Supplementary Materials for:
Motion Guided Region Message Passing for Video Captioning

Shaoxiang Chen and Yu-Gang Jiang[*]
Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
Shanghai Collaborative Innovation Center on Intelligent Visual Computing
{sxchen13, ygj}@fudan.edu.cn

More Ablation Results

#	RRA	MGCMP	ATGD	B	M	R	C
0	✗	✗	✗	37.4	27.0	58.8	42.3
1	✓	✗	✗	37.5	26.9	58.9	43.1
2	✓	✓	✗	40.9	28.4	61.2	49.9
3	✓	✓	✓	41.7	28.9	62.1	51.4
4	✓	✗	✓	38.4	27.4	59.8	44.5
5	✗	✗	✓	38.2	27.5	59.7	43.6
obj	✗	✗	✗	34.0	25.2	56.4	36.7

Table 1. Results of more ablation studies on MSR-VTT. We supplement more results to the Table 5 in the main paper (#4, #5, and #obj).

As shown in Table 1, we supplement more results to the Table 5 in the main paper:

#4 and #5: We apply ATGD to more model variants. Comparing #4 and #1, #5 and #0, we can conclude that ATGD is a generally effective caption decoder, and can be applied on top of different video feature encoders to obtain better captioning performance than two-layer LSTMs (which is used in most methods).

#obj: We use detected object features in the baseline model, but the features are spatially mean-pooled as in #1. As can be seen, the performances of #obj is worse than #0 (grid) and #1 (region). We conjecture the reason is that the object features can ignore some background regions because of extracting information only from object bounding box areas, while grid features (#0) contains all the information of a feature map and region features (#1) can capture the background information by the soft attention maps. This information loss can be compensated by the MGCMP, that is why the performances of our model with object and region features are close in the Table 4 of the main paper.

More Implementation Details

We provide more implementation details of the MGCMP here. As in [1], we apply a MLP after the message calculation outputs and the message updating outputs. The MLP is applied region-wise, and consists of two fully connected layers with `[input, output]` dimensions configured as `[IN1, OUT1]` and `[IN2, OUT2]`, where `OUT1=IN2=2×IN1` widens the feature representation, `IN1=OUT2` keeps the output dimension as the same as inputs and we can introduce a residual connection from the inputs to the MLP outputs. Each MLP layer has ReLU activation, and the final outputs are layer-normalized. Also, as indicated in [1], using multiple heads in the attention block is helpful, so we split the attentions in MGCMP into 4 heads for better performance. Following previous work, we use beam search with a beam size of 5 during inference. Our method is implemented with Pytorch 1.6.0 and training is done on two RTX 2080 Ti GPUs.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 1

[*] Corresponding author.