QUERY BASED VIDEO SUMMARIZATION

TEAM MEMBERS

Dr. Muhammad Rafi
muhammad.rafi@nu.edu.pk

Alishba Subhani
k200351@nu.edu.pk

Mannahil Miftah
k200234@nu.edu.pk

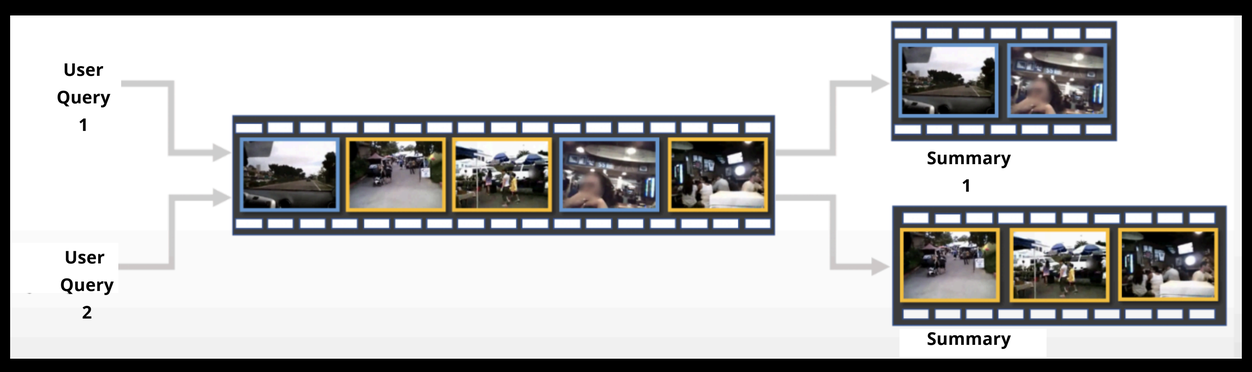
Muhammad Umer
k200427@nu.edu.pk

01. Introduction

Video summarization is the process of 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.

02. Objective

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 according to the query provided by the user.

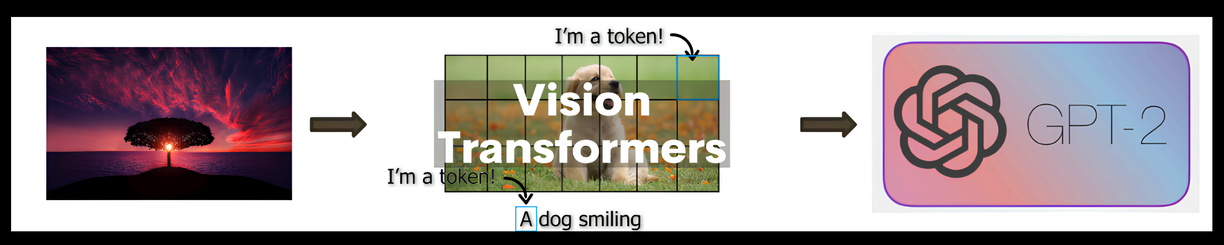


03. Methodology

The video summarization process entails two primary tasks: image captioning and retrieving frames similar to a user query.

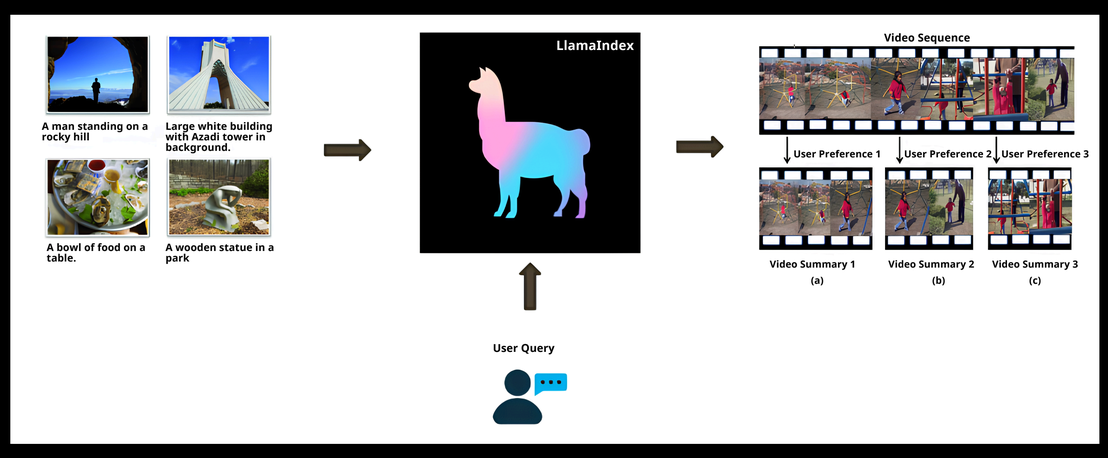
Image Captioning

For image captioning, a combination of pretrained models, Vision Transformer (ViT) and GPT-2, is employed. ViT serves as the encoder for image feature extraction, while GPT-2 acts as the decoder to translate image embeddings into natural language using its tokenizer.



Similar Frame Retrieval

In parallel, for frame retrieval, captions are generated for every 15th frame using the trained model and stored in LlamaIndex. When a user query is made, frames similar to the input query are retrieved, and preceding and succeeding 4-second segments are selected alongside them to maintain coherence within the summary.



RESULTS

We achieved F1-score of 47.7 and recall of 57.3, compared to baseline of 44.8 recall.

FUTURE WORK

- Integrating multimodal approaches that consider both visual and textual information simultaneously
- Development of a captioning model which takes video as its input

With fine-tuning, your model stands on the shoulders of giants.
~ Creator of Keras