Supplementary Material:
Syntax-Aware Action Targeting for Video Captioning

Qi Zheng  Chaoyue Wang  Dacheng Tao
UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering,
The University of Sydney, Darlington, NSW 2008, Australia
{qi.zheng, chaoyue.wang, dacheng.tao}@sydney.edu.au

In this supplementary material, Section 1 summarizes the details of our SAAT model. Section 2 discusses the influence of hyper-parameter $\lambda$ in our loss function. Section 3 presents more qualitative examples and some failure cases of our model to provide more insight on the proposed model.

1. Implementation Details

For self-attended scene representation, we exploit one head with one encoding layer, where the size of hidden units is 512. For the captioner, we use 1-layer LSTMs with 512 hidden units. The embedding size of one-hot encoded-words is 512, and the embedding is randomly initialized. Input features from C2D, C3D and the detector are projected to 512-dim. The size of hidden units used in the syntax-attention layer is 100. All the projection layers are followed by a ReLU activation layer with a dropout probability of 0.5. Similar to [2], the training stops when no higher CIDEr score is achieved in the following 50 successive epochs on the validation set. For comparison, the weight parameter $\lambda$ is set to 1.0 and 2.0 on MSVD dataset and MSR-VTT dataset respectively in our paper, and more analysis about $\lambda$ is available in the following section. The proposed method is implemented under PyTorch [1] framework with Python3.

2. Sensitive Analysis

In our model, the weight parameter $\lambda$ is used to balance the loss from the prediction of syntax components and that from the generation of captions. Here we provide the performance of different settings on the test set of MSR-VTT dataset in Fig. 1. It can be seen that without the penalty on the prediction of syntax components, i.e., $\lambda = 0$, the SAAT module performs poorly, especially on BLEU@4 and CIDEr scores. The highest CIDEr score of 49.2 and BLEU@4 score of 40.6 are achieved by $\lambda = 0.5$, the highest ROUGE score of 60.9 is achieved by $\lambda = 2.0$, and the highest METEOR score of 28.3 is achieved by $\lambda = 2.5$.

When $\lambda$ keeps increasing after that, the performance decreases a bit and then stays the same. It indicates that action-guide from syntax components benefits the captioning process, and our model is robust to extreme $\lambda$ settings.

3. Additional Results

In this section, we present more qualitative comparison of the proposed SAAT model and the Baseline. Fig. 2 and Fig. 3 show additional results that our model describes the action in video clips more accurately than the Baseline. Multiple cases demonstrate that the proposed model can describe the action in video clips more accurately than the Baseline, e.g., running on the field vs. playing football, laying in bed vs. kissing each other, surfing in the water vs. playing a video game, pouring oil vs. cooking. Besides, the proposed model also alleviates the case where action is missing in generated descriptions, e.g., the last video clip in Fig. 2.

Fig. 4 provides two failure cases where the proposed model performs worse than the Baseline, i.e., fails to exactly
Figure 2. Qualitative comparison between the Baseline and our SAAT model by examples from the validation set and the test set of MSR-VTT and MSVD datasets. Three frames are shown for each video clip. 3--5 human annotated descriptions are listed for illustration. Text in blue highlights the subject in a sentence. Words in green and red show the predicted action by Baseline and by SAAT, respectively.
Figure 3. Qualitative comparison between the Baseline and our SAAT model by examples from the validation set and the test set of MSR-VTT and MSVD datasets. Three frames are shown for each video clip. 3-5 human annotated descriptions are listed for illustration. Text in blue highlights the subject in a sentence. Words in green and red show the predicted action by Baseline and by SAAT, respectively. 

GT: {'a woman is surfing', 'a woman is skating in a sea water', 'a girl is surfing and a guy is riding a bike on mountains', 'a girl rides the wave', 'a woman surfing in the ocean'}
Baseline: 'a person is playing a video game'
SAAT: 'a man is surfing in the water'

GT: {'a bunch of men are eating steamed buns', 'a group of people are eating food at a table', 'a group of people are eating chinese food', 'a man is eating and describing soup dumplings'}
Baseline: 'a man is cooking food'
SAAT: 'a man and a woman are eating food'

GT: {'a car drives around at night', 'a man drives a car', 'a man drives a mazda on a track', 'he is driving a car', 'a red car is driving on a road at dark', 'a man zooms around in a new car commercial'}
Baseline: 'a man is talking about a car'
SAAT: 'a man is driving a car'

GT: {'a man is eating a lot of food', 'a man is eating some snacks and noodles and ice creams', 'a man is eating many different foods', 'a man is eating food', 'a man is eating gluttonously'}
Baseline: 'a man is singing'
SAAT: 'a man is eating a glass of food'

GT: {'a woman is placing makeup on her face', 'a woman is putting on makeup', 'the woman is applying makeup', 'a woman is applying makeup to her face', 'a woman is placing makeup on her face'}
Baseline: 'a woman is talking'
SAAT: 'a woman is putting on makeup'

GT: {'a man is mixing pizza dough', 'a man is mixing flour in a bowl', 'a man is stirring dough ingredients in a bowl', 'someone is mixing meat', 'a person is mixing the flour'}
Baseline: 'a person is cooking'
SAAT: 'a man is mixing rice in a bowl'
describe the content in video clips. We analyze that the first case is caused by the ambiguity of linguistic words, i.e., explaining and talking, where the action can be described by different words. The second case shows the situation where more than one predominant actions exist in one video clip, where further extension to the proposed model is expected to deal with such cases.

References
