MI_TJU at TRECVID 2023: Medical Video Question Answering

Zibo Xu, Weizhi Nie, Qiang Li, Ning Xu, Yingchen Zhai, Zimu Lu, and Anan Liu

Tianjin University, Tianjin 300072, China
{xzb6666,weizhinie,liqiang,ningxu,zhaiyingchen,luzimu}@tju.edu.cn,
anan0422@gmail.com

Abstract. We participate in Video Corpus Visual Answer Localization (VCVAL), which includes two subtasks Video Retrieval (VR) and Temporal Segment Prediction (TSP). We trained our system through official youtube videos and eventually submitted one result, denoted by "run-1". Among the five submissions from three teams, our performance ranked first in two subtasks. For the video retrieval task, we mainly relies on matching video transcripts to the question to retrieve relevant videos. For the temporal segment prediction task, we combine visual and transcript features to accurately locate answers. We can attribute the effectiveness of our system to the robust fusion of the multimodal features. The highest ranking in each indicator underscores the value of our methodology, highlighting the importance of combining multi-modal features. This experience has given us insight into the power of a multimodal approach in handling complex VQA tasks.

1 Introdution

The Video Corpus Visual Answer Localization VCVAL task comprises two subtasks: video retrieval and temporal segment prediction [1, 2]. The goal of video retrieval is to identify the relevant videos in a given collection that are consistent with the medical questions [4]. Once the appropriate video is identified, the next task is to pinpoint the time period in the video. In these segments, answers to medical questions are provided, or relevant medical information is visually displayed. This task requires efficient identification of relevant video and precise positioning of information fragments to cover the integrated objectives of video retrieval and time fragment prediction [7].

2 Method

2.1 Video Retrieval

The training data is a structional videos from YouTube, it covers various aspects, such as the use of medical instruments, handling injuries, and providing care.

Unlike typical YouTube videos, medical instructional videos are distinct due to their educational nature, featuring extensive subtitles, a slower pace, and limited interactivity. From a visual perspective, medical instructional videos typically maintain a slow pace, focusing on the actions or explanations within a specific area by the instructor. These videos rarely exhibit sudden intrusions of objects or significant variations.

Therefore, we use the video transcript to summarize the main information of the video. The applicability of video transcripts can be attributed to the fact that medical instructional videos usually require the instructor to introduce or explain a medical procedure, often consisting of several sub-steps. Thus, by extracting the corresponding transcripts, the primary information from the videos can be obtained. In cases where videos lack subtitles, we utilize *YouTubeTranscriptApi* to convert the spoken words of the video characters into text.

Therefore, we view the video retrieval task as a scoring task between video transcripts, and questions. As shown in fig. 1 (left), we extract the transcripts features for each video and all question features. Then, we match the transcripts features with the question features, sort it by semantic correlation, and get a set of candidate videos. By extracting the corresponding transcripts, it‘s easy to get the topic of the medical instruction video, so visual features are not imperative during this process [5, 6].

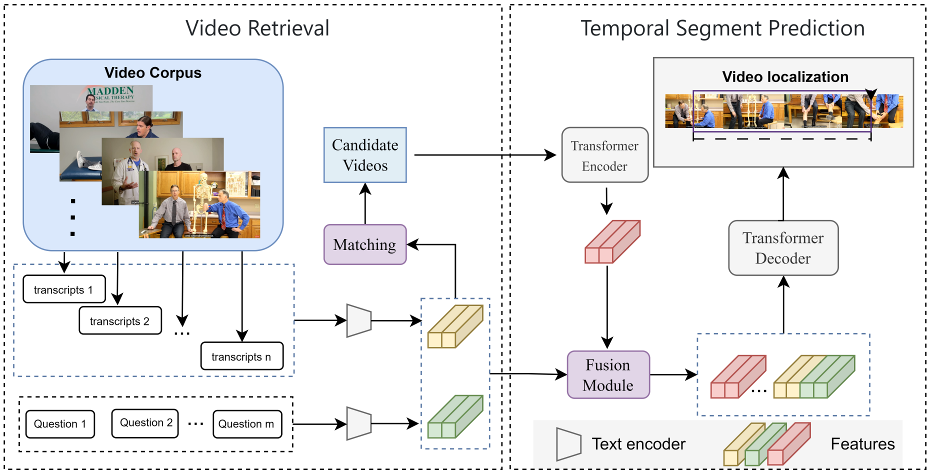

Fig. 1. Structure of two subtasks. For video retrieval, we extract and match the feature of each video transcripts and question text to get the candidate video related to the question. For temporal segment prediction, the extracted text features are fused with visual features, and the time interval where the predicted answer is located is finally obtained.

2.2 Temporal Segment Prediction

In the temporal segment prediction task, we used a pre-trained visual encoder to extract video features, as well as a pre-trained text encoder to extract questions and subtitle features for each video [3]. The extracted video, subtitles, and question features are fused into multimodal representations through the cross-modal attention mechanism. Finally, we use a feedforward network to locate the range of answers.

As shown in fig. 1 (right). the visual encoder helps to capture visual information in the video content, and the text encoder is used to process the language information in the questions and transcripts. This multi-modal fusion allows us to make associations between videos, questions, and transcripts to more accurately target the time period of answers. The advantage of this method is that it can use the information of different modes comprehensively and improve the performance of the model in the temporal segment prediction task.

3 Experiment

3.1 Results

In the video retrieval task, our team has demonstrated exceptional performance, achieving the highest scores among three competing teams in all five of our submissions. As shown in Table 1, our system has excelled across various evaluation metrics, including MAP, R@5, R@10, P@5, P@10, and nDCG. Notably, our outstanding performance is highlighted in terms of R@5 and R@10, where we effectively covered the majority of relevant videos. Additionally, our system achieved high precision in P@5 and P@10. Furthermore, our system displayed outstanding nDCG scores when ranking relevant videos, indicating that our results are not only of high quality but also excel in terms of ranking. These outcomes underscore our team's exceptional capabilities and performance in the field of video retrieval.

Table 1. The results of the video retrieval task.

Team	Run ID	MAP	R@5	R@10	P@5	P@10	nDCG
UNCWAI	run-2.json	0.1839	0.1903	0.1903	0.29	0.145	0.2858
VPAI	run-1.json	0.2427	0.2489	0.2489	0.31	0.155	0.3804
UNCWAI	run-1.json	0.3669	0.2221	0.3654	0.395	0.3575	0.5094
UNCWAI	run-3.json	0.3669	0.2221	0.3654	0.395	0.3575	0.5094
MI_TJU	run-1.json	**0.404**	**0.3549**	**0.4132**	**0.545**	**0.3625**	**0.5448**

In the temporal segment prediction task, our performance has been exceptional as shown in Table 2. We have demonstrated the ability to efficiently localize answers, regardless of whether the IoU threshold is relatively lenient (0.3) or stringent (0.7). Specifically, we achieved an outstanding IoU score of 67.5 at

IoU=0.3, showcasing our system's precision in locating answers even under the more relaxed threshold. At IoU=0.7, we also achieved a commendable IoU score of 50, indicating our system's proficiency in precise answer localization under stricter conditions. Furthermore, our average IoU (mIoU) stands at 55.24, underscoring our system's consistent ability to accurately localize answers across various conditions.

Table 2. The results of the temporal segment prediction.

Team	Run ID	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
UNCWAI	run-1.json	10	7.5	0	9.32
UNCWAI	run-3.json	25	10	5	15.78
UNCWAI	run-2.json	42.5	32.5	22.5	31.37
VPAI	run-1.json	57.5	35	25	39.97
MI_TJU	run-1.json	**67.5**	**62.5**	**50**	**55.24**

4 Conclusions

In this paper, we propose an efficient video retrieval method and a novel medical video localization method based on cross-modal representations. Different from other teams' methods, our approach combines linguistic features with visual features and achieves satisfactory results in two subtasks. Building on the success of this method, our team aims to extend its application to visual question-answering tasks, exploring in depth its effectiveness in tasks related to text generation. Through this exploration, we hope to understand the diversity and potential impact of our approach beyond the field of video localization.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant U21B2024.

References

1. George Awad, Keith Curtis, Asad A. Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Eliot Godard, Lukas Diduch, Yvette Graham, Georges Quénot: "TRECVID 2023 - A series of evaluation tracks in video understanding," Proceedings of TRECVID 2023 (2023).
2. Alan F. Smeaton, Paul Over, Wessel Kraaij: "Evaluation campaigns and TRECVid," in Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR '06), Santa Barbara, California, USA, 2006, pages 321–330. ACM Press. DOI: http://doi.acm.org/10.1145/1178677.1178722.
3. WonJun Moon, Sangeek Hyun, Sang-shin Paldal-gu Suwon-city Park, Dongchan Park, Jae-Pil Heo: "Query-Dependent Video Representation for Moment Retrieval and Highlight Detection," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pages 23023–23033. URL: https://api.semanticscholar.org/CorpusID:257757326.

4. Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, Ioannis Patras: "Video Summarization Using Deep Neural Networks: A Survey," Proceedings of the IEEE, vol. 109, no. 11, 2021, pages 1838–1863. DOI: 10.1109/JPROC.2021.3117472.
5. Hao Zhang, Aixin Sun, Wei Jing, Joey Tianyi Zhou: "Span-based Localizing Network for Natural Language Video Localization," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 2020, Online. Publisher: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.acl-main.585, pages 6543–6554.
6. H. Zhang, A. Sun, W. Jing, L. Zhen, J. T. Zhou, R. S. M. Goh: "Natural Language Video Localization: A Revisit in Span-based Question Answering Framework," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. DOI: 10.1109/TPAMI.2021.3060449.
7. Wojciech Kusa, Georgios Peikos, Oscar Espitia, Allan Hanbury, Gabriella Pasi: "Dossier at medvidqa 2022: Text-based approaches to medical video answer localization problem," in Proceedings of the 21st Workshop on Biomedical Language Processing, 2022, pages 432–440.