Video Summarization Using Transformer Models

Vinayak Agarwal (20BCT0318)
Harshita Rajput (20BCE0752)
Kartikeya Rawat (20BCE0641)
Dev Rishi (20BCE0965)
Dhruv Singh (20BCI0318)

Introduction

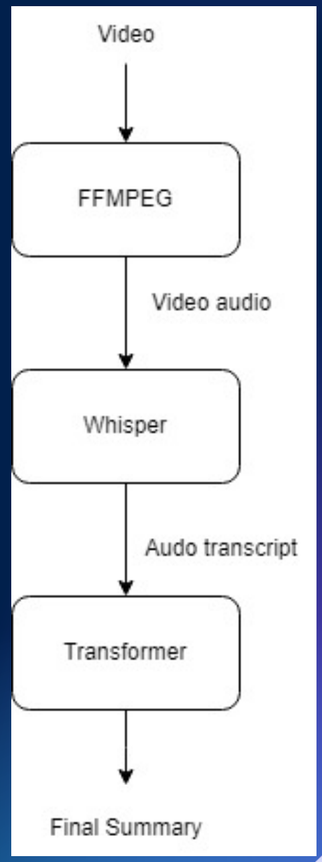
- There is a growing problem of video fatigue due to overload of online video content
- Therefore, there is a need for video summarization to provide concise yet informative summaries
- This would save users' time and will focus on key information
- Also it is an opportunity to apply state-of-the-art transformer models

Overview

- Audio extraction using FFmpeg
- Speech recognition with Whisper API
- Summarization using transformer model (e.g. GPT-3, BERT)
- Leverage Hugging Face Transformers library
- Parameter tuning to optimize summary quality

Methodology

- We use a video in the .mkv format and the **FFmpeg** library to transcode it into audio.
- We then use the **Whisper STT model by OpenAI** to convert the audio file to text based on the GPT2 Transformer model.
- We use the **Billsum dataset**, an excellent dataset for summarization tasks.
- We use the transformers-based **T5-small tokenizer** to preprocess the dataset
- We train our model on the dataset for 20 epochs and evaluate it using the **ROGUE score**



Results

- Whisper AI for Audio-to-Text:

Robust performance in accurately transcribing audio recordings.
Effective handling of various accents and noisy environments.

- Transformer-Based Summarization:

Leveraged state-of-the-art natural language processing techniques.
Generated coherent and concise abstractive summaries of transcribed text.

- Performance Assessment:

Evaluated summarization model using standard metrics like ROUGE.
Results indicated high coherence and relevance in generated summaries.

- Synergy of Technologies:

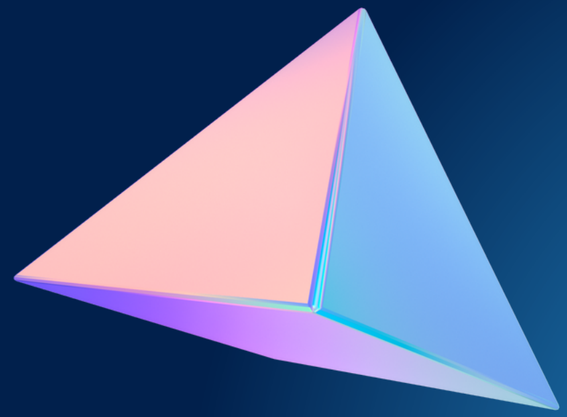
Combined use of Whisper AI and transformer-based model showcased powerful synergy.
Holds promise for automating tasks that require efficient extraction of key information from spoken content.

Results

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	No log	2.608489	0.145200	0.046300	0.119800	0.119300	19.000000
2	No log	2.470221	0.168300	0.059000	0.137300	0.137200	19.000000
3	No log	2.392994	0.198200	0.082400	0.163600	0.163500	19.000000
4	No log	2.349102	0.222100	0.103600	0.185600	0.185600	19.000000
5	No log	2.320213	0.228100	0.108100	0.190300	0.190500	19.000000
6	No log	2.293399	0.225500	0.107200	0.189100	0.189200	19.000000
7	No log	2.272245	0.227200	0.111000	0.191300	0.191300	19.000000
8	No log	2.251587	0.228400	0.111900	0.192200	0.192400	19.000000
9	2.598000	2.238634	0.230100	0.113300	0.194000	0.194200	19.000000
10	2.598000	2.227638	0.229000	0.113400	0.193600	0.193800	19.000000
11	2.598000	2.216547	0.227800	0.114500	0.193900	0.194000	19.000000
12	2.598000	2.205566	0.228100	0.114700	0.194000	0.194200	19.000000

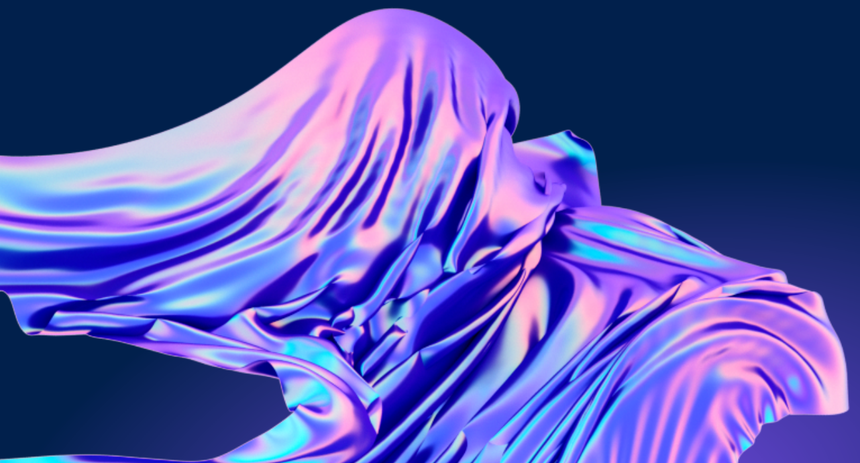
Rouge Scores

Conclusion

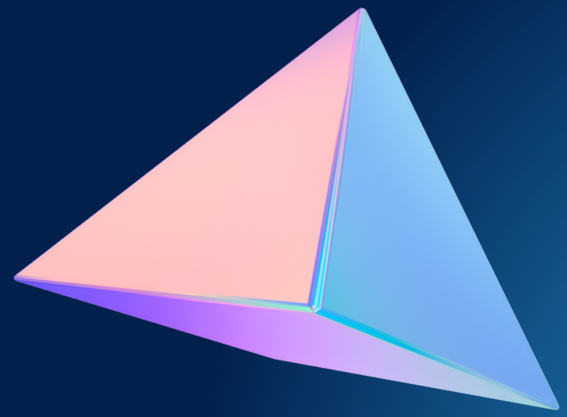


- Project demonstrated effective use of transformers for video summarization
- Whisper + transformer model was robust and produced high quality output

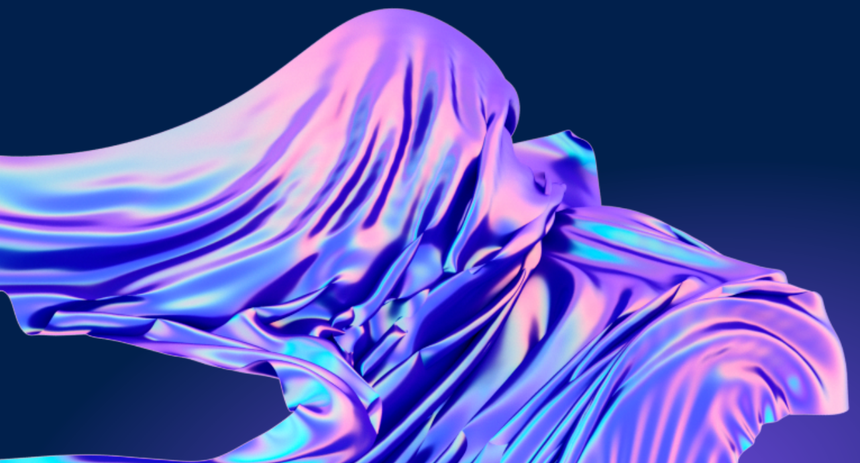
Continued progress in this direction can enable more efficient video understanding



References



- https://colab.research.google.com/drive/1dBLRmXsR0MwvMhTPOovlq88hb7pvoQt1?usp=sharing#scrollTo=IAKSe7mdie8S
- https://arxiv.org/abs/2212.04356
- https://huggingface.co/



Thank You