MedVidQA 2023

Long Qi

Zhejiang University

Abstract

With the development of artificial intelligence, AI-aid learning is getting increasingly popular. Extracting essential information from instructional videos is one of them. Facing the challenge of locating the timestamp of instructional medical videos to answer specific medical questions, our research group proposes several applicable Natural Language Processing models. I propose a three-stage model based on text-to-text model. At the first stage, text similarity is calculated between questions and video subtitles to select the most related videos that probably contain answer to the question. At the second stage, a text-to-text conditional generation model T5 is fine-tuned to generate textual answer to the question based on video subtitles. 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 can reach 0.5877 on test dataset. It is an innovative temptation to apply "text-to-text" language model for answer locating and medical instruction extraction.

Challenge

Many people prefer instructional medical videos to teach or learn how to accomplish a particular medical task with a series of step-by-step procedures in an effective and efficient manner. With an aim to provide visual instructional answers to consumers' first aid, medical emergency, and medical educational questions, given a medical query and a collection of videos, the TRECVID task on medical video question answering requires participants to retrieve the appropriate video from the video collection and then locate the temporal segments (start and end timestamps) in the video where the answer to the medical query is being shown, or the explanation is illustrated in the video, given a medical query and a collection of videos. [1]

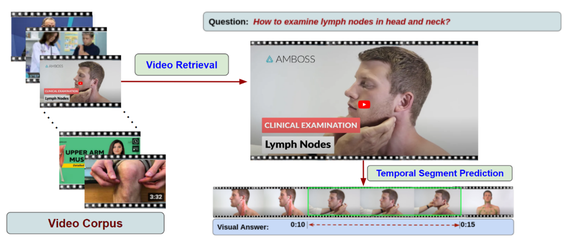

Figure 1:MedVidQA 2023 Challenge [1]

My Work

- Design a video subtitle question answering NLP pipeline to locate answer timestamp
- Fine-tune T5-Large Model to handle the challenge
- Compare effectiveness of different models
- Experiment to test the performance of the model

For the proposed NLP model :

- Extend "text-to-text" generation strategy [2] to wider applications.
- Fine-tune T5 for medical question answering.
- Innovatively build connection between text and video time segment.

My Model

For train pipeline:
Given questions and related videos, T5 model is fine-tuned to generate answer text from the video subtitles. The training loss is the difference between generated text and label text.
For test pipeline:
Given a question and a video corpus, at the first stage, text similarity is calculated between questions and video subtitles and the highest subtitles are selected. At the second stage, text answer is generated by fine-tuned T5 model. At the third stage, similarity between generated text and video subtitle and a match between subtitle and time stamp are used to generate the final result.

Framework

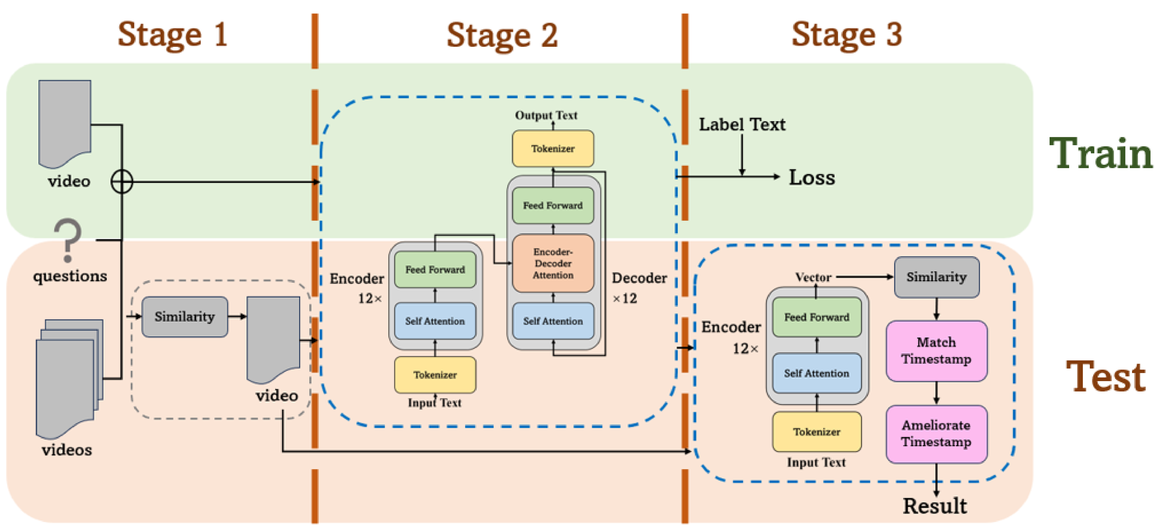

Figure 2:Three stage model proposal including train and test pipelines. Stage 1-Video Selection, Stage 2-Answer Text Generation, Stage 3-Timestamp Location.

Experiment

Training:
MedVidQA 2023 official training dataset is used for training. The base model is t5-large, which is downloaded from huggingface.co. The model is trained for 30 epoches, with loss reducing from above 1800 to below 0.7.
Testing:
MedVidQA 2023 official validation dataset is used for testing. The criterion to show performance is video timestamp IOU [3]. The model has reached 0.5877 with

$$IOU = \frac{prediction - and - gold}{prediction - or - gold} \quad (1)$$

Generating submission file:
MedVidQA 2023 official testing dataset is used to make a submission. 3 stage strategy is used strictly for each of the 40 questions. Top 20 related videos are selected for each question in the first stage.

Results

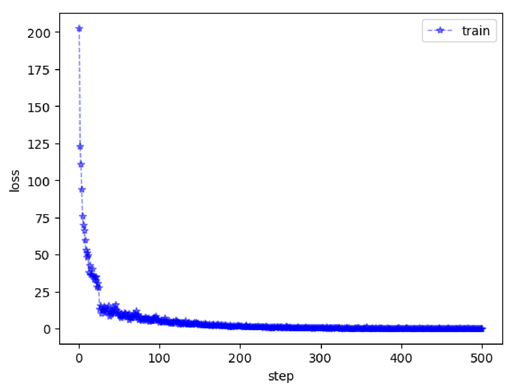

Figure 3:Training loss with respect to each step, each step contains 100 samples.

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

Table 1:IOU tests on different models.

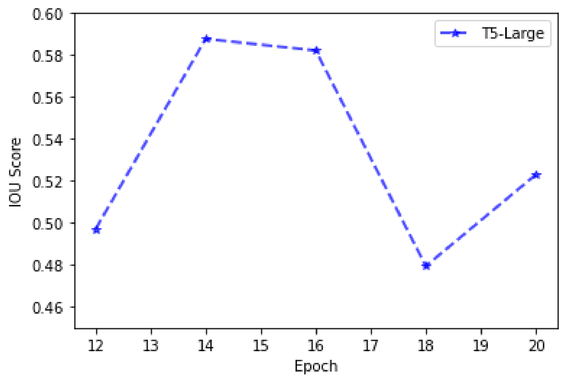

Figure 4:IOU Score on validation dataset with respect to training epoch

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. A three-stage model is proposed to do the MedVidQA 2023 task and reaches a good performance of average IOU 0.5877.

Future Work

- Multi-modal. Since it is video question answering, a combination of computer vision and NLP method may be better than NLP only.
- Pretrained model. Since the challenge is within medical area, a T5 model pretrained on large medical dataset can improve the performance.

References

1. D. Gupta and D. Demner-Fushman, TRECVID Task on Medical Video Question Answering MedVidQA 2023, https://medvidqa.github.io/
2. C. Raffel, N. Shazeer, et al, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2020, arXiv:1910.10683v3 [cs.LG].
3. D. Gupta and D. Demner-Fushman, Overview of the MedVidQA 2022 Shared Task on Medical Video Question-Answering.

Acknowledgements

Thanks Applied AI Lab, University of North Carolina Wilmington and GEARS Program. Thanks my instructor Dogan Gulustan. Thanks my teammates. Owen Paul helped with stage 1: video selection. Xie Zhifei helped with speech recognition and dataset building. Cui Xingyu helped with experiment and model comparison.

Contact Information

- Email: qi.21@intl.zju.edu.cn
- Phone: +86 18926820176