A Two-Stage Chinese Medical Video Retrieval Framework with LLM

Ningjie Lei[1], Jinxiang Cai[1], Yixin Qian[1], Zhilong Zheng[1], Chao Han[1,3], Zhiyue Liu[2,3], and Qingbao Huang[1,3(✉)]

[1] School of Electrical Engineering, Guangxi University, Nanning, Guangxi, China
[2] School of Computer, Electronics and Information, Guangxi University, Nanning, Guangxi, China
[3] Guangxi Key Laboratory of Multimedia Communications and Network Technology, Nanning, China
{leiningjie,keeplucky}@st.gxu.edu.cn
qbhuang@gxu.edu.cn

Abstract. With the increasing popularity of online videos, research on video corpus retrieval (VCR) has made significant progress. However, existing VCR models have not performed well in the medical field due to the unique characteristics of medical VCR task. Specifically, the open-ended queries used in medical VCR are more challenging compared to image-caption style queries, and the long duration of medical videos poses a great burden on model retrieval efficiency. To address these challenges, we propose a two-stage framework based on GPT-3.5 and cross-modal contrastive global-span (CCGS) for medical video VCR (termed GPT-CMR). In the first stage, we leverage the powerful natural language processing capabilities of the large language model (LLM) GPT-3.5 to improve retrieval efficiency. In the second stage, we use CCGS model to further enhance retrieval accuracy. Additionally, we developed a CCGS-VCR Analyzer to leverage the characteristics of the CCGS model's output without additional training costs. According to the official result, our method achieve first place in Track 2 of the NLPCC 2023 Task 5 competition. Experiments show that our method has retrieval efficiency and accuracy far exceeding the official baseline.

Keywords: Video corpus retrieval · Large language model · Cross-modal contrastive global-span

1 Introduction

In recent years, the rise of online videos has fundamentally changed the way people acquire knowledge and access information [12,24]. However, in the case of medical videos, individuals often lack the necessary medical expertise to effectively navigate the vast array of resources available on the internet. Therefore, it is highly meaningful to explore a video retrieval system that can assist people in efficiently and accurately obtaining targeted medical videos.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
F. Liu et al. (Eds.): NLPCC 2023, LNAI 14304, pp. 211–220, 2023.
https://doi.org/10.1007/978-3-031-44699-3_19

Video Corpus Retrieval (VCR) is a complex task that requires a deep understanding of both language and video. The majority of current VCR datasets, such as MSVD [2], ActivityNet [7] and MSR-VTT [25] consist of short video clips accompanied by a few queries. These queries are often in the form of image captions like "a dog is running in the grass." However, in the medical video domain, the goal of Video Corpus Retrieval is to retrieve target videos based on open-ended queries like "How can I ease my neck pain?" This demands models with a more profound comprehension of the videos. Medical videos also tend to be longer, with an average duration of 388.68 s in the proposed medical video dataset by [6]. Open-ended problems and extended video duration require models with robust overall capabilities. Models that prioritize efficient reasoning, such as the dual-tower structure utilized in [17], may not be well-suited for medical video retrieval. While span-based models like [13] have shown good performance, they suffer from low reasoning efficiency.

To address the aforementioned challenges, we present a two-stage retrieval-rerank framework aimed at improving both reasoning efficiency and accuracy. In the first stage, we utilized large language model [1] GPT-3.5 due to its excellent performance in natural language processing tasks to generate video summaries based on video subtitles. As the length of the summary is much shorter than that of the subtitles, we can use pretrained language models [4,9,11,21] like RoBERTa [16] for efficient retrieval. However, due to the loss of local key text information during the process of subtitles conversion into summarizes, further reranking was necessary to improve retrieval accuracy. In the second stage, we made the following enhancements to the cross-modal contrastive global-span (CCGS) model [13] : (1) We designed a CCGS-VCR Analyzer for the VCR task that leverages the characteristics of the CCGS model's output without training cost. This CCGS-VCR Analyzer utilizes a simulated annealing algorithm [10] to weigh the position and quantity sequences to obtain the final prediction sequence; (2) To address the scarcity of training samples in medical video datasets, we employed projected gradient descent(PGD) [19] for adversarial training to improve model robustness.

In summary, our contributions include:

- To balance efficiency and accuracy for Chinese medical long video retrieval, we designed a two-stage retrieval-rerank framework using GPT-3.5 and CCGS. To the best of our knowledge, we are the first to attempt using large language model to assist with the retrieval of Chinese medical videos.
- In this study, we propose a novel CCGS-VCR Analyzer without training cost specifically designed for the VCR task that leverages the output characteristics of the CCGS model. The results of the ablation experiments demonstrate the effectiveness of the CCGS-VCR Analyzer.
- Our solution achieve first place in Track 2 of the NLPCC 2023 Task 5 competition, with significantly improved retrieval accuracy and efficiency compared to the official baseline.

2 Related Work

With the growing popularity of online videos, VCR task has emerged as a crucial research topic in the field of multimodal learning. With the expansion of pre-training data such as Laion-400m [23] and Laion-5b [22] and the emergence of contrastive learning [8], multimodal pre-training models [3,14,15] such as CLIP [20] have gained prominence in video retrieval due to their robust image-text matching capabilities. Typically, clip-based method [18] utilize the CLIP model for image-text encoding, followed by cosine similarity calculation to generate the output. One advantage of this approach is that visual features can be computed beforehand and stored as vectors, enabling efficient inference by encoding the query and computing cosine similarity with the visual feature vectors. However, although effective for short videos and image-caption queries, this method may not be suitable for lengthy medical videos with open-ended queries.

Li et al. [13] developed the CCGS method for medical video retrieval to tackle the challenge of lengthy medical videos. The CCGS method first extracts features to obtain positive and negative pairs of video samples, and then feeds them into a language model along with their corresponding positive subtitle samples to extract text features. The resulting text feature pairs are then fused with visual feature pairs through cross-modal fusion, which generates a Global-span Matrix that is used for prediction purposes. Although CCGS achieve good results on the MedVidCQA dataset [6], its inference efficiency is limited by the long input subtitles. In the domain of natural language processing, text retrieval tasks usually follow a retrieval-rerank two-stage framework, which includes initial sorting using simple methods like the dual-tower model proposed by [5] in the retrieval stage to ensure efficiency and more complex methods in the rerank stage to improve retrieval accuracy. Works within such a framework have demonstrated good retrieval efficiency and accuracy. Additionally, previous studies like [1] have showcased the remarkable ability of large language models for text-related tasks such as Text Summarization.

3 The Proposed Approach

3.1 Task Definition

Formally, Video Corpus Retrieval(VCR) task comprises a set of queries $Q = \{q_1, q_2, q_3, q_4, ..., q_k\}$ and a video corpus $V = \{v_1, v_2, v_3, ..., v_4\}$, where k denotes the total number of queries and n represents the number of videos in the video corpus. Each query q corresponds to a unique video v, while each video may correspond to multiple queries. The primary objective of the VCR task is to accurately identify the specific video v that corresponds to each query q.

3.2 Method

Figure 1 illustrates our two-stage framework comprising retrieval and rerank stages. The retrieval stage employs GPT-3.5 and RoBERTa [16] for preliminary

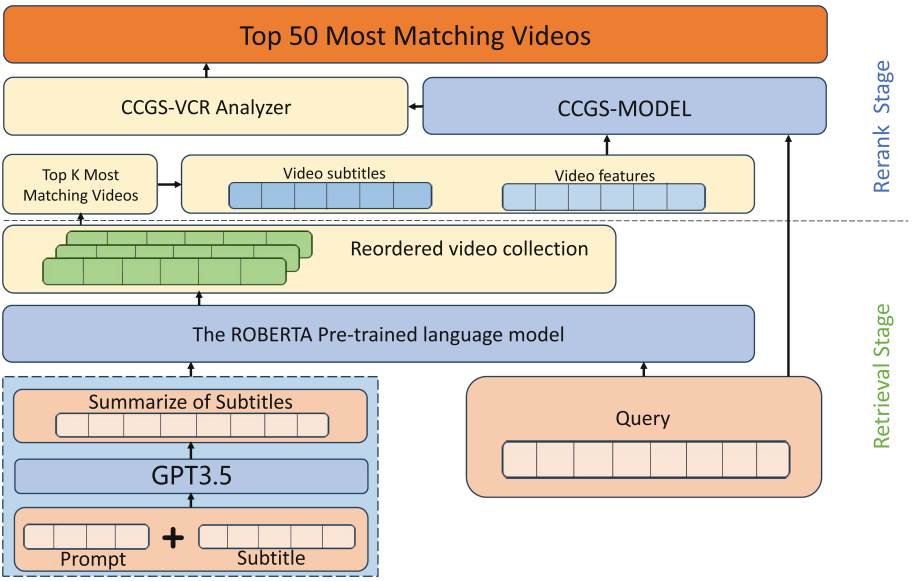


Fig. 1. The overall framework of GPT-CMR.

retrieval ranking, while the rerank stage utilizes the CCGS model [13] in conjunction with an CCGS-VCR Analyzer that we developed to selectively reorder the results obtained from the retrieval stage.

Retrieval Stage. We utilized GPT-3.5 to generate a video summary from the subtitles of a single video. To do so, we used a manually crafted prompt that instructed the model to retain key information for ease of retrieval tasks. We then combined this prompt with the video subtitles and input it into GPT-3.5 to produce the video summary. The output of this process is a new video corpus $V = \{v_1 : s_1; v_2 : s_2; v_3 : s_3; ...; v_n : s_n\}$, where $v_i$denotes the video ID and s_i represents the corresponding video summary. Subsequently, we concatenated each query q_i in the set $Q=\{q_1,q_2,q_3,q_4,...,q_k\}$ with every video summary s_i in V to construct the $Input_i=\{q_i + s_1,\ q_i + s_2,\ q_i + s_3,\ ...,\ q_i + s_n\}$ for RoBERTa [16]. This input was employed to generate $P_i=\{v_1 : p_1,\ v_2 : p_2,\ v_3 : p_3,\ ...,\ v_n : p_n\}$, where p_i signifies the probability score of each video in V given the query q_i. Finally, we sorted the P_i in descending order based on P_i values, which enabled us to rank the videos in V according to their relevance to the queries in Q.

Rerank Stage. After the first stage, each query can retrieve a list of video ID sorted in descending order by their relevance score. For instance, let $R_i = \{v_7,\ v_1,\ v_{15},\ ...,\ v_q\}$ represent such a list. In this stage, we select the top k videos from R_i and feed them, along with their corresponding subtitles and video features, into the CCGS model [13]. To improve robustness against adversarial attacks, we incorporate PGD [19] perturbations into the text embedding layer of the CCGS model during training.

CCGS-VCR Analyzer. Considering that the CCGS model [13] is a span-based model, we have designed a CCGS-VCR Analyzer for the VCR task based on the

characteristics of the CCGS model's output. For example, let's assume that the original prediction obtained from CCGS is $Predict_{orig}$=[$video_1 : segment_1$, $video_3 : segment_2$, $video_3 : segment_{24}$, $video_3 : segment_{15}$, $video_2 : segment_{28}$]. For the assessment of $Predict_{orig}$, a comprehensive analysis can be conducted from two perspectives: positional order and segment count. Positional order refers to the relative position of a video within $Predict_{orig}$. A higher position indicates greater relevance to the query. As such, we can generate $Predict_{pos}$=[$video_1$, $video_3$, $video_2$] by organizing the videos in $Predict_{orig}$ based on their positional order. Similarly, the frequency of appearance of a video within $Predict_{orig}$ is indicative of its relevance to the query. Thus, we can generate $Predict_{num}$=[$video_3$, $video_1$, $video_2$] by organizing the videos in $Predict_{orig}$ based on their frequency of appearance.

Finally, we will calculate the weighted sum of $Predict_{pos}$ and $Predict_{num}$ to obtain the final prediction, which is defined as:

$$Prediction = \alpha * Predict_{pos} + \beta * Predict_{num} \tag{1}$$

where the parameters α and β are obtained through a simulated annealing algorithm [10] with the overall metric on the validation set as the optimization target.

4 Experiments

4.1 Dataset and Evaluation

We utilized Chinese Medical Instructional Video Question Answering(CMIVQA) dataset, which was released by NLPCC 2023 Shared Task 5, to assess the effectiveness of our method. Table 1 presents the composition of the training and testing datasets in CMIVQA, as well as the average video length. During training, we randomly selected four hundred samples from the training set to use as the validation set. To evaluate the system's performance, we used R@1, R@10, R@50, MRR, and overall value as metrics, where overall is the sum of R@1, R@10, R@50, and MRR. In addition, to explore reasoning efficiency, we will also calculate the average reasoning time for each query, which is the total reasoning time divided by the number of queries.

4.2 Experimental Settings

All of our experiments were conducted on a single NVIDIA 3090 GPU. The training process consisted of retrieval and rerank stages, and the detailed parameters can be found in Table 2. In terms of experimental settings, we considered both retrieval accuracy and efficiency. To investigate the best retrieval accuracy, we selected the top 150 of the retrieval stage as the input for the rerank stage in the comparative and ablation experiments. To evaluate the impact of our framework on retrieval accuracy while improving efficiency, the retrieval stage's top k values were set to top 5, top 10, top 20, top 50, top 100, and top 150, respectively. Retrieval accuracy and average search time were then computed on the test set.

Table 1. Composition of the CMIVQA Dataset.

Dataset	Videos	QA pairs	Vocab Nums	Question Avg. Len	Video Avg. Len
Train	1,228	2,937	3125	17.16	263.3
Test	200	492	2171	17.81	242.4

Table 2. Main hyper-parameter setting

Retrieval stage		Rerank stage	
Epoch num	100	Epoch num	30
Batch size	32	Batch size	1
Optimizer	AdamW	Optimizer	AdamW
Learning rate	3e-6	Learning rate	1e-5
Weight decay	0.01	PGD ϵ/α	1/0.3

4.3 Results and Discussions

To assess the efficacy of our framework, we conducted ablation experiments and compared our method with the official baseline and CCGS [13]. The detailed experimental results are documented in Table 3. We carried out a comparison between our framework and the baseline model in terms of retrieval efficiency and retrieval accuracy under the condition of top k=top 10 to demonstrate its superiority, as presented in Fig. 2. Moreover, we investigated the influence of retrieval efficiency on retrieval accuracy using our method by testing different values of top k separately, and the results are outlined in Table 4 and Fig. 3.

Table 3. Comparison experiment and ablation experiment results. The experimental results of the official baseline were obtained from official, while the experimental parameters of CCGS were consistent with GPT-CMR's reranking stage, except for the absence of PGD.

Model	R@1	R@10	R@50	MRR	Overall
Official baseline	0.3943	0.5366	0.6423	0.4412	2.0144
CCGS	0.4012	0.8024	0.8696	0.4991	2.5723
GPT-CMR(Ours)	**0.5764**	**0.8391**	**0.9431**	**0.6710**	**3.0296**
W/o Analyzer	0.5163	0.8374	0.9431	0.6323	2.9290

From the results, it is evident that our proposed GPT-CMR framework outperformed all other models in every metric. When compared to the official baseline, GPT-CMR presented significant improvements across all metrics. Furthermore, when compared to CCGS, GPT-CMR demonstrated enhancements across all metrics, especially with 17.52% increase in R@1, indicating superior retrieval accuracy performance.

The last row of Table 4 presents the ablation experiment results of GPT-CMR after removing CCGS-VCR Analyzer. We used simulated annealing algorithm [10] to optimize the overall value as the optimization objective on the validation set and determined the weights in formula (1) to be α =0.9 and β=0.1. From the ablation experiment results, it can be seen that removing CCGS-VCR Analyzer led to a significant decrease in both R@1 and MRR, indicating the effectiveness of CCGS-VCR Analyzer.

Furthermore, we evaluated the retrieval efficiency and accuracy of GPT-CMR, CCGS, and the official baseline by setting top k as top 10, and the comparative results are depicted in Fig. 2. As shown in Fig. 2(a), although the retrieval accuracy decreased after reducing top k from top 150 to top 10, GPT-CMR still outperformed all baseline models in overall retrieval accuracy. Additionally, as seen in Fig. 2(b), the retrieval time of GPT-CMR was significantly faster than that of all baseline models.

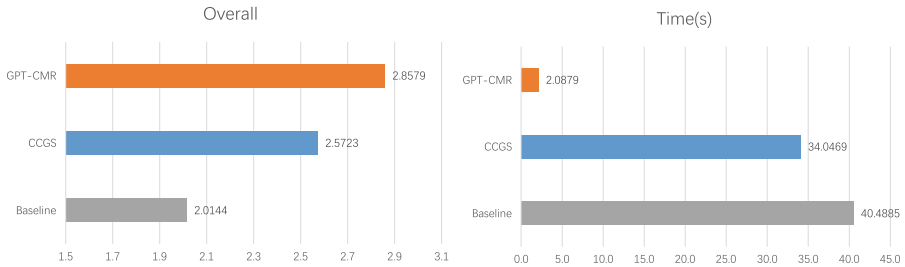
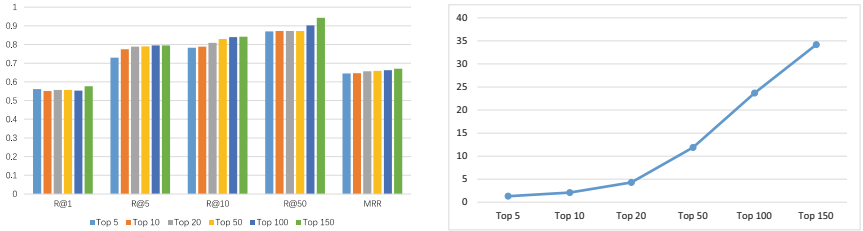


Fig. 2. Experimental results of retrieval efficiency time. Figure 2(a) shows the comparison of the Overall metrics between GPT-CMR and other models under the top 10 condition, with the horizontal axis indicating the Overall value. Figure 2(b) shows the comparison of retrieval time, with the horizontal axis representing the average retrieval time of queries in seconds.

Table 4. The retrieval metrics and retrieval time results of GPT-CMR under different top k conditions.

Top k	R@1	R@5	R@10	R@50	MRR	time(s)
top 5	0.5609	0.7296	0.7825	0.8699	0.6455	1.3076
top 10	0.5508	0.7744	0.7886	0.8719	0.6465	2.0879
top 20	0.5569	0.7886	0.8089	0.8719	0.6568	4.2988
top 50	0.5565	0.7906	0.8292	0.8719	0.6590	11.8995
top 100	0.5535	0.7947	0.8321	0.9021	0.6627	23.6831
top 150	0.5764	0.7947	0.8391	0.9429	0.6710	34.1863



(a) (b)

Fig. 3. The experimental results under different top k conditions. Figure 3(a) shows the statistics of retrieval accuracy under different values of k, while Fig. 3(b) illustrates the trend of retrieval time under different values of k.

From the experimental results in Table 4, it can be observed that the increase of k value has a small impact on R@1 before k=150. However, the overall metrics including R@5, R@10, R@50 and MRR show an increasing trend as k value increases. This trend is more evident in Fig. 3(a) where there is a slow but steady growth. As for retrieval efficiency, it is clear that the average retrieval time significantly increases with an increase in k value. This is evident from the line chart in Fig. 3(b) showing an overall increasing trend.

Taking the comparison between top 10 and top 100 as an example, we can observe that top 100 only increases by 0.0027% in terms of R@1 compared to top 10, and MRR also only improves by 0.0162%. However, the retrieval time for top 100 is 11.343 times that of top 10. This phenomenon indicates that the improvement in retrieval accuracy brought about by the increase in top k is not proportional to the decrease in retrieval efficiency.

In addition, in terms of the R@5 metric, there is a significant improvement of 4.48% in accuracy when comparing top 10 to top 5. However, the retrieval time for top 10 is only 0.7803 s longer than that of top 5. Therefore, selecting a reasonable top k can achieve decent retrieval accuracy on the basis of fast retrieval efficiency.

5 Conclusion

In this paper, we propose a two-stage framework called GPT-CMR (Chinese Medical Video Retrieval with CCGS and GPT-3.5) for medical VCR task. To improve the efficiency of medical video retrieval, we use the large language model GPT-3.5 in the first stage to generate video summaries based on the video subtitles, which are then used for initial retrieval. In the second stage, we employ CCGS [13] in conjunction with CCGS-VCR Analyzer to rerank the top k retrieval results obtained in the first stage. This process yields the final retrieval results. Comparative experiments demonstrate the superiority of GPT-CMR in terms of retrieval accuracy and time efficiency. Ablation experiments also confirm the effectiveness of the CCGS-VCR Analyzer. Furthermore, the retrieval

accuracy and efficiency of GPT-CMR can be impacted by varying values of top k. To address this, we conducted an analysis of GPT-CMR's performance using different top k values. Our data analysis indicates that GPT-CMR can maintain excellent retrieval accuracy while sustaining efficient retrieval performance.

We believe that there are some directions for future exploration. Firstly, there is a need to explore more effective ways of leveraging large language models to improve the performance of the VCR task. While we have attempted to generate video summaries to assist in VCR, other techniques such as keyword-based retrieval can also be explored. Secondly, given that the GPT-CMR model heavily relies on textual information, there is a pressing need to investigate how audio and visual information can be integrated more effectively to assist in VCR tasks. Finally, although GPT-CMR has demonstrated notable progress in terms of retrieval efficiency compared to baseline models, it still falls short of meeting the requirements of practical applications. As such, further research is necessary to enhance retrieval efficiency.

Acknowledgments. This work was supported by the Guangxi Natural Science Foundation (No. 2022GXNSFAA035627), Guangxi Natural Science Foundation Key Project (Application No. 2023JJD170015), National Natural Science Foundation of China (62276072), Guangxi Scientific and Technological Bases and Talents Special Projects (guikeAD23026213 and guikeAD23026230), Innovation Project of Guangxi Graduate Education, and the Open Research Fund of Guangxi Key Laboratory of Multimedia Communications and Network Technology.

References

1. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
2. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 190–200 (2011)
3. Chen, Y.-C., et al.: UNITER: universal image-text representation learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12375, pp. 104–120. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58577-8_7
4. Dong, L., et al.: Unified language model pre-training for natural language understanding and generation. In: Advances in Neural Information Processing Systems 32 (2019)
5. Gao, W., et al.: Deep retrieval: learning a retrievable structure for large-scale recommendations. arXiv preprint arXiv:2007.07203 (2020)
6. Gupta, D., Attal, K., Demner-Fushman, D.: A dataset for medical instructional video classification and question answering. Sci. Data **10**(1), 158 (2023)
7. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: a large-scale video benchmark for human activity understanding. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 961–970. IEEE (2015)
8. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1), 2 (2020)

9. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, vol. 1, p. 2 (2019)
10. Kirkpatrick, S., Gelatt, C.D., Jr., Vecchi, M.P.: Optimization by simulated annealing. Science **220**(4598), 671–680 (1983)
11. Lewis, M., et al.: Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880 (2020)
12. Li, B., Weng, Y., Sun, B., Li, S.: Towards visual-prompt temporal answering grounding in medical instructional video. arXiv preprint arXiv:2203.06667 (2022)
13. Li, B., Weng, Y., Sun, B., Li, S.: Learning to locate visual answer in video corpus using question. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)
14. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900. PMLR (2022)
15. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: vision and language representation learning with momentum distillation. Adv. Neural. Inf. Process. Syst. **34**, 9694–9705 (2021)
16. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
17. Lu, W., Jiao, J., Zhang, R.: Twinbert: distilling knowledge to twin-structured compressed bert models for large-scale retrieval. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 2645–2652 (2020)
18. Luo, H., et al.: Clip4clip: an empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
19. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. Stat. **1050**, 4 (2019)
20. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
21. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
22. Schuhmann, C., et al.: Laion-5b: an open large-scale dataset for training next generation image-text models. Adv. Neural. Inf. Process. Syst. **35**, 25278–25294 (2022)
23. Schuhmann, C., et al.: Laion-400m: open dataset of clip-filtered 400 million image-text pairs. In: NeurIPS Workshop Datacentric AI. No. FZJ-2022-00923, Jülich Supercomputing Center (2021)
24. Weng, Y., Li, B.: Visual answer localization with cross-modal mutual knowledge transfer. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)
25. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: a large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5288–5296 (2016)