NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES

FYP-II FINAL REPORT

ON

QUERY BASED VIDEO SUMMARIZATION

SUBMITTED BY

ALISHBA SUBHANI 20K-0351
MANNAHIL MIFTAH 20K-0234
MUHAMMAD UMER 20K-0427

SUPERVISOR

Dr. Muhammad Rafi

Abstract

Generic video summaries are the concise versions of lengthy videos which provide full information about the content of the whole video. However, user specific video summaries are required to produce a summary customized to the user's needs. Existing models for this task do not harness the powers of pretrained huge models like blip, ViTs and LLMs. An approach was implemented based on a combination of GPT2 and ViT to produce dense scene captioning of the videos. To retrieve desired frames with respect to the user query, LLAMA is used because of its efficiency in information retrieving tasks. This model is finetuned on videos from ego4d dataset which is the world's largest egocentric (first person) video ML dataset and benchmark suite. A subset comprising 44 videos is selected from this repository. Experimentation yielded promising results with F1 score and recall being 47.7 and 57.3 respectively through comparing the frames with ground truth summaries.

Contents

1 INTRODUCTION

1.1 Motivation

Video content has become increasingly prevalent across various platforms, resulting in a massive influx of videos that users encounter daily. As the volume of video data grows, the need for effective and efficient video summarization techniques becomes paramount. Because it has become almost impossible for one to keep up with the latest trends, and knowledge acquisition through videos is now an extremely time consuming task. Therefore, there must be a solution to revolutionize the way of interacting with rich and diverse video content and provide the user with such a summary that fulfills his personalized criteria. This is possible by generating video summaries through user written queries.

1.2 Problem Description

Video summarization involves condensing the essential information from a video while maintaining its meaningful context. The generated summary should have the following attributes: coverage, sparsity representativeness, diversity and cohesiveness [2]. This requires generating a summary consisting of significant shots, sequences, and events from the original video, which, being far shorter than the source video, guarantees the original content's comprehension, integrating relevant elements according to the user specified query while simultaneously limiting recurrence and maintaining the coherence amongst the shots. This problem statement addresses the challenge of developing a robust video summarization system, which produces summaries customized to the user's demands utilizing deep learning techniques.

1.3 Problem Statement

The goal of this project is to create an end-to-end deep learning solution for a query based video summarization, where the system automatically generates concise yet informative summaries of input videos pertaining to the query provided by the user. The summarization process should capture the most important scenes, actions, events, and content from the video while maintaining coherence and avoiding redundancy and fulfilling all the aspects of user provided context.

Figure 1: Generation of two different summaries from a same video based on the individual queries of users.

1.4 Gap Analysis

Existing video summarization techniques revolve around developing deep learning models from scratch and training those models on the custom dataset. This approach produced significant results, but the sudden boom in generative AI and consequent development of huge models trained on immense amounts of data superseded this approach. The reason behind is the extreme amount of computational resources put into development of such models. Models developed from scratch cannot beat the performance of these pre-trained architectures, hence to fully benefit ourselves from them is finetuning them on our custom data. This way we can leverage their remarkable performance on our custom dataset.
Moreover, there are a substantially low number of existing approaches which incorporate generation of video summaries as per the user requirements. We provide a one-click solution which harnesses the capabilities of Vision Transformer and GPT-2, and also generate user curated video summaries.

2 LITERATURE REVIEW

Video Summarization has been a prevalent topic of research in the Computer Vision domain of Artificial Intelligence. There have been multiple approaches towards this problem's solution. A brief review of related literature is presented below.

2.1 UniMD: Towards Unifying Moment Retrieval and Temporal Action Detection

The authors propose UniMD, a unified framework that tackles both moment retrieval and temporal action detection tasks.The UniMD framework consists of a shared encoder that extracts features from input videos, task-specific heads for moment retrieval (MR) and temporal action detection (TAD), and multi-task learning to optimize both MR and TAD losses simultaneously. The model is trained using a multi-task loss function that combines moment retrieval and action detection losses. The UniMD model achieves competitive performance on the Charades and Charades-STA benchmarks, with a new state-of-the-art (SOTA) result of 63.98 in R1@50 for moment retrieval and significant improvement in temporal action detection when incorporating CLIP features. On the ActivityNet (ANet) and ANet-Caption benchmarks, the co-trained UniMD model sets new SOTA results, achieving 39.83 mAP and 60.29 mAP@50 in temporal action detection and 80.54 R5@50 in moment retrieval. On the Ego4D dataset the model achieves state-of-the-art results, outperforming existing methods in both Temporal Action Detection (TAD) and Moment Retrieval (MR) tasks by achieving the result of 44.8 in R1@50 for moment retrieval [15].

2.2 CLIP-It! Language-Guided Video Summarization

The authors proposed CLIP-It, a novel framework that tackles both generic and query-focused video summarization tasks in a single model. It uses a language-guided multimodal transformer to score video frames based on their importance and correlation with a user-defined query or automatically generated caption. The model can

be trained with or without ground-truth supervision and achieves state-of-the-art results on standard video summarization datasets (TVSum and SumMe) and a query-focused dataset (QFVS), demonstrating strong generalization capabilities. The CLIP-It method achieves an average F1 score of 54.55 on the QFVS dataset, outperforming the best baseline by 10%, and demonstrates its ability to generate different summaries for the same video based on different input user queries. The qualitative results show that the method can effectively identify relevant frames in the video that match the input query, such as "Book and chair" or "Sun and tree", and generate summaries that are conditioned on the user's interests [16].

2.3 Shot selection based Video Summarizing techniques

[13] [14] approaches the video summarization by detecting shot boundaries or extracting keyframes [11], reducing redundancy, and the scene changes method. These methods do not generate summaries according to the user's interest. One of the methods of summarizing videos works by detecting shots by measuring the transition between successive frames.

2.4 Supervised Video Summarizing Techniques

According to the studies in [1] and [2], video summarization is treated as a seq-to-seq problem which produces the keyshot/keyframe sequence. The understanding of heterogeneous inter-dependency between video frames plays an important role in generating accurate video summaries. LSTMs incorporated with encoder and decoder frameworks were used in both studies to model variable range dependency. Since LSTMs are known to solve the vanishing gradient problems posed by RNNs, both [1] and [2] achieved state-of-the-art results. But this solution comes with one shortcoming that all the input frames are given the same importance regardless of the output shots to be predicted. This is because all the necessary information is encoded in one single context vector. To aid this problem, both studies simply ignored the temporal structure of the video.

2.5 Attention Based Video Summarizing Techniques

Basavarajaiah M and Sharma P [4] proposed approaches which aim to model frames' dependencies using variants of Transformer Network's self attention mechanism. One of the approaches [5] calculates regression of the frames' importance scores by combining a soft self attention mechanism with a two-layer fully connected network. Liu et al [7] present a hierarchical based approach that first identifies the shot-level candidate key-frames, and then uses a multi-head attention model to evaluate and select key-frames for the summary. Li et al [6] enhance self-attention mechanism's training pipeline by adding a step that computes attention values in order to increase the visual content diversity of the summary. These computed attention values are then used to determine the frames' importance. A variation of architecture from [5] is proposed by Ghauri el al [8], incorporating video content's additional representations. In addition to using typical CNN-based features from the pool5 layer of GoogleNet [9] trained on ImageNet, an Inflated 3D ConvNet [10] trained on Kinetics is also employed. These different sets of features are processed through self-attention mechanism, and the outputs from this are then combined to form a common embedding space for video frames' representation.

2.6 Tabular Representation of Literature Review

RESEARCH PAPER	METHODOLOGY	RESULTS/KEY FINDINGS
UniMD: Towards Unifying Moment Retrieval and Temporal Action Detection (2024)	The methodology involves a transformer-based architecture that takes as input a video and a query, and outputs a set of relevant moments and their corresponding action labels. The model is trained using a multi-task loss function that combines moment retrieval and action detection losses.	• Ego4D: R1@50 is 44.8% for MR. • ActivityNet (ANet) and ANet-Caption: R5@50 is 80.54% for MR. • Charades and Charades-STA: R1@50 is 63.98% for MR.
CLIP-It! Language-Guided Video Summarization (2021)	A language-guided multimodal transformer is used to score video frames based on their importance and correlation with a user-defined query or automatically generated caption.	Achieved an average F1 score of 54.55% on the QFVS dataset, outperforming the best baseline by 10%
Summarizing Videos using Concentrated Attention and considering the uniqueness and diversity of the video frames (2022)	Supervised video summarization with VASNet-based soft self-attention mechanism.	Proposed an approach based on VASNer architecture with a soft self-attention mechanism. Trained the model in an unsupervised manner, considering frame dependencies and producing block diagonal sparse attention matrix.
Unsupervised video summarization framework using keyframe extraction and video skimming (2021)	Generalized unsupervised video summarization using clustering and video skimming.	Explored a framework for extracting vision keyframes through clustering and video skimming. Utilized techniques like VGG16 and K-means algorithm for feature extraction and clustering.

Summarization Via Multiple Feature Sets with Parallel Attention (2021)	Proposed Multi-Source Visual Attention (MSVA) model combining image and motion features through parallel self-attention mechanisms. Utilized pre-trained models (GoogleNet for image features & I3D for motion features). Applied linear layers and activation functions for intermediate fusion.	Achieved best performances with intermediate fusion architecture. Outperformed other methods, including VASNet, on benchmark datasets.
Query-controllable Video Summarization (2020)	Proposed a three-module method for query-controllable video summarization. Video Summary Controller: Takes user query, produces vector representation using bag of words. Video Summary Generator: Composed of a CNN structure and multi-modality features fusion module. Feature Fusion Module: Combines frames_based features with input text-based query features to get relevance scores. Video Summary Output Module: Generates video summary based on relevance score prediction vector.	Enables query-controllable video summarization by incorporating user queries.
Summarizing videos with attention (2019)	Introduced VASNet sequence-to-sequence model architecture. Replaced LSTM encoder-decoder with soft self-attention mechanism and two-layer fully connected network for regression. Employed CNN feature vectors for input	Outperformed previous methods, including LSTM based architectures. Soft self-attention mechanism demonstrated effectiveness in capturing dependencies.

	representation & applied knapsack algorithm for keyframe selection.	
Video Summarization by Learning from Unpaired Data (2019)	Unsupervised video summarization by learning mapping from raw videos to summaries. Adversarial objective used for distribution matching.	Proposed a formulation for video summarization from unpaired data. Model learned mapping to generate summaries with similar distribution to the set of video summaries.
Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks (2019)	Unsupervised video summarization with self-attention mechanism and feature selector.	Integrated self-attention mechanism in Generator-Discriminator architecture. Used feature selector for better frames' importance estimation. Modeled long-range dependencies with multi-head self-attention.
Video summarization by visual co-occurrence (2015)	Co-summarization using Maximal Biclique Finding (MBF) algorithm.	Identifies key visual concepts by finding shots that co-occur frequently across videos. Utilized MBF algorithm for sparse co-occurring patterns.

3 DATASET

Ego4D is a large-scale, multi-modal dataset for egocentric (first-person) video understanding. It consists of over 3,000 hours of video recordings from 700 unique individuals, capturing daily activities like cooking, cleaning, and socializing. The dataset includes:

1. 3,000+ hours of egocentric video.
2. 40,000+ annotated episodes (short video clips).
3. 12,000+ hours of audio narration.
4. 2.5 million+ annotated objects, actions, and events.

The Ego4D dataset has been limited to the use case of cleaning and laundry videos for this project. Total 44 videos of varying length (3 to 30 minutes) were downloaded and used for training the captioning model.
The data contains the description of frames, frame number and time stamp along with the video.

```
narrations['062f1e55-67d8-4cdc-89c6-7fb361a9b0f9']
```

```
[{'time': 0.0320312,
  'frame': 0,
  'text': '#C C places a picture frame on a wall.'},
 {'time': 0.9191312,
  'frame': 27,
  'text': '#C C removes the picture frame from the wall.'},
```

Figure 2: Sample Annotation

3.1 Significance

Instead of benchmark datasets of video summarization; TVSUM and SUMME, Ego4D dataset was used for this project because query focused video summarization requires the natural language representations of videos in order to produce video summaries guided by the user query which is also given in natural language.

4 METHODOLOGY AND ARCHITECTURE

The process of video summarization entails two primary tasks: image captioning and retrieving frames similar to a user query. For this purpose, the following steps were performed:

1. A given video (30 FPS) is broken down into its constituent frames.
2. All those frames were selected for training whose descriptions were given in the dataset.
3. These frames are saved as numpy arrays and PIL Images.
4. A hugging face dataset is created which contains frames and their respective descriptions.
5. A combination of two large pretrained models (ViT and GPT-2) is finetuned over this custom dataset.
6. For each video, a description is generated every 15th frame (2 frames' descriptions generated per second).
7. The set of these generated descriptions along with their respective frame numbers are made into documents of LlamaIndex.
8. These nodes are stored into VectorStoreIndex which acts as a retriever.
9. The retriever is then queried with the user query, and it outputs top 60 frames similar to the user query.
10. The overlapping frames are eliminated to avoid repetition of frames in a summary.
11. For every similar frame preceding and succeeding 4-second segments are selected alongside them to maintain coherence within the summary.

12. The generated summaries are evaluated against human generated ground truth. The evaluation metrics used are recall and F1-score.

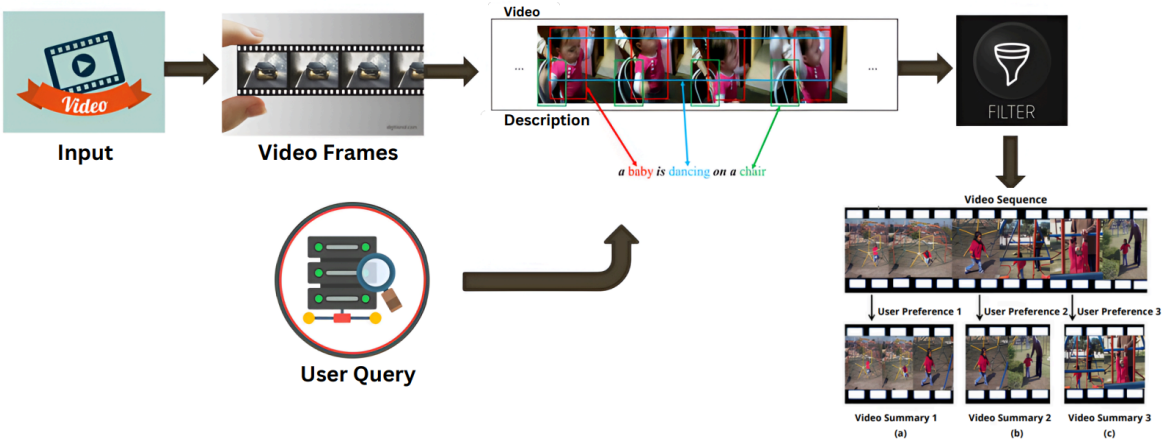


Figure 3: High Level Framework

4.1 Models Explanation

4.1.1 Image Captioning by finetuning ViT-GPT2

The decision to incorporate pretrained models like ViT and GPT-2 into our video summarization task was driven by their widespread acceptance and proven effectiveness. ViT was employed as the encoder to comprehend the visual content within the video and generate image representations, while GPT-2 served as the decoder to convert this information into natural language, producing frame-level descriptions by decoding the embeddings generated by ViT. Initially, the video was segmented into frames, and then the visual features of each frame were extracted using ViT's AutoFeatureExtractor, resulting in image embeddings, i.e., pixel values, for each frame. Simultaneously, GPT-2's AutoTokenizer was used to tokenize the descriptions of each frame. These image representations, along with their corresponding caption embeddings as labels, were fed into the GPT-2 model. These caption embeddings guided the learning process of GPT-2 to generate relevant frame descriptions. Consequently, descriptions of video frames were generated and saved with the respective frame number, to retrieve frames pertinent to the user query by leveraging the similarity between the generated frame-level descriptions and the user query.

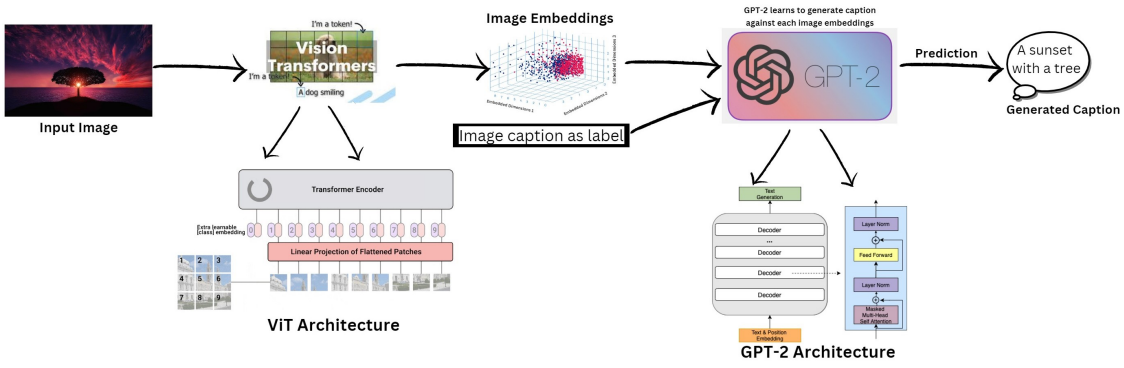


Figure 4: Image Captioning Model

4.1.2 Similar Frame Retrieval using LlamaIndex

Descriptions of video frames are converted into documents by extracting frame numbers and corresponding generated text. Each text is then encapsulated within a document along with its associated metadata, which includes the frame number. Subsequently, a LangchainNodeParser is employed to parse nodes from the documents, utilizing a RecursiveCharacterTextSplitter. During this parsing process, the embeddings model from hugging face is utilized to generate embeddings for documents list, thereby constructing structured nodes. These nodes are then used to construct a VectorStoreIndex, which organizes the embeddings in a manner conducive to efficient similarity search operations. The VectorStoreIndex is converted into a retriever with a specified parameter (20, in our case) for the number of similar items to be retrieved with respect to the given user query.

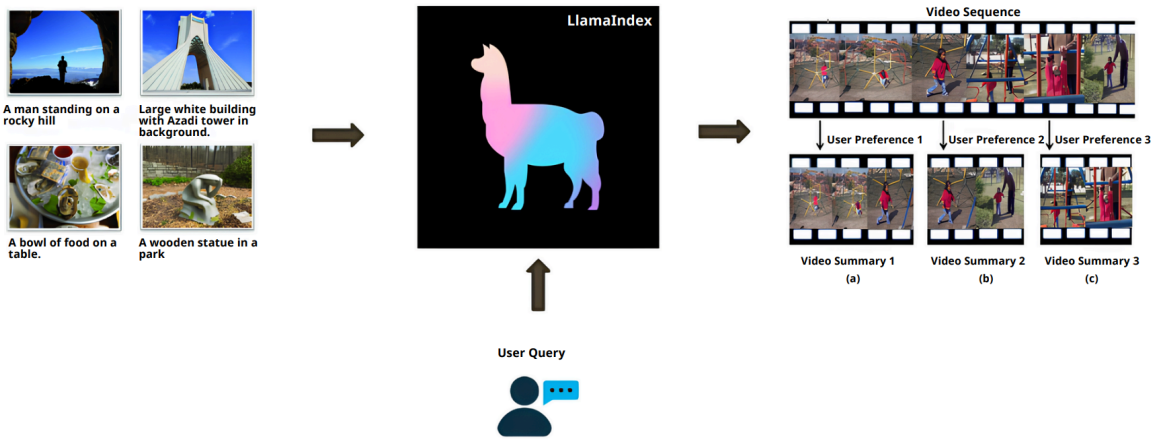


Figure 5: Frame Retrieval

5 EVALUATION

The criteria of the efficiency and performance of any model is done through evaluation matrices. The evaluation matrices give an insight of the design of the model and are used to compare different approaches in order to determine the best model. Different evaluation matrices are used in different problems. The best evaluation matrices for image captioning are given below:

5.1 Rouge

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score measures how similar a candidate document is to a collection of reference documents. In this project, Rouge is used during finetuning of the ViT-GPT-2 model. It evaluates the model's performance by calculating the similarity between the generated caption and input frame's caption.

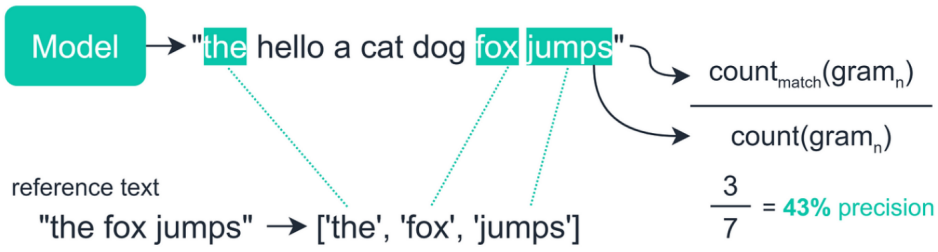


Figure 6: Rouge Formula

Scores obtained for 2 epochs during training are as follows:

Rouge1	Rouge2	Rougel	Rougelsum
48.994500	27.722300	48.317200	48.403600
54.334000	34.359200	53.537800	53.532900

Figure 7: Results during training

5.1.1 Rouge Indexing

1. ROUGE-1 scores are excellent around 0.5.
2. For ROUGE-2, scores above 0.4 are good, and 0.2 to 0.4 are moderate.
3. ROUGE-L scores are good around 0.4.

In the light of the indexing shown above, we can infer that our model is producing very relevant frame descriptions.

5.2 Bleu Score

The BiLingual Evaluation Understudy (BLEU) is a metric used to evaluate machine-translated text. It ranges from zero to one and indicates how similar the machine-translated text is to a set of reference translations. If machine translation is shorter than the reference translation, a Brevity penalty is applied to the precision score.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Figure 8: Bleu Score Formula

On 1040 testing images bleu score was calculated 0.512 value was achieved.

5.2.1 Bleu Score Indexing

BLEU score is a precision based measure that ranges from 0 to 1, with higher values indicating better prediction. While a perfect score of 1 is not possible to achieve, a score higher than 0.3 is generally considered good.
According to the indexing above, our bleu score lies in a very good range.

5.3 Difference between Rouge & Bleu Score

Bleu measures precision: how many words (and/or n-grams) from the machine generated summaries appear in the human reference summaries.
Rouge measures recall: how many words (and/or n-grams) from the human reference summaries appear in the machine generated summaries.

5.4 Recall & F1 Score

For evaluating the machine generated video summaries, commonly used metrics are recall and F1-score. True positive, false positive and false negatives, in this case, would be:

1. True Positives = Frames in both ground truth and machine summary.
2. False Positives = Frames in machine summary but not in ground truth.
3. False Negatives = Frames in ground truth but not in machine summary.

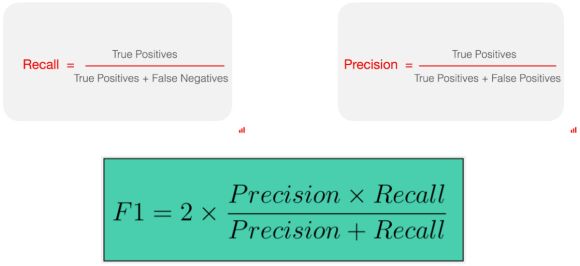


Figure 9: Precision, Recall & F1 Score Formulae

Video Summaries were generated for 5 videos and evaluated against human generated ground truth summaries. The obtained recall and F1-score was 57.36 and 47.73, an increase of 13% is observed on this dataset.

Following is the evaluation table of 5 videos:

S.No	F1-Score	Recall
1.	41.74	54.64
2.	71.34	70.50
3.	19.79	35.02
4.	60.06	75.17
5.	45.74	51.48
Average	**47.73**	**57.36**

6 RESULTS

6.1 Image Captioning Predictions

This video consists of two household chores; Washing dishes & Chopping vegetables.

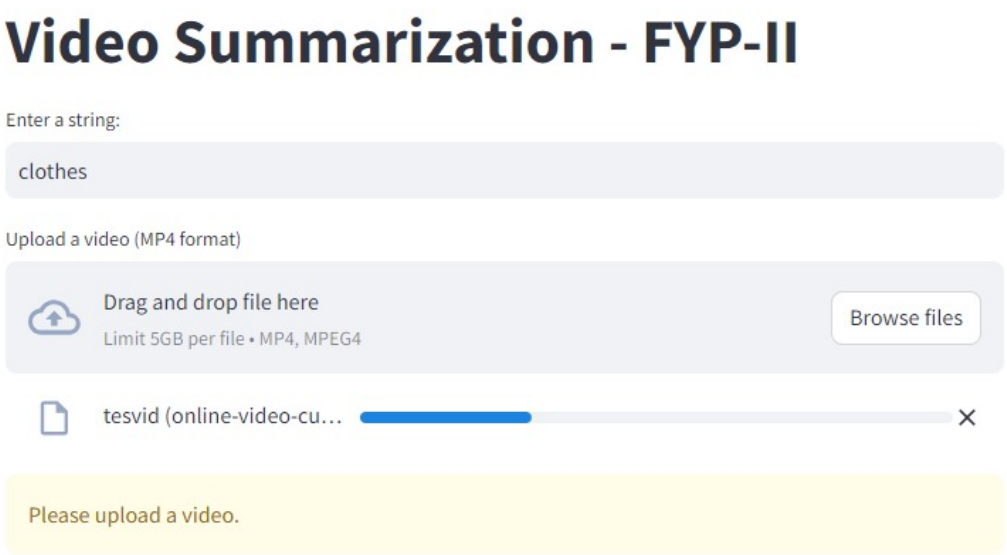

```
"35070": [{"generated_text": "#C C picks the plastic container in the bag"}],
"35085": [{"generated_text": "#C C picks a plastic container from the bathroom sink"}],
"35100": [{"generated_text": "#C C picks up a plastic container from the bathroom sink"}],
"35115": [{"generated_text": "#C C picks up a plastic bag of food from the counter with her right hand."}],
"35040": [{"generated_text": "#C C picks up a container of liquid washing-soap from the kitchen counter with her"}],

"34935": [{"generated_text": "#C C picks up the garlic from the chopping board with his right hand."}],
"34950": [{"generated_text": "#C C drops the chopped bell pepper into the plate on the counter with his right hand."}],
```

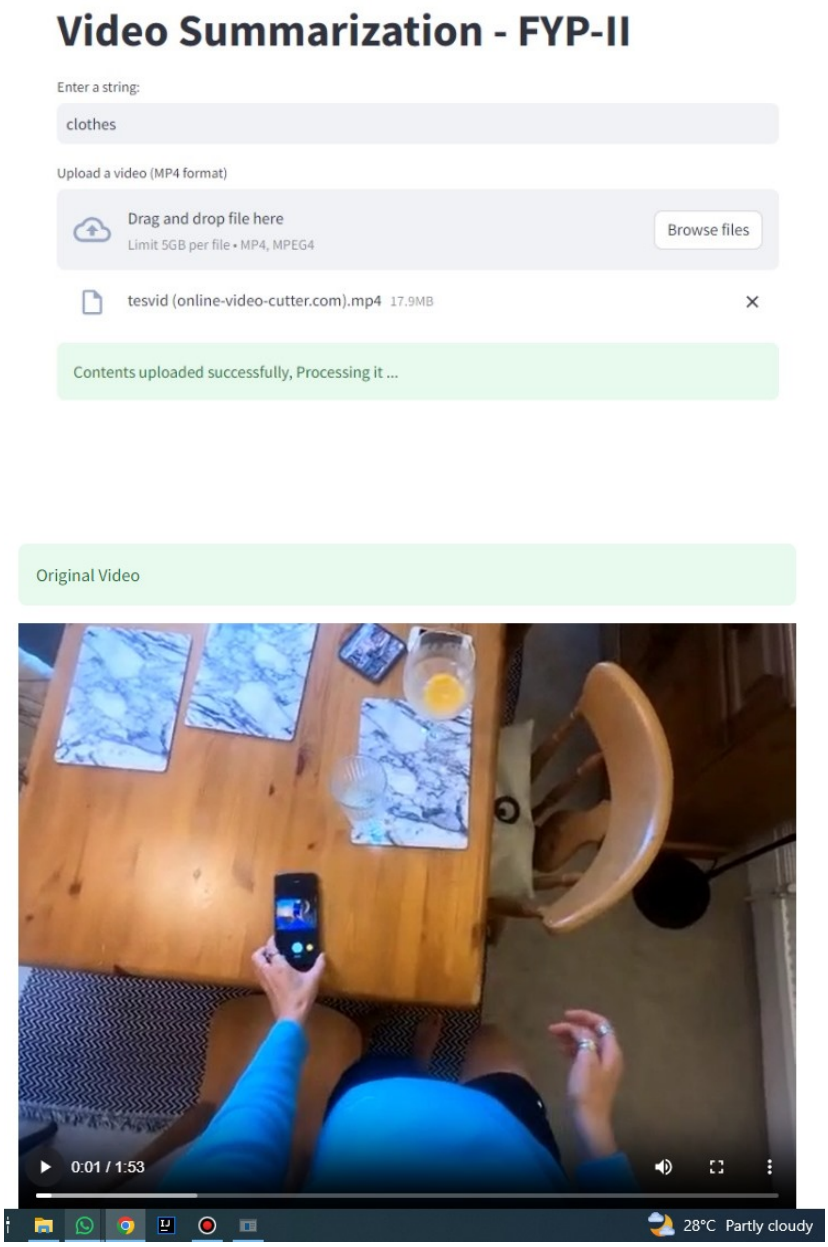

Figure 10: A chunk from image captioning prediction

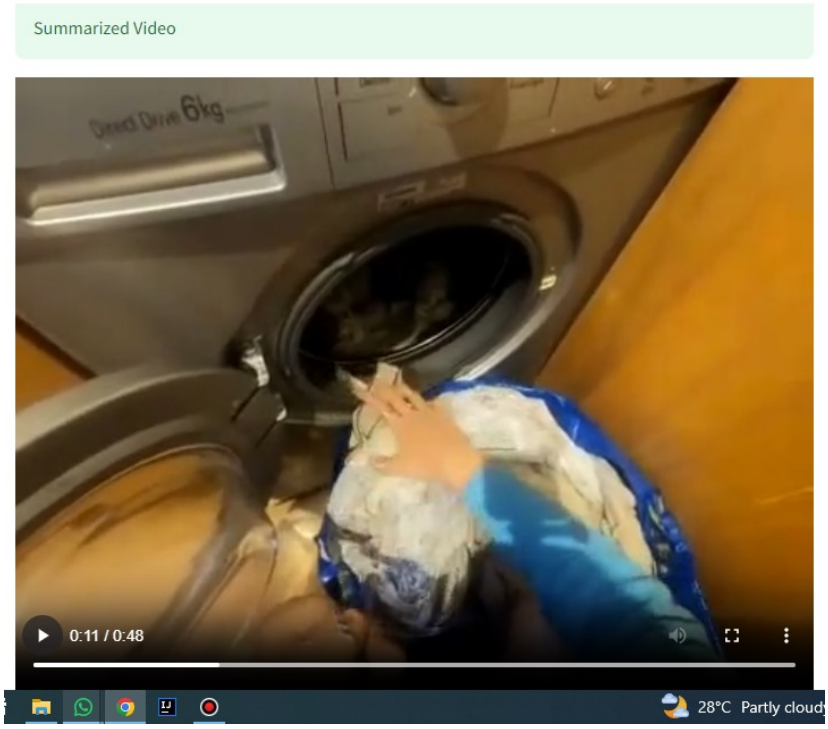
As it can be seen that these two chores are being predicted by the image captioning model.

7 GUI USING STREAMLIT

GUI takes an input video and a user query and displays the summarized video.




Video Summarization - FYP-II
Enter a string:
clothes
Upload a video (MP4 format)
Drag and drop file here
Limit 5GB per file • MP4, MPEG4
Browse files
tesvid (online-video-cutter.com).mp4 17.9MB
Contents uploaded successfully, Processing it ...
Original Video
0:01 / 1:53
28°C Partly cloudy



8 FUTURE WORK

1. The current implementation of the image captioning model takes frames as its inputs. These frames initially need to be saved as Images. This overhead should be minimized by altering the architecture in such a way which takes a whole video as its input, and produces frame-level descriptions without the frames being saved as images. In short, the image captioning model should be turned into a video captioning model. This would also require the model to maintain the context of some previous frames to generate a caption of the current frame accordingly.

2. The developed model is already a multimodal model i.e considering the visual features present in the video and transforming that into text, but more can be achieved if the model learns to consider the textual information present within a video frame. This type of video summarization would be very beneficial in educational sectors, where there are video frames loaded with textual information. Audio features can also be incorporated in further research.

3. The model is trained only on cleaning and laundry videos. To further broaden this umbrella, a dataset (videos with their frame level descriptions) must be prepared for that domain.

4. Object detection using YOLO can identify the objects present in the video frames. This will help in retrieving frames having the objects similar to those specified by the user in his query.

9 COMPARISON AND CONCLUSION

Paper	Model	Dataset	F1 Score	Recall
CLIP-It! Language-Guided Video Summarization	Language-guided multimodal transformer to score video frames based on their importance and correlation with a user-defined query or automatically generated caption.	Egocentric dataset	49.98	47.91
UniMD: Towards Unifying Moment Retrieval and Temporal Action Detection	UniMD framework consists of a shared encoder for feature extraction, task-specific heads for moment retrieval (MR) and temporal action detection (TAD), and multi-task learning to optimize both MR and TAD losses simultaneously.	Ego4D	-	44.8
ViT-GPT2 Model (Our Model)	ViT & GPT-2 for image captioning. Whereas, LlamaIndex for similar frame retrieval.	Ego4D	**47.73**	**57.36**

Hence, in query focused video summarization fine-tuning of huge models is leveraged in order to save computational resources and achieve accurate results of these pre-trained models. Multiple summaries of same videos are generated against different queries and their performance is evaluated by human supervision which yielded quite promising results.

10 REFERENCES

[1] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman, " Video summarization with long short-term memory," in Proc. Eur. Conf. Comput. Vis., 2016, pp. 766–782.

[2] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1–10.

[3] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko, "Sequence to sequence-video to text," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 4534–4542.

[4] A. Vaswani et al., "Attention is All You Need," in 31st Int. Conf. on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[5] J. Fajtl et al., "Summarizing Videos with Attention," in Asian Conf. on Comp. Vision 2018 Workshops. Cham: Springer Int. Publishing, 2018, pp. 39–54.

[6] P. Li et al., "Exploring Global Diverse Attention via Pairwise Temporal Relation for Video Summarization," Pattern Recognition, vol. 111, no. 107677, 2021.

[7] Y.-T. Liu et al., "Learning Hierarchical Self-Attention for Video Summarization," in 2019 IEEE Int. Conf. on Image Processing. IEEE, 2019, pp. 3377–3381.

[8] J. Ghauri et al., "Supervised Video Summarization Via Multiple Feature Sets with Parallel Attention," in 2021 IEEE Int. Conf. on Multimedia and Expo. CA, USA: IEEE, 2021, pp. 1–6.

[9] C. Szegedy et al., "Going Deeper with Convolutions," in 2015 IEEE Conf. on Comp. Vision and Pattern Rec., pp. 1–9.

[10] J. Carreira et al., "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in 2017 IEEE Conf. on Comp. Vision and Pattern Rec., pp. 4724–4733.

[11] Tiwari V, Bhatnagar C (2021) A survey of recent work on video summarization: approaches and techniques. Multimed Tools Appl 80(18):27187–27221.

[12] Hu W, Xie N, Li L, Zeng X, Maybank S (2011) A survey on visual content-based video indexing and retrieval. IEEE Trans Syst Man Cybern C 41(6):797–819.

[13] Money AG, Agius H (2008) Video summarisation: a conceptual framework and survey of the state of the art. J Vis Commun Image Represent 19(2):121–143.

[14] Basavarajaiah M, Sharma P (2021) Gvsum: generic video summarization using deep visual features. Multimed Tools Appl 80(9):14459–14476.

[15] Yingsen Zeng Yujie Zhong Chengjian Feng Lin Ma Meituan Inc. (2024). "UniMD: Towards Unifying Moment Retrieval and Temporal Action Detection."

[16] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell University of California, Berkeley (2021). "CLIP-It! Language-Guided Video Summarization."