute to making new content more discoverable by surfacing content from creators that people might not have otherwise come across. This can increase the user's exposure to a variety of content , which helps diversify their platform experience.Finally, since they eliminate the need for manual content curation, video recommendation algorithms can also serve to improve the platform's overall effectiveness. By doing this, resources may be made available for other crucial duties like content filtering, platform optimization, and user support.

The amount of user-generated data that can become enormous as video platforms grow makes it challenging for recommendation systems to scale efficiently. Apart from what users search, what they upload is an important factor as well and if used efficiently can be used to provide users with proper recommendations. Matching uploaded videos with already presented videos in the database is not feasible. Therefore we propose a framework that can help in providing similar content to users with very lesser computational requirements. Apart from this problem we are also trying to solve another important problem that is selection of important frames in preprocessing tasks with the help of genetic algorithm based method.

Our significant contributions to this project include the use of genetic algorithms to select significant frames in addition to the standard pre-processing steps, global and motion-based feature extraction, a GRU-based model combined with an attention mechanism, and a recommendation system that uses the generated descriptions as input and suggests related content to the user. The Tf-Idf model, sentence transformers, graph generation with page rank are all combined in the recommender system's architecture.

The following summarizes the format of the rest of the paper.In section II, we go through most of the recent work that has been done in the domain of video captioning and attention mechanism in video clips. After describing the proposed work and model architecture in Section III, we demonstrate the results of our experiments which is then followed by the analysis in Section IV. In Section V, future work is discussed and we conclude with the inferences obtained.

## 1.2 Motivation

The primary motivation behind this project is the creation of a framework that can address a wide range of issues, including data loss when choosing frames because most methodologies rely on choosing every Kth frame. Creating recommender systems more scalable by lowering the number of operations necessary for making recommendations, as it becomes incredibly difficult to match the content of different videos as any application scales. Generating descriptions also has practical applications because they can be read out to persons who have vision issues, allowing them to more easily access the information in videos. These descriptions can also be incorporated into a teaching tool that aids language learners in relating simple tasks to descriptions of those activities in the language they are trying to learn.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Background and Related Works

In paper [1] In order to increase the popularity of videos, authors suggest a brand-new hybrid tagging approach based on multi-modal content analysis, specifically textual semantic analysis and deep learning-driven video content identification. A brand-new method called TF-SIM is suggested for ranking the list of potential keywords for a video. On the one hand, the algorithm takes into account a keyword's frequency by figuring out how many times the keyword appears in the list of potential keywords; The algorithm, on the other hand, applies the NGD formula to determine how semantically similar the keyword is to the original video keywords.

In paper [2] The primary contribution of this research is the showing of improved end performance at a number of manipulation tasks utilising a Genetic Algorithm (GA) to find DDPG and HER parameter values that lead to improved performance at these tasks more quickly. The research found a non-linear relationship between learning algorithm parameters and task performance and learning rate.Instead, depending on the settings of the RL parameter values, the success rate can vary greatly.

This study [3] The proposed algorithm is tested against the most recent peer competitors, which include eight manually designed CNNs,seven automatic + manually tuned CNNs, and five automatic CNN architecture design algorithms, on widely used benchmark image classification datasets. The proposed algorithm starts to work by receiving a set of predefined CNN building blocks, the population size, as well as the maximum generation number for the GA and the image classification dataset. Through a series of evolutionary processes, the algorithm eventually finds the best CNN architecture to classify the given image dataset.

In paper [4] User data from video viewing in these regions is gathered and kept on a dedicated cloud server. One of these clouds is designated as the aggregator and is in charge of compiling the training parameter files. Other cloud servers transmit the parameter files to the aggregator after receiving the requests. After some processing, the aggregator then sends the new parameter files to these cloud servers.

[5] authors have proposed a LSTM based encoder-decoder architecture.It is a combination of different CNNs,2D-CNN and 3D-CNN.In this model,encoder is a bidirectional LSTM and decoder is a two-layer LSTM.In this paper [6]the proposed model extract similar images from the frames of the input video using cosine similarity. In the second step,Adversarial inference is used for adding the auxiliary data from the image dataset as input to generate suitable captions for videos. It helps to generate proper captions for the videos.The model consists of a generator that gives the proper sentences along with auxiliary captions.Finally,a hybrid discriminator selects the final caption as input.In this paper [7]The authors have surveyed different neural networks and their accuracy in generating caption for videos.They have analysed RNN,ANN and CNN and have concluded that CNN is the best performer. The proposed model [8] makes use of the functionalities of Deep Learning and Natural Language Processing. This paper is for people who are visually impaired or suffer from short sightedness.In this paper [9]The authors have focused on the algorithmic overlap between image and video captioning,along with audio, The model focuses on various aspects like automatic metadata generation for photos; picture and video search engine indexing; different robot vision systems.

## 2.2 Outcome of Literature Review

From the above literature review we learn about the general structure of framework that can be used to generate descriptions and also the video recommending systems . Most models use either a LSTM based or GRU encoder and decoder. This allows to map input of varying length to the output of varying length.Encoding refers to conversion of data into desired format. The two-dimensional vector will be transformed into the output sequence, which will be the English sentence, by the decoder.In order to predict the English term, it is also constructed with RNN layers and dense layers. The encoder decoder model performs better thanks to the attention method. The idea behind this approach is to utilize the most crucial input sequence elements in a flexible way. We also learnt that results also depends on the way we extract features from videos. The more enriched extracted features are better the model performs.We learned that genetic algorithms improve the results if used while selecting the frames

since selecting frames at equal distances maya result in the loss of significant information.

## 2.3 Problem Statement

Design a framework based on genetic algorithm, GRU along with attention mechanism and hybrid information retrieval method for recommendation of videos based on previous uploads of users.

## 2.4 Objectives of the Project

(1) Implementation of Genetic Algorithm based method for frames selection

(2) Extraction of global and motion based features from selected frames

(3) Implementation of encoder-decoder based on Gated recurrent units

(4) Implementation of attention mechanism

(5) Evaluation of obtained descriptions

(6) Bench-marking performance and analysing logs of error rate with epochs.

(7) Implementation of hybrid information retrieval method comprising of TF-IDF, sentence transformers and graph construction with page rank

(8) Web based user interface that shows users with the recommendations based on their past uploads and enables users to search for video

# CHAPTER 3

# PROPOSED METHODOLOGY

## 3.1 Frames Selection Using Genetic Algorithm

Genetic algorithm can be used for the selection of essential frames. Each video can be represented as the binary string where each bit represents the frames and bit set to one implies the selection of that frame. We expect offspring to be better than their parents in terms of the frames selected. For this we need to have a fitness function with the help of which we can evaluate the parents.

Fitness function for this process is designed in the following way, we calculate the histogram of each frame and find the difference of each frame's histogram with that of other frames. If the difference is lesser than certain threshold, add it to score. Higher the score, the higher are the chances of the similarity of the current frame with the other frames. Thus making it less importance. Factor of distance is also considered. We need dissimilar frames that are at a lesser distance to capture more data. Therefore net fitness value becomes proportional to the inverse of distance and also logarithmic reciprocal of above mentioned score. Logarithm is used for scaling score.

Inititally we start with the randomly generated strings each comprising of 28 ones as here we intend to take 28 frames for the next steps of framework.

First step is selection of parents for this fitness values of all 4 parents are calculated then by using the roulette wheel selection two parents are selected and crossover and mutation operators are applied on them.

Crossover : for crossover, a random point is selected and interchanging is performed about that point. There is a possibility that the resulting string may contain more or less number of ones than required; therefore we need to perform the mutation to bring a number of ones equal to the required number.

Mutation : If the number of ones is equal to the required number of ones than no mutation is performed. If the number of ones exceeds the required number of ones we perform the merging operation. It refers to the merging of adjacent segments until a number of ones become equal to the required number of one. If the number of ones is

less than the required number of ones then randomly zeroes are picked and changed to one.

After the mutation operations both the strings are saved for the next generation this entire process is repeated once again since we need 4 strings in any of the generation. This continues till the fixed number of iterations. After fixed iterations a random candidate from current generation is taken following which frames are picked and are sent for the next steps. General flow is shown in figure 3.1.1

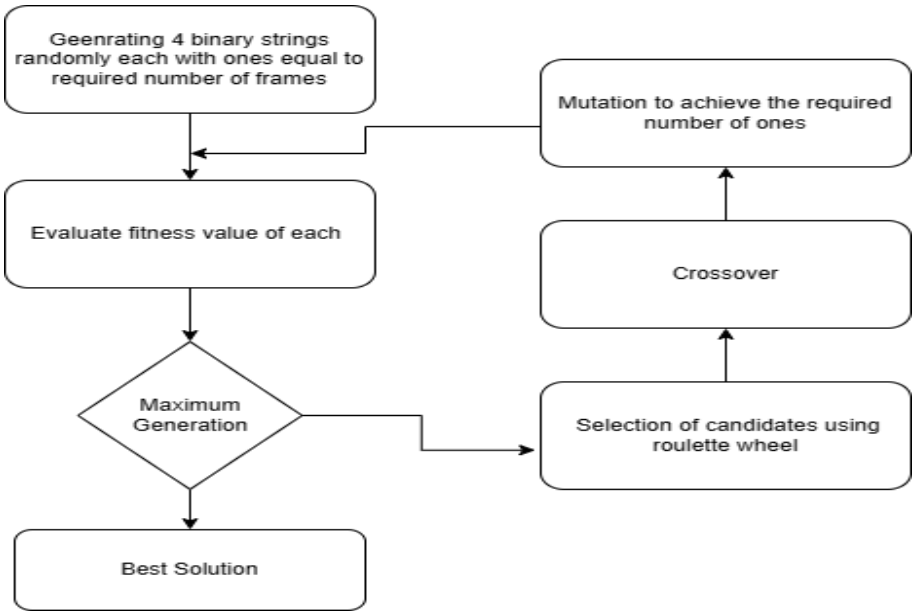


Figure 3.1.1: Flow Diagram For Genetic Algorithm

## 3.2 Pre-processing

### 3.2.1 Feature Extraction

In order to make computation easier we use 28 frames that are evenly spread from each video. These frames after the featuers extraction will be fed to the encoder as the input.We extract three different kinds of features: motion, local, and global. The global features stand in for the global entities or the most important things in the videos. The motion features are used to record time-level events in the video clips, such as running, falling, etc. Thus, the motion features are applied on a window

of frames. The local features are utilised to capture a variety of areas in the frames, such as the local items that are present in the video recordings and are helpful for captioning.

1) Global Features: From each video clip, 28 equally spaced frames were extracted. If the video contained less than 28 frames, the last frame was padded at the end of the video to make the total number of frames as 28. The frames were then passed through InceptionV3model one-by-one for feature extraction pre-trained on imagenet dataset.

2) Motion Features: Motion features were calculated on the video clip frames using a fixed window size of 16 frames using a C3D model pre-trained on Sports-1M dataset. Once the motion features were extracted for each window, 28 equally spaced windows were selected.

3) Local Features: For local features, 28 equally-spaced frames are taken. The output of the of the pretrained Faster RCNN network has been used to extract features from 28 equally spaced frames from the input video.

### 3.2.2 Text Pre-processing

Pre processing is performed on the corpus that includes removing punctuations, removing spaces and special characters.Sentences with less that 6 words are eliminated from the corpus and also the sentence with more that 10 words are eliminated.To indicate the beginning and end of sentences, the tokens ⟨BOS⟩ and ⟨EOS⟩ are added to each phrase, accordingly.To ensure that each of these sentences of the same length when a batch of them is created, a token ⟨pad⟩ is added, this also ensures the speeding up computation for the batch.

## 3.3 Framework

Fig. 3.2.1 shows the overall framework followed in this study.We are considering MSVD dataset here.MSVD dataset contains Youtube clips and their descriptions for this study.There are approximately 1,970 videos in the MSVD dataset. After we recieve frames of all the videos present in dataset next step is to extract features from these videos. Here we extract three types of features. Three types of features are

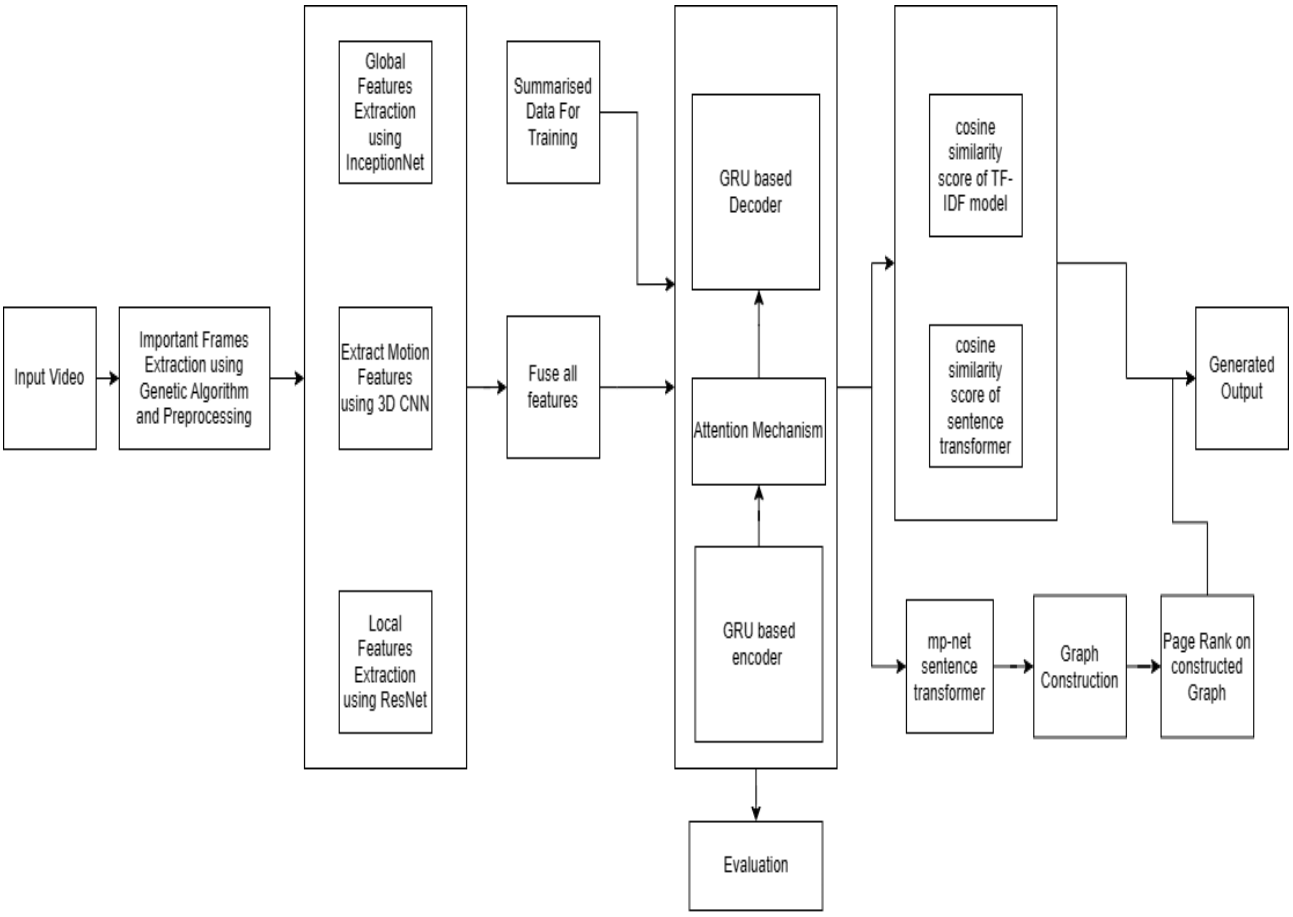


Figure 3.3.1: Flow Diagram

motion features, global features and local features. The global features represent the global entities or the most significant objects present in the video clips. The motion features are applied on a window of frames.

Then all the features are concatenated and fed to encoded network.Attention mechanism will be employed in encoder model and its output will be initial state for our decoder.

### 3.3.1 Encoder-Decoder Model

As shown in Figure 3.2.2 the encoder model consists of an Gated recurrent units model which we feed with input that we have recieved by the concatenation of extracted features i.e local,motion and global features.The encoder input to the Gated recurrent

units is of dimension 28 * 10240. Where 28 is the number of frames and 10240 is the length of concatenated features.The input of attention layer is the output of this layer. The attention layer will provide weights to important frames. The output of attention layer that is (28,512) dimension vector is passed to decoder Gated recurrent units. The output of this decoder model is fed to dense layer. This layer gives output of set word limit and these words are then converted to the sentence.The output of the decoder Gated recurrent units is fed to the final Dense layer via a dropout of 0.3. The dense layer generates a classifier output of pre determined number of words which can further be converted into a sentence

Due to the fact that GRUs only have two gates and LSTMs have three gates, they are computationally more effective than Gated recurrent unitss. The primary distinction between this network and the Gated recurrent units model is the absence of the additional layer used in the encoder after the attention layer. This reduces the amount of trainable parameters, making the model computationally lightweight. We have also compared results of both the models with and without using frames selection using genetic algorithm

### 3.3.2 Attention Mechanism

The important entities in the video clips are described in the global features. The events that take place throughout a variety of frames in the video recordings are described by motion characteristics. Important region level features are extracted from the video frames using local features. Local features help the spatial attention mechanism by giving particular region-level features more weight. However, the three characteristics are all necessary for temporal attention. Through Temporal attention mechanism we recieve information of the events occuring in the video with respect to the time. This will help us in finding important actions/verbs in the final generated captions. Certain frames are given higher priority over the other while generating the captions. Output generated from the attention mechanism are fed again to the Gated recurrent units model i.e encoder and than its output is fed as one of the input to the decoder which is also a Gated recurrent units model. Spatial attention is inspired from the visual attention of humans. Certain regions of the frames carry more important information which is helpful in generating accurate captions. The

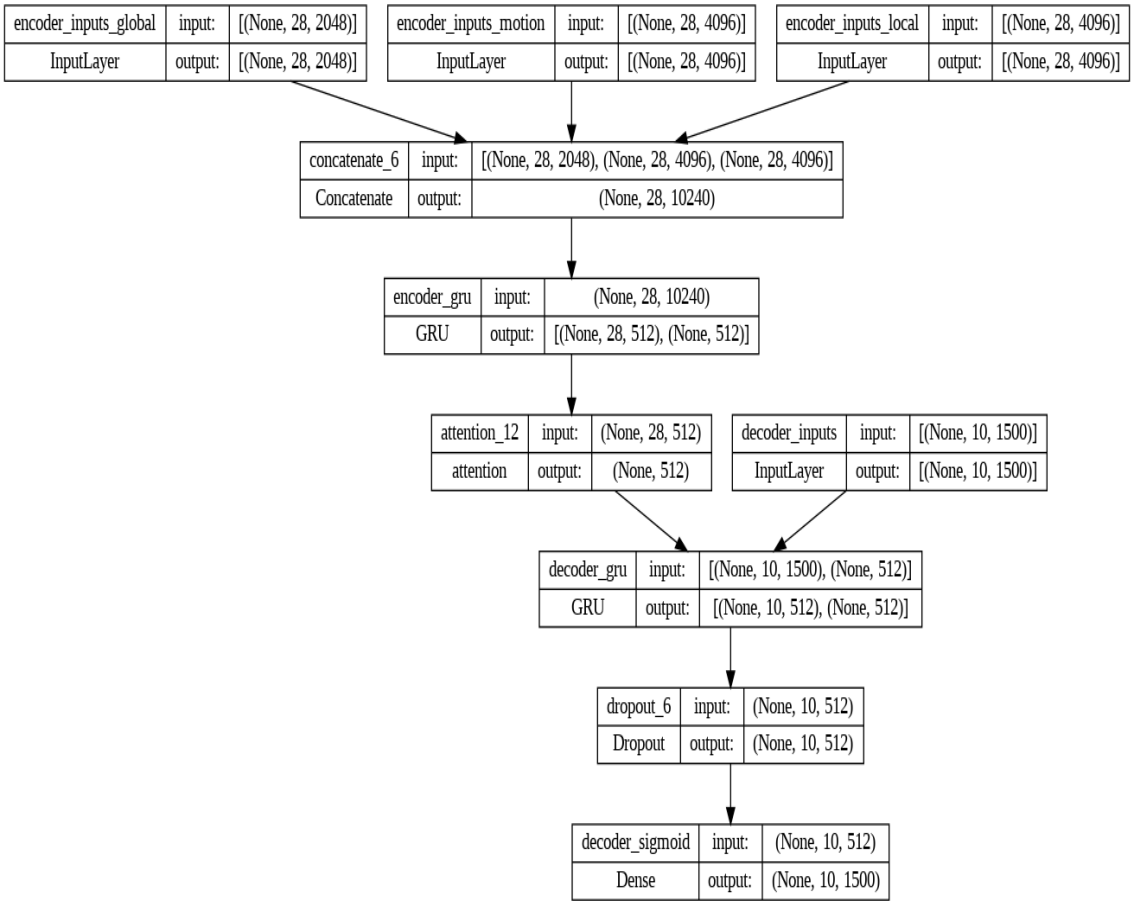


Figure 3.3.2: GRU based Encoder-Decoder Model

spatial attention mechanism captures these regions.Let the features extracted for a video be fg = fg1, fg2, ..., fgk where k is the number of frames and fgi denotes the feature vector of the ith frame. The attention mechanism applies weights to each of these features and the dynamic weighted sum of each of the features are added

In this method Output of encoder Gated recurrent units - 1 of dimension (Batch-Size * 28 * 512) is passed to attention layer.

Let say input vector is x: Following calculation is performed on it

$$e = K.tanh(K.dot(x, W) + b)$$

$$a = K.softmax(e)$$

$$output = x * a$$

here w and b are weight and biases. Output of dimension (BatchSize * 28 * 512) is returned which is passed to decoder Gated recurrent units

## 3.4 Recommender System

The generated output from the encoder-decoder model serve as query to the recommender system and all the videos descriptions are used as inputs to the recommender system. It comprises of combination of 3 models which are as follows : TF-IDF based, sentence transformer based and page rank based. Scores from all three methods are normalised and added and ranking list is generated accordingly. A statistical technique called TF-IDF is used to assess a document's word relevance in relation to a corpus of documents. It calculates a word's frequency within a document and devalues words that appear often throughout all documents. A word's TF-IDF score in a document indicates how significant it is to that document.

On the other hand, sentence transformer-based approaches are built on deep learning models that have been trained to recognise the meaning of sentences and carry out tasks like sentence classification, question-answering, and text production. Sentence Transformer models embed sentences in a high-dimensional space where related sentences are kept close to each other using a transformer architecture and pre-trained language models.

For finding score using page rank based method we can follow steps below:

I) we can create a graph where each sentence is represented by a node and the edges show how the sentences relate to one another. The edges can be defined in a variety of ways, such as by co-occurrence, similarity, or the semantic connection between sentences, here we are using sentence transformer to obtain the embeddings.

II) After the graph is built, we can determine each node's PageRank rating. Each node in the graph receives a score from the iterative PageRank algorithm based on inbound links from other nodes.Nodes with a high score are generally thought to be more significant and likely to be related to the query.

All three scores arrays are normalised and normalised values are added. Final rank list is obtained by adding all three scores. Based on the scores top 10 video suggetions are shown to the users. These results can be evaluated using standard

information retrieval methods like precision, recall etc.

## 3.5 Client Server Architecture

For developing a real-life application, we have used the following tech stack, for user interface, we have used the open-source frontend library ReactJs, and we have used the flask framework for the backend. To expose our local server to the web, we have used tunnel which provide an unique URL that can be used to fetch responses for the requests form the client-side application. Our client side of the application have three modules that are search - where videos are shown based on queries of users, upload - user can upload video based on which recommendations will be shown to the users and recommendations feed module - which shows recommendations to the users.

On server side various endpoints are created which are accessed through various routes. Endpoints created are for receiving video uploaded by the user, serving videos based on the user's search query and serving videos on the recommender feed.

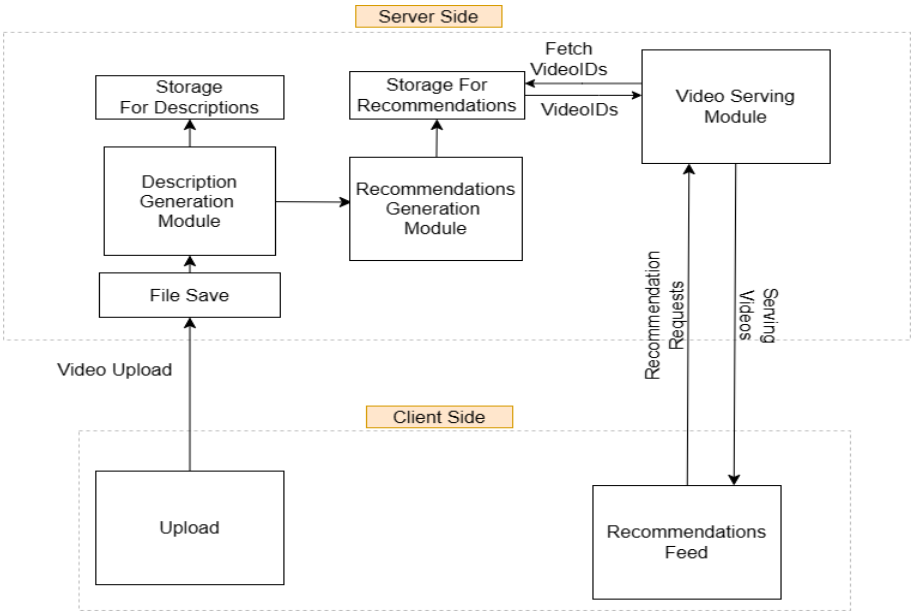


Figure 3.5.1: Flow Diagram For Recommendation Module

All the functionalities of Fig. 3.5.1 and Fig. 3.5.2 are from above sections of methodology.

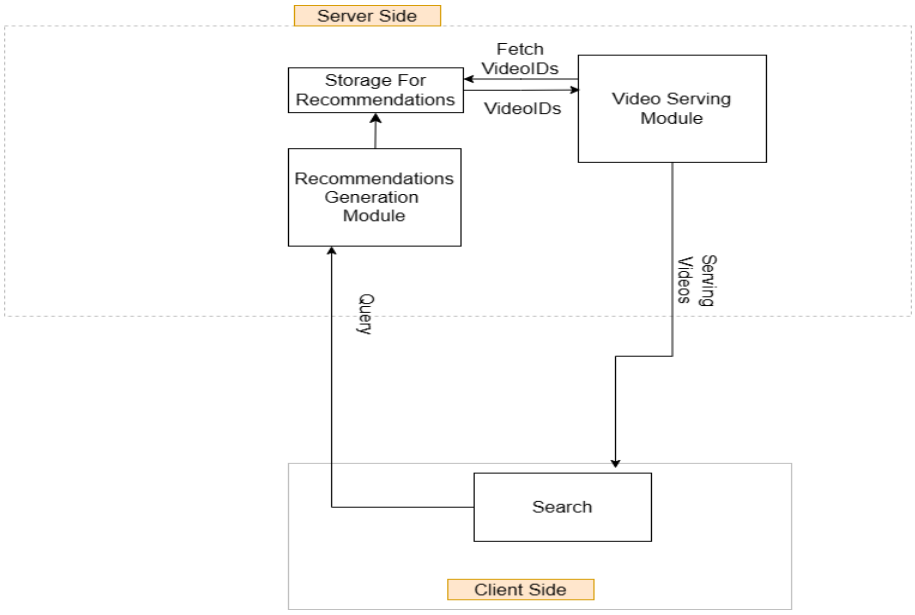


Figure 3.5.2: Flow Diagram For Search Module

## 3.6 Evaluation

Two evaluation metrics are to be considered for the evaluation of the descriptions generated. Two evaluation metrics are METEOR and BLEU-4. A statistic called METEOR is based on the weighted harmonic mean of recall and precision of generated captions that are captured as unigrams, with recall having a higher weight than precision. In order to determine scores, BLEU4 compares the generated captions to a corpus of reference texts before averaging the scores. Finding how well the machine-generated text matches the human translation of the video is the central tenet of BLEU-4.

For the evaluation of our recommender and search system, we have used two metrics which are as follows - recall which is the number of relevant videos retrieved out of the total relevant videos, and precision, which is the number of relevant videos retrieved out of total videos retrieved. This will help us in assessing and improving our system as we want more relevant information to be served to the users of our application

# CHAPTER 4

# RESULTS AND ANALYSIS

## 4.1 Feature Extraction Analysis

After extracting the frames from the video clips using the genetic algorithm, each frame was converted into the dimension of (224, 224, 3).

1) Global Features: From each video clip, 28 equally spaced frames were obtained using genetic algorithm . If the video contained less than 28 frames, the last frame was padded at the end of the video to make the total number of frames as 28. The frames were then passed through InceptionV3 model one-by-one for feature extraction. The combined output of the GlobalAveragePooling2D layer having dimension of (28, 2048) is considered as the feature vector.

2) Motion Features: Motion features were calculated on the video clip frames using a fixed window size of 15 frames using a C3D model pre-trained on Sports-1M dataset. Once the motion features were extracted for each window, 28 equally spaced windows were selected. The fc6 layer output of the C3D model was considered as the feature vector. Thus, considering all 28 windows, the dimensions for the motion features for each video came out to be (28, 4096).

3) Local Features: From 28 evenly spaced frames, the top-n objects are chosen to represent local features. The top-8 features from 28 evenly spaced frames of the input video were extracted using the output of the fc7 layer of the Faster RCNN network. The length of each feature vector is 4096. Consequently, each video's retrieved features are of the dimension (28, 4096).

## 4.2 Encoder-Decoder Model Training

The Google Colab environment, which has an Intel(R) Xeon(R) CPU running at 2.30GHz, 13 GB of RAM, and 12 GB of Nvidia K80 GPU, was used for the model training. Every description in the training data is viewed by the model as a different input. As a result, many data points may share the same feature vector as the input but have distinct description. With a training split ratio of 0.85, the train data is

separated into train and validation. The remaining 100 videos in the MSVD dataset are utilised for testing the trained models, while the additional 1870 videos have been taken into account for training and validation. The learning rate has been set to 0.0003 when using the Adam optimizer. Additionally, each word in has been represented by a classifier that uses the sigmoid activation function. Graphs of the description's loss against epoch for the model and accuracy against epochs were produced using the entire word pool. Graphs are shown in Fig. 4.2.1 and 4.2.2

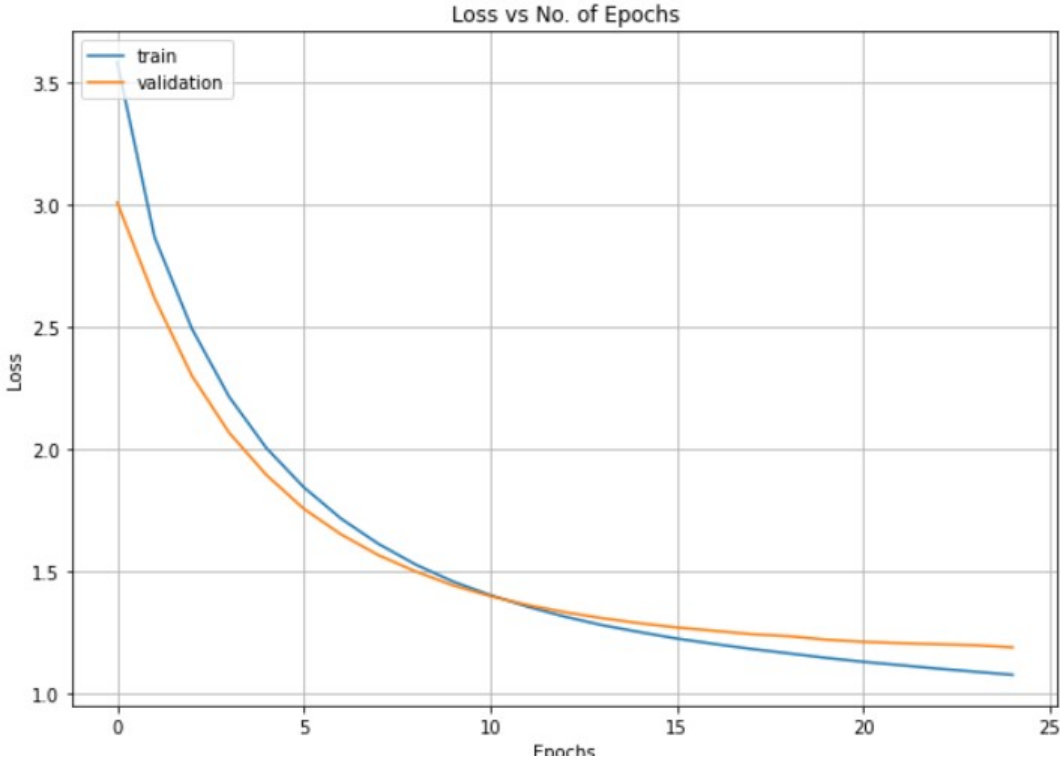


Figure 4.2.1: Loss vs No. of Epochs

## 4.3 Evaluation Score And Outputs

LSTM encoder-decoder network and GRU based encoder-decoder network's performances were evaluated using both the cases considering genetic algorithm based frames selection and other one where K equal spaced frames are selected. All the results are shown in the table 4.3.1. It can be observed from the table that the GRU along with the genetic algorithm based frames selection and attention mechanism

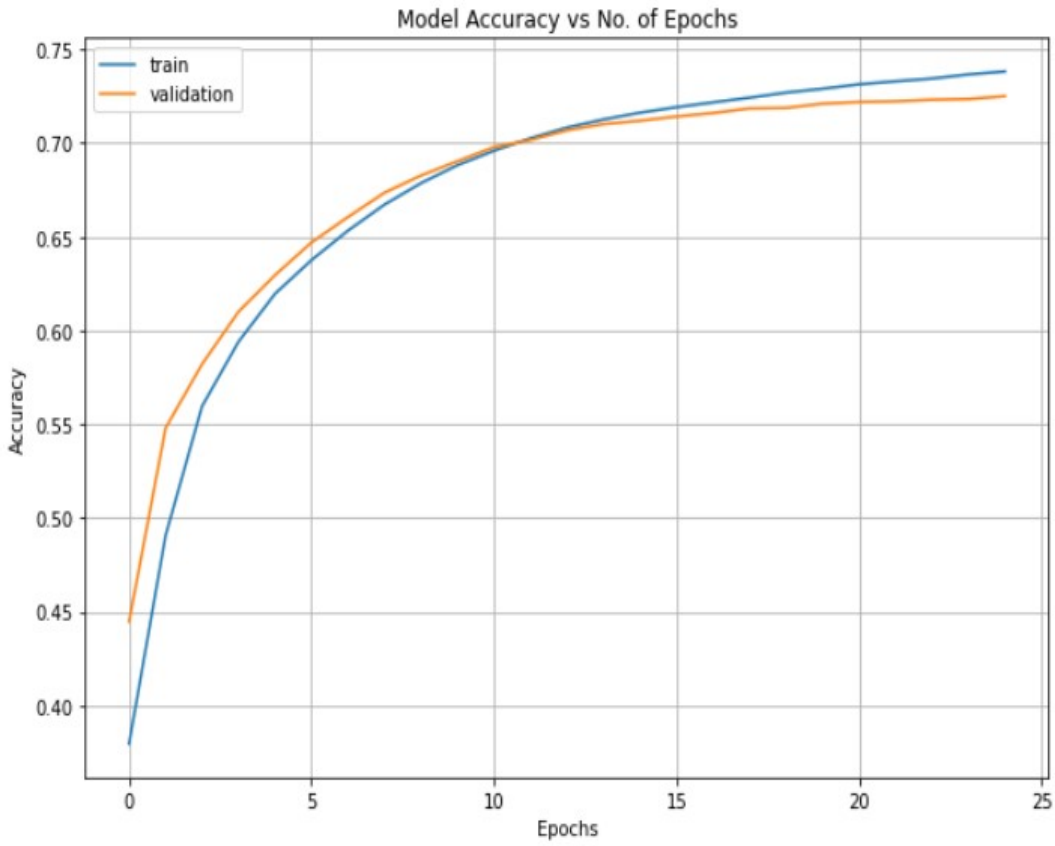


Figure 4.2.2: Model Accuracy vs No. of Epochs

performed best among all with BLEU-4 score of 0.386 whereas LSTM model without attention mechanism performed worst among all with the score of 0.262. To determine scores, BLEU- 4 compares the generated descriptions to a corpus of reference texts before averaging the results. Finding how well the machine-generated text matches the human translation of the video is the central tenet of BLEU-4. BLEU-4 does not consider recall in calculations.BLEU -4 is geometric mean of precisions. BLEU -4 does not consider semantic similarities as well. Table 4.3.2 shows the METEOR scores obtained on various models with and without GA and Table 4.3.3 shows the comparisions with various models. Table 4.3.4 and Table 4.4.5 shows the evaluation of recommender and search system. Recommender system is evaluated after uploading 10 videos and taking out average for precision@k and recall@k whereas search system is evaluated for 10 unique queries and taking out average for precision@k and recall@k.

| Methods | BLEU 4 Score (Without GA) | BLEU 4 Score (With GA) |
|---|---|---|
| GRU based encoder-decoder with attention mechanism | 0.371 | 0.396 |
| LSTM based encoder-decoder with attention mechanism | 0.305 | 0.318 |
| GRU based encoder-decoder without attention mechanism | 0.291 | 0.315 |
| LSTM based encoder-decoder without attention mechanism | 0.262 | 0.276 |

Table 4.3.1: BLEU-4 score of various models

| Methods | METEOR Score (With GA) | METEOR Score (Without GA) |
|---|---|---|
| GRU based encoder-decoder with attention mechanism | 0.299 | 0.287 |
| LSTM based encoder-decoder with attention mechanism | 0.252 | 0.241 |
| GRU based encoder-decoder without attention mechanism | 0.261 | 0.255 |
| LSTM based encoder-decoder without attention mechanism | 0.225 | 0.212 |

Table 4.3.2: METEOR score of various models

| Models | BLEU-4 Score | METEOR Score |
|---|---|---|
| GRU based encoder-decoder with attention mechanism (Proposed) | 0.396 | 0.299 |
| TA [7] | 0.419 | 0.296 |
| S2VT [6] | 0.332 | 0.288 |
| LSTM-E [9] | 0.453 | 0.310 |
| LSTM-YT [7] | 0.3119 | 0.27 |

Table 4.3.3: Comparisions With Various State Of The Art Models

| Value Of K | Precision@K | Recall@K |
|---|---|---|
| 10 | 0.70 | 0.411 |
| 20 | 0.70 | 0.82 |
| 30 | 0.50 | 0.88 |

Table 4.3.4: Evaluation Of Recommender System

| Value Of K | Precision@K | Recall@K |
|---|---|---|
| 10 | 0.70 | 0.50 |
| 20 | 0.55 | 0.79 |
| 30 | 0.36 | 0.79 |

Table 4.3.5: Evaluation Of Results Of Search

## 4.4 Outputs Of User Interface

Fig. 4.4.1 Shows the upload page of the application where users can upload videos. Fig.4.4.2 shows my uploads page, where users can see all his/her uploaded videos from past along with their descriptions generated by our model. Fig. 4.4.3 and Fig. 4.4.4 shows our recommendations page, where recommendations are shown based on uploads of users. Fig. 4.4.5 shows the search page where the search result for the query is shown.

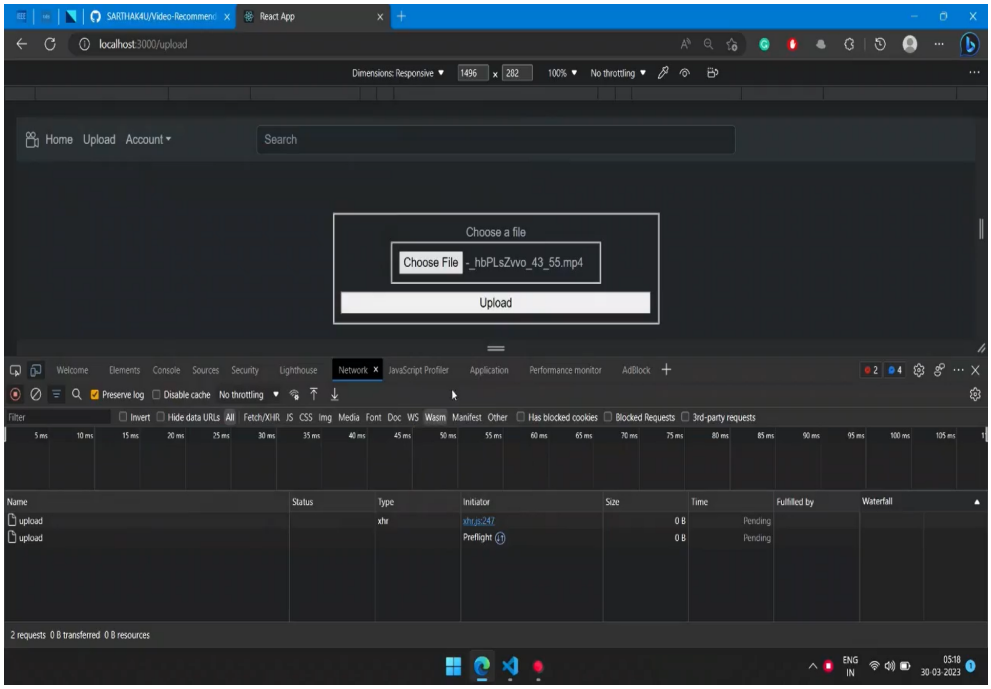


Figure 4.4.1: Uploading Of Videos

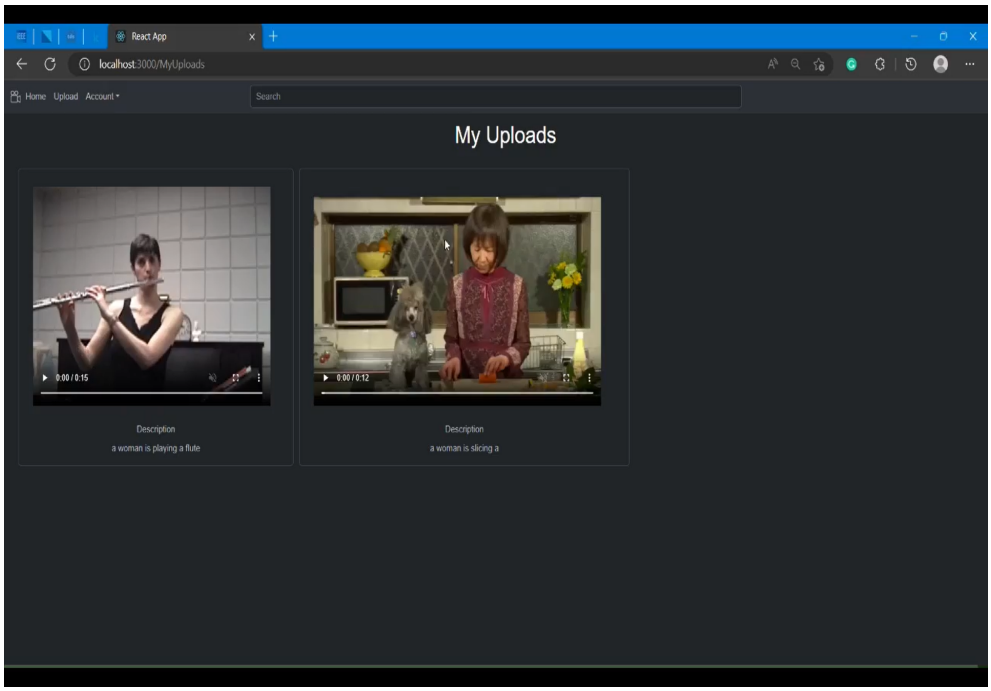


Figure 4.4.2: My Uploads Page

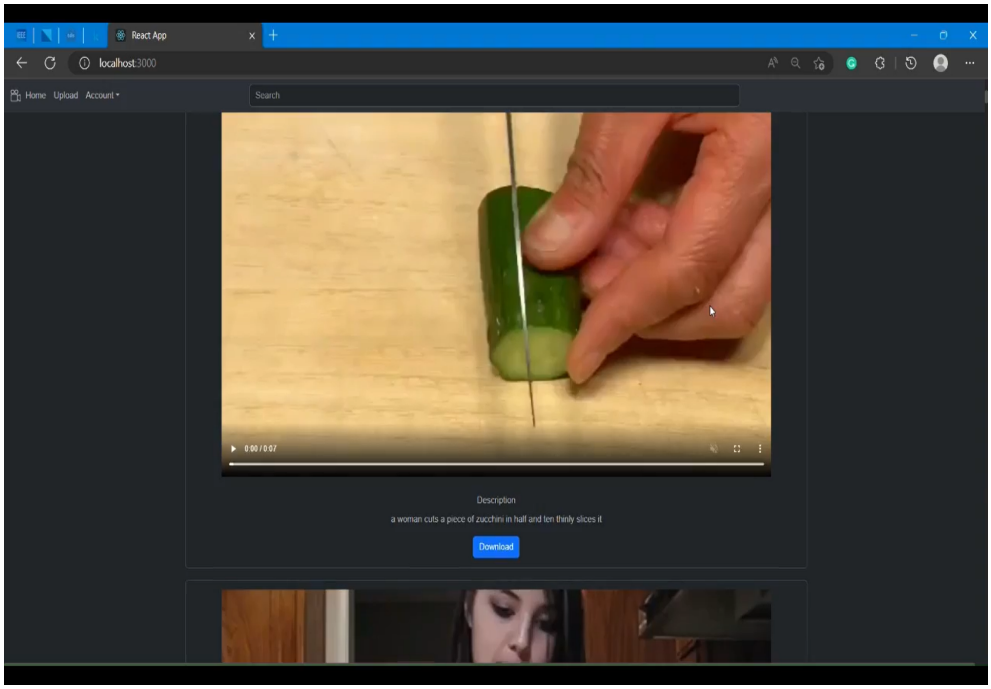


Figure 4.4.3: Recommendations

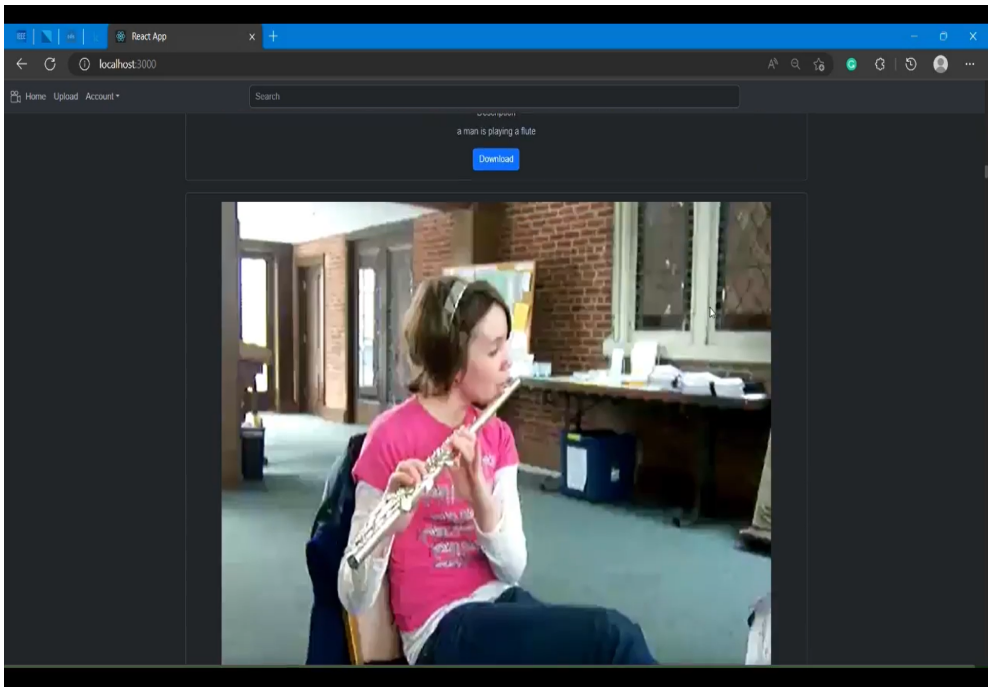


Figure 4.4.4: Recommendations

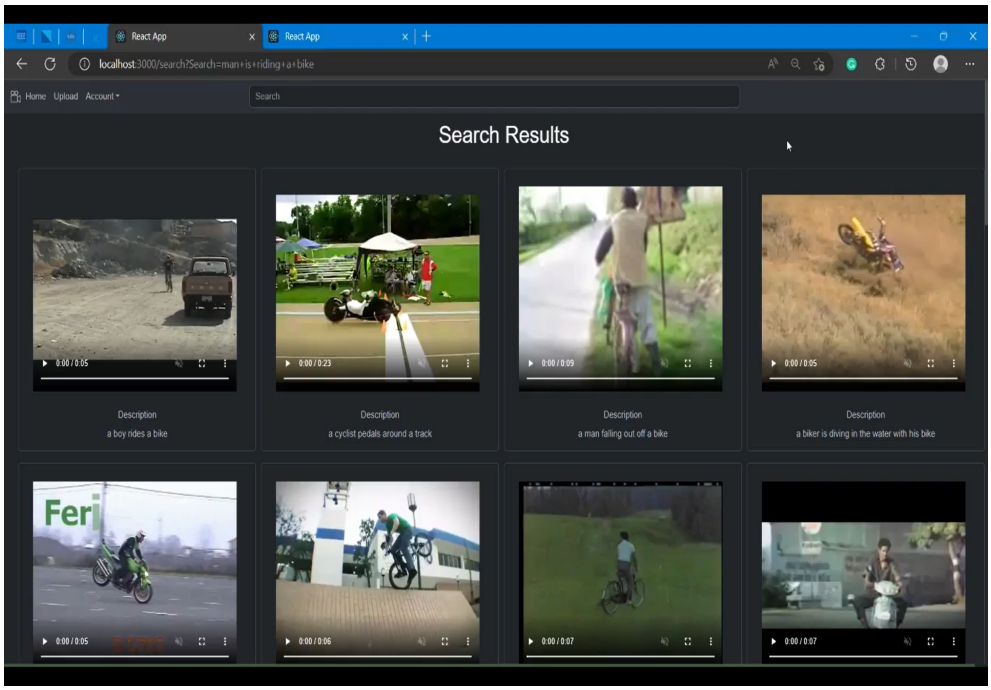


Figure 4.4.5: Search Result For A Query

# CHAPTER 5

# CONCLUSIONS

The proposed model consists of GRU based encoder and decoder. Since GRU has two gates and LSTM has three, GRU uses lesser memory and is faster than LSTM. The proposed framework makes use of the novel idea of using genetic algorithms to select the video frames which contain the most important information. GRU is combined with an Attention mechanism which improves the performance of encoder-decoder model human tendency to selectively focus on key areas of photos and videos to gain insight is the main inspiration for the attention mechanism. Objective of attention mechanism is to use most important parts of sequence provided as input in flexible manner. In this paper, a different feature extraction method is used. Instead of using one pre-trained feature extractor model, we are using models to extract global features, local features, and motion features. separately.These features are extracted from the frames which are selected by using genetic algorithms. Since they contain the most significant information, and the results are improved. Output of GRU based encoder-decoder model is fed as an input to recommender which consists of three models which are TF-IDF based, sentence transformer based and page rank based. Based on the added scores of these three models, top 10 video suggestions are given as output. The main contribution is the usage of genetic algorithms which reduces loss and increases efficiency.GRU based model is also helpful in faster computation and easy implementation. We have also designed a client server architecture using the frontend library ReactJS and flask for the backend. Our application is capable of showing recommendations of videos based on their uploads and also enables users to search for the videos. All the endpoints of the applications are functionalities obtained by implementing methodology mentioned earlier.

# REFERENCES


[1] R. Zhou, D. Xia, J. Wan and S. Zhang, "An Intelligent Video Tag Recommendation Method for Improving Video Popularity in Mobile Computing Environment," in IEEE Access, vol. 8, pp. 6954-6967, 2020, doi: 10.1109/ACCESS.2019.2961392.

[2] A. Sehgal, H. La, S. Louis and H. Nguyen, "Deep Reinforcement Learning Using Genetic Algorithm for Parameter Optimization," 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy, 2019, pp. 596-601, doi: 10.1109/IRC.2019.00121.

[3] Y. Sun, B. Xue, M. Zhang, G. G. Yen and J. Lv, "Automatically Designing CNN Architectures Using the Genetic Algorithm for Image Classification," in IEEE Transactions on Cybernetics, vol. 50, no. 9, pp. 3840-3854, Sept. 2020, doi: 10.1109/TCYB.2020.2983860.

[4] S. Duan, D. Zhang, Y. Wang, L. Li and Y. Zhang, "JointRec: A Deep-Learning-Based Joint Cloud Video Recommendation Framework for Mobile IoT," in IEEE Internet of Things Journal, vol. 7, no. 3, pp. 1655-1666, March 2020, doi: 10.1109/JIOT.2019.2944889.

[5] Raj, Amir Hossain, et al. "Deep Learning Based Video Captioning in Bengali." 2021 26th International Conference on Automation and Computing (ICAC). IEEE.

[6] Vaishnavi, J., and V. Narmatha. "Video Captioning based on Image Captioning as Subsidiary Content." 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT). IEEE, 2022.

[7] Agrawal, Rishi. "A Survey of Video Captioning Methods." 2021 5th International Conference on Information Systems and Computer Networks (ISCON). IEEE, 2021.

[8] Sehgal, Smriti, Jyoti Sharma, and Natasha Chaudhary. "Generating image captions based on deep learning and natural language processing." 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO). IEEE, 2020.

[9] Yenugula, Swapna, et al. "Automatic Image and Video Captioning Production using Deep Learning." 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). IEEE, 2022.

## TIMELINE OF THE B.TECH.(IT) MAJOR PROJECT

1. August - Topic Selection And Literature Review
2. September 2022 - Global and Motion Features Extraction And Basic Implementation of Encoder
3. October 2022 - Implementation Of Decoder, Attention Mechanism and Evaluation
4. November 2022 - Hyperparameter Optimisation and exploring information retrieval techniques
5. January 2023 - Implementation Of Gentic Algorithm Based Frames Selection And GRU Based Encoder-Decoder
6. February 2023 - Implementation of hybrid information retrieval method for recommendation of videos
7. March 2023 - Implementation of UI and other recommendation techniques