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 λ 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 λ is set to 1.0 and 2.0 on MSVD dataset and MSR-VTT dataset respectively in our paper, and more analysis about λ 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 λ 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 $\lambda = 2.5$.

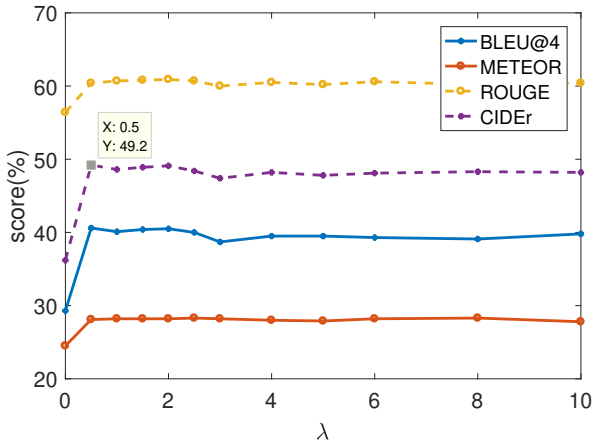


Figure 1. Sensitivity of λ in the SAAT model w.r.t BLEU@4, METEOR, ROUGE and CIDEr scores on the test set of MSR-VTT dataset.

When λ 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 λ 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

GT: {'a person is playing a type of puzzle game', 'a man is Using a Mechanical arm to play a game on a screen', 'person is controlling the machine on the computer monitor', 'a guy tests out a new bionic arm using a computer'}
Baseline: 'a man is **talking** about something'
SAAT: 'a person is **playing** a video game'

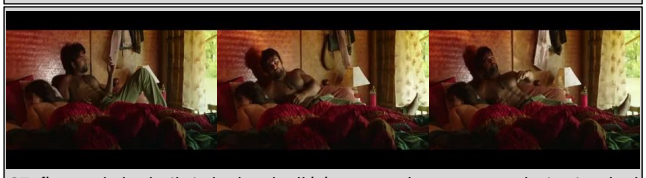
GT: {'a man holding another man s arm', 'people are fighting in travel', 'a clip from movie shows people holding on', 'people on the plane are about to fly out', 'a clip from movie shows people holding on'}
Baseline: 'a man and a woman are **having** a conversation'
SAAT: 'a man is **fighting** with a woman'

GT: {'a kid is playing in the sand', 'a child digging in the sand with a toy shovel', 'a kid is digging a beach sand and put in bucket', 'a child is demonstrating the word digging by digging in the sand'}
Baseline: 'a man is **talking** to a man'
SAAT: 'a man is **playing** with a toy'

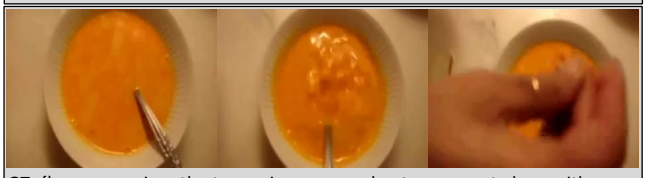
GT: {'a girl feeds and touches a horse outside', 'a girl is feeding a horse', 'a little girl is petting a horse', 'kids are playing with horse in grass', 'two girls playing with horse'}
Baseline: 'a woman is **talking** about a horse'
SAAT: 'a person is **playing** with a horse'

GT: {'video game character gets into a car', 'a man is running across a dark street and entering a car and driving it into a garage in the video game clip', 'a person is riding in a car in a video game'}
Baseline: 'a person is **playing** a video game'
SAAT: 'a man is **driving** a car in a video game'

GT: {'a football player runs the ball down the field past the other players', 'a football player runs through and past his competitors', 'linemen running for a touchdown in the nfl'}
Baseline: 'people are **playing** football'
SAAT: 'a football player is **running** on the field'



GT: {'a couple lay lazily in bed and talk', 'a man and a woman are laying in a bed and talking', 'a man and woman are laying in bed together talking', 'a man and woman are lying in bed together'}
Baseline: 'a man and a woman are **kissing** each other'
SAAT: 'a man and a woman are **laying** in bed'



GT: {'a person mixes the turmeric power and put some root along with some oil', 'a woman is adding some ingredients and mixed them together in a big bowl', 'a woman mixing some spices into a bowl of soup'}
Baseline: 'a woman is **showing** how to make a dish'
SAAT: 'a person is **mixing** a dish in the kitchen'

GT: {'a child jumps outside', 'a child is performing a dance routine in a yard', 'a girl is leaping across her front lawn', 'a girl jumps in her front yard', 'a woman is jumping in her front yard'}
Baseline: 'a girl is **playing** with a ball'
SAAT: 'a girl is **jumping** on a trampoline'

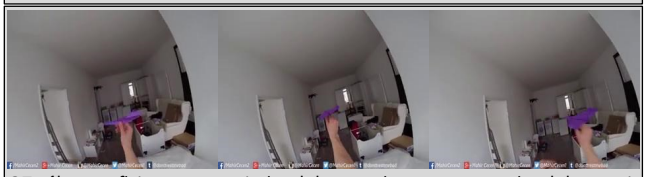
GT: {'there are many soldiers who fighting on the battle ground', 'in the war the soldiers are fighting against each other to win the war', 'warriors killing their enemies and fight against them'}
Baseline: 'a scene from a movie scene from a movie'
SAAT: 'people are **fighting** in a plane'

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.

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 crowd of people are gathered around and someone is hitting a kid with a sack', 'a group of young people fight and a child is hurt', 'several people fight in a small village', 'small children are fighting each other'}
Baseline: 'a man and a woman are **kissing** each other'
SAAT: 'a man and a woman are **fighting** with each other'

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'

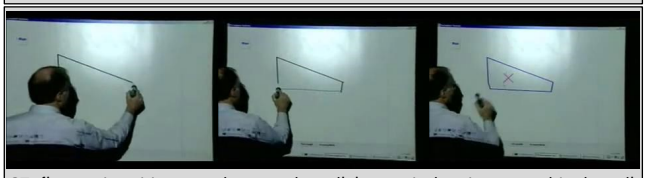


GT: {'a man flying a paper airplane', 'a man throws a paper plane', 'a man is flying a paper airplane', 'a person is throwing a plane', 'a person threw a paper airplane'}
Baseline: 'a person is **playing** a video game'
SAAT: 'a man is **throwing** a paper airplane'

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: {'someone pours liquid from a plastic container into a ziploc bag containing meat pieces', 'a person is pouring sauce into a bag of meat', 'a man is pouring marinade from a bowl into a bag'}
Baseline: 'a man is **cooking**'
SAAT: 'a man is **pouring** oil into a plastic container'

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'

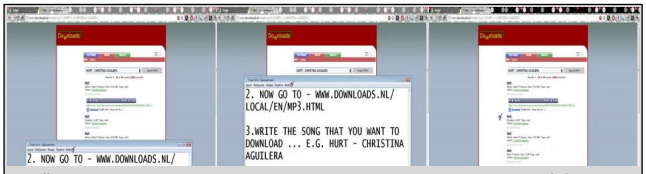


GT: {'a man is writing on a dry erase board', 'a man is drawing on a white board', 'a man is writing on the board', 'the teach drew a geometrical shape on the board', 'a man is writing on a school board'}
Baseline: 'a woman is **talking** on the phone'
SAAT: 'a woman is **writing** on a computer'

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 meal', 'a person is mixing the flour'}
Baseline: 'a person is **cooking**'
SAAT: 'a man is **mixing** rice in a bowl'

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 clip showing a computer screen which contains instructions', 'a person explaining how to download a song', 'a person is showing what to do on their computer', 'a video showing how to download songs from youtube'}
Baseline: 'a person is **explaining** something'
SAAT: 'a man is **talking** about a computer program'

GT: {'a girl chatting with someone on phone', 'a girl in bedroom knocking on the wall and texting', 'a girl is texting on her cell phone while laying in bed', 'a girl knocks on the wall then continues texting'}
Baseline: 'a girl is **texting** on her phone'
SAAT: 'a girl is **laying** in bed and knocking on the wall'

Figure 4. Failure examples of the proposed model from 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.

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

[1] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NeurIPS Workshop Autodiff*, pages 1–4.

[2] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. Controllable video captioning with pos sequence guidance based on gated fusion network. In *ICCV*, 2019.