

**PROJECT REPORT**

Query based Educational Video Retrieval using BERT model

By

Ezhilan

National Institute of Technology, Puducherry

***Abstract*—** **Query-based educational video summarization is a sophisticated technique designed to generate concise, relevant summaries of videos based on user-defined queries. This project introduces a system that takes a query or question as input and produces a summarized video by extracting sliding window segments that match the query's keywords which is found in the video's subtitles. By leveraging state-of-the-art natural language processing models, such as BERT (Bidirectional Encoder Representations from Transformers) and Transformers, the system accurately identifies and aligns the input query with relevant video content. The process begins with subtitle processing to identify best matches to the query. BERT is employed to ensure semantic context understanding between the query and subtitle content, enhancing the precision of matches. Following this, a Sentence transformer further improves the accuracy of segment extraction. The resulting video summary is composed of the most pertinent clips, each annotated with subtitles, thereby offering a coherent and contextually accurate summary. This approach significantly enhances efficiency and accuracy over traditional video summarization methods, providing a powerful tool for various applications, including education, media, and content management. By integrating advanced deep learning models, this project demonstrates an intuitive and user-centric method for navigating and extracting information from extensive video content. The system not only addresses the challenges of traditional video summarization but also presents a scalable solution capable of handling diverse and complex queries, ultimately offering a more streamlined and effective way to access valuable video information.**

***Keywords—*** *Educational video retrieval, query-based, BERT, Sentence transformers,Cosine Similarity, Subtitles*

1. INTRODUCTION

Along with the advance of digital devices and platforms, video is now one of the most desired data types for consumers. Although the large information capacity of videos might be beneficial in many aspects, e.g., informative.and entertaining, inspecting the videos is time-consuming, so that it is hard to capture the desired moments. In contrast, our approach circumvents this computational overhead by not directly processing the video itself. Instead, we leverage the accompanying captions to retrieve the best matching segments, thereby significantly reducing the computational burden. This is particularly effective in dialogue-driven content where the textual component is rich in information and closely aligned with the user's query. Query-based educational video summarization addresses this challenge by generating concise video extraction tailored to user-defined queries. This project aims to develop an advanced system that leverages state-of-the-art natural language processing models, such as BERT (Bidirectional Encoder Representations from Transformers) and Transformers [4], to deliver highly relevant video lecture summaries based on the user's query. By focusing on the captions, our method ensures computational efficiency while maintaining high accuracy in retrieving relevant content. This project demonstrates a novel solution that utilizes recent technological advancements in NLP to provide a computationally efficient method for educational video summarization. Our approach is particularly well-suited for dialogue-driven content, where the accuracy and relevance of the summaries can significantly enhance the user experience by quickly delivering the desired information.

1. LITERATURE SURVEY

*A. Paper work*

Query-oriented micro-video summarization is a burgeoning field that addresses the need for efficient video content retrieval tailored to user queries.

A notable study in 2024, titled Query Oriented Micro Video Summarization [1], leverages the QMV-Kwai dataset, which comprises micro-videos from the Kwai platform associated with multiple user queries. This research generates a video summary as an output for each micro-video based on user queries. In order to generate the output summary, The visual and textual features are extracted using pre-trained models like CLIP and BERT, representing the video's content accurately.

Key entities are identified and highlighted using a Multi-Layer Perceptron (MLP), which scores words based on their relevance to user queries. These important words are then highlighted and combined with visual and textual features.This visual and textual features combinedly create a unified multi-modal representation. An entity-aware multi-modal learning module improves the summary by matching key information across different types of data using a cross-attention mechanism. The decoder, using the pre-trained BART model, generates the summary by focusing on the most relevant parts of the input and building the summary word by word. The system's performance is evaluated using metrics like BLEU, ROUGE, and CIDEr, ensuring that the summaries are both relevant and informative, meeting users' needs for accurate information from video content.

In contrast, the 2023 study "Query-based video summarization with multi-label classification network"[2] uses datasets like UTE, TVSum, and ARS to integrate user preferences via textual queries into video summarization. Traditional approaches have largely ignored the user query in generating very diverse and representative summaries. The authors in this study formulate the task as a multi-label classification problem and use a pre-trained ResNet feature extractor along with the MLC-SUM algorithm.It uses a multi-layer perceptron to predict the relevance of video segments and refines such predictions with a cross-correlation matrix of labels, which helps improve the accuracy by considering the relationships between different labels. This method ensures that the selected video segments are more closely aligned with what the user is looking for. After that, the Top-K method is used to choose video segments based on weighted relevance scores. The quality of the results will be measured with precision, recall, F1 score, and also with Intersection over Union(IOU),which sheds light on how similar the segments are with the generated text. However, this approach considered fixed time segments of the video.To handle this problem,in this study used the sliding window concept on the video segment to find the best match time segments.

Another 2023 study, "Query-Dependent Video Representation for Moment Retrieval and Highlight Detection"[3], addresses both moment retrieval (MR) and highlight detection (HD).The paper is oriented to the task of Highlighting certain moments in videos according to a text description and detecting the most related fragments in the video. In particular, transformer-based architectures, such as Moment-DETR and UMT, made good progress in the integration of multimodal data for MR tasks. QD-DETR was proposed for generating representations of videos based on queries. It used the architecture of a cross-attention transformer encoder. Negative pair learning strengthens model distinction between relevant and irrelevant pairs. A saliency token is introduced to make the processes of highlight detection and moment retrieval within this model effective, thus showcasing the power of sophisticated deep learning methodologies in enhancing video content analysis.

These studies collectively highlight the advancements in query-based video summarization and retrieval, demonstrating how state-of-the-art natural language processing and machine learning techniques can improve the relevance and efficiency of video summarization, tailored to user-defined queries.

*B. Related work*

In the process of its evolution, a wide variety of techniques has been applied toward query-based video retrieval for efficiency and accuracy in the retrieval of relevant video content. The early approaches were mainly keyframe extraction[2] and video skimming-based techniques. Keyframe extraction is a technique where some representative frames are selected that can make up an abstraction of the content of the video, and video skimming creates a concise summary through key segment extraction. These methods provided a simple approach toward the creation of video summaries and information retrieval but were computationally efficient and limited by the accuracy of relevance. Recent developments utilized deep learning in query-based video summarization. A trained neural network can leverage to understand and extract semantic features from the video better, hence providing an accurate and contextually relevant retrieval. Though deep learning-based methods are more performing, they typically entail much more computational resources, in particular for long videos. Another line of work that has recently emerged is tries to overcome the above limitations by extracting representation from the video subtitles rather than the actual video content. This approach slides over subtitles to find the best matches to a query for retrieving relevant segments. This can massively alleviate computational overhead, as textual processing generally involves fewer computing resources than that of a video. The subtitles-based retrieval works for long videos for which the other methods are prohibitively time-consuming. It has been demonstrated that in the case of educational videos, query-based retrieval can be very effectively performed by using subtitles. Rich text content is high in educational videos and closely relates to topics and user-generated queries. With a focus on subtitles, it takes the most detailed and structured information from the educational videos and returns a more correct and contextually relevant answer. Therefore, such an approach not only enhances the efficiency of the retrieval process but also yields an enhanced user experience with more accurate and meaningful results. While more traditional keyframe extraction methods and deep learning-based summarization have pioneered query-based video retrieval, the shift towards subtitles based methods can have more computational efficiency for lengthy educational videos. Building on these recent advances, this paper furthermore highlights that the potential of subtitles in improving the retrieval of educational video content lies there in.

Our approach uses BERT sentence transformers [4] to understand the semantic and contextual nuances of the text. BERT's ability to convert both the query and the segment subtitles into the same sized embeddings allows for effective comparison. We specifically use only the encoder part of BERT to generate the text embeddings, ensuring that the model can comprehend and match the context of the query with the relevant subtitles segments accurately.

*C. Dataset*

This research uses the **EDUVSUM** dataset [5], which only contains educational videos and timestamp based subtitles. To improve its suitability for this study, varieties of queries for every video, along with answers/summaries provided by domain experts. Each query-answer pair is annotated with a relevance score, reflecting the alignment of the answer with the query based on the video transcript. These enhancements provide a neat benchmark dataset, allowing for effective evaluation of query-based educational video retrieval methods. Benchmark dataset is named as **EDUVRSUM**.

1. METHODOLOGY

The proposed approach for query-based video retrieval involves a series of systematic steps aimed at efficiently identifying and extracting relevant video segments based on user queries.

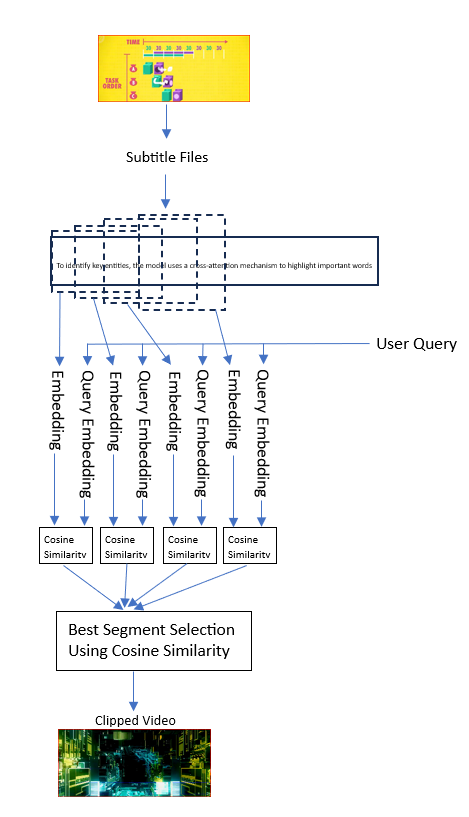
1. *Overview*

At first, for every video, time-framed subtitles are converted into a format that contains start time, end time and the respective subtitle. The subtitle file is in the format of WebVTT. The size and interval of the sliding window are in units of time. A sliding window is moved onto the subtitle with a predetermined interval and size till the end of the subtitle file of a video. An encoder part of BERT, which is essential for generating meaningful text embeddings. At a particular time, the user query as well as a segment of the subtitle from the sliding window are converted into a text embedding of equal size with the help of BERT Sentence transformer. Subsequently, cosine similarity between the embeddings is calculated. This process is repeated until the selection best subtitle segment with the higher cosine similarity is found.

Finally, Ffmpeg[6] is utilized to clip the identified video segments, ensuring that users can directly access the specific parts of the video that match their query.

1. *Data preparation*

To prepare data for our query-based video retrieval, we first extract video subtitles in WebVTT format, which includes precise timestamps for each segment. Using the webvtt library, we parse these time-framed subtitles and convert the start and end times into timedelta objects. This conversion allows for accurate manipulation and segmentation of the video text data, making it easier to manage and retrieve specific segments based on user queries. This process ensures that the video subtitles are accurately aligned with their respective timestamps for effective retrieval.Once the subtitles are parsed, they are split into segments based on a predefined window size and step size. The window size determines the duration of each segment, while the step size controls the overlap between consecutive segments. By iterating through the video captions and extracting text within each segment's time interval, we create a series of text segments. These segments form the basis for our subsequent analysis and retrieval process.

Next, we process the segmented subtitles to create embeddings using BERT sentence transformers. Each text segment, along with the user query, is converted into an embedding that captures its semantic and contextual meaning. This conversion is crucial because it enables us to compare the query with the text segments using cosine similarity. By doing this, we can identify the most relevant segments based on their content, ensuring efficient and accurate query-based video retrieval. This method ensures that the system retrieves the most contextually appropriate video segments for any given user query.Fig. 1. Workflow Diagram

1. *Embedding Generation*

For embedding generation, we use a pre-trained BERT model, which is really good at creating high-quality sentence embeddings. With the Sentence Transformers library, we turn subtitles and user queries into dense vector representations. First, we preprocess the text to make sure it's ready for embedding. Then, we use the BERT model to encode the text into fixed-size dense vectors that capture the text's meaning. We use these embeddings to compute cosine similarity scores between the user query and subtitles segments. This helps us match relevant video content accurately. BERT's ability to understand and represent detailed textual meanings is key for finding the right video segments, improving video content analysis and retrieval.

1. *Similarity Computation*

To compare the embeddings of subtitles and user queries, we employ cosine similarity, a metric that measures the cosine of the angle between two vectors in a multidimensional space. This approach is particularly effective for determining the similarity between text embeddings, as it focuses on the orientation of the vectors rather than their magnitude.

Cosine similarity ranges from -1 to 1, where 1 indicates that the vectors are identical in orientation, 0 indicates orthogonality (no similarity), and -1 indicates diametrically opposite vectors. The calculation involves the dot product of the vectors divided by the product of their magnitudes, formalized as:

*Cosine similarity (A , B) = A.B / |A||B|*

*WP.CITE(1)*

In our methodology, we start by generating embeddings for both the user query and the subtitles segments using a pre-trained BERT model Sentence Transformer. These embeddings are dense vectors that capture the meaning of the text. After obtaining the embeddings, we calculate the cosine similarity between the user query embedding and each subtitles segment embedding. This similarity score shows how closely the meanings of the two texts match. The equation (1) represents this in mathematical form.

High cosine similarity scores mean that the subtitles segment is very similar in meaning to the user query, helping us find the most relevant video segments. This method uses BERT's powerful contextual embeddings along with cosine similarity to measure how close the meanings are, making it easier to accurately and efficiently retrieve the right video content.

1. *Video clipping*

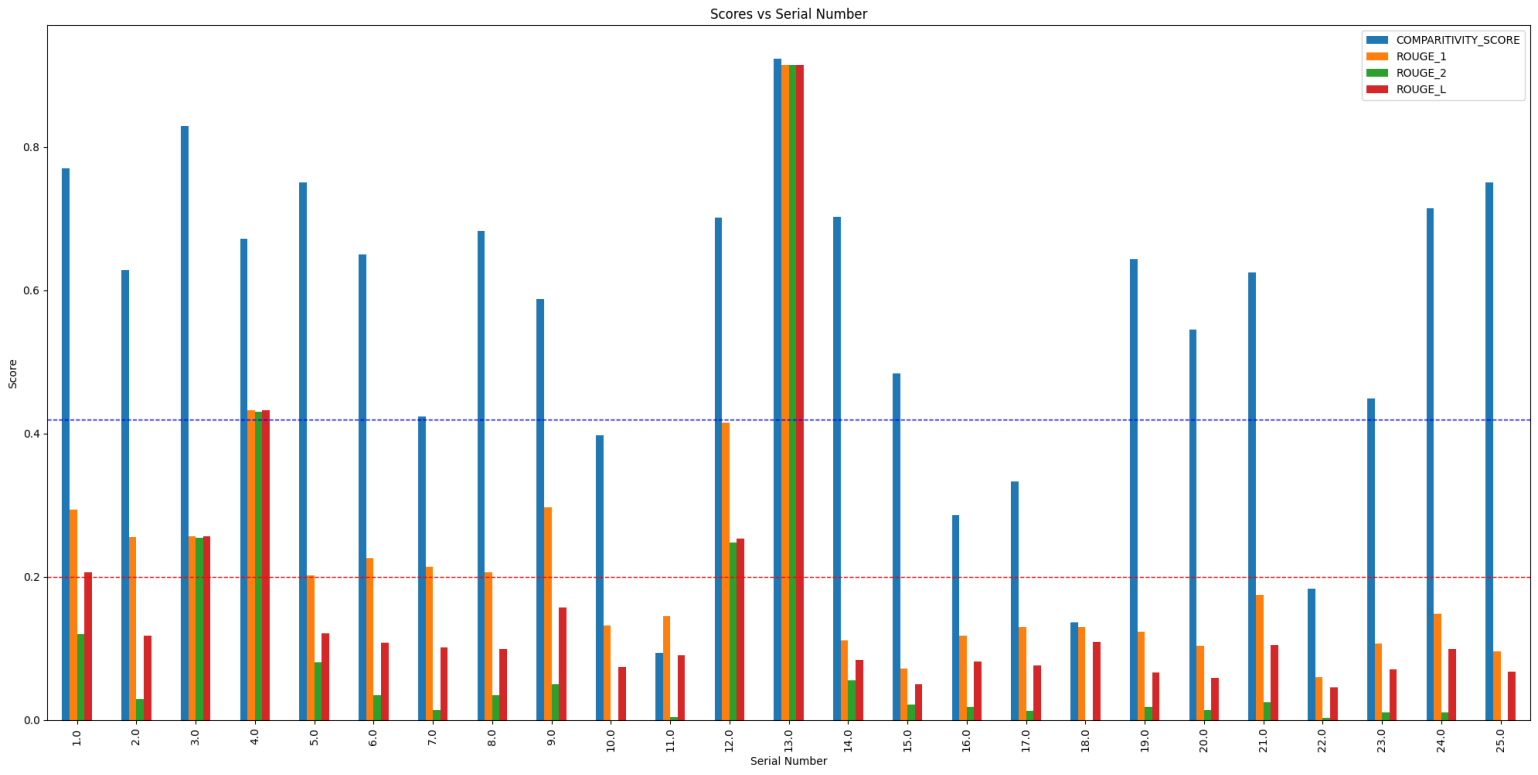
The retrieval process finds and gets the most relevant video parts by comparing user queries and subtitles embeddings. We pick the best subtitles segment which has the highest similarity scores. To do the Video clipping, we use FFmpeg, a strong multimedia tool. For each best subtitles segment, we find its start and end times and use FFmpeg to cut that part from the video. We specify the input file, start and end times, and output file to accurately copy the video and audio. This method lets us extract exact video content that matches the user's query.

Fig. 2. Bar graph between various queries vs scores (upto 25 queries)

1. EXPERIMENTAL RESULTS
2. *Dataset*

For evaluation, we created a benchmark dataset called the Educational Video Query Relevancy Dataset (EDUVRSUM). This dataset is derived from the existing **EDUVSUM** [5] dataset by adding several new columns: query, relevance score, Domain expert answer, and summary.

The additional columns provide enhanced information for evaluating the relevance of video content to specific queries.

The EDUVRSUM dataset includes 10 videos and their subtitles, for which we formulated 53 queries and corresponding answers which are actually given by the domain experts. The relevance score is a novel metric we introduced to classify the relationship between the query and the video subtitles:

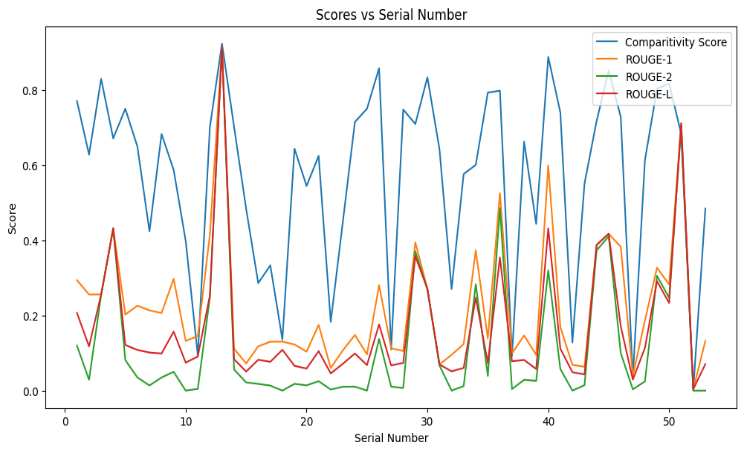
• Relevance = 1 (Fully Relevant): The query is fully relevant and its answer exists in the video.

• Relevance = 0 (Not Relevant): The query is not relevant and does not exist in the video.

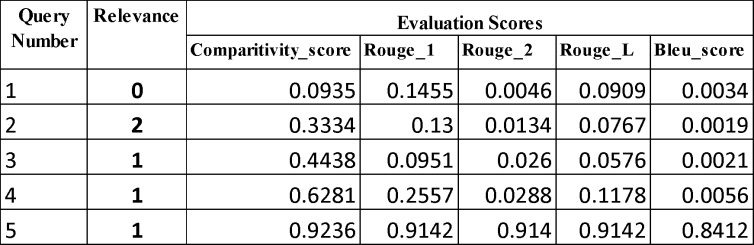
• Relevance = 2 (Partially Relevant): The query is relevant to the video's topic but not explicitly covered and its answer partially exists in the video.

Answers for queries with a relevance score of 1 are answers extracted directly from the subtitles by domain experts. For queries with a relevance score of 0, answers are provided by domain experts. For queries with a relevance score of 2, answers are partially derived from the subtitles with additional information from other sources, as these queries are under the advanced topics. This dataset helps check how well the video retrieval system can find the right video segment based on user queries. This way, we can see how good the system is performing for educational content videos.

1. *Evaluation Metrics*

In this study, we employed several evaluation metrics to assess the performance of our video retrieval system using the EDUVRSUM dataset, which contains various queries and their corresponding actual answers. The evaluation process involves calculating comparative scores, ROUGE scores, and BLEU scores to compare the actual answer and retrieved segment text. The comparative score is calculated using cosine similarity between the embeddings of the retrieved segment text and the actual answer. This metric helps in evaluating how closely the retrieved segment matches the actual answer in terms of semantic meaning. The process involves generating embeddings for both texts using a pre-trained BERT model sentence transformer and then computing the cosine similarity between these embeddings.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [7] is a set of metrics for evaluating automatic summarization and machine translation. We use ROUGE-1, ROUGE-2, and ROUGE-L scores to compare the overlap of unigrams, bigrams, and the longest common subsequences between the retrieved text and the actual answer. The ROUGE scores provide a measure of the quality and relevance of the retrieved segments by assessing the overlap of words and phrases.

BLEU (Bilingual Evaluation Understudy) [8] is another metric commonly used in machine translation and text generation. It measures the precision of n-grams in the retrieved text compared to the actual answer. We calculate the BLEU score using the sentence-level BLEU with smoothing to handle shorter texts. This metric evaluates the accuracy of the retrieved segments by considering the overlap of word sequences.

* **Segment Retrieval:** We first parse the VTT files and split the subtitles into segments using a sliding window of window size of three minutes and a step size of one minute. For each query, we calculate cosine similarity scores between the query embedding and the segment embeddings, sorting the results to identify the most relevant segments text.
* **Comparative Score Calculation:** For the top retrieved segment, we calculate the comparative score by computing the cosine similarity between the best segment embedding and the actual answer embedding.
* **Summary Evaluation:** We use the ROUGE scorer to calculate ROUGE-1, ROUGE-2, and ROUGE-L scores by comparing the retrieved text with the actual answer. Additionally, we compute the BLEU score for the retrieved text with the actual answer.
* **Data Compilation:** The calculated scores, along with the relevance information and other details, are compiled into a data frame for analysis.

1. *Results*

As a result from various evaluation metrics, including the comparative score (cosine similarity based on BERT embeddings) and ROUGE scores. The study concluded that the actual answer and retrieved segment texts achieved a

Fig. 3. Line graph between various queries vs scores

comparative scores of above 0.43 are the answers that are fully relevant and exist in the video Fig. 2 .The actual answers are given by domain experts in the EDUVRSUM dataset The threshold has been selected from the 25th quantile approach of the comparative score. A comparative score above 0.2 indicates partial relevance, where the answer exists partially, while scores below 0.2 indicates that the answer does not exist in the video. The ROUGE scores also exhibit similar trends in score similarity Fig. 3.

1. RELEVANCE SCORE SUMMARY TABLE

We illustrated these findings with a bar plot of 25 queries and their respective scores in "Fig. 2". "Table 1: Score Summary Table" presents a summary of the scores (Comparative, ROUGE-1, ROUGE-2, ROUGE-L, and BLEU) for queries with relevance scores of 0, 1, and 2, providing a clear overview of the performance across different metrics. Additionally, line graphs can illustrate trends and patterns for all queries and scores in the EDUVRSUM dataset in "Fig. 3".

We clearly see that, the maximum score of 0.92 for a relevant query which is matching with actual answer and a minimum score of 0.09 for not relevant query.

1. CONCLUSION

This study aims to improve students' learning by helping them to find specific topics in video without unnecessary contents. Instead of processing whole videos, we use subtitles to quickly find the best segment of educational videos that match user queries, saving computing power. Users can upload a video and its subtitles to get personalized clips based on their questions.This research has provided insights into state-of-the-art techniques in video summarization within the educational domain.

In the future, we want to improve our system to work with videos that are not just dialogues and that show emotions. We also plan to create summaries that combine information from multiple video segments for single queries, to give a complete understanding of their answers. Currently, our system has trouble finding topics that are spread throughout the video, as it mainly picks the best-matching parts.

1. REFERENCES
2. Jia M, Wei Y, Song X, Sun T, Zhang M, Nie L. Query-Oriented Micro-Video Summarization. IEEE Trans Pattern Anal Mach Intell. 2024 Jun;46(6):4174-4187. doi: 10.1109/TPAMI.2024.3355402. Epub 2024 May 7. PMID: 38236680.
3. Hu, W., Zhang, Y., Li, Y. *et al.* Query-based video summarization with multi-label classification network. *Multimed Tools Appl* **82**, 37529–37549 (2023). https://doi.org/10.1007/s11042-023-15126-1
4. W. Moon, S. Hyun, S. Park, D. Park and J. Heo, "Query - Dependent Video Representation for Moment Retrieval and Highlight Detection," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023 pp. 23023-23033.

doi: 10.1109/CVPR52729.2023.02205

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
2. Ghauri, J.A., Hakimov, S., & Ewerth, R. (2020). Classification of Important Segments in Educational Videos using Multimodal Features. *ArXiv, abs/2010.13626*.
3. Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal*, *2006*(146), 10.
4. Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
5. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.