Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
Volume XVIII, Issue I, Jan-June 2024

YOUTUBE VIDEO SUMMARIZER: A WEB BASED APPLICATION FOR CONCISE VISUAL AND TEXTUAL SUMMARY

Gauri Mandar Puranik[1], Nidhi Kamath[2],Nakshatra Akhadkar[3],Gargi Dusane[4]

[1]Assistant Professor at GCOERC Nashik, [2]Student at GCOERC Nashik , [3]Student at GCOERC Nashik ,[4]Student at GCOERC Nashik

ABSTRACT

In today's world, YouTube has become a treasure trove of entertainment and knowledge, offering a vast array of content with varying degrees of quality and information. The problem we face today is selecting the most appropriate content for our needs. Our system is designed to assist our users in swiftly identifying the most suitable content, thereby saving them valuable time that might otherwise be spent on clickbait videos and leading thumbnails. The system utilizes the power of Python, along with its frameworks, libraries, and Natural Language Processing (NLP) capabilities to generate both textual and visual summaries. These summaries give them a quick and accurate overview of the original content, saving them precious time and increasing their productivity.

Keywords— Framework, Knowledge, Libraries, NLP, Python, YouTube.

[1] INTRODUCTION

In today's digital age, YouTube serves as a vast reservoir of information, entertainment, and learning. However, navigating through lengthy videos can be time-consuming, which is where the" YouTube Video Summarizer" steps in. This web-based application is your shortcut to efficient knowledge extraction. It can swiftly transform extensive video content into concise visual and textual summaries, allowing you to grasp the essential information in a fraction of the time. What makes this project special is its commitment to user-friendliness, accessibility, and cost-efficiency. Whether one is a student, a lifelong learner, or a busy professional, this tool empowers you to make the most of your time in the digital age, ensuring that valuable insights are just a click away.

[2] MOTIVATION

The project idea is to conquer or fill the gaps in the existing models or systems developed for the summarization of text and videos. To develop an architecture that is simple, lightweight and runs on minimum hardware specification. To provide users with an enhanced user experience through a user-friendly UI. To develop two modules one for text summarization and the other for visual content summarization and provide our users with a choice to select between them. The motivation behind the" YouTube Video Summarizer: A web application for complete visual and textual summary using Python and NLP" is to target college and school going students and learners who rely majorly on YouTube for their doubts, academics, concepts. To enable them to increase their time efficiency and not falling victims to clickbait videos with attractive thumbnails but no content .

[3] RELATED WORK

The work on this research paper started with a literature survey on several previously developed systems and methodologies. The literature survey has highlighted several key themes in the field of video summarization. Summarization using Deep Neural Networks have a high potential for creating visual and textual summaries. Some of the papers referred are as follows:

ANIQA DILAWARI, MUHAMMAD USMAN GHANI KHAN ASOVS : ABSTRACTIVE SUMMARIZATION OF VIDEO SEQUENCES IEEE ACCESS MARCH 20, 2019.

Joint end-to-end model is used for deep neural network to generate natural language description and abstractive textual summarization of an input video sequence .They formulate multitask feature learning framework using CNN (Convolutional Neural Net work) that extracts persons and its attributes (age, gender, emotion, etc), objects, scenes and actions to generate a multi-line video description. Which is then passed onto RNN(Recurrent Neural Network) for text processing .RNNs are the elementary structure of deep learning which is capable of storing information and learning sequential data to foresee the forthcoming number of but, these save information in memory only for a certain time period. Thus they have used LSTM(Long Short Term Memory). An LSTM unit comprises of a memory cell that is able to store information for longer duration of time and specifically used the bidirectional LSTM architecture for video description generation. Both forward and backward information can be processed to capture contextual information from the data .In context the local features such as human attributes like age , gender , emotion object and action are mined using CNN which is then fed into bidirectional RNN for textual generation which is then passed onto attention based LSTM.[1]

Siddhartha, Prashu Pandey, Ansh Saxena, YouTube Transcript Summarizer International Journal of Research in Engineering and Science (IJRES) . 6-5-2023

In this paper, the system automatically generates a summary from the transcript of the video. The proposed model exploits the various NLP techniques and tools for extracting information, implementation and testing on YouTube Transcript dataset. The model builds a Chrome extension with the help of various APIs like the Python API, Flask API , YouTube API and various frontend languages like HTML, CSS, and JavaScript. User downloads the Chrome extension. •User opens a YouTube video and clicks on the summarize button of the extension. The event triggers an HTTP request to the backend to give the transcript for the given YouTube id. The response in the transcript of the video. The model performs the text summarization and displays a summarized text to the user.[2]

P. Vijaya Kumari ,M. Chenna Keshava, C. Narendra, P. Akanksha, K. Sravani YouTube Transcript Summarizer Using Flask and NLP Journal of Positive School Psychology, 2022.

In this paper, a flask backend is used, which receives API calls from users and gives responses with summarized text. The client sends a request to the Flask backend server, which in turn asks for subtitles from the YouTube server using the YouTube API. The extracted subtitles are given to the CNN model on the backend server. And the summarized text will be received by the client. In the application, the user is allowed to download the summarized text in different file formats and also translate the transcript text into different languages provided.[3]

Yudong Jiang, Kaixu Cui, Bo Peng, Changliang Xu, Comprehensive Video Under standing: Video summarization with content-based video recommender de sign, 2019.

In this paper, the system deploys a scalable deep neural network. It establishes scene and action recognition for untrimmed videos. The system also uses multi-task learning and augmented reality to prevent overfitting of the model. The video summarization network contains three subnetworks SegNet, VideoNet and HighlightNet . The dataset consists of 100 videos. The goal of the Video Summarization framework is to select important frames. The multi-tasking framework enhances the robustness of the model. The summarized video is then generated by concatenating frames with high scores.[4]

Cheng Huang, Hongmei Wang A Novel Key-frames Selection Framework for Com prehensive Video Summarization , IEEE Transactions on Circuits and Systems for Video Technology (Volume: 30, Issue: 2, February 2020)

The proposed method utilizes CapsNet features of the frame sequences to represent the spatiotemporal information and generate inter-frames motion curve. A transition effects detection (TED) method is proposed for quickly automatic shot segmentation on this curve. Then the self-attention model is able to learn the inner laws of each shot to select the key frames. Finally the content summarization and motion information summarization of the video are aggregated by processing those key frames. They extract important information from the frames in a video using a special technique called Capsule Networks (CapsNet). Then create a curve that shows how things move across frames in the video. This helps to find where one scene or shot ends and another begins by spotting changes in the motion curve. Then figure out which frames in each scene are the most important.[5]

Jiatong Li, Ting Yao , Qiang Linga, Tao Mei, Detecting shot boundary with sparse coding for video summarization ,Neurocomputing 266 ,2017

In this paper, the summarization takes place by learning a dictionary from the origi nal content. The dictionary is a set of rules or visual patterns the system will follow to understand the video. The feature reconstruction loss ensures that the dictionary is used completely and the resulting video is similar to the original video. The sparsity control ensures that the content is as simple as possible and the model does not enter into overfitting. The important frames and features are represented as the video follows the learned dictionary. Shot boundaries are detected based on the feature representation. Shot boundaries are grouped and selected and a concise result is generated.[6]

Siyu Huang, Zhongfei Zhang, Fei Wu, JunWei Han,User-Ranking Video Summariza tion with Multi-Stage Spatio-Temporal Rep- resentation, IEEE TRANSAC TIONS ON IMAGE PROCESSING, 2018

They have made use of deep learning techniques like 2D Convolutional Neural Net work(CNN),1D CNN's to extract features and LSTM(Long Short Term Memory to per form spatial and temporal analysis. The system extracts visual features like age, gender from video frames which serve as input for neural network stages .These frames go through three stage neural network architecture. User ranking mechanism is used to improve the quality of the generated summaries .Log weighted scoring is used for identifying the most important parts from the videos. The summaries formed provide salient features which his relevant to the videos .The 1D-CNNs capture the short-term temporal context in a local range around current frame, while the LSTMs capture the long-range context of the sequence.[7]

Mengjuan Fei, Wei Jiang, Weijie Mao, A novel compact yet rich key frame creation method for compressed video summarization, Multimedia Tools and Applications volume 77, 2017.

They have proposed key frame generation method that performs an optimum programming function to optimally select what object activity is displayed on the generated key frame. Then, they attempt to automatically and seamlessly stitch the selected object activities on an image. Based on moving-object-outlier-detection, they perform a matting technique on the selected objects outliers and splice them into the final key frame. Through such a matting approach, visually unpleasant seams are avoided and a natural key frame is generated. An effective summary should maintain the point of interest, diversity. Using Mutual Information(MI) which detects significant changes the original video is segmented into shots. By analysing motion vectors static objects and moving background is detected in the shot. Suitable shot creates key frames while maintaining moving and static background. KNN(K- nearest neighbour matting based stitching is used to stitch the objects in the frame to avoid visual inaccuracy. [8]

[4] METHODOLOGIES OF PROBLEM SOLVING AND EFFICIENCES

Summarization using Deep Neural Networks is highly potential for creating visual and textual summaries .*Convolutional Neural Networks (CNNs)* : CNNs are types of deep learning models that have multiple layers and slide filters through them for detecting patterns and features. They can identify scene changes and select frames at the start of each scene. They can create a visual summary by selecting important keyframes and arranging them in a sequence. *Recurrent Neural Network (RNNS)* : RNNs are different from CNNs as they have a dynamic structure and memory of previous inputs. LSTM, i.e., long short-term memory, is a type of RNN that specializes in maintaining a hidden state that serves as memory. The principle behind RNN is to process each element in a sequence while preserving an internal state of memory from previous input.
Natural Language Processing : The NLP techniques help in understanding and generating human language. It provides us with various features and tools for text preprocessing, keyword extraction, text analysis, transcription, etc. The subset, i.e., Natural Language Understanding (NLU), helps in deeper understanding of human text, provides semantic meanings of sentences, and captures various entities of video like names, places, etc.

Efficiency : The efficiency of RNNs and LSTM on various datasets consisting of 20–25 user videos comes out to be 40–60 percent. The LSTM performs better than the CNN model as the LSTM has a higher degree of complexity, i.e., they have a hidden state for memory, so they perform well in identifying various patterns. The efficiencies of NLP and NLU approaches largely depend on dataset inputs, text preprocessing, and model complexity.

[5] SYSTEM ARCHITECTURE

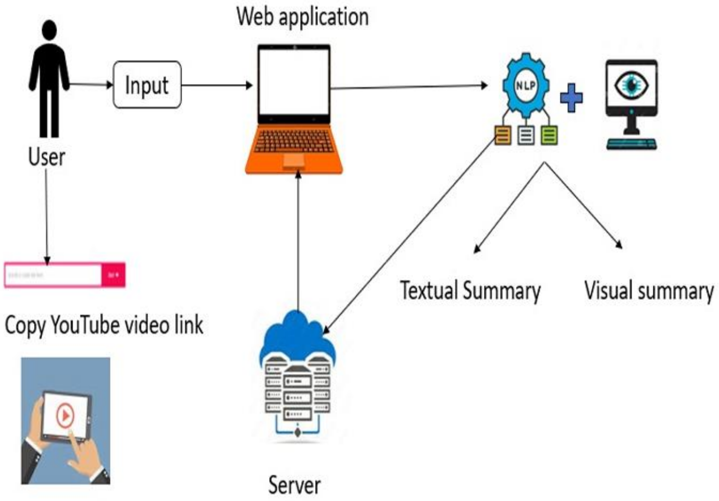


Fig 1 System Architecture

The user copies the link of the desired YouTube and gives it as input to our system's frontend. The user then chooses between textual and visual summaries. The system with the help of Python libraries fetches the details of the video and the transcript for it. The backend consisting of Bert Model the system performs text summarization and visual summarization. Based on the user's choice, the desired output is then displayed to the user. The final summarized video also gets downloaded on the system.

[6] PROPOSED ARCHITECTURE

Utilizing pre-trained BERT Model :
BERT Model is a bidirectional model that reads the sentence from both the ends and understands the context of the words in the sentence. The architecture of BERT is based on top of transformers. It is a self -attention model .It leverages powers of supervised pre-training and then supervised fine-tuning.

Data Cleaning:
The datasets of YouTube reviews , comments undergoes data wrangling which includes missing data handling ,normalization and transformation to make it suitable for training.

Text Pre-Processing :

The input text is converted into tokens suitable for training BERT model. Then the tokens are fed to BERT model. The model selects the important words and then generates a precise and concise textual summary .

Post Processing:
For visual summary, the audio transcripts will be analyzed and processed. The audio transcripts will be trimmed along with the video transcripts and at the end all the trimmed clips will be concatenated together generating the visual summary.

Flask Integration :
The framework is used for rendering templates and routing , defining endpoints to handle HTTP requests from users .It incorporates BERT summarization logic .It renders the output to user based on his / her choice.

Workflow :

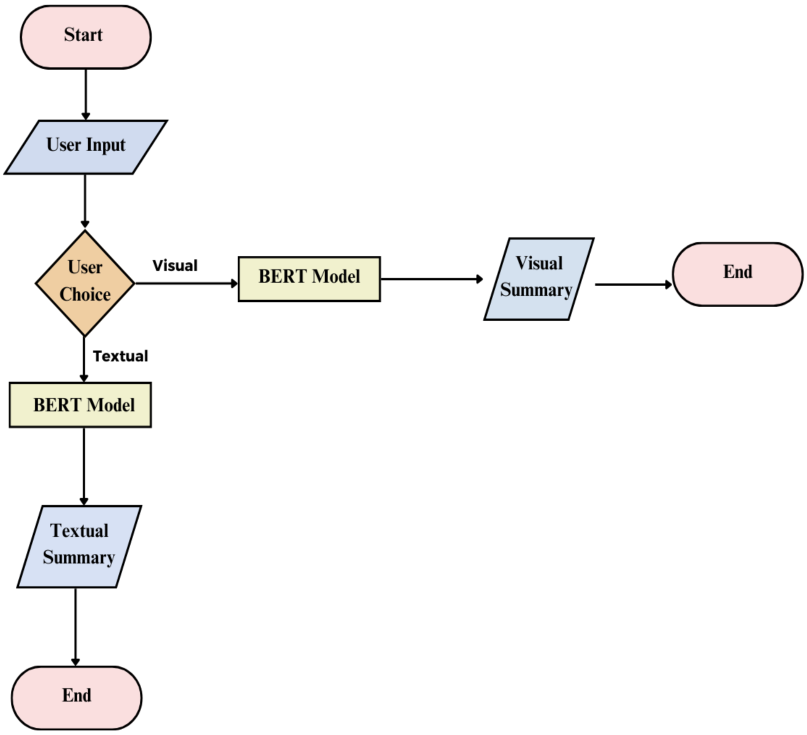
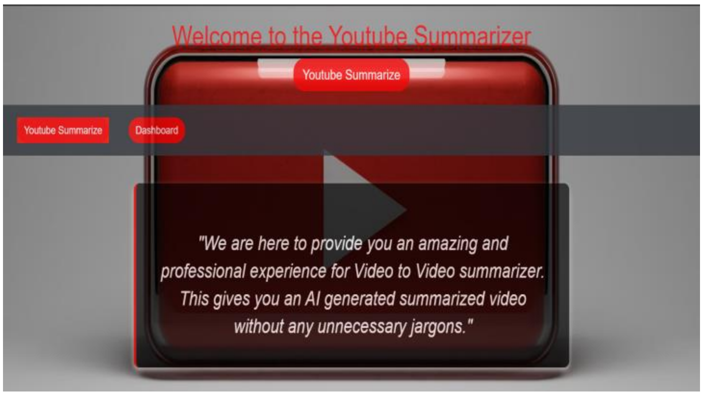


Fig 2 Workflow

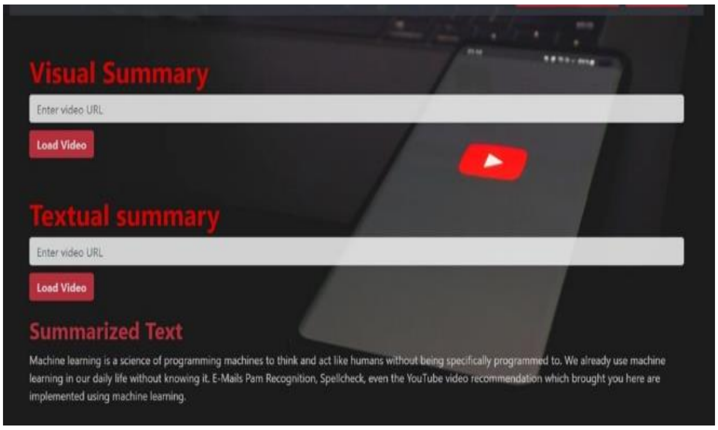
[7] RESULTS

We have created a user friendly UI which is interactive .Our platform features a sleek homepage that effortlessly guides users to their desired destinations, along with a dynamic dashboard .The accuracy of the summarized content is 89 percent .We have provided the system with many YouTube videos of 10 to 12 mins which the system has summarized to 2-3 mins and has generated a rich summary for our users.



"We are here to provide you an amazing and professional experience for Video to Video summarizer. This gives you an AI generated summarized video without any unnecessary jargons."

Fig 3 Homepage



Machine learning is a science of programming machines to think and act like humans without being specifically programmed to. We already use machine learning in our daily life without knowing it. E-Mails Pam Recognition, Spellcheck, even the YouTube video recommendation which brought you here are implemented using machine learning.

Fig 4 Dashboard with textual summary

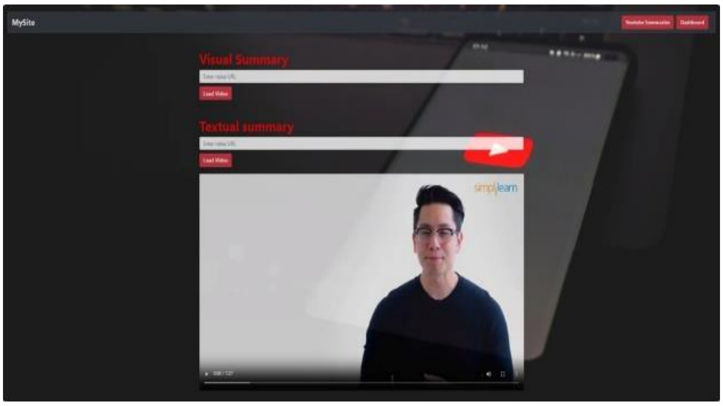


Fig 5 Dashboard with visual summary

[8] APPLICATION

E-learning : The 'YouTube Video Summarizer' project is a valuable asset in the education and e-learning domain. Its primary purpose is to provide condensed versions of lengthy educational videos, saving time and improving content comprehension for learners. Users can select between textual or visual summaries, while educators can streamline their content, particularly

valuable in distance education. *Corporate Training :* IT companies can use this system to create training modules for freshers.
Content creators : They can use this tool to create user friendly and time friendly videos and address users' time constraints.
Product Reviews : The system can be used to summarize various product reviews enabling enhanced shopping experience for users.

[9] LIMITATIONS

Internet connectivity : Users need an good Internet connectivity.
Transcript Dependency : Our textual summary relies on captions of the YouTube video.

[10] CONCLUSION

The YouTube Video Summarizer leverages the power of the BERT model and generates a rich textual summary and visual summary. It empowers our users by giving them a choice between textual and visual summary. The summary is precise and saves our users' precious time and prevents them from falling victims to clickbait videos and irrelevant content. The system is built using Python to ensure its robust functioning and fast results.

REFERENCES

[1] Aniqa Dilawari, Muhammad Usman Ghani Khan, ASoVS: Abstractive Summariza tion of Video Sequences, IEEE Access March 20, 2019.

[2] Siddhartha, Prashu Pandey, Ansh Saxena, Youtube Transcript Summarizer International Journal of Research in Engineering and Science (IJRES) .May 6-5,2023

[3] P. Vijaya Kumari, , M. Chenna Keshava, C. Narendra, P. Akanksha, K. Sravani, Youtube Transcript Summarizer Using Flask And Nlp Journal of Positive School Psychology , 2022 .

[4] Yudong Jiang, Kaixu Cui, Bo Peng, Changliang Xu, Comprehensive Video Under standing: Video summarization with content-based video recommender design ,2019 IEEE/CVF International Conference on Computer Vision Workshop.

[5] Cheng Huang, Hongmei Wang A Novel Key-frames Selection Framework for Com prehensive Video Summarization , IEEE Transactions on Circuits and Systems for Video Technology (Volume: 30, Issue: 2, February 2020).

[6] Jiatong Li, Ting Yao , Qiang Linga, Tao Mei,Detecting shot boundary with sparse coding for video summarization,Neurocomputing 266 ,2017.

[7] Siyu Huang, Zhongfei Zhang, Fei Wu, JunWei Han, User-Ranking Video Summariza tion with Multi-Stage Spatio-Temporal Rep- resentation, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. XX, NO. X, 2018.

[8] Mengjuan Fei, Wei Jiang, Weijie Mao, A novel compact yet rich key frame creation method for compressed video summarization, Multimedia Tools and Applications volume 77, 2017.