ICDAR2023 NewsVideoQA Competition Technical Report

ZhuangZhuang Cai

GammaLab, Pingan, China
caizhuang588@pingan.com.cn

Abstract. This paper presents a news video question-answering method based on OCR, layout analysis, object tracking, and ASR technologies. The method utilizes OCR technology to recognize text in video frames, uses layout analysis to merge paragraphs, employs object tracking algorithms to remove duplicate text in video frames, and finally uses ASR technology to transcribe speech in video clips. The OCR de-duplicated text and ASR text are concatenated to form the context for an extractive question-answering task. Our method achieved competitive results in the ICDAR2023 NewsVideoQA competition, demonstrating the effectiveness of using OCR and ASR technologies for news video question-answering.

Keywords: Question answering · OCR · ASR · News video understanding.

1 Introduction

ICDAR2023 NewsVideoQA is a competition that focuses on news video question answering[2], where the system needs to answer questions based on a given news video. In this competition, we proposed a basic method that utilizes OCR, layout analysis, object tracking, text deduplication, and ASR technologies to extract text information from video clips, and then uses it to build the context for extractive question answering tasks.

2 Methodology

Our method consists of the following five parts, with serial or parallel connections between different algorithms.

2.1 OCR Text Recognition

We use OCR technology to track and identify text in video frames. Specifically, we use the OpenCV library to extract individual frames from each video clip, and apply scene text detection methods[3] to detect text blocks. Then, we use text recognition algorithms[1] to recognize each text block, and save the coordinate positions of each text block.

2.2 Layout Analysis

After OCR processing, we obtain the original pixel segmentation results of text detection. We use algorithms such as connected component analysis, dilation, and erosion from the OpenCV library to merge text lines. The resulting paragraph text is composed of multiple text lines sorted vertically and horizontally according to their bounding box positions.

2.3 Object Tracking

Based on the results of layout analysis, we use the target tracking algorithm Deep SORT[6, 5] to track each paragraph box and assign the same ID number to the boxes that belong to the same target.

After recognizing and tracking all text segments in the video clip, we concatenated them based on their assigned IDs to obtain the OCR text of each video clip. We also removed duplicate text segments to ensure that the resulting OCR text is concise and accurate.

2.4 ASR Text Recognition

Besides OCR text, we also utilized ASR technology to obtain speech information from each video clip. To accomplish this, we employed OpenAI's Whisper [4] for transcribing the audio in each video clip.

To enhance the richness of contextual speech information, we concatenated the current video clip with the three preceding and following video clips before transcribing the audio. This allowed us to capture not only the speech in the current video clip but also the surrounding context.

After transcribing the audio for all four video clips, we concatenated the resulting ASR text to obtain the final text representation of the video clip.

2.5 Concatenation and Fine-tuning

To form the context for the extractive question-answering task, we concatenated the OCR text and ASR text. In order to answer judge questions, we added a "yes or no" sentence in front of the aforementioned context, which served as the answer for the extractive question-answering task.

We then fine-tuned the *deepset/bert-large-uncased-whole-word-masking-squad2* pre-trained model from the Huggingface library[7]. We use multiple training strategies to obtain multiple algorithm models.

3 Results

Our approach achieved competitive results in the ICDAR2023 NewsVideoQA competition. Applying model ensemble techniques, which involved simple voting among the top 6 best-performing models, we further improved the performance to an accuracy of 51.68% and an ANLS score of 66.83% on the same validation set.

4 Conclusion

In this competition, we proposed a vanilla approach that leverages OCR, layout analysis, target tracking, and ASR technologies to obtain text information from video clips, and then use it to form the context for the extractive question-answering task. Our approach achieved competitive results in the ICDAR2023 NewsVideoQA competition, demonstrating the effectiveness of using OCR and ASR technology for video question answering.

References

1. Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H.: What is wrong with scene text recognition model comparisons? dataset and model analysis (2019)
2. Jahagirdar, S., Mathew, M., Karatzas, D., Jawahar, C.V.: Watching the news: Towards videoqa models that can read (2022)
3. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization (2019)
4. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022)
5. Wojke, N., Bewley, A.: Deep cosine metric learning for person re-identification. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 748–756. IEEE (2018). https://doi.org/10.1109/WACV.2018.00087
6. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3645–3649. IEEE (2017). https://doi.org/10.1109/ICIP.2017.8296962
7. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Huggingface's transformers: State-of-the-art natural language processing (2020)