Video Captioning via Hierarchical Reinforcement Learning: Supplementary Material

Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, William Yang Wang

University of California, Santa Barbara
{xwang,wenhuchen,jiawei_wu,yfwang,william}@cs.ucsb.edu

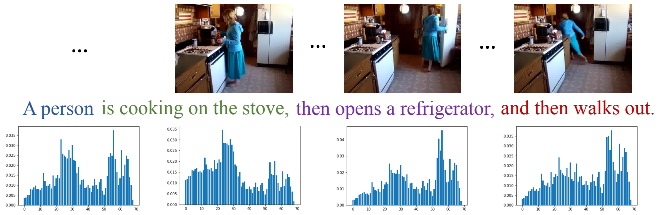

Figure 1: A visualization demo of the attentions. Different text segments were in different colors, and the associated attentions were provided below the corresponding segments. We also showed the keyframe in the top row, which was selected from the most noticeable area for each segment.

A. Attention Visualization

Fig. 1 demonstrated a visualization example where the associated attentions of the learned text segments over video frames were plotted. Clearly, when generating different text segments, the HRL model attended to different temporal frames. For example, when the model was producing the segment *is cooking on the stove*, the first halve of the video, which contained the action *cooking*, played a more important role with larger attention values.

B. Qualitative Examples on MSR-VTT

In the main paper, we showed some generated results on Charades Captions dataset. Here we demonstrated more qualitative examples on MST-VTT dataset in Figure 2.

Particularly, Example (a) and (b) revealed that our HRL method was able to capture more details of the video content and generate more fine-grained descriptions. For example, our HRL model provided both the event (*a group of people are dancing*) and the scene (*on the beach*) in Example (a) while the other baseline methods failed to depict where the event is happening. Example (c) (d) (e) and (f) further illustrated the correctness and accuracy of our HRL results. For instance, in Example (c), only the result of our HRL method described the video correctly. The ground truth caption was *a group of men are racing around a track* and our result was *a group of people are running on a track*. While both the XE-baseline and RL-baseline captioned by mistake the video with *a group of people are playing a game* and *a man is playing a football game* respectively. Apparently, compared with the results of the baseline methods, our results were more accurate and descriptive in general.

C. Network Architecture

In this section, we illustrate the exact architecture used for the experiments (see Figure 2 in the main paper).

Encoders For both datasets, we sampled each video at $3fps$ and used ResNet-152 [4] (pretrained CNN model on ImageNet) to extract frame features without fine-tuning. Then the 2048-dim frame features were projected to 512-dim. The low-level encoder was a Bi-LSTM with hidden size 512, and the high-level encoder was an LSTM with hidden size 256.

Worker The worker network consisted of a worker LSTM with hidden size 1024, an attention module similar to the one proposed by Bahdanau *et al.* [1], a word embedding of size 512, and a projection module (Linear → Tanh → Linear → SoftMax) that produced the probabilities over all tokens in the vocabulary.

Manager The manager network was composed of a manager LSTM with hidden size 256, an attention module, and a linear layer that projected the output of the LSTM into latent goal space.

Internal Critic The internal critic was also an RNN network, which contained a GRU [3], a built-in word embedding, a linear layer, and a Sigmoid function. The hidden size of the GPU and the word embedding size were both 128 for MSR-VTT and 64 for Charades Captions.

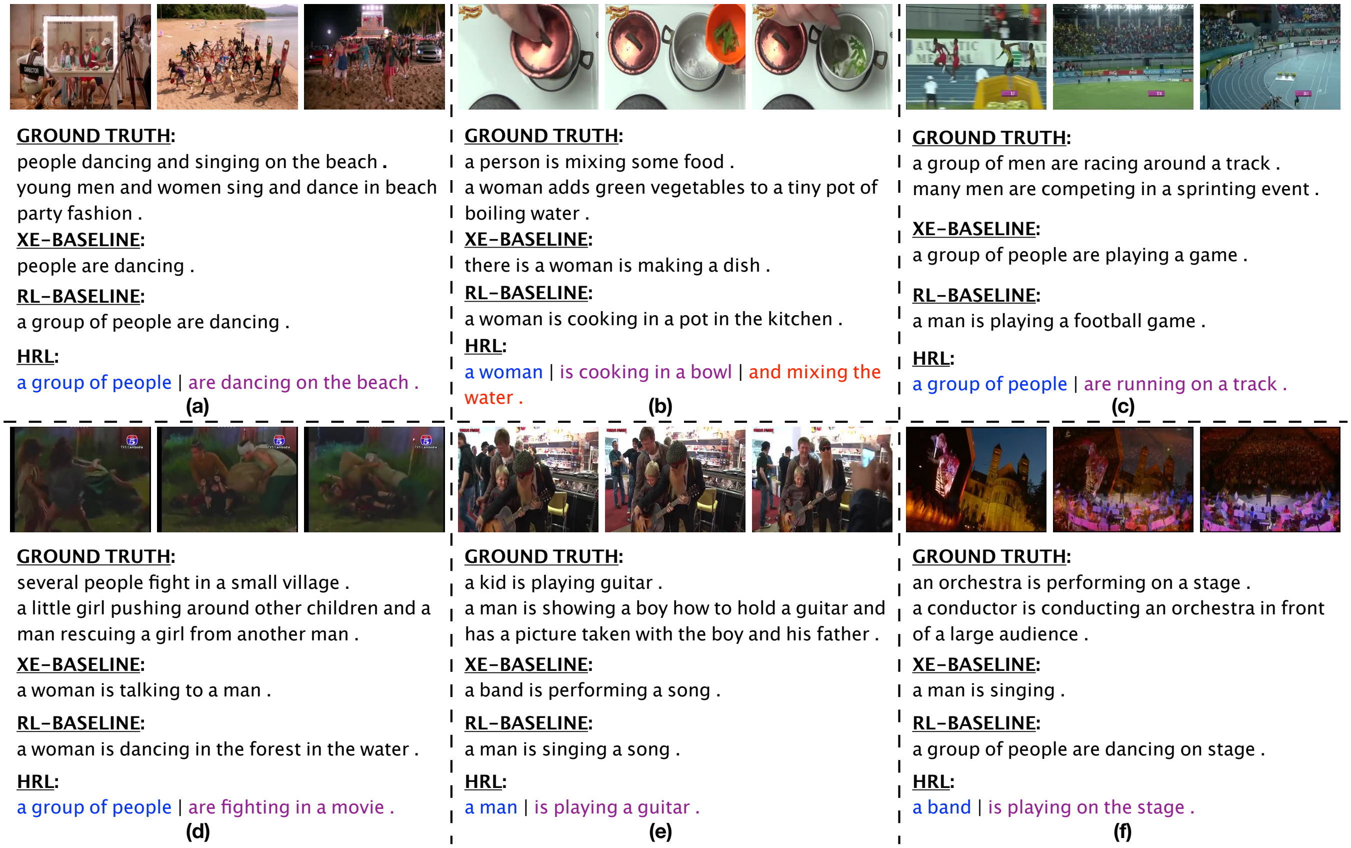
GROUND TRUTH:
people dancing and singing on the beach .
young men and women sing and dance in beach party fashion .
XE-BASELINE:
people are dancing .
RL-BASELINE:
a group of people are dancing .
HRL:
a group of people | are dancing on the beach .
(a)
GROUND TRUTH:
a person is mixing some food .
a woman adds green vegetables to a tiny pot of boiling water .
XE-BASELINE:
there is a woman is making a dish .
RL-BASELINE:
a woman is cooking in a pot in the kitchen .
HRL:
a woman | is cooking in a bowl | and mixing the water .
(b)
GROUND TRUTH:
a group of men are racing around a track .
many men are competing in a sprinting event .
XE-BASELINE:
a group of people are playing a game .
RL-BASELINE:
a man is playing a football game .
HRL:
a group of people | are running on a track .
(c)
GROUND TRUTH:
several people fight in a small village .
a little girl pushing around other children and a man rescuing a girl from another man .
XE-BASELINE:
a woman is talking to a man .
RL-BASELINE:
a woman is dancing in the forest in the water .
HRL:
a group of people | are fighting in a movie .
(d)
GROUND TRUTH:
a kid is playing guitar .
a man is showing a boy how to hold a guitar and has a picture taken with the boy and his father .
XE-BASELINE:
a band is performing a song .
RL-BASELINE:
a man is singing a song .
HRL:
a man | is playing a guitar .
(e)
GROUND TRUTH:
an orchestra is performing on a stage .
a conductor is conducting an orchestra in front of a large audience .
XE-BASELINE:
a man is singing .
RL-BASELINE:
a group of people are dancing on stage .
HRL:
a band | is playing on the stage .
(f)

Figure 2: Qualitative comparison with the baseline methods on MSR-VTT dataset. For each video example, we listed two ground truth captions, the generated result by XE-baseline (cross entropy), the result by RL-baseline (policy gradient), and the result by our HRL method (hierarchical reinforcement learning). In our HRL results, different segments were in different colors and separated with "|".

D. Training Details

All the hyperparameters were tuned on the validation set, including the dimension sizes in Sec. C. Moreover, we adopted Dropout [5] with a value 0.5 for regularization. All the gradients were clipped into the range [-10, 10]. We initialized all the parameters with a uniform distribution in the range [-0.1, 0.1]. For MSR-VTT dataset, we used a fixed step size of 50 for the encoder LSTMs and a maximum length of 30 for the captions. For Charades Captions dataset, they were set to 150 and 60 respectively.

To train the cross-entropy (XE) models, Adadelta optimizer [6] was used with batch size 64. The learning rate was initially set as 1 and then reduced by a factor 0.5 when the current CIDEr score didn't surpass the previous best for 4 epochs. Schedule sampling [2] was employed to train the XE models. When training the RL and HRL models, we used the pretrained XE models to warm start and then continued training them with a learning rate 0.1. The discounted factors of the Manager and the Worker were both 0.95. At test time, we used beam search of size 5.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1

[2] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015. 2

[3] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 1

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014. 2

[6] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 2