DMT_CUC_01 for AVS

T Wang, Z Liu

School of Information and Communication Engineering, Communication University of China, Beijing 100024, China; tiantian_wong@cuc.edu.cn

This article describes the method of extracting features based on the CLIP pre-training model of the DMT_CUC_01 team participating in the trecvid avs task in 2021.

Our method relies on the large-scale pre-training model Contrastive Language-Image Pre-training(CLIP)[1] to extract the features of the video frame, and adopts a multi-level coding method to capture the time-related features of multiple time scales; for the query sentence, we use the same kind of CLIP model to encode. Then the two modality encoding are mapped to the same public space[2], finally the training of the model is completed by calculating the marginal ranking loss by cosine similarity[3]. In the training stage, we use two public datasets of Microsoft, MSVD and MSR-VTT, predict results return and submit for the 2021 query given in TRECVid AVS competition. Here is the framework of our model in Fig 1.

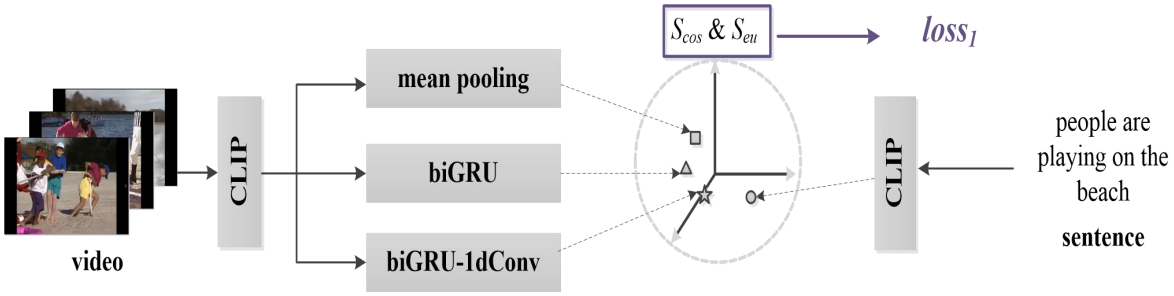

Fig 1. framework of model

We first use the method of extract frames at equal intervals, extracting one frame of the video every 0.5s, and sending all the video frames to the CLIP model to extract frame-level features, then we refer to the dual encoding method to use a three-level encoding idea respectively, to the video frame-level features perform three parallel operations of mean pooling, biGRU and biGRU-1d conv, while the text encoding is obtained through CLIP. As a result we use three video encoding calculate cosine similarity for margin rank loss with one text encoding respectively and take the average as the final score to trains our model.

We use two public datasets MSVD and MSRVTT in video-text retrieval field by Microsoft. The performance results in testing data of these two data sets are shown in the following Table1. Our submission results are used by checkpoints obtained from MSRVTT training.

Table 1. The result of MSVD and MSR-VTT

Evaluation	R@1	R@5	R@10	Medr	Meanr	Recall sum	mAP
MSVD	37.4	68.0	79.3	2.0	14.9	184.6	0.512
MSR-VTT	12.7	34.9	48.4	11.0	49.1	96.0	0.238

After that, we downloaded the official test video by extracting at equal intervals to complete the task test and submit the results.

We submitted a run result: Run ID: M_M_C_D_DMT_CUC_01.21_1, Class: M - Manually-assisted, Training type: D (Any non-IACC training data), Task: M, Novelty: C, Priority: 1. Run result is in the Fig 2.

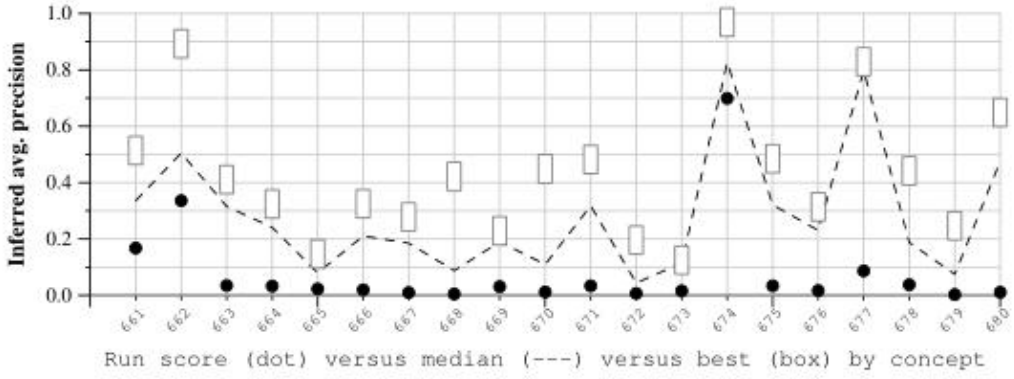

Fig 2. The progress runs submitted in 2021 against 10 selected topics

Due to resource constraints, we did not download all the AVS test video datasets, but used proportionally downloaded partial data sets to return our results. Therefore, the performance of our experimental results is not good. At the same time, our model is not an end-to-end model, so there is a certain calculation error in the test return time.

In the future, we will continue to carry out research on tasks related to avs to achieve end-to-end network results for matching from multiple scales. On the basis of completing the overall matching of video content and text content, we can achieve the entity or action in the video screen and match with the text.

Reference

[1] Radford A , Kim J W , Hallacy C , et al. Learning Transferable Visual Models From Natural Language Supervision[J]. 2021.

[2] Dong J.; Li X.; Xu C.; Ji S.; He Y.; Yang G.; Wang X. Dual encoding for zeroexample video retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15-20 June 2019; pp. 9346–9355.

[3] Faghri F.; Fleet D.J.; Kiros J.R.; Fidler S. VSE++: Improved Visual-Semantic Embeddings. In Proceedings of the 29th British Machine Vision Conference (BMVC'18), Northumbria University, UK, 3-6 September 2018.