Video Captioning via Hierarchical Reinforcement Learning

CVPR2018

Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, William Yang Wang University of California, Santa Barbara

Reporter:ziyang

Motivation

- *it is still very challenging to caption a video containing multiple fine-grained actions with a detailed description.*

Caption #1: A woman offers her dog some food.

Caption #2: A woman is eating and sharing food with her dog.

Caption #3: A woman is sharing a snack with a dog.

Caption: A person sits on a bed and puts a laptop into a bag. The person stands up, puts the bag on one shoulder, and walks out of the room.

Contributions

- first to consider HRL in image caption.
- state-of-the-art.

Model

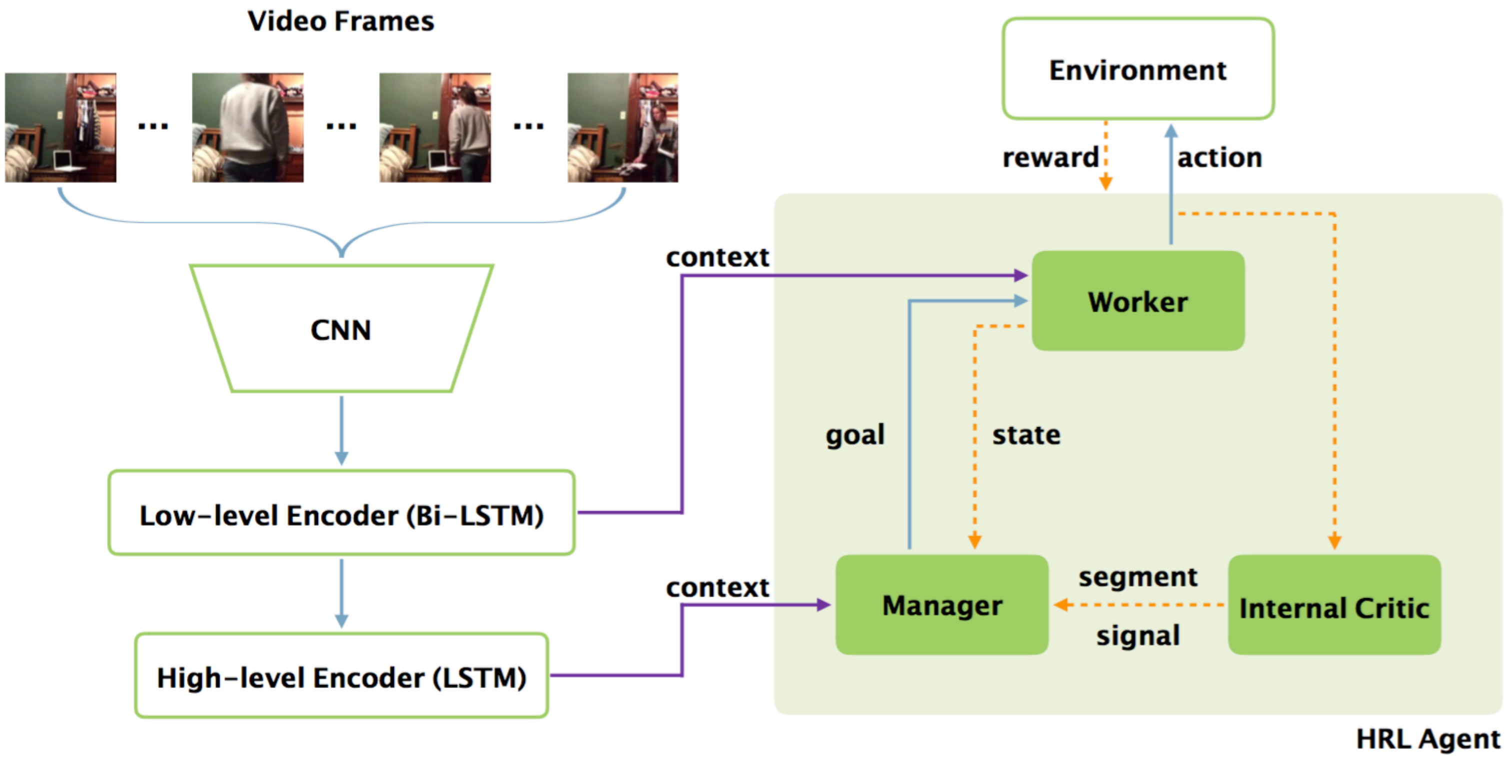

Figure 2: Overview of the HRL framework for video captioning. Please see Sec. 3.1 for explanation.

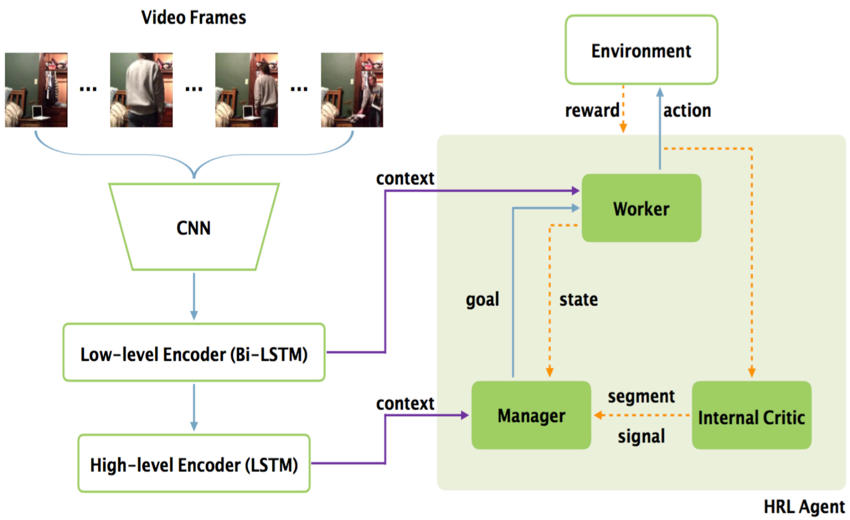

Figure 2: Overview of the HRL framework for video captioning. Please see Sec. 3.1 for explanation.

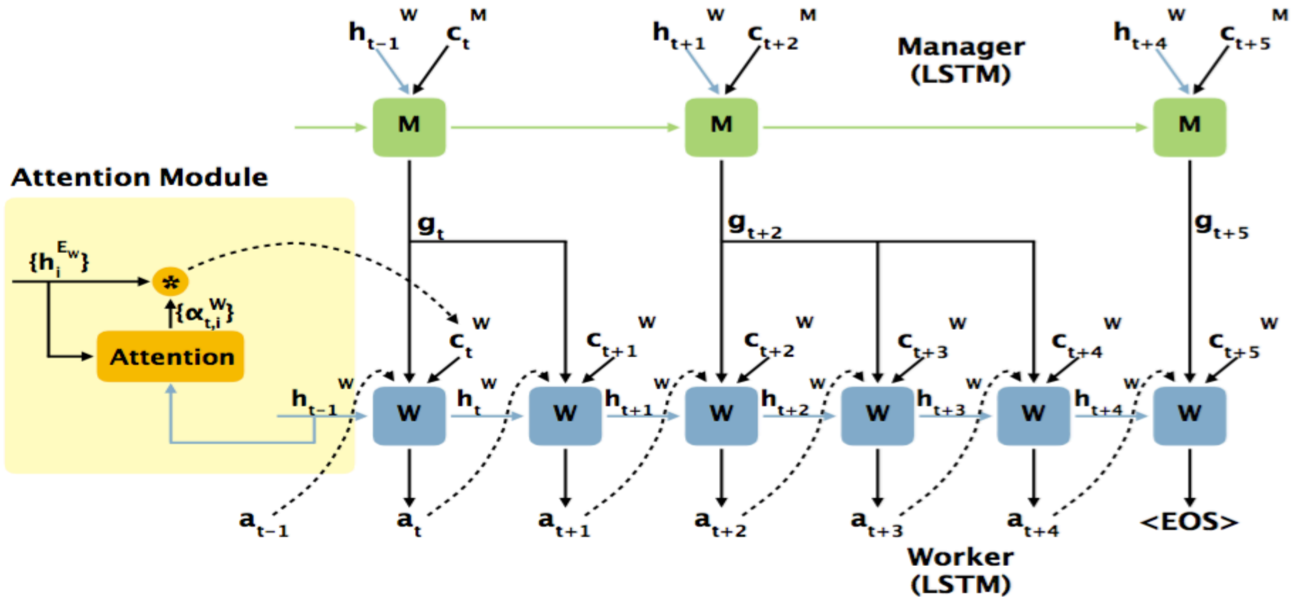

Figure 3: An example of the unrolled HRL agent in the decoding stage (from time step t to $t+5$). The yellow region shows how the attention module is incorporated into the encoder-decoder framework.

Manager

$$h_t^M = S^M(h_{t-1}^M, [c_t^M, h_{t-1}^W])$$

$$g_t = u_M(h_t^M)$$

Worker

$$h_t^W = S^W(h_{t-1}^W, [c_t^W, g_t, a_{t-1}])$$

$$x_t = u_W(h_t^W)$$

$$\pi_t = SoftMax(x_t)$$

Internal Critic

$$h_t^I = RNN(h_{t-1}^I, a_t)$$

$$p(z_t) = sigmoid(W_z h_t^I + b_z)$$

Policy Learning

Stochastic Worker Policy Learning:

$$\nabla_{\theta_w} L(\theta_w) \approx -(R(a_t) - b_t^w)\nabla_{\theta_w} \log \pi_{\theta_w}(a_t) \qquad (14)$$

Deterministic Manager Policy Learning:

$$L(\theta_m) = -\mathbb{E}_{g_t}[R(e_t)\pi(e_{t,c}; s_t, g_t = \mu_{\theta_m}(s_t))] \qquad (17)$$

$$\nabla_{\theta_m} L(\theta_m) = -R(e_{t,c})\nabla_{g_t} \log \pi(e_{t,c})\nabla_{\theta_m}\mu_{\theta_m}(s_t) \quad (19)$$

$$\nabla_{\theta_m} L(\theta_m) =$$

$$-(R(e_{t,c}) - b_t^m)[\sum_{i=t}^{t+c-1} \nabla_{g_t} \log \pi(a_i)]\nabla_{\theta_m}\mu_{\theta_m}(s_t) \qquad g_t = \mu_{\theta_m}(s_t)$$

$$(22)$$

Reward Definition

- adopt delta CIDEr score as the immediate reward.
- f (x) = CIDEr(sent + x) CIDEr(sent)

Worker Discounted Return:

$$R(a_t) = \sum_{k=0}^{\infty} \gamma^k f(a_{t+k})$$

Worker Discounted Return:

$$R(e_t) = \sum_{n=0}^{\infty} \gamma^n f(e_{t+n})$$

Training

apply the cross-entropy loss optimization to **warm start** both the worker and the manager simultaneously, where the manager is completely treated as the latent parameters.

Algorithm 1 HRL training algorithm

Require: Training pairs <video, GT caption>

Randomly initialize the model parameters θ
Load the pretrained CNN model and internal critic
for iteration=1,M **do**
    Randomly sample a minibatch
    if Train-Worker **then**
        Disable the goal exploration
        Run a forward pass to get the sampled caption $a_1 a_2 \ldots a_T$
        Calculate $R(a_t)$ for each a_t
        Freeze the manager
        Update the worker policy using Equation 14
    else if Train-Manager **then**
        Initialize a random process $\mathcal{N}$ for goal exploration
        Run a forward pass to get the greedily decoded caption $e_1 e_2 \ldots e_n$
        Calculate $R(e_t)$ for each e_t
        Freeze the worker
        Update the manager policy using Equation 22
    end if
end for

Experiment

Method	BLEU@4	METEOR	ROUGE-L	CIDEr
Mean-Pooling	30.4	23.7	52.0	35.0
Soft-Attention	28.5	25.0	53.3	37.1
S2VT	31.4	25.7	55.9	35.2
v2t_navigator	40.8	28.2	60.9	44.8
Aalto	39.8	26.9	59.8	45.7
VideoLAB	39.1	27.7	60.6	44.1
XE-baseline	**41.3**	27.6	59.9	44.7
RL-baseline	40.6	28.5	60.7	46.3
HRL (Ours)	**41.3**	**28.7**	**61.7**	**48.0**

Table 1: Comparison with state of the arts on MSR-VTT dataset.

Method	B@1	B@2	B@3	B@4	M	R	C
XE-baseline	55.0	36.4	23.6	15.0	18.7	39.0	16.7
RL-baseline	57.6	41.4	28.0	**18.8**	17.7	39.8	21.6
HRL-16	**64.4**	**44.3**	**29.4**	**18.8**	**19.5**	**41.4**	23.2
HRL-32	64.0	43.4	28.4	17.9	19.2	41.0	21.3
HRL-64	61.7	43.0	28.8	**18.8**	18.7	31.2	**23.6**

Table 2: Results on Charades Captions dataset. We reported BLEU (B), METEOR (M), ROUGH-L (R) and CIDEr (C) scores of our HRL method and two baselines for comparison.

some thoughts

- Why HRL is successful？
- Generate sentence segment by segment
- Limit