UNCWAI at MedVidQA 2023: T5 Model for Video Temporal Segment Prediction

Long Qi*, Owen Deen†, ZhiFei Xie‡, Xingyu Cui§, Gulustan Dogan¶
*Zhejiang University, Email: qi.21@intl.zju.edu.cn
†University of North Carolina Wilmington, Email: opd4736@uncw.edu
‡Beijing Institute of Technology, Email: xie_bit@163.com
§Nankai University, Email: 2012197@mail.nankai.edu.cn
¶University of North Carolina Wilmington, Email: dogang@uncw.edu

***Abstract*—In this paper, we present our solution to the MedVidQA 2023 Task 1: Video Corpus Visual Answer Localization. We used the training and testing datasets provided by the MedVidQA 2023 competition. For our run-1: we utilized a subtitle-questions cosine similarity score to rank the videos and then implemented a T5 model. For our run-2: we adjusted our ranking system to return the top three subtitle-answer similarity from the outputs of our BigBird model. For run-3: we used methods almost identical to run-1 except the T5 model was adjusted under different constraints. We found that run-1 had the best performance in leveraging both the selection of the relevant videos and IOU score. Although, we did notice that the IOU in run-2 could be stronger on some queries, yet the ranking did not encompass a broader selection of all of the possible relevant videos. The results of run-3 were of lower performance when compared to our previous runs since the fine-tuning of the T5 model was not at an optimal level. One of the issues we had with solving this task were the capabilities of our GPUs and the length of the training time. Our datasets were quite large and this put significant strain on our models.**

***Index Terms*—Timestamp Localization, Text Generation, T5 Model, Natural Language Processing, Instructional Videos, Subtitle Fragment Localization**

I. Introduction

Within the medical domain, an intriguing problem is to retrieve instructional content from a collection of videos. The goal of the TRECVID MedVidQA 2023 Task 1 [1] was given a medical query, locate the most relevant videos from the corpus, and then return the temporal segments of the given answer. Instructional medical videos have been popular to learn how to perform a particular medical task with a series of straightforward step-by-step procedures. These medical instructional videos have continued to grow in popularity due to the ease of the information for the user and visual-aided feedback. Compared to only Natural Language instructional content, the videos provide more robust information which is valuable in the medical domain. However, an issue with these medical instructional videos is that their duration could be long and distract the user from the answer for their query. Thus, our work can help the user save time and effort by being able to find the procedure timestamps directly. We also incorporate a ranking system, where if several videos contain answers to the user's medical query then we show the most relevant and helpful videos.

The main contributions of our research group are as follows:

- We developed a three-stage method based on a text-to-text model. We used subtitle-question similarity scores to rank the input video subtitles.
- We fine-tuned a base T5 model [2] for medical question answering and built connection between output text and video temporal segments.

II. Related Works

We first researched the existing methods used last year in the MedVidQA 2022 tasks. In particular, our first concern was methods for the localization of the answer timestamps. The methods proposed in [3] used pre-trained language models, namely RoBERTa and MPNet, with two approaches for answer localization: multi-ouput and peak detection. For our research, we were primarily concerned with the first method of answer localization from [3]. Their multi-output technique relied on normalizing the input lengths of the dataset. Then, for each video the subtitles were split into equal sized bins and with the maximum and median text-question cosine similarity calculated. We choose to implement a similar strategy for the answer localization in our T5 model architecture.

We also looked at the models of the top placing teams from the MedVidQA 2022 competition given in [4]. We found that the BigBird [5] model had promising results in the 2022 competition, so we decided to try a similar approach as mentioned in our Experiments section. In particular, the research team from VPAI [6] used a two-stage monomodal BigBird language model along concating video title and subtitles for text encoding.

III. Experiments

A. Dataset

The video corpus was provided by the MedVidQA 2023 competition [1]. We used the python library, *youtube-transcript-api*,[1] to find the subtitles for the given video identifications from the video corpus. Due to time constraints, video identifications that were unable to produce a subtitle transcript using the aforementioned library were discarded.

[1]For more information, see youtube-transcript-api.

B. Methods

We used the Pytorch [7] module to implement our models. The run-1 utilized the three-stage model based on a text-to-text model. At the first stage, text similarity is calculated between questions and video subtitles to select the most related videos that should contain relevant answers to the selected questions. At the second stage, a text-to-text conditional generation model T5 is fine-tuned to generate textual answers for the question based on video subtitles inputs. At the last stage, embedding cosine similarity is calculated to locate the subtitle fragment and subsequently locate the timestamp of that fragment. Experiments have been carried out to compare different generation models and test the model's effectiveness on solving the task. The average IOU (intersection over union) can reach up to 0.5877 on our test dataset.

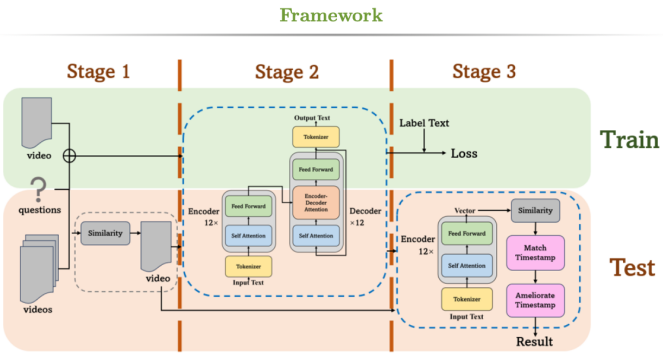

Fig. 1. Stage 1-Video Selection, Stage 2-Answer Text Generation, Stage 3-Timestamp Location.

In the training pipeline, the T5 model is fine-tuned to generate answer text from video subtitles, utilizing questions and their associated videos. The training process involves calculating the training loss, which represents the disparity between the generated answer text and the label text.

In the test pipeline, which involves answering questions with a video corpus, several stages are employed. Initially, the text similarity between the question and video subtitles is assessed to identify the most relevant subtitles. Then, a textual answer is generated through the use of a fine-tuned T5 model. In the final stage, the generated text is compared for similarity with the video subtitles, and a match between the subtitle and a timestamp is employed to produce the ultimate result. This multi-stage process enables the effective retrieval of answers in response to specific questions within the context of a video corpus.

Our base model was a T5-large which was available from Hugging Face [2], which we choose after the comparison in the figure below.

Models	IOU=0.3	IOU=0.5	IOU=0.7	avgIOU
T5-Small	0.45	0.3	0.2	0.35
T5-Large	0.8	0.5	0.5	0.5877

Fig. 2. IOU tests on different models.

We looked at the 40 questions over the top-20 selected relevant videos from the subtitle-question cosine similarity score.

The results of our run-1 T5 model is given below.

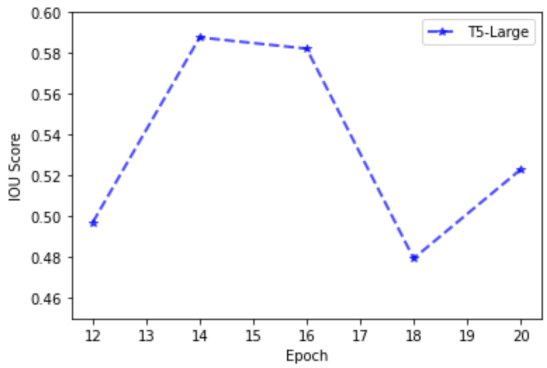

Fig. 3. IOU score on validation dataset with respect to training epoch.

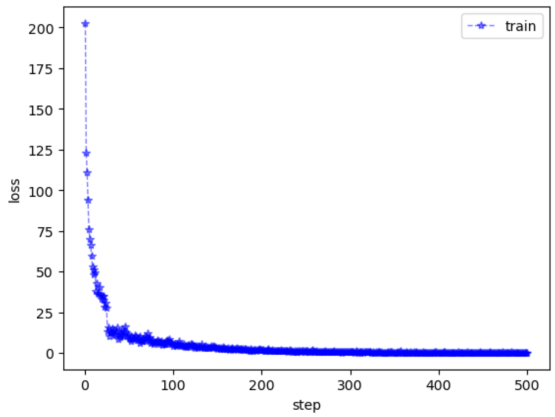

Fig. 4. Training loss with respect to each step, each step contains 100 samples.

IV. Conclusion

To address the video question answering challenge, conditional generation NLPs have high effectiveness. Treating timestamp extraction or question answering problem as a "text-to-text" problem first is of great benefits. Although, in the contest of video question answering, adopting a multi-modal approach may be more advantageous. These multi-modal models would combine both NLP techniques with Computer Vision such that we could extract text displayed on the screen. The addition of the visual components along with the language comprehension capabilities should enhance the overall performance.

Further, we noticed that the utilization of a pre-trained model is highly effective. Particularly in the domain of medical education due to the high specialization of the content. In our run-1 case, the employment of a T5 model that has been pre-trained on a large medical dataset can significantly improve our system's performance. The pre-training process equips the model with domain-specific knowledge and language patterns,

enabling it to better comprehend and generate medical information. This leads to more accurate and contextually relevant answers to medical queries posed within the video corpus.

V. Acknowledgement

We would like to thank NSF project 1053575 - XSEDE: eXtreme Science and Engineering Discovery Environment, for the computational resources used in our project. We would also like to thank NC State University for hosting the GEARS summer program.

References

[1] G. Awad, K. Curtis, A. A. Butt, J. Fiscus, A. Godil, Y. Lee, A. Delgado, E. Godard, L. Diduch, D. Gupta, D. D. Fushman, Y. Graham, , and G. Quénot, "Trecvid 2023 - a series of evaluation tracks in video understanding," in *Proceedings of TRECVID 2023*, NIST, USA, 2023.

[2] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[3] W. Kusa, G. Peikos, O. Espitia, A. Hanbury, and G. Pasi, "Dossier at medvidqa 2022: Text-based approaches to medical video answer localization problem," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 432–440, 2022.

[4] D. Gupta and D. Demner-Fushman, "Overview of the medvidqa 2022 shared task on medical video question-answering," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 264–274, 2022.

[5] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," 2021.

[6] B. Li, Y. Weng, F. Xia, B. Sun, and S. Li, "Vpai_lab at medvidqa 2022: a two-stage cross-modal fusion method for medical instructional video classification," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 212–219, 2022.

[7] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.