Leveraging Replay Feedback To Improve Video Transcript Summarisation

Track: Research

Varun Khurana
IIIT Delhi
varun19124@iiitd.ac.in

Harsh Kumar
IIIT Delhi
harsh19043@iiitd.ac.in

1 Introduction

Attention is a fundamental cognitive process that allows humans to selectively focus on relevant information while filtering out distractions. In today's fast-paced world, attention is a valuable commodity, particularly when it comes to consuming large volumes of online video content. With the rise of social media and video sharing platforms, the amount of video content being uploaded and viewed daily has exploded, making it increasingly difficult for users to find the most relevant and informative content quickly.

In recent years, there has been growing interest in developing attention-aware video summarisation techniques that mimic the way humans selectively attend to video content. These techniques aim to identify the most salient moments in a video by detecting key visual, auditory, and semantic cues that are likely to capture human attention.

YouTube has a new feature that has been designed to help people identify the most important parts of the video they are watching. This "Most Replayed" feature is essentially a temporal graph that can be used to locate and watch the most salient segments of a video, the parts that have been played by other users the most. By leveraging recent advances in deep learning and natural language processing, our approach aims to automatically generate high-quality, attention-driven summaries of video content that are both informative and engaging. The attention factor is incorporated by the use of most replayed graph data from Youtube API.

2 Problem Statement

We formally define the problem statement as follows: Given a video transcript $V = \{v_1, v_2, ..., v_N\}$ which is a time sequence of N tokens, a time sequence of replay scores R, a replay feedback-aware video transcript summarisation model $F(V, R)$ generates an extractive video transcript summary $S = \{s_1, s_2, ..., s_M\}$ which is a time sequence of the top-M (most relevant) tokens where $M << N$.

3 Motivation and Impact

Nowadays, the consumption of online video content has increased exponentially, especially due to the improvements in technology and wider access to cost-effective internet connectivity. The consumer is spoiled for choice. Therefore, to enable a user to gauge in a glance which video would best serve their requirements, we propose that a short, crisp and accurate textual summary of the video contents would prove to be a vital functionality. It would be useful for improving the user journey and ease of navigation.

Our hypothesis is that by integrating replay information provides crucial information about the most salient parts of a video. This feedback can be used to improve traditional deep learning-based video transcript summarisation models as it indicates the timestamps where viewers pay most attention in a video.

Therefore, even with lesser model parameters, the proposed model can achieve nearly the same performance, thereby saving computational cost and time.

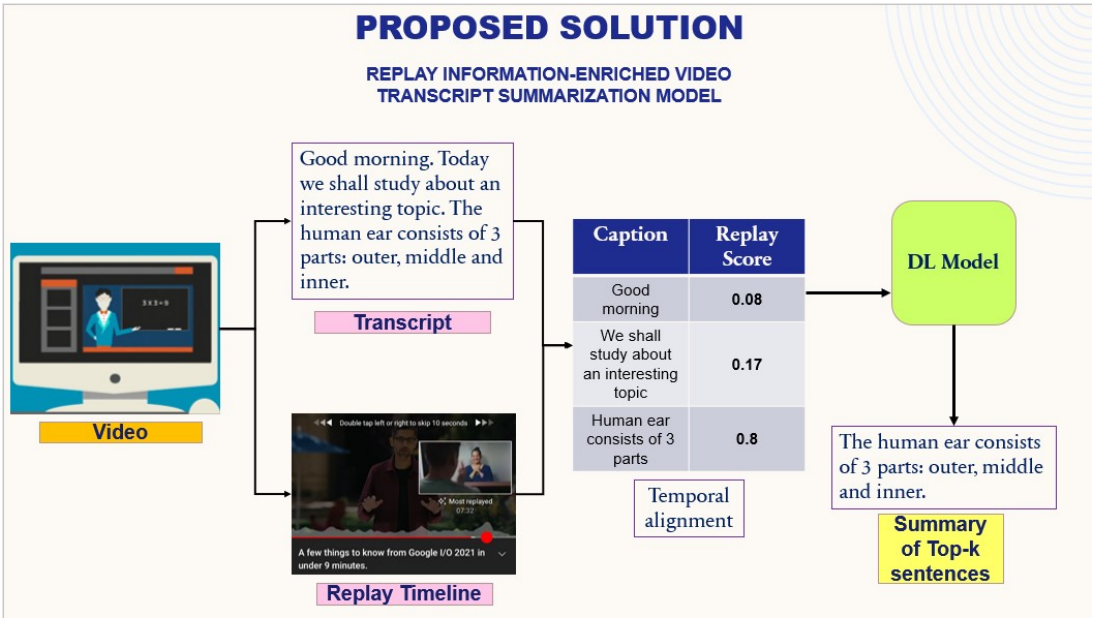

Figure 1: A schematic representation of the proposed replay information-enriched video transcript summarisation model

4 Related Work

Video Summarisation and Video Transcript Summarization: Video Summarisation as a task entails generating a short synopsis summarising the video content by selecting the most informative parts in chronological order. Recent works employ deep learning techniques to achieve good performance (Feng et al., 2020). With the advent of Vision Transformer (ViT) (Dosovitskiy et al., 2020), transformers have become popular for video summarisation (Feng et al., 2022). Our task is summarizing transcript of videos, rather than the video themselves. (Taskiran et al., 2006) proposed a method to automatically generate video summaries using transcripts obtained by automatic speech recognition. They divided the full program into segments based on pause detection and derive a score for each segment, based on the frequencies of the words and bigrams it contains. Then, a summary was generated by selecting the segments with the highest score to duration ratios while at the same time maximizing the coverage of the summary over the full program.

Text Summarisation: Text summarization is a well explored area. Text summarization can be done in two major ways: abstractive and extractive. Abstractive Text Summarization is the task of generating a short and concise summary that captures the salient ideas of the source text. The generated summaries potentially contain new phrases and sentences that may not appear in the source text. Extractive summarization aims at identifying the salient information that is then extracted and grouped together to form a concise summary.

(Liu and Lapata, 2019) showcases how BERT can be usefully applied in text summarization and propose a general framework for both extractive and abstractive models. They introduced a novel document-level encoder based on BERT which is able to express the semantics of a document and obtain representations for its sentences.

(Hovy, 2005) describes the design, implementation, and performance of various summarization systems. It describes the stages of automated text summarization as topic identification, interpretation, and summary generation, each having its sub stages.

5 Contributions

We summarise our contributions as follows:

- We propose to develop a novel video transcript summarisation model that takes

Model	rouge-1	rouge-2	rouge-4	rouge-l	rouge-w-1.2	rouge-s4	rouge-su4
Pegasus	0.284238	0.093417	0.055448	0.187455	0.074975	0.085250	0.119377
MT5	0.214794	0.049672	0.010313	0.124378	0.044798	0.040740	0.070726

Table 1: Baseline Results

that produces high quality extracive summaries of the contents in the video.

- We shall leverage the replay scores as feedback to improve the text summarisation model, thereby achieving better performance than conventional deep learning models.
- We shall create a computationally efficient deep learning model that supports faster inference.
- We scrape a dataset of YouTube videos across different domains such as "news", "education", "podcasts", "documentaries", etc., using the YouTube API. It contains the transcript and replay scores aligned with respect to timestamps. This dataset will be helpful for future work in this area also.

6 Data Collection and Processing

- We prepared a list of video content domains for which we wish to perform the transcript summarisation task such as news, education, podcast, documentary, etc. We plan to expand this list of domains and include multiple languages in future work.
- We identified a list of about 10 videos for each domain, each video being around 8-10 minutes in length.
- We scraped those videos for which captions and replay information was available using the YouTubeAPI[1].
- Captions often lack proper punctuation, specially if they're auto generated. To overcome this, we have added generated punctuation marks to our captions. (Kiss and Strunk, 2006)
- The video captions and normalised replay scores $\in [0, 1]$ were aligned with respect to time so that they can be used together in the text summarisation model.

7 Baseline Results

The dataset we have used for this experiment has been curated by our team. Thus, we don't have ground truth summaries for the video transcripts. To overcome this, we have considered summary generated by distilbart-cnn-12-6[2] as the ground truth. This enables us to train the proposed models in a supervised fashion.

For the baseline results, we use a few popular text summarisation models, namely Pegasus (Zhang et al., 2020) and MT5 (Xue et al., 2020)

- **Pegasus**: This model is based upon a transformer encoder-decoder architecture on large text corpora. It was pretrained with a masking objective function wherein important sentences which may be masked from the input document are generated.
- **MT5**: It is a multilingual text-to-text transfer transformer. It is capable of working well even with multilingual data. It was pretrained on Common-Crawl based dataset covering more than 100 languages.

As can be observed from Table 1, Pegasus (Zhang et al., 2020) performs consistently better on all metrics. This can be attributed to its training on a wide variety of summarisation tasks including news, science, stories, instructions, emails, etc.

[1] https://github.com/Benjamin-Loison/YouTube-operational-API

[2] https://huggingface.co/sshleifer/distilbart-cnn-12-6

8 Evaluation

For text summarisation, the degree of similarity between the gold standard summary (oracle summary) and the most relevant sentences from the original document predicted by the model is measured by ROUGE score (Lin, 2004). It stands for Recall-Oriented Understudy for Gisting Evaluation. It counts the number of overlapping units such as n-grams, word sequences or word pairs between the predicted and ground truth summary. Rouge-N computes the overlap of N-grams, while Rouge-L measures the longest common subsequence.

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

Suru Feng, Yuxiang Xie, Yingmei Wei, Jie Yan, and Qi Wang. 2022. Transformer-based video summarization with spatial-temporal representation. In *2022 8th International Conference on Big Data and Information Analytics (BigDIA)*, pages 428–433.

Xuming Feng, Lei Wang, and Yaping Zhu. 2020. Video summarization with self-attention based encoder-decoder framework. In *2020 International Conference on Culture-oriented Science & Technology (ICCST)*, pages 208–214.

Eduard Hovy. 2005. 583 Text Summarization. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32:485–525.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders.

C.M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E.J. Delp. 2006. Automated video program summarization using speech transcripts. *IEEE Transactions on Multimedia*, 8(4):775–791.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.