Synthesizing A Concise Summary By Combining Content From Video Lecture And Reading Material

Team Members:

1. Michael Soebroto
2. Mohnish Sai Prasad
3. Ritesh Kumar

1. Topic:
   - **GAI for Education** - Summarization of a Set of Education Videos and Learning Materials

2. **Problem Statement**: Synthesize a concise summary by combining the important portions of both the reading material and a lecture video based on that material
   - This makes it easier for the user (student) to go through the material in less time and they don't have to watch the entire lecture video or read the complete reading material.
   - Previous approaches have been using Extractive methods to summarize content which often leads to less coherent summaries when compared to summaries created using Abstractive methods. Also, to the best of our knowledge there hasn't been any work on synthesizing a summary based on both lecture videos and the reading material associated with it.

3. Literature review:
   - Benedetto, I. et al., [1] have proposed a pipeline to summarize content from a video lecture. The pipeline is as follows: Speech-to-Text, Punctuation restoration and tokenization, extractive summarization, and abstractive summarization. The main model used for abstractive summarization is BART, pre-trained on the SAMSum dataset.
   - Mahapatra, D. et al., [2] propose MMToC and phrase cloud as two video summarization tools. Further they also propose Kenlists as a tool to put together a coherent set of video clips, and a course curation tool using the above. They have also designed a search that reranks YouTube search results according to the user profile.
   - Benedetto, I. et al., [3] have proposed a text summarizer based on facebook/bart-large-cnn model, which summarizes on smaller sections of size 3500 words. It follows a 5 step process to summarize the video. They use a text similarity algorithm and assign a score from 0 to 1 based on its similarity to the corresponding subtitles.
   - Tyagi, T. et al., [4] proposed the idea of tackling video summarisation in two phases: the first phase involves performing speech-to-text conversion for generating respective transcripts for

input videos while the second phase involves performing Extractive Text Summarization to summarize the text generated by extracting the important information.

- S. S. Thomas et. al., [5] present an approach to create a video summarization that is a precise representation of the video content. First, the approach finds out the salient activities that are taking place in the video. Second, the frames with the salient activities are stitched to form a single frame. Third, the summarized frame over multiple video shots obtained by the approach gives superior retrieval performance.
- Chau H. etal., [8] propose a method of concept extraction from adaptive textbooks, textbooks that are meant to adapt to the learners reading goals. Using automatic keyphrase extraction, they create a supervised feature-based machine learning method (FACE) to automatically extract concepts from digital textbooks.
- E. Apostolidis et. al., [9] present a survey on the landscape of the video summarization field utilizing deep learning models. In this paper, they examine the different deep learning approaches that have been used along with evaluating methods and performance comparison between the methods
- Ke et. al., [10] gives an approach towards video summarization utilizing a GAN model to identify the beginning and end of key clips from a video and transforming them into a more condensed video. In this approach, they utilize the bullet screen qualities of the BiliBili platform in order to identify these key points to be discovered by their model

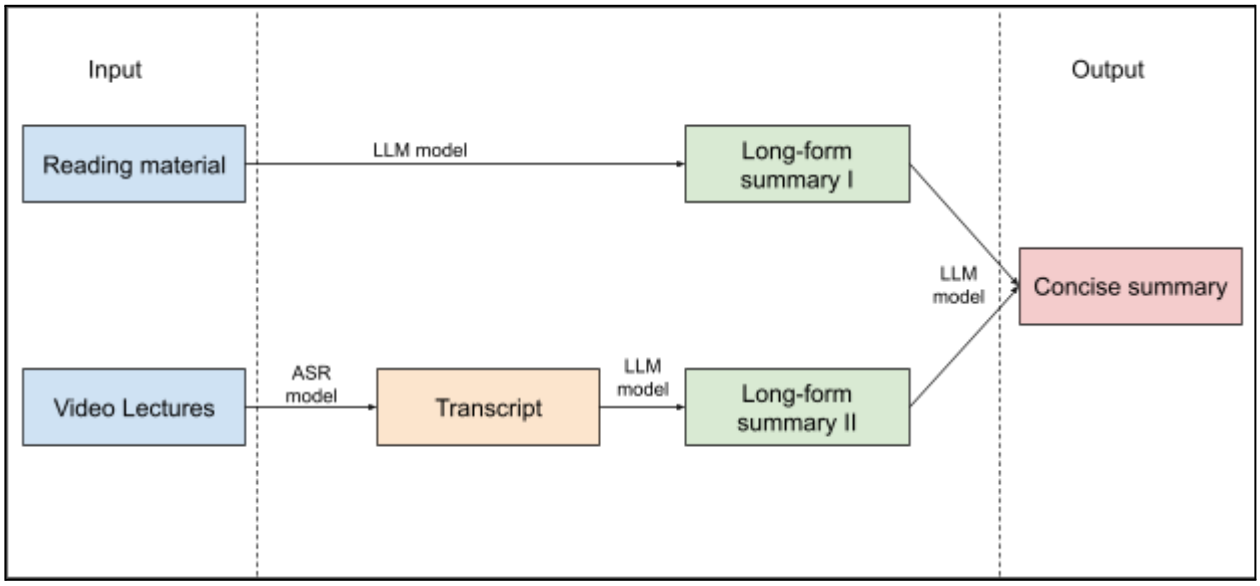
4. Methodology:

- Main approach: We will be taking in two forms of input: a textual input (main reference book for the course, and any other forms of notes), and a video input (a lecture video for example). Our output will be a concise summary of the input given, tailor-made from the student's perspective.

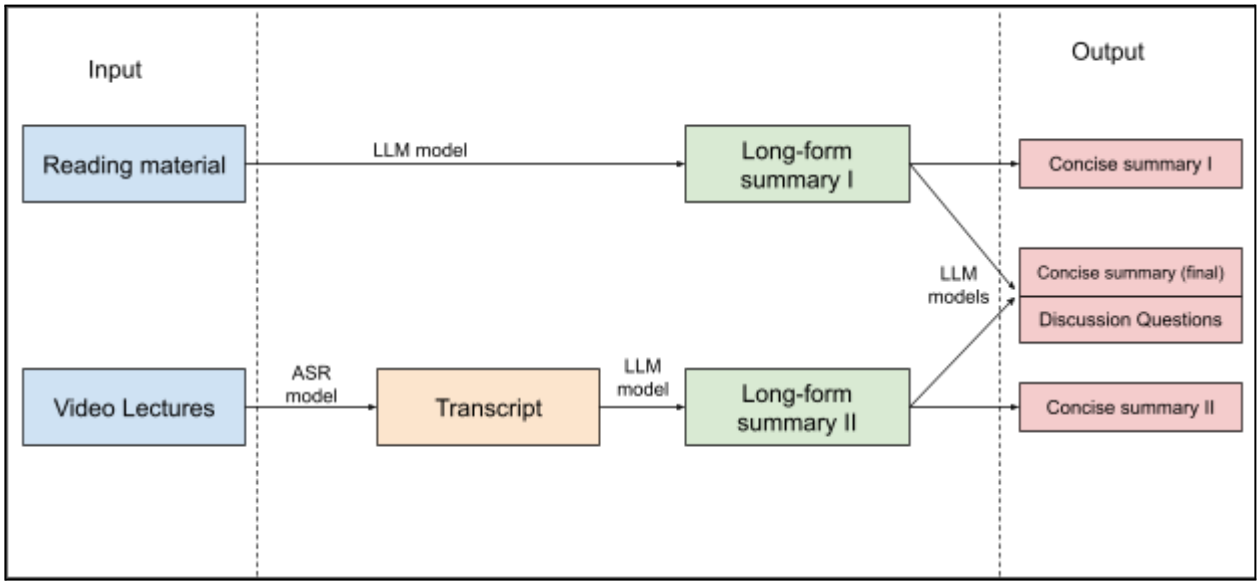
    i Step 1: We will convert the textual input to a long-form summary (a longer version of the summary. Not as concise as the final summary will be, but definitely shorter than the input material). The idea is to use a trained LLM for this step.

    ii Step 2: We will convert the video input to its transcript with the help of a speech-to-text translator (a generative model-based ASR). And from the transcript, we will extract a long-form summary similar to the one generated in Step 1.

    iii Step 3: From the two long-form summaries generated in Steps 1 and 2 we will generate a concise summary. For this purpose, we might use the same LLM model as above.



- Along with the concise summary we are planning to add these 2 features:
    - i Add an option to return the summary of each of the materials separately if the student requests for it. The returned summary could in the long-form manner (in which case we return the intermediate summary as is) or in the concise manner (we pass the long-form summaries through the LLM model separately)
    - ii We could add an option to generate thought-provoking questions to help students with deeper understanding of the materials. We pass both the long-form summaries through an LLM model trained exclusively for generating questions and get the output. We could even generate a small quiz to check the students understanding in the topic.



- Additional tasks: These are some improvements that we have planned to incorporate if time permits:
    - i To handle video lectures without a speaker: In the main approach we take in only the video lectures which can be converted to a transcript for our summary generation. For

the cases without a speaker (for example, if it is just a slideshow) we will use a keyframe extractor to select only the important snapshots of the video and then use a generative model (for example the CLIP model) to describe the keyframe snapshots in the form of text. With the collection of text from the video, we will generate a transcript (which will require another LLM model).

ii Adding images in the summary: As of now we will be giving the output in the form of text. But the students will find it useful if we can attach images along with the textual summary. We will do this in 2 stages:

1 Extract the images from the source materials - Pick all the images present in the reading materials, and select the key frames in the video lectures, and filter those images which are relevant to the text present in the combined concise summary (we will use an image describer for this purpose, a generative model).

2 Generate images from the summary text: Use a text-to-image based Generative AI model (like DALL-E) to create images for the summarized text and combine both.

5. Implementation:

- The pipeline of the process will be implemented in the Python language in Google Colab (to assist in collaborative work between us teammates). We will be using various AI libraries, notably from OpenAI, to assist us in this project.
- For the LLM models used in this project we intend to do two steps:

i Use a pre-trained model (like ChatGPT) with zero-shot learning. This is to set a benchmark on the performance of the summarisation, and the pipeline as a whole. This process is quick, and while ChatGPT is widely claimed to be very efficient in summarisation, it is known to fabricate content at times (Julian Tyson [6]) and is biased or inaccurate especially in the academic line (Harrer, S. [7])

ii Use a variety of models (pre-trained or vanilla) and train it with a dataset. The dataset we will be using is mentioned in the Evaluation section. We will compare the performance of each of these models and select the best out of them. If the performance varies drastically between models for each of the use-cases in the flow, then we will resort to using different models for each use-case.

6. Evaluation:

- The EDUVSUM dataset contains labeled data which we can use to do our preliminary testing on

i This dataset includes its own test set so we can test the accuracy precision and recall of our models on this set

    ii. If time allows, we could also create our own test dataset and check if the outcomes are what should be expected

- We could also create some samples and conduct a survey about the quality of the content.

7. Conclusion:

- We propose a system that can produce a concise summary of a particular lecture video and the reading material associated with it. For a baseline system we plan to use Zero Shot Learning, by using a pretrained model (from Hugging Face) to summarize the content.
- Further, we plan to fine tune this model based on availability of training data (video transcripts + reading material). We plan to use the EDUVSUM dataset, which contains about 98 videos with human generated summaries, to test our system.
- The significance of our project is that since students usually have a very limited attention span, making the course content to be shortened can have positive effects on the learning process. A student can choose to skip certain lectures just by looking at the synthesized summary and not have to watch the entire lecture or read the entire reading material. This way students can save time and learn about content that they're actually interested in.

8. Timeline, resources, and responsibilities:

<u>Resources:</u>

- Software: Google Colab, Google Cloud
- Datasets: EDUVSUM
    - A dataset of educational videos with subtitles for 3 top e-learning platforms
    - 98 videos with ground truths annotated by a user with an academic background in computer science

<u>Responsibilities:</u>

- Each group member will be responsible for a segment of the work.
- They will also be responsible for testing their own part and then at the end will collaboratively test the full model
- Models
    - Michael - Model for making summary from reading material
    - Mohnish - Model for making summary from video material
    - Rithesh - Take a list of summaries and creates a condensed summary from them

<u>Timeline:</u>

- Week 1: Forming the team and Brainstorming ideas
- Week 2: Creating the project flow, formulating the problem statement, and exploring the dataset
- Weeks 3-4: Building the flow and using zero-shot learning to make the benchmark.
- Weeks 5-7: Building and fine-tuning models to fit the problem, involving testing and comparing with the baseline performance.
- Week 8: Testing and fixing the entire workflow.
- Week 9: (Buffer) Improvements / extra features

- Week 10: Final touches and presentation

9. References:

[1] Benedetto, I., La Quatra, M., Cagliero, L., Canale, L., & Farinetti, L. (2023). Abstractive video lecture summarization: applications and future prospects. *Education and Information Technologies*, 1-21.

[2] Mahapatra, D., Mariappan, R., Rajan, V., Yadav, K., & Roy, S. (2018, April). Videoken: Automatic video summarization and course curation to support learning. In *Companion Proceedings of the The Web Conference 2018* (pp. 239-242).

[3] Benedetto, I., Farinetti, L., Canale, L., & Cagliero, L. (2021). *Video lectures summarization* (Doctoral dissertation, MA thesis. July 2021. url: https://webthesis. biblio. polito. it/19175/(cit. on p. 47)).

[4] T. Tyagi, L. Dhari, Y. Nigam and R. Nagpal, "Video Summarization using Speech Recognition and Text Summarization," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-7, doi: 10.1109/INCET57972.2023.10169901.

[5] S. S. Thomas, S. Gupta and V. K. Subramanian, "Context Driven Optimized Perceptual Video Summarization and Retrieval," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 10, pp. 3132-3145, Oct. 2019, doi: 10.1109/TCSVT.2018.2873185.

[6] Julian Tyson (2023). Shortcomings of ChatGPT. Journal of Chemical Education 2023 100 (8), 3098-3101. DOI: 10.1021/acs.jchemed.3c00361

[7] Harrer, S. (2023). Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. eBioMedicine, 90, 104512. https://doi.org/10.1016/j.ebiom.2023.104512

[8] H. Chau, I. Labutov, K. Thaker, D. He, and P. Brusilovsky, “Automatic Concept Extraction for Domain and Student Modeling in Adaptive Textbooks,” International Journal of Artificial Intelligence in Education, vol. 31, no. 4, pp. 820–846, Dec. 2021, doi: 10.1007/s40593-020-00207-1.

[9] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris and I. Patras, "Video Summarization Using Deep Neural Networks: A Survey," in Proceedings of the IEEE, vol. 109, no. 11, pp. 1838-1863, Nov. 2021, doi: 10.1109/JPROC.2021.3117472.

[10] F. Ke, P. Li, and W. Lu, “Video Summarization by DiffPointer-GAN,” in Neural Information Processing, T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, Eds., Cham: Springer International Publishing, 2021, pp. 605–614.