WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Video Transcript Summarization Using Bert

[1]M Iswarya, [2]P Sai Krishna, [3]K Naveen, [4]M Ganesh, [5]M Yasin

[1,2,3,4,5]Department of CSE, GMR Institute of Technology, Rajam, Andhra Pradesh, India

ABSTRACT:

Throughout the years, the Internet and social networking sites have experienced a significant expansion in multimedia, notably video material. The amount of video data created has expanded at an exponential pace over time. Yet, many people are unable to read the entire information owing to time limits. The goal of video summarization is to create video summaries that incorporate the most significant and relevant material from a video stream in a concise manner. The summarising technique can remove a significant quantity of information while maintaining the most important information in the article. Our model has received the video, and the speech will be retrieved. The model then translates audio to text by chunking an audio file, which is handy for processing big files. The voice is then automatically transformed into text using automated speech recognition, which is a quick and simple approach to work with audio files. It transforms audio clips to text. Lastly, the BERT algorithm is applied to text summarization. Transformer, an attention mechanism that learns the contextual relationships between words in a summarised text, is used in the summarising process.

Keywords: BERT Algorithm, Summarization, Automatic Speech Recognition, Transformer, Text Chunking.

1. Introduction:

Every day, a massive amount of video recordings is created and circulated online. When we can't discover the relevant information, we're seeking for, our efforts may be futile. It is both frustrating and time-consuming to seek for videos that contain the information we are looking for. It has gotten increasingly difficult to watch videos that are longer than planned. For example, there are countless videos online where the speaker covers a specific subject in depth, but it might be difficult to understand the message the speaker is attempting to convey to the audience until we watch the entire video. When a vast volume of data is provided at once, including irrelevant stuff, extracting meaningful information becomes a time-consuming and hard effort.

Video transcript summary is one answer to this issue. This approach includes automatically constructing a summary of the video's content based on its transcript, allowing viewers to easily identify the information they want. Natural language processing and machine learning algorithms are used in video transcript summarising to detect essential phrases, ideas, and concepts in the video, reducing hours of film into a short summary. Viewers may save time and discover the information they need without having to watch the full video by using video transcript summary.

Video transcript summary is a strong approach for condensing lengthy videos into a few essential points or key takeaways. This is especially handy for videos that include a lot of important information but are too long to see all at once. Video transcript summarization may swiftly distil the important aspects of a video into a succinct summary by utilising algorithms and techniques such as natural language processing and speech recognition. This summary may be used for a number of things, such as automatic video transcript search, video indexing, and so on.

This technology has the potential to revolutionize the way we consume video content, making it more efficient, accessible, and convenient. As technology continues to improve, we can expect to see more advanced video transcript summarization tools that are better able to understand and summarize the complex nuances of human language and behaviour, further enhancing the benefits of this technology. However, there are also challenges to overcome in implementing video transcript summarization technology. Accurately summarizing a video requires not only identifying key phrases and concepts but also understanding the context in which they are used. Despite these challenges, video transcript summarization is more useful.

2. Literature Survey:

Ma et al. proposes a topic-aware text summarization method that uses a pre-trained BERT model to encode documents and topics. The method is designed to generate summaries that are both informative and focused on the topics of interest. Firstly, Topic Extraction extract the topics of interest from the input document. This can be done using various methods, such as TF-IDF or LDA. In this paper, the authors use a pre-trained topic extractor based on BERT to identify the most relevant topics in the document. Secondly, Once the topics have been extracted, they are encoded using a pre-trained BERT model. The authors use a BERT-based topic model that is trained on a large corpus of documents to encode the topics into fixed-length vectors. Then next, the input document is also encoded using a pre-trained BERT model. The document is tokenized and fed into the BERT model to obtain a sequence of hidden representations, which are then used to generate a fixed-length vector that represents the document. To generate a topic-aware representation of the document, the authors combine the topic and document encodings. Final step is to generate a summary using the topic-aware encoding. The authors use

a sequence-to-sequence model with attention to generate summaries that are focused on the relevant topics. The attention mechanism is designed to pay more attention to the topic-aware encoding when generating summary tokens that are related to the topics [1].

Alomari et al. provides a comprehensive overview of the current state-of-the-art techniques for abstractive text summarization using deep reinforcement and transfer learning. The paper aims to help researchers and practitioners understand the latest advancements in the field and provides an extensive review of the literature in this area. The paper begins with an introduction to the problem of text summarization and explains the difference between extractive and abstractive summarization. The authors then describe the challenges associated with abstractive summarization, including the need to generate coherent and grammatically correct summaries while preserving the most important information from the source text. Next, the paper reviews the existing approaches for abstractive text summarization, including classical methods such as topic modeling and clustering, as well as more recent deep learning-based approaches. The authors provide a detailed description of the deep learning architectures used for abstractive summarization, including sequence-to-sequence models, attention-based models, and transformer-based models. The paper then focuses on the use of reinforcement learning for abstractive summarization. The authors explain the basics of reinforcement learning and how it can be used to train models that generate high-quality summaries. They describe different reinforcement learning algorithms, such as Q-learning and policy gradient methods, and explain how they can be applied to text summarization. Finally, the paper discusses the use of transfer learning for abstractive summarization. The authors explain how pre-training models on large corpora of text can improve the performance of summarization models, and they review the most common pre-training techniques, including BERT, GPT, and T5. Overall, the paper provides a comprehensive overview of the current state-of-the-art techniques for abstractive text summarization using deep reinforcement and transfer learning [2].

Ghosh et al. proposes a method to automatically identify off-topic concepts in video lectures and link them to relevant video lecture segments. Firstly, Video Lecture Segmentation process is to segment the video lecture into smaller video segments based on the lecture's slide changes or the speaker's pauses. Secondly, Text Extraction extract the text from each of these video segments using Optical Character Recognition (OCR) or speech-to-text conversion techniques. And then, Once the text has been extracted from the video segments, the paper applies a topic modeling technique to identify the topics covered in each segment called topic modeling. The paper uses Latent Dirichlet Allocation (LDA), a popular probabilistic topic modeling technique, to identify the topics covered in each video segment. And then it uses the identified topics to identify off-topic concepts in the video lecture called Topic Identification. If a segment contains a significant amount of text related to a topic that is not relevant to the lecture's main topic, it is considered off-topic. Finally, the paper links the off-topic concepts to relevant video segments. Overall, the paper's approach involves segmenting the video lecture into smaller parts, extracting text from each segment, identifying the topics covered in each segment, identifying off-topic concepts based on the identified topics, and linking off-topic concepts to relevant video segments [3].

Aliakbarpour et al. proposes a new method for abstractive text summarization that uses a combination of deep neural networks and attention mechanisms to improve the readability and saliency of the generated summaries. The proposed method consists of three main components: The encoder-decoder model is used to generate the summary by encoding the input text into a fixed-length vector and then decoding it into the summary. The attention mechanism is used to focus on the relevant parts of the input text during the encoding and decoding process. The auxiliary attention mechanism is used to further improve the attention mechanism by incorporating additional information about the input text. The encoder-decoder model is based on a sequence-to-sequence architecture with an LSTM-based encoder and decoder. The encoder takes the input text as a sequence of word embeddings and outputs a fixed-length vector representing the input text. The decoder takes this vector as input and generates the summary as a sequence of words. The attention mechanism is used to weigh the importance of each word in the input text when generating the summary. The attention weights are computed based on the similarity between the encoded input text and the decoder hidden state at each time step. The auxiliary attention mechanism is used to further improve the attention mechanism by incorporating additional information about the input text. Specifically, it uses an additional set of attention weights to focus on the salient and relevant parts of the input text. These attention weights are computed based on the importance of each sentence in the input text, which is determined by a sentence saliency model. The proposed method was evaluated on the CNN/Daily Mail dataset, which contains news articles and their corresponding summaries. The results showed that the proposed method outperformed several state-of-the-art methods in terms of both ROUGE and human evaluation metrics [4].

Kota, Bhargava Urala, Alexander Stone, Kenny Davila, Srirangaraj Setlur, and Venu Govindaraju proposes a process for automatically summarizing whiteboard lectures by detecting and representing different content regions. Firstly, the video is pre-processed by converting it into a sequence of frames and resizing each frame to a fixed size. The frames are also converted from RGB to grayscale and contrast normalization is applied. Secondly, Content region detection: A content region detection algorithm is used to identify different regions in each frame of the video. The algorithm is based on the idea that regions containing content are likely to have a high variation in pixel intensity. The algorithm first applies a Laplacian of Gaussian filter to each frame to highlight regions with high intensity variation. It then thresholds the resulting image to identify regions with high intensity variation. These regions are classified as content regions and are used for further processing.

Content region representation: Each content region is represented using a feature vector that captures its visual characteristics. The feature vector is based on the histogram of oriented gradients (HOG) descriptor, which captures the distribution of gradient orientations in the region. The HOG descriptor is computed for each content region and the resulting feature vector is used to represent the region.

Summary generation: A summary of the video is generated by selecting representative content regions from the video. The selection is based on a clustering algorithm that groups similar content regions together. The clustering algorithm is based on the k-means algorithm and it groups the content regions based on their feature vectors. The resulting clusters represent different topics or concepts in the video. The most representative content region from each cluster is selected to form the summary.

Postprocessing: The summary is postprocessed to remove redundant content regions and to ensure coherence and consistency. The postprocessing step involves analysing the relationships between the selected content regions and removing any that are redundant or do not contribute to the overall coherence of the summary [5].

Al-Azani and El-Alfy proposes a methodology for sentiment analysis of videos using a combination of textual, auditory, and visual information. The first step involves collecting the necessary data for sentiment analysis. This includes videos, their associated transcripts, and any available metadata such as captions, timestamps, and speaker information. And then data is collected, that is then pre-processed to remove any irrelevant information and to prepare it for analysis. This involves steps such as text cleaning, speech-to-text conversion, and video segmentation. Next, various features are extracted from the pre-processed data. These include textual features such as sentiment scores, topic models, and word embeddings; auditory features such as speech rate, pitch, and energy; and visual features such as facial expressions, gestures, and body language. The extracted features from different modalities are then combined or fused to form a unified feature set. This is done using various fusion techniques such as concatenation, summation, and weighted averaging. The fused feature set is then used to perform sentiment analysis on the videos. This involves using various machine learning techniques such as support vector machines, neural networks, and decision trees to classify the sentiment of the video into positive, negative, or neutral. The final step involves evaluating the performance of the proposed approach. This is done by comparing the results obtained from the proposed approach with those obtained from existing approaches. The performance is evaluated using various metrics such as accuracy, precision, recall, and F1 score [6].

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed proposes a detailed overview of the different approaches and techniques used in automatic text summarization. The paper provides an introduction to automatic text summarization, its importance, and its applications in various fields such as news, research papers, and social media. The paper discusses different types of summaries such as extractive, abstractive, and mixed summarization. Extractive summarization involves selecting important sentences or phrases from the original text and presenting them as a summary. Abstractive summarization involves generating a summary that is not simply a subset of the original text but captures the main idea in a new way. The paper provides an overview of various extractive summarization techniques such as frequency-based methods, graph-based methods, and machine learning-based methods. Frequency-based methods involve selecting sentences or phrases based on their frequency in the original text. Graph-based methods involve representing the text as a graph and selecting important nodes based on their centrality. The paper discusses various abstractive summarization techniques such as natural language generation, text-to-text generation, and encoder-decoder models. Natural language generation involves generating a summary using a template-based approach or a rule-based approach. Text-to-text generation involves mapping the original text to a summary text using a sequence-to-sequence model. Encoder-decoder models involve encoding the original text using a neural network and then generating a summary using a decoder neural network. The paper discusses various evaluation metrics used to measure the quality of automatic text summarization such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit ORdering) measure. It also suggests future directions for research in this area such as exploring the use of deep learning models for summarization and developing new evaluation metrics for summarization [7].

Wang Qicai, Peiyu Liu, Zhenfang Zhu, and Lindong Zhang proposes a text summarization model that utilizes BERT word embedding and reinforcement learning to generate abstractive summaries. Firstly, the paper uses the CNN/Daily Mail dataset for training and evaluation. The dataset consists of news articles and corresponding summaries. The authors pre-process the data by cleaning and tokenizing the text. Secondly, the paper uses BERT to encode the input text. BERT is a pre-trained neural network that can generate high-quality word embeddings. And next, the paper proposes a reinforcement learning-based model for abstractive summary generation. The model consists of an encoder, a decoder, and a reinforcement learning component. The encoder takes the input text and generates a sequence of hidden states. The decoder takes the hidden states and generates the summary sequence. The reinforcement learning component provides feedback to the decoder to improve the quality of the generated summary. The authors use policy gradient optimization to train the reinforcement learning component. The paper evaluates the proposed model using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics. ROUGE measures the overlap between the generated summary and the reference summary in terms of n-gram matches. The authors compare the performance of their model with other state-of-the-art summarization models and demonstrate that their model achieves better results. The paper analyses the results of the evaluation and discusses the strengths and limitations of the proposed model. The proposed model outperforms other state-of-the-art summarization models and demonstrates the effectiveness of the proposed approach [8].

Mridha et al. provides an overview of the state-of-the-art techniques and challenges in automatic text summarization. This paper involves generating a shorter version of a longer document while preserving the most important information. This paper discusses the two main types of summarization: extractive and abstractive. Extractive summarization involves selecting and combining important sentences from the original document, while abstractive summarization involves generating new sentences that convey the essence of the original document. Techniques for extractive and abstractive summarization are proposed. The paper discusses various techniques for extractive summarization, including graph-based methods, clustering-based methods, and supervised learning-based methods. The paper discusses various techniques for abstractive summarization, including deep learning-based methods, rule-based methods, and hybrid methods. The challenges of abstractive and extractive summarization, such as generating grammatically correct and coherent sentences, are also discussed. The paper discusses the challenges of evaluating the quality of summarization. Common evaluation metrics, such as ROUGE and BLEU, are introduced, and their limitations are discussed. The paper concludes with a discussion of the challenges and future directions in automatic text summarization. Some of the challenges include dealing with multi-document summarization, summarizing non-textual data such as images and videos, and summarizing in low-resource languages. The paper suggests that future research should focus on developing more accurate and efficient summarization techniques, as well as improving the evaluation metrics for summarization [9].

Suleiman and Awajan provides a comprehensive survey of deep learning approaches to abstractive text summarization, including an overview of different datasets and evaluation measures used in this field. The paper begins with an introduction to text summarization and its importance in various applications.

The authors explain that extractive summarization involves selecting and combining sentences from the original text, while abstractive summarization involves generating new sentences that capture the essence of the original text. Next, the paper provides a detailed overview of various deep learning approaches that have been used for abstractive text summarization, including sequence-to-sequence models, attention-based models, and reinforcement learning-based models. The authors discuss the strengths and weaknesses of each approach. The paper then discusses some of the popular datasets that have been used for training and evaluating abstractive text summarization models, including CNN/Daily Mail, Gigaword, and DUC. The authors explain the characteristics of each dataset and highlight some of the challenges associated with using them for abstractive summarization. The paper also discusses various evaluation measures that have been used to evaluate the performance of abstractive text summarization systems, including ROUGE, BLEU, and METEOR. Finally, the paper concludes with a discussion of some of the key challenges faced by researchers in developing accurate and effective abstractive text summarization systems, including the need for large amounts of training data, the difficulty of capturing the semantic meaning of the original text, and the challenge of generating coherent and fluent summaries [10].

3. Methodology:

The amount of video data produced nowadays is constantly rising. But, seeing the entire film is essential to identify the material on which the speaker is most focused. The user loses a lot of time since he or she does not know which movie to watch or pick for his or her study. The suggested paradigm allows the audience to read the synopsis before watching the video. The proposed model accepts video as input from the user and provides a summary of the video content that was uploaded. The model is based on three major steps. They are:

1. Audio extraction
2. Audio to text conversion
3. Text summarization

3.1 Audio Extraction

The technique of extracting audio from a video that the user has submitted is known as audio extraction. In this model, we utilised the MoviePy package to extract audio. MoviePy audio extraction is a quick and straightforward technique to extract audio from video files. MoviePy is a Python library for working with video files and performing various operations such as audio extraction, trimming, concatenating, and converting between different video formats. It also has several audio-related features, such as audio extraction, audio manipulation, and audio visualisation.

3.2 Audio to Text Conversion

The model's next step is to translate audio to text. The main concept is to break up a lengthy audio recording into shorter pieces or chunks, then transcribe each fragment into text independently. This method enables for more efficient audio data processing and can assist increase transcription accuracy. Often, translating audio to text includes multiple processes. Using the Pydub package, the audio file is first broken into tiny pieces or chunks ranging in duration from a few seconds to a minute or two.

Pydub is a well-known Python library that aids in the manipulation of audio files. It has a simple interface for importing, editing, and saving audio files in a number of formats such as MP3, WAV, and others. It is built on top of the FFmpeg library, which has a robust collection of audio and video processing features. This means Pydub can operate with a broad variety of audio formats and can easily perform sophisticated audio processing jobs. Automatic Speech Recognition, or ASR for short, is a cutting-edge technology that transcribes human voice into written text using machine learning and artificial intelligence. In this model, we generate a Python Recognizer instance using the SpeechRecognition package.

3.3 Text Summarization

Text summarization is a method of condensing information. Although human text summary is a solid way for retaining the sense of a document, it can be a time-consuming effort. Another approach is to employ automated text summarising (ATS). ATS allows multiple applications to be hooked into computers to provide data summaries. Hence, text summary provides a well-explained and proper interpretation of a long paragraph by focusing on the key sections and retaining the entire context. Automatic text summarization is a way of examining, interpreting, and extrapolating data from human language in natural language processing (NLP). For Automated text summarization, we employed BERT in our model. BERT is a powerful language system. It can be used for text summarization.

4. Conclusion:

Video transcript summarization has the potential to greatly improve the efficiency and effectiveness of video content consumption, making it an excellent choice for anyone looking to save time and improve their understanding of video information. In this study, audio from a video provided by the user is extracted and translated into text using an Automatic speech recognition model that provides the best level of confidence during transcription. The retrieved speech is summarized using BERT. To compare summarization performance, the ROUGE metric is utilized. Despite improvements in text

summarization, the technology still has severe limitations. Because of their length, lengthier papers need more processing. In future research, we can improve the model byemploying better summarization models.

5. References:

[1] Ma, T., Pan, Q., Rong, H., Qian, Y., Tian, Y., & Al-Nabhan, N. (2021). T-bertsum: Topic-aware text summarization based on bert. IEEE Transactions on Computational Social Systems, 9(3), 879-890.

[2] Alomari, A., Idris, N., Sabri, A. Q. M., & Alsmadi, I. (2022). Deep reinforcement and transfer learning for abstractive text summarization: A review. Computer Speech & Language, 71, 101276.

[3] Ghosh, K., Nangi, S. R., Kanchugantla, Y., Rayapati, P. G., Bhowmick, P. K., & Goyal, P. (2022). Augmenting video lectures: Identifying off-topic concepts and linking to relevant video lecture segments. International Journal of Artificial Intelligence in Education, 32(2), 382-412.

[4] Aliakbarpour, H., Manzuri, M. T., & Rahmani, A. M. (2022). Improving the readability and saliency of abstractive text summarization using combination of deep neural networks equipped with auxiliary attention mechanism. The Journal of Supercomputing, 78(2), 2528-2555.

[5] Kota, B. U., Stone, A., Davila, K., Setlur, S., & Govindaraju, V. (2021, January). Automated whiteboard lecture video summarization by content region detection and representation. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 10704-10711). IEEE.

[6] Al-Azani, S., & El-Alfy, E. S. M. (2020). Enhanced video analytics for sentiment analysis based on fusing textual, auditory and visual information. IEEE Access, 8, 136843-136857.

[7] El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165, 113679.

[8] Wang, Q., Liu, P., Zhu, Z., Yin, H., Zhang, Q., & Zhang, L. (2019). A text abstraction summary model based on BERT word embedding and reinforcement learning. Applied Sciences, 9(21), 4701.

[9] Mridha, M. F., Lima, A. A., Nur, K., Das, S. C., Hasan, M., & Kabir, M. M. (2021). A survey of automatic text summarization: Progress, process and challenges. IEEE Access, 9, 156043-156070.

[10] Suleiman, D., & Awajan, A. (2020). Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges. Mathematical problems in engineering, 2020.