Methodology for Short and Meaningful Video Content Summarization

Abstract

This paper presents a comprehensive methodology for creating short and meaningful video content summaries. The processing stages include extraction, transcription, preprocessing, and summarization. Each stage is elaborated to ensure that the resulting summaries are as accurate and relevant as possible.

Keywords: Video Summarization, Whisper ASR, TF-IDF, Large Language Models, Natural Language Processing

1 Methodology

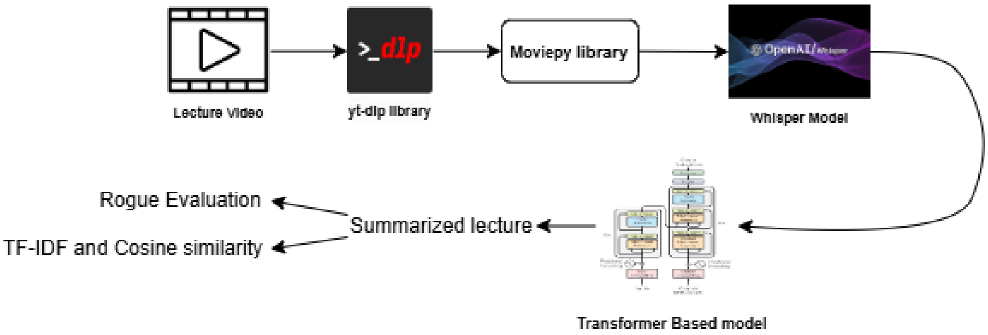

Fig. 1 Overview of the Video Content Summarization Methodology

For the creation of short and meaningful video content summaries in this project, the processing stages were extraction, transcription, preprocessing, and summarization. Each stage in the design was elaborated to make the summaries resulting from it as accurate and relevant as possible.

1.1 Video Extraction and Transcription

First, the video content was downloaded using the `yt-dlp` library. Then, audio extraction was performed by the `moviepy` library. The extracted audio was further processed into subparts using the Whisper ASR model by OpenAI and returned as text. Each segment contained start and end timestamps so that this alignment could be possible; the summarized content and the original video timeline would align [1].

1.2 Preprocessing of Transcriptions

Transcription texts were preprocessed to improve the summarization quality [2]. Preprocessing included:

- **Lowercasing and Cleaning**: All text was normalized to lowercase, and punctuation was removed to normalize the text [2].

This preprocessed text was stored along with the corresponding timestamps to ensure generating time-aligned summaries.

1.3 Summarization Techniques

Summarization can be performed in two ways: traditional TF-IDF-based summarization [2] or LLM-based extractive summarization [3]. In the latter case, a pre-trained model determines the key segments based on how structures of language are compared and then summarizes the relevant points across many chunks of text.

1.3.1 TF-IDF Approach

The TF-IDF vectorizer calculates the term frequency-inverse document frequency score of words in the text [4]. The sentences with the higher scores, which carry higher information, were chosen according to this technique. It was chosen to help in the output results of the summarized video [5].

1.3.2 Extractive Summarization Based on LLM

Extractive summarization based on a large language model was carried out for each segment. Initial summary generation was performed, followed by TF-IDF-based cosine similarity between the summary and the original transcript in order to filter sentences using a set similarity threshold that guarantees a high level of relevance to the original content [3].

These methods generated summaries that retain substantial information but with fewer words from the original transcript.

2 Model Selection

The two main models used for this work include OpenAI's Whisper for transcription and the Large Language Model for extractive summarization.

2.1 Whisper ASR Model

Since the model accuracy is high in transcription speech to text, Whisper was chosen [1]. It supports most languages and performs well on most accents with background noises; hence, the generated transcripts are reliable. With Whisper's output of segment-by-segment transcription along with timestamps, the summarization was created aligned with the original video timeline.

2.2 Transformer-Based Model

The chosen transformer-based model was BERT2BERT transformer. It was chosen due to its benchmarks against other transformers while considering that the pipeline would run on the free Google Colab plan [6].

2.3 Large Language Model

An LLM, pre-trained for natural language understanding and generation, was used for extractive summarization [3]. Considering the fact that the model has been trained with complete contextual understanding and can hence locate sections of text relevant to the subject in discussion, it is expected that the model would be very well-suited for extractive summarization. The model allows for customizable generation, adjusting factors such as length penalty and beam search in order to optimize the output for brevity and clarity.

2.4 TF-IDF Vectorizer

Complementing the LLM was an approach based on TF-IDF to help present to the model a statistical look at word importance [4]. The TF-IDF vectorization helped identify those sentences that unequivocally contained terms with high significance relative to the document. The TF-IDF vector space supported the summarization by LLM, which was more complex and based on language understanding, by a very simple frequency-based technique [2].

3 Assessment Evaluation

An evaluation was carried out on two main fronts: the correctness of transcription and the quality of summary. Testing the performance of summarization models therefore involved both quantitative and qualitative methods.

3.1 ROUGE Metric of Summarization Quality

The ROUGE metric was used to determine the overlap between the generated summaries and reference summaries [7]. The ROUGE scores, including ROUGE-1, ROUGE-2, and ROUGE-L, helped measure the accuracy of the generated summaries by comparing them to manually created ones. A higher ROUGE score meant closer alignment with the reference, indicating better performance.

3.2 Cosine Similarity for Relevance

Besides generating the summary, cosine similarity was measured between the summary and the original text using TF-IDF vectors [4]. The metric provided an understanding of how well the summary grasped the highlights of the original transcript. For example,

Table 1 ROUGE Scores Across TRY2 Experiments for Each Video using Extractive transformer

	TRY2 (Removing Stop Words)			TRY2 (Without Removing Stop Words)		
Video	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Definition of Information	0.4517	0.1966	0.2273	0.4615	0.3032	0.3317
Recoverable Schedules	0.3330	0.1350	0.1759	0.2576	0.1745	0.1919
Introduction to OS	0.4595	0.2751	0.3031	0.3263	0.2497	0.2593
Introduction to ML	0.5456	0.3219	0.3632	0.4341	0.3008	0.3286

Table 2 ROUGE Scores Across TRY3 and TRY4 Experiments for Each Video

	TRY3 (abstractive transformer)			TRY4 (LLM)		
Video	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Definition of Information	0.0	0.0	0.0	0.5230	0.2189	0.3006
Recoverable Schedules	0.2374	0.0290	0.1151	0.2688	0.1395	0.1773
Introduction to OS	0.4357	0.1313	0.1786	0.4330	0.3420	0.3564
Introduction to ML	0.2005	0.0595	0.1002	0.6667	0.4604	0.4872

summaries whose similarity scores exceed a threshold, such as 0.4, would be considered to retain enough relevance, ensuring that no important information disappears after summarization.

3.3 Contextual Coherence Manual Review

Apart from the quantitative metrics, a manual review was performed to ensure that the summaries were contextually coherent and that they actually represented the intent of the content. This step materially helped in the evaluation of LLM-based summaries, as sometimes a language model provides relevant but completely off-target output from the original content. The review ensured that the summaries were meaningful and added value to the abstract form of video content.

References

[1] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022)

[2] Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval.* Cambridge University Press, Cambridge (2008)

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)

[4] Salton, G., McGill, M.J.: *Introduction to Modern Information Retrieval.* McGraw-Hill, New York (1983)

[5] Rajaraman, A., Ullman, J.D.: *Mining of Massive Datasets*. Cambridge University Press, Cambridge (2011)

[6] Rothe, S., Narayan, S., Severyn, A.: Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics* **8**, 264–280 (2020)

[7] Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)