Dialogue-to-Video Retrieval

Chenyang Lyu(✉), Manh-Duy Nguyen, Van-Tu Ninh, Liting Zhou,
Cathal Gurrin, and Jennifer Foster

School of Computing, Dublin City University, Dublin, Ireland
{chenyang.lyu2,manh.nguyen5,van.ninh2}@mail.dcu.ie,
{liting.zhou,cathal.gurrin,jennifer.foster}@dcu.ie

Abstract. Recent years have witnessed an increasing amount of dialogue/conversation on the web especially on social media. That inspires the development of dialogue-based retrieval, in which retrieving videos based on dialogue is of increasing interest for recommendation systems. Different from other video retrieval tasks, dialogue-to-video retrieval uses structured queries in the form of user-generated dialogue as the search descriptor. We present a novel dialogue-to-video retrieval system, incorporating structured conversational information. Experiments conducted on the AVSD dataset show that our proposed approach using plain-text queries improves over the previous counterpart model by 15.8% on R@1. Furthermore, our approach using dialogue as a query, improves retrieval performance by 4.2%, 6.2%, 8.6% on R@1, R@5 and R@10 and outperforms the state-of-the-art model by 0.7%, 3.6% and 6.0% on R@1, R@5 and R@10 respectively.

Keywords: Dialog-based retrieval · Dialogue search query · Conversational information

1 Introduction

The aim of a video retrieval system is to find the best matching videos according to the queries provided by the users [5,8,20,25,26]. Video retrieval has significant practical value as the vast volume of videos on the web has triggered the need for efficient and effective video search systems. In this paper, we focus on improving the performance of video retrieval systems by combining both textual descriptions of the target video with interactive dialogues between users discussing the content of the target video.

Previous work on video retrieval applied a CNN-based architecture [12,16,18] combined with an RNN network [3] to handle visual features and their time-series information [2,30,32]. Meanwhile, another RNN model was employed to embed a textual description into the same vector space as the video, so that their similarity could be computed in order to perform the retrieval [2,26,32]. Due to the huge impact of the transformer architecture [29] in both text and

C. Lyu, M.-D. Nguyen and V.-T. Ninh—Contributed equally.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
J. Kamps et al. (Eds.): ECIR 2023, LNCS 13981, pp. 493–501, 2023.
https://doi.org/10.1007/978-3-031-28238-6_40

image modalities, this network has also been widely applied in the video retrieval research field, obtaining improvements over previous approaches [4,9,13,17,22].

Current video retrieval research, however, mainly focuses on plain text queries such as video captions or descriptions. The need to search videos using queries with complex structures becomes more important when the initial simple text query is ambiguous or not sufficiently well described to find the correct relevant video. Nevertheless, there are only a few studies that focus on this problem [23,24]. Madusa et al. [23] used a dialogue, a sequence of questions and answers about a video, as a query to perform the retrieval because this sequential structure contains rich and detailed information. Specifically, starting with a simple initial description, a video retrieval model would return a list of matching videos from which a question and its answer were generated to create an extended dialogue. This iterative process continued until the correct video was found. Unlike the model of Maeoki et al. [24] which applied a CNN-based encoder and an LSTM [14] to embed data from each modality and to generate questions and answers, Madusa et al's system, ViReD [23], applied Video2Sum [28] to convert a video into a textual summary which can be used with the initial query to get the generated dialogue with the help of a BART model [19].

In this paper, we focus on a less-studied aspect of video retrieval: dialogue-to-video retrieval where the search query is a user-generated dialogue that contains structured information from each turn of the dialogue. The need for dialogue-to-video retrieval derives from the increasing amount of online conversations on social media, which inspires the development of effective dialogue-to-video retrieval systems for many purposes, especially recommendation systems [1,11,33]. Different from general text-to-video retrieval, dialogue-to-video uses user-generated dialogues as the search query to retrieve videos. The dialogue contains user discussion about a certain video, which provides dramatically different information than a plain-text query. This is because during the interaction between users in the dialogue, a discussion similar to the following could happen "A: *The main character of that movie was involved in a horrible car accident when he was 13.* B: *No, I think you mean another character.*". Such discussion contains subtle information about the video of interest and thus cannot be treated as a plain-text query.

Therefore, to incorporate the conversational information from dialogues, we propose a novel dialogue-to-video retrieval approach. In our proposed model, we sequentially encode each turn of the dialogue to obtain a dialogue-aware query representation with the purpose of retaining the dialogue information. Then we calculate the similarity between this dialogue-aware query representation and individual frames in the video in order to obtain a weighted video representation. Finally, we use the video representation to compute an overall similarity score with the dialogue-aware query. To validate the effectiveness of our approach, we conduct dialogue-to-video experiments on a benchmark dataset AVSD [1]. Experimental results show that our approach achieves significant improvements over previous state-of-the-art models including FiT and ViReD [4,23,24].

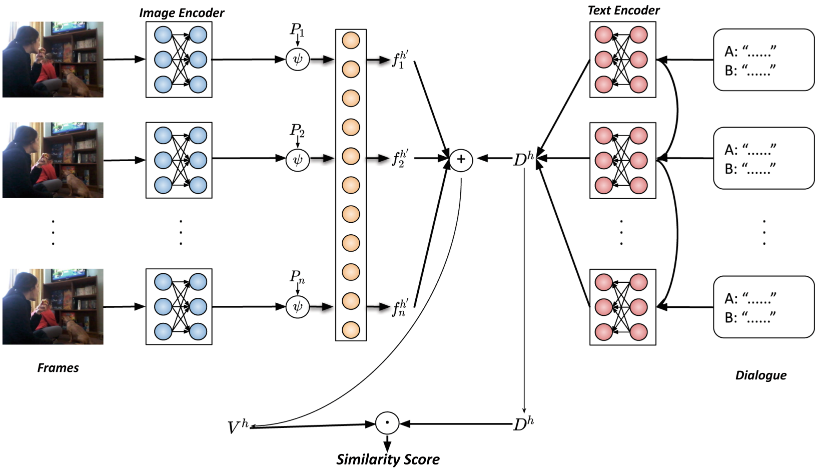

Fig. 1. The architecture of our proposed approach.

2 Methodology

In this section, we describe how our dialogue-to-video retrieval system works. Our retrieval system consists of two major components: 1) a ***temporal-aware video encoder*** responsible for encoding the image frames in video with temporal information. 2) a ***dialogue-query encoder*** responsible for encoding the dialogue query with conversational information. As shown in Fig. 1, our model receives video-query pairs and produces similarity scores. Each video consists of n frames: $V = \{f_1, f_2, ..., f_n\}$ and each dialogue query is composed of m turns of conversation: $D = \{d_1, d_2, ..., d_m\}$.

In the *video encoder*, we encode each frame f_i to its visual representation f_i^h. Then we incorporate temporal information to the corresponding frame representation and feed them into a stacked MULTI-HEAD-ATTENTION module, yielding temporal frame representation $f_i^{h'}$. In the *dialogue-query encoder*, we sequentially encode D by letting $d_i^h = \text{TEXT-ENCODER}(d_{i-1}^h, d_i)$ in order to produce a dialogue-history-aware dialogue representation. We then obtain the final dialogue-query representation by fusing all d_i^h: $D^h = g(d_1^h, ..., d_m^h)$ where g represents our fusion function. After obtaining D^h, we use it to calculate similarities with each frame $f_i^{h'}$, which are then used to obtain a video representation V^h based on the weighted summation of all $f_i^{h'}$. Finally, we obtain the dialogue-to-video similarity score using the dot-product between D^h and V^h.

2.1 Temporal-Aware Video Encoder

Our *temporal-aware video encoder*, which is built on Vision Transformer [7] firstly encodes each frame f_i to its visual representation:

$$f_i^h = \textsc{Image-Encoder}(f_i) \tag{1}$$

Then we inject the positional information of the corresponding frame in the video to the frame representation and feed it to the Multi-Head-Attention module:

$$f_i^{h'} = \textsc{Multi-Head-Attention}([f_1^p, ..., f_n^p]) \tag{2}$$

where f_i^p is the frame representation with positional information $f_i^p = \psi(f_i^h, p_i)$ and p_i is the corresponding positional embedding. Practically, we add *absolute* positional embedding vectors to frame representation as in BERT [6]: $f_i^p = f_i^h + p_i$. Finally, we obtain the temporal-aware video representation $V^{h'} = \{f_1^{h'}, ..., f_n^{h'}\}$.

2.2 Dialogue-Query Encoder

The dialogue-query encoder is responsible for encoding the dialogue-query $D = \{d_1, d_2, ..., d_m\}$:

$$d_i^h = \textsc{Text-Encoder}(d_{i-1}^h, d_i) \tag{3}$$

where Text-Encoder is a Transformer-based encoder model [6,27,29] in our experiments. Then we fuse all d_i^h to obtain a dialogue-level representation D^h for the dialogue-query:

$$D^h = g(d_1^h, ..., d_m^h) \tag{4}$$

2.3 Interaction Between Video and Dialogue-Query

To calculate the similarity score between each V and D, we firstly compute the similarity scores between dialogue-query D^h and each frame $f_i^{h'}$. Then we obtain a weighted summation of all frames $f_i^{h'}$ as the video representation V^h:

$$V^h = \sum_{i=1}^{n} c_i f_i^h \tag{5}$$

$$c_i = \frac{e^{\phi(D^h, f_i^h)}}{\sum_{j=1}^{n} e^{\phi(D^h, f_j^h)}} \tag{6}$$

The final similarity score is obtained by dot-product between D^h and V^h: $s = D^h(V^h)^T$

2.4 Training Objective

We perform in-batch contrastive learning [10,15]. For a batch of N video-dialogue pairs $\{(V_1, D_1), ..., (V_N, D_N)\}$, the dialogue-to-video and video-to-dialogue match loss are:

$$L_{d2v} = -\frac{1}{N} \sum_{i=1}^{N} \frac{e^{D_i^h (V_i^h)^T}}{\sum\limits_{j=1}^{N} e^{D_i^h (V_j^h)^T}} \tag{7}$$

$$L_{v2d} = -\frac{1}{N} \sum_{i=1}^{N} \frac{e^{D_i^h (V_i^h)^T}}{\sum\limits_{j=1}^{N} e^{D_j^h (V_i^h)^T}} \tag{8}$$

The overall loss to be minimized during the training process is $L = (L_{d2v} + L_{v2d})/2$.

3 Experiments

3.1 Dataset

We conduct our experiments on the popular video-dialogue dataset: AVSD [1].[1] In AVSD, each video is associated with a 10-round dialogue discussing the content of the corresponding video. We follow the standard dataset split of AVSD [1,24], 7,985 videos for training, 863 videos for validation and 1,000 videos for testing.

3.2 Training Setup

Our implementation is based on CLIP [27] from Huggingface [31]. CLIP is used to initialize our Image-Encoder and Text-Encoder. For performance and efficiency consideration, we employ ViT-B/16 [27] as our image encoder.[2] We train our system with a learning rate of 1×10^{-5} for 10 epochs, with a batch size of 16. We use a maximum gradient norm of 1. The optimizer we used is AdamW [21], for which the ϵ is set to 1×10^{-8}. We perform early stopping when the performance on validation set degrades. We employ R@K, Median Rank and Mean Rank as evaluation metrics [1].

3.3 Results

We present our experimental results on the test set of AVSD [1] in Table 1, where we also show the results of recent baseline models including: 1) LSTM [24], an LSTM-based interactive video retrieval model; 2) FiT [4], a Transformer-based

[1] https://video-dialog.com.
[2] https://openai.com/blog/clip/.

Table 1. Experimental results on AVSD dataset

	Use dialogue	R@1	R@5	R@10	MedRank	MeanRank
LSTM [24]	✓	4.2	13.5	22.1	N/A	119
FiT [4]	✗	5.6	18.4	27.5	25	95.4
FiT + Dialogue [4]	✓	10.8	28.9	40	18	58.7
ViReD [23]	✓	24.9	49.0	60.8	6.0	30.3
D2V + Script	✗	21.4	45.9	57.5	9.0	39.8
D2V + Summary	✗	23.4	48.5	59.1	6.0	33.5
D2V + Dialogue	✓	**25.6**	**52.1**	**65.1**	**5.0**	**28.9**

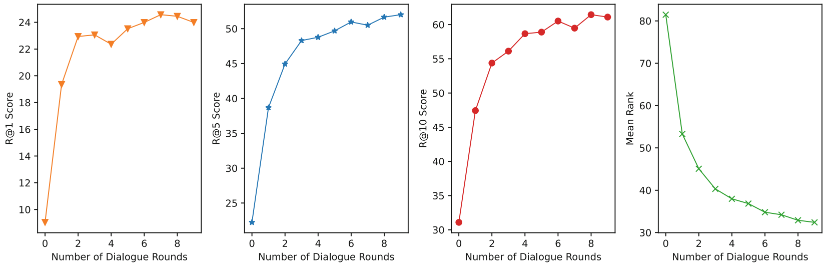

Fig. 2. Effect of dialogue rounds

text-to-video retrieval model using the video summary as the search query; 3) FiT [4] + Dialogue, the FiT model with dialogue in AVSD [1] as the search query[3]; 4) ViReD [23], a video retrieval system based on FiT and CLIP [27] using the dialogue summary as the initial query and model-generated dialogue as an additional query. In Table 1, our model is named D2V (**D**ialogue-**to**-**V**ideo). We also include the results of our system using the video caption (script in AVSD dataset) – D2V+Script – and the dialogue summary (summary in AVSD dataset) as the search query – D2V+Summary.

The results in Table 1 show that our proposed approach, D2V, achieves superior performance compared to previous models. First, D2V+Script with plain-text video caption input outperforms its counterpart FiT by a large margin (15.8 R@1 improvement) and even obtains significant improvements (by 10.6 R@1) over FiT using dialogue as input. That shows the effectiveness of our proposed model architecture. Second, D2V+Dialogue significantly outperforms D2V+Script and D2V+Summary by 3.2 R@1 and 2.2 R@1 respectively, which demonstrates the benefit of incorporating dialogue as a search query. The results in Table 1 show that the dialogue does indeed contain important information about the video content and demonstrates the plausibility of using dialogue as a search query.

[3] We concatenate all the rounds of dialogue as plain text to serve as the search query.

Effect of Dialogue Rounds. We investigate the effect of dialogue rounds on the retrieval performance. The results on the validation set of AVSD are shown in Fig. 2, where we use a varying number of dialogue rounds (from 1 to 10) when retrieving videos. We observe a consistent improvement with an increasing number of dialogue rounds. The results show that with more rounds of dialogue, we can obtain better retrieval performance. The improvement brought by increasing the dialogue rounds is more significant especially in the early stage (when using 1 round of dialogue versus 3 rounds).

4 Conclusion

In this paper, we proposed a novel dialogue-to-video retrieval model which incorporates conversational information from dialogue-based queries. Experimental results on the AVSD benchmark dataset show that our approach with a plain-text query outperforms previous state-of-the-art models. Moreover, our model using dialogue as a search query yields further improvements in retrieval performance, demonstrating the importance of utilising dialogue information.

Acknowledgements. This work was funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning (18/CRT/6183). We thank the reviewers for their helpful comments.

References

1. Alamri, H., et al.: Audio-visual scene-aware dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
2. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: ICCV (2017)
3. Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 (2015)
4. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: a joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021)
5. Cheng, X., Lin, H., Wu, X., Yang, F., Shen, D.: Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. arXiv:2109.04290 (2021)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics, June 2019. https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
7. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2020)

8. Dzabraev, M., Kalashnikov, M., Komkov, S., Petiushko, A.: MDMMT: multidomain multimodal transformer for video retrieval. In: CVPR (2021)
9. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 214–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_13
10. Gao, T., Yao, X., Chen, D.: SimCSE: simple contrastive learning of sentence embeddings. In: Empirical Methods in Natural Language Processing (EMNLP) (2021)
11. He, F., et al.: Improving video retrieval by adaptive margin. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1359–1368 (2021)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Hezel, N., Schall, K., Jung, K., Barthel, K.U.: Efficient search and browsing of large-scale video collections with vibro. In: Þór Jónsson, B., et al. (eds.) MMM 2022. LNCS, vol. 13142, pp. 487–492. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98355-0_43
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
15. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. In: Empirical Methods in Natural Language Processing (EMNLP) (2020)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
17. Le, T.-K., Ninh, V.-T., Tran, M.-K., Healy, G., Gurrin, C., Tran, M.-T.: AVSeeker: an active video retrieval engine at VBS2022. In: Þór Jónsson, B., et al. (eds.) MMM 2022. LNCS, vol. 13142, pp. 537–542. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98355-0_51
18. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
19. 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. Association for Computational Linguistics, July 2020. https://doi.org/10.18653/v1/2020.acl-main.703. https://www.aclweb.org/anthology/2020.acl-main.703
20. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: video retrieval using representations from collaborative experts. arXiv:1907.13487 (2019)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=Bkg6RiCqY7
22. Luo, H., et al.: CLIP4Clip: an empirical study of CLIP for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
23. Madasu, A., Oliva, J., Bertasius, G.: Learning to retrieve videos by asking questions. arXiv preprint arXiv:2205.05739 (2022)
24. Maeoki, S., Uehara, K., Harada, T.: Interactive video retrieval with dialog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 952–953 (2020)
25. Miech, A., Laptev, I., Sivic, J.: Learning a text-video embedding from incomplete and heterogeneous data. arXiv:1804.02516 (2018)
26. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: ICMR (2018)

27. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
28. Song, Y., Chen, S., Jin, Q.: Towards diverse paragraph captioning for untrimmed videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11245–11254 (2021)
29. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, USA, pp. 6000–6010. Curran Associates Inc. (2017). http://dl.acm.org/citation.cfm?id=3295222.3295349
30. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4534–4542 (2015)
31. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, October 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.6. https://aclanthology.org/2020.emnlp-demos.6
32. Yang, X., Zhang, T., Xu, C.: Text2Video: an end-to-end learning framework for expressing text with videos. IEEE Trans. Multimedia **20**(9), 2360–2370 (2018)
33. Zheng, Y., Chen, G., Liu, X., Sun, J.: MMChat: multi-modal chat dataset on social media. In: Proceedings of the 13th Language Resources and Evaluation Conference. European Language Resources Association (2022)