Supplementary Material for "Unifying Event Detection and Captioning as Sequence Generation via Pre-Training"

Qi Zhang, Yuqing Song, and Qin Jin*

School of Information, Renmin University of China
{zhangqi1996, syuqing, qjin}@ruc.edu.cn

1 DVC Results on Charades-STA Dataset

To verify the generalization ability of our proposed model, we carry out additional experiments on the Charades-STA [1] dataset. It contains about 10K videos of daily indoor activities, with 12,408 video-sentence pairs for training and 3,720 pairs for testing. We adopt C3D as the visual feature extractor.

As shown in Table 1, when pre-training for the event detection, Ours-mutual model achieves better event detection results than Ours-single model, which improves the average recall from 41.85% to 42.50% and average precision from 24.29% to 24.76%, indicating that the event detection task can benefit from the event captioning task. Meanwhile, we also report the results of PDVC [2] on the Charades-STA dataset with the official released code[1]. Experimental results show that our model outperforms PDVC significantly on the recall and the final F1 score for the event detection.

Table 1. Event detection results on the Charades-STA testing set. Ours-single: pre-training model with MEFM task, Ours-mutual: pre-training model with MLM, MVM and MEFM tasks. * denotes the results are run by ourselves.

Method	Recall					Precision					F1
	0.3	0.5	0.7	0.9	avg	0.3	0.5	0.7	0.9	avg	
PDVC*[2]	47.76	34.90	20.47	7.01	27.54	54.60	39.20	19.79	5.99	29.89	28.66
Ours-single	86.34	55.09	22.50	3.47	41.85	50.65	31.36	12.88	2.27	24.29	30.74
Ours-mutual	87.62	55.49	22.43	4.44	42.50	50.90	31.95	13.72	2.45	24.76	31.29

Figure 1 shows the event captioning results of Ours-single and Ours-mutual models. Ours-mutual model is shown to outperform Ours-single model by a large margin on all the metrics, which improves the BLEU@4 score from 1.06 to 1.79, METEOR score from 14.67 to 15.26 and CIDEr score from 28.79 to 32.89. It

* Corresponding author

[1] https://github.com/ttengwang/PDVC

Table 2. Dense video captioning results on the Charades-STA testing set. * denotes the results are run by ourselves.

Method	with GT event proposals			with detected event proposals
	B@4	M	C	SODA$_{mr}$
PDVC*[2]	1.56	14.68	34.09	5.37
Ours-mutual	1.79	15.26	32.89	5.56

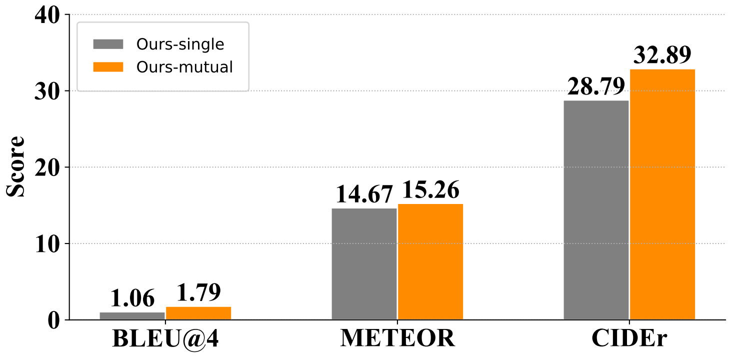
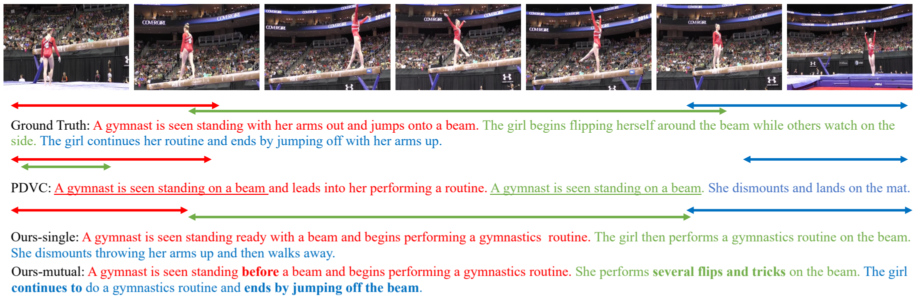


Fig. 1. Event captioning results on the Charades-STA testing set. Ours-single: pre-training model with MLM and MVM tasks, Ours-mutual: pre-training model with MLM, MVM and MEFM tasks.

demonstrates that the event captioning task can also benefit from the event detection task based on our proposed framework. The dense video captioning results of our final model (Ours-mutual) and the compared PDVC method are shown in Table 2. Our model outperforms the PDVC especially when captioning with the automatically detected event proposals.

2 Qualitative Results

Figure 2 shows a dense captioning example by our models and another state-of-the-art model PDVC. The event proposals generated by our model are more accurate with less redundancy. With more diversified events, our models generate more coherent and diverse descriptions covering the full video content, while the PDVC model repeats the event of "standing on a beam". Furthermore, enhancing the model with inter-task associations between event detection and captioning (Ours-mutual) helps generate more accurate descriptions than Ours-single (shown in bold).



Ground Truth: A gymnast is seen standing with her arms out and jumps onto a beam. The girl begins flipping herself around the beam while others watch on the side. The girl continues her routine and ends by jumping off with her arms up.

PDVC: A gymnast is seen standing on a beam and leads into her performing a routine. A gymnast is seen standing on a beam. She dismounts and lands on the mat.

Ours-single: A gymnast is seen standing ready with a beam and begins performing a gymnastics routine. The girl then performs a gymnastics routine on the beam. She dismounts throwing her arms up and then walks away.

Ours-mutual: A gymnast is seen standing **before** a beam and begins performing a gymnastics routine. She performs **several flips and tricks** on the beam. The girl **continues to** do a gymnastics routine and **ends by jumping off the beam**.

Fig. 2. Qualitative results on the ActivityNet Captions dataset. The underlined words indicate redundant captions, and the words in bold are more accurate descriptions generated by Ours-mutual model. Ours-single: pre-training model with only MLM task, Ours-mutual: pre-training model with both MLM and MEFM tasks.

References

1. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. pp. 5267–5275 (2017)
2. Wang, T., Zhang, R., Lu, Z., Zheng, F., Cheng, R., Luo, P.: End-to-end dense video captioning with parallel decoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6847–6857 (2021)