Open-book Video Captioning with Retrieve-Copy-Generate Network
(Supplementary Material)

1. Additional Ablation Analysis

We summarize the results of training the model with different numbers of retrieved sentences on MSR-VTT in Tab.1. We have obtained a basically consistent conclusion in the second question of Sec.4.2 in the main paper. In the training phase, a moderate number (3 in MSR-VTT) of sentences are conducive to the generation; too much ones will introduce some noises instead. In addition, by comparing the results between using the fixed and jointly trained retrievers, we find that the advantage of the latter is not obvious. The reason may be that the video semantics in MSR-VTT are relatively scattered, and only a few sentences related to the video can be found, which leads to the small gain of joint training.

As shown in Tab.2, we supplement the results of our proposed model RCG using more features in two datasets. For the Video-Text Retrieval task, compared with the state-of-the-art method HGR [1], our retrieval model can achieve comparable results, which proves the effectiveness of our proposed Bi-encoders architecture based retriever. From the results of the Video-Text Retrieval and the Video Captioning, we can see that our RCG is sensitive to the retrieval model. Therefore, how to improve the performance of the retrieval model, and how to adapt it to our two-stage caption model are the next steps in the future work.

2. Additional Qualitative Examples

As mentioned in the last question in Sec.4.2 of the main paper, our proposed RCG can improve the performance of zero-shot video captioning compared with the baseline model, but the performance is not satisfactory. We make some examples to illustrate what caused the success and failure of the model.

Table 1. Performance of video-text retrieval and video captioning tasks using fixed retrievers trained by more different features.

Table 2. Performance of training the model with different numbers of retrieved sentences on MSR-VTT. The model is tested via top-15 sentences. Fixed denotes whether the retriever is fixed or jointly trained.

References

[1] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In CVPR 2020, pages 10635–10644, 2020.
Figure 1. Examples of Zero-shot Video Captioning. M and V denote the sources of sentences are MSR-VTT and VATEX. The whole model is pre-trained on VATEX training set, then the model is tested on MSR-VTT test set as zero-shot videos. Since the model has never seen these videos before, the performance is very poor, so the retriever pre-trained on two datasets is plugged to replace the original retriever to achieve the retrieval of content-related sentences. This realizes the transformation of knowledge and the expansion of model memory.
Figure 2. More examples of proposed RCG method on VATEX dataset.
| Video Clips | Video Snaps of Retrieved Sentences |
|-------------|----------------------------------|
| **Ground-Truth** | 1-a man is teaching  
2-a man is teaching math by using a white board  
3-a man is writing  |
| **Retrieved Sentences** | 1-a man stands in front of a classroom and writes a lesson on the whiteboard  
2-a man giving a physics lecture with an equation on a whiteboard  
...  
15-a man is standing in front of a whiteboard with math problems on it as he explains how to do them  |
| **RCG** | a man is writing on a dry erase board and explaining what he is doing |

(a) **Good case**

| Video Clips | Video Snaps of Retrieved Sentences |
|-------------|----------------------------------|
| **Ground-Truth** | 1-a man is folding a piece of yellow paper  
2-a person is folding paper  
3-a plane is made out of yellow paper  |
| **Retrieved Sentences** | 1-how to fold a yellow paper  
2-a person folding a yellow paper  
...  
15-a person is folding a piece of yellow paper  |
| **RCG** | a woman is folding a piece of yellow paper into a triangle |

(b) **Bad case caused by the bad generator**

| Video Clips | Video Snaps of Retrieved Sentences |
|-------------|----------------------------------|
| **Ground-Truth** | 1-a woman washes her hands dries them with a paper towel and then wipes off a countertop  
2-a woman wearing scrubs washes her hands with coffins around her  
3-a woman is dipping her hands in water and then drying them off with a paper towel  |
| **Retrieved Sentences** | 1-man in grey shirt is preparing something in the kitchen  
2-a man in a white shirt prepares things in a kitchen  
...  
15-man in white uniform is making something in a kitchen  |
| **RCG** | a woman is standing in front of a kitchen counter and she is washing a piece of clothing |

(c) **Bad case caused by the bad retriever**

Figure 3. Examples of proposed RCG method on MSR-VTT dataset.