Controllable Video Captioning with POS Sequence Guidance Based on Gated Fusion Network

—Supplementary Material

Paper ID 1039

Bairui Wang[1*] Lin Ma[2†] Wei Zhang[1†] Wenhao Jiang[2] Jingwen Wang[2] Wei Liu[2]
[1]School of Control Science and Engineering, Shandong University [2]Tencent AI Lab
{bairuiwong, forest.linma, cswhjiang, jaywongjaywong}@gmail.com
davidzhang@sdu.edu.cn wl2223@columbia.edu

In this appendix, we add some technical details mentioned in the submitted ICCV2019 manuscript, entitled as "Controllable Video Captioning with POS Sequence Guidance Based on Gated Fusion Network" with paper ID as 1039, and present extra ablation experiments on the ActivityNet 1.3 [1]. Specifically, we first introduce the self-critical sequence training [4] method, which is employed for training our model. Then we analysis the experimental results on the ActivityNet 1.3. And Finally, more supplementary qualitative results of our model on MSR-VTT dataset are illustrated.

1. Reinforcement Learning

In the submitted manuscript, we intend to directly train the captioning models guided by the evaluation metrics, specifically the CIDEr [6] in this work, instead of the cross-entropy losses. However, such a evaluation metric is discrete and non-differentiable, which makes the network difficult to be optimized with traditional methods. As such, we employ the reinforcement leaning (RL) [5] which has been widely used in both image captioning [4, 3] and video captioning [2, 7]. We resort to the self-critical sequence training [4], an excellent REINFORCE-based algorithm, that is specializing in processing the discrete and non-differentiable variables and first purposed for boosting image captioning.

1.1. REINFORCE Algorithms

In our case, the videos and words can be considered as the *environment*, and the proposed captioning model is considered as the *agent* that interacts with the *environment*. The parameters θ_{gen} of the model define the policy $\pi_{\theta_{gen}}$ which takes an *action* to predict a word followed by the updating of the *state*, that is the hidden states, the memory cells, and other learnable parameters of the captioning model.

The *reward* is obtained when a sentence is generated, which denotes the score of language metric CIDEr in this work. The model is trained by minimizing the negative expected reward:

$$\mathcal{L}_{RL}(\theta_{gen}) = -\mathbb{E}_{\mathbf{S}^k \sim \pi_{\theta_{gen}}} \left[r\left(\mathbf{S}^k\right) \right], \tag{1}$$

where $\mathbf{S}^k$ denotes the sentence sampled by the model for the k_{th} video in the dataset. Subsequently, the $r\left(\mathbf{S}^k\right)$ denotes the reward of the sentence, that is CIDEr in our work. Using the REINFORCE algorithm, the gradient of the non-differentiable reward function in Eq. (1) can be obtained as follows:

$$\nabla_{\theta_{gen}} \mathcal{L}_{RL}\left(\theta_{gen}\right) = -\mathbb{E}_{\mathbf{S}^k \sim \pi_{\theta_{gen}}} \left[r\left(\mathbf{S}^k\right) \cdot \nabla_{\theta_{gen}} \log \pi_{\theta_{gen}}\left(\mathbf{S}^k\right) \right]. \tag{2}$$

*This work was done while Bairui Wang was a Research Intern with Tencent AI Lab.
†Corresponding authors.

For each sample from the training set, as $\mathcal{L}_{RL}\left(\theta_{gen}\right)$ is generally estimated with a single sample from π_θ, the Eq. (2) can be represented as follows:

$$\nabla_{\theta_{gen}}\mathcal{L}_{RL}\left(\theta_{gen}\right) \approx -r\left(\mathbf{S}^k\right)\cdot\nabla_{\theta_{gen}}\log\pi_{\theta_{gen}}\left(\mathbf{S}^k\right). \tag{3}$$

However, estimating the gradient with a single sample will inevitably result in a high variance. To solve this issue, a baseline reward b is usually used to generalize the policy gradient without influencing the expected gradient, if b is not a function of $\mathbf{S}^k$ [5], which can be represented as follows:

$$\nabla_{\theta_{gen}}\mathcal{L}_{RL}\left(\theta_{gen}\right) \approx -\left(r\left(\mathbf{S}^k\right)-b\right)\cdot\nabla_{\theta_{gen}}\log\pi_{\theta_{gen}}\left(\mathbf{S}^k\right). \tag{4}$$

With the chain rule, the gradient of the loss function can also be written as:

$$\nabla_{\theta_{gen}}\mathcal{L}_{RL}\left(\theta_{gen}\right) = \sum_{t=1}^{n}\frac{\partial\mathcal{L}_{RL}\left(\theta_{gen}\right)}{\partial u_t}\frac{\partial u_t}{\partial\theta_{gen}}, \tag{5}$$

where u_t denotes the item that be input to the Softmax function at the t_{th} time step of the description generator, that is $\left(\mathbf{W}_s^{(D)}h_t^{(D2)}+\mathbf{b}_s^{(D)}\right)$ in Eq. (9) of the submitted manuscript. Using REINFORCE, the estimate of the gradient of $\frac{\partial\mathcal{L}_{RL}(\theta_{gen})}{\partial u_t}$ with the baseline is given by [8] as follows:

$$\frac{\partial\mathcal{L}_{RL}\left(\theta_{gen}\right)}{\partial u_t} \approx \left(r\left(\mathbf{S}^k\right)-b\right)\left(\pi_{\theta_{gen}}\left(s_t^k\right)-1_{s_t^k}\right), \tag{6}$$

where s_t^k denotes the t_{th} word in the sentence for the k_{th} video.

1.2. Self-critical Sequence Training

Based on REINFORCE algorithm, the self-critical sequence training is first proposed by Rennie *et al.* for image captioning, and greatly improves the performance [4].

Instead of learning another reward network for estimating the baseline reward b, Rennie *et al.* utilize the reward obtained by the current model under the inference algorithm in testing stage as the baseline reward b, that is $b = r\left(\hat{\mathbf{S}}\right)$, where $\hat{\mathbf{S}}$ is the sentence generated with the greedy search strategy by the current model. Thus, the estimate of the gradient in Eq. (6) can be rewritten as:

$$\frac{\partial\mathcal{L}_{RL}\left(\theta_{gen}\right)}{\partial u_t} \approx \left(r\left(\mathbf{S}^k\right)-r\left(\hat{\mathbf{S}}\right)\right)\left(\pi_{\theta_{gen}}\left(s_t^k\right)-1_{s_t^k}\right). \tag{7}$$

From aforementioned description, it can be observed that if a sample policy results a higher $r\left(\mathbf{S}^k\right)$ than the baseline $r\left(\hat{\mathbf{S}}\right)$, such a policy is encouraged by increasing the probability of the corresponding word. Conversely, those sample strategies with low rewards will be suppressed. The self-critical sequence training reduces the gradient variance as well as trains the model more effectively as only a forward propagation is required for the baseline estimating.

2. Experiments on ActivityNet 1.3

2.1. ActivityNet 1.3

The ActivityNet 1.3 dataset [1] is a large scale benchmark with the complex human activities for high-level video understanding, including temporal action proposal, action detection, and dense video captioning. There are 20,000 untrimmed long videos, with each has multiple annotated events with starting and ending time as well as the associated caption. It contains 10,024 videos for training, 4,926 for validation, and 5,044 for testing. For the training and validation data, we construct the video-sentence pairs by extracting the labeled video segments indicated by the starting and ending time stamps, as well as their associated sentences. As the ground-truth annotations of the testing split are for temporal action proposal task instead of video captioning, we simply validate our model on the validation split.

Model	B@4	M	R	C
EncDec+F (IR+M)	4.4	9.5	20.2	32.4
EncDec+CG (IR+M)	4.9	9.8	22.4	36.7
Ours (IR+M)	**5.0**	**10.2**	**22.9**	**37.3**
EncDec+F (I3D+M)	4.6	9.6	21.0	34.3
EncDec+CG (I3D+M)	**5.0**	10.2	22.9	37.2
Ours (I3D+M)	**5.0**	**10.3**	**23.0**	**38.4**

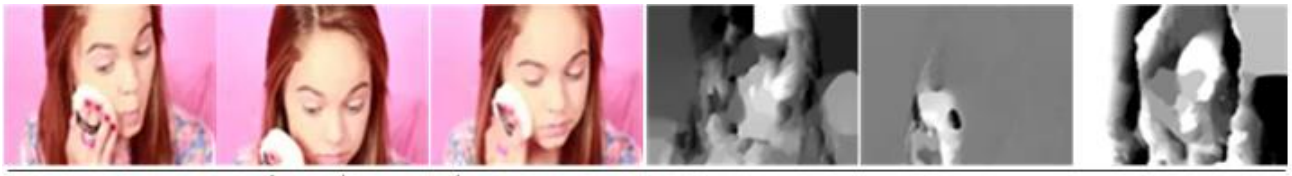
Table 1. Performance comparisons with different baseline models on the validation split of ActivityNet 1.3 in terms of BLEU@4 (B@4), METEOR (M), ROUGE-L (R), and CIDEr (C) scores (%). Methods of the same name but different text in the brackets indicates the same method with different feature inputs. IR and I3D denote the visual content features extracted from RGB frames by Inception_ResNet_V2 and I3D networks, respectively, and M denotes the motion feature extracted from optical flows by I3D network.

2.2. Ablation Studies

We conduct the ablation studies on ActivityNet 1.3 to further verify the effectiveness and reliability of our proposed model. The ablation experimental results are shown in Table. 1.

Taking IR and M as inputs, we find that the EncDec+CG (IR+M) outperforms the EncDec+F (IR+M) on all the evaluation metrics. For example, EncDec+CG (IR+M) is 4.3% higher than EncDec+F (IR+M) on CIDEr score, which is a significant improvement on ActivityNet 1.3. Once again, it demonstrates the proposed gated fusion network can particularly explore the relationships between different features, *e.g.*, the IR and M in this experiment, and merge them in an effective way. While compare ours (IR+M) with EncDec+CG (IR+M), a further improvement brought by global POS information is observed. The similar performance improvements can also be obtained when taking I3D and M as inputs. It indicates that the proposed gated fusion network for video captioning with the global POS guidance is effectively and generalizable.

3. More Qualitative Samples

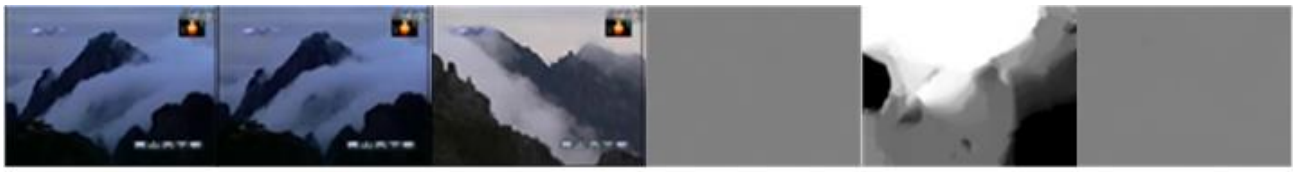


EncDec-F: A woman is putting on makeup.
Ours: A woman is showing how to apply makeup.
[POS]: ART NOUN VERB VERB NOUN VERB.
GT: A woman is showing how to do her makeup.

EncDec-F: A woman is talking about a recipe.
Ours: A man is making a dish in the kitchen.
[POS]: ART NOUN VERB VERB ART NOUN PRT ART VERB.
GT: A person is adding sauerkraut to his plate of food.

EncDec-F: A woman is cooking.
Ours: A person is cutting a piece of vegetables in a kitchen.
[POS]: ART NOUN VERB VERB ART NOUN PREP NOUN AUX ART NOUN.
GT: Someone chopping vagitable on chopping board.

EncDec-F: A man in a suit is talking to a man.
Ours: A man speaks to a camera for a video segment.
[POS]: ART NOUN VERB PREP NOUN CONJ PREP ART'NOUN.
GT: A man is interviewing another man.



EncDec-F: A person is playing a video game.

Ours: There is a big mountain is moving on the screen.

[POS]: ADV ART NOUN VERB VERB PREP ART NOUN.

GT: Fog rolls over the top of a mountain range while the sun is either rising or setting.

EncDec-F: People are playing sports.

Ours: A group of people are playing basketball in a stadium.

[POS]: ART NOUN ART NOUN VERB VERB AUX ART NOUN.

GT: School teams play basketball and soccer.

EncDec-F: A person is using a computer program.

Ours: A person is showing how to use a computer program.

[POS]: ART NOUN VERB VERB AUX VERB ART NOUN.

GT: A person explains how to use computer software.

EncDec-F: A person is cooking a dish in a pan.

Ours: A person is cooking a dish in a pan and adding ingredients to the pan.

[POS]: ART NOUN VERB VERB ART NOUN PREP ART NOUN.

GT: A person doing a cooking show and mixing the ingredients for the recipe.

EncDec-F: A man in a kitchen cutting a small knife.

Ours: A man is cutting a piece of meat on a cutting board.

[POS]: ART VERB VERB VERB ART NOUN PREP ART VERB NOUN.

GT: A man is preparing beef on a cutting board in a kitchen.

EncDec-F: A car is shown.

Ours: A man is talking about the features of a car.

[POS]: ART NOUN VERB VERB PREP ART NOUN.

GT: A car is being shown along with a list of its features while a person talks about them.

EncDec-F: A woman in a blue shirt is talking to a woman.

Ours: Kids are watching a video.

[POS]: ART NOUN VERB VERB ART NOUN.

GT: Kids are watching the incident activities through live cam.

EncDec-F: A baseball game is being played.

Ours: A baseball player is hitting the ball in the ground.

[POS]: ART NOUN NOUN VERB ART NOUN PREP ART NOUN.

GT: A batter hitting the ball during a ball game.

4. More Controllable Samples

Original Description:
[POS]: ART NOUN VERB VERB PREP ART NOUN.
Ours: A man is talking about something.

Controlling Description: add adjectives
[POS]: ART ADJ NOUN VERB VERB ART NOUN.
Ours: A man in a suit and glasses is talking.

Original Description:
[POS]: ART NOUN VERB VERB PREP ART NOUN PREP ART NOUN.
Ours: A man is talking with a woman.

Controlling Description: add adjectives
[POS]: ART ADJ NOUN VERB VERB ART NOUN.
Ours: An old man is talking about something.

Original Description:
[POS]: ART NOUN VERB VERB PREP ART NOUN.
Ours: A woman is talking to the camera.

Controlling Description: add adjectives
[POS]: ART ADJ NOUN VERB VERB ART NOUN PREP ART NOUN.
Ours: A woman in a blue shirt is cooking in a kitchen.

Original Description:
[POS]: ART NOUN VERB VERB AUX ART NOUN PREP ART NOUN.
Ours: A person is running on the beach.

Controlling Description: generate 'THERE BE'
[POS]: ADV VERB ART NOUN VERB VERB PREP ART NOUN.
Ours: There is a woman is running on the beach.

Original Description:
[POS]: ART NOUN VERB VERB PREP ART NOUN.
Ours: Some photos are shown in a move.

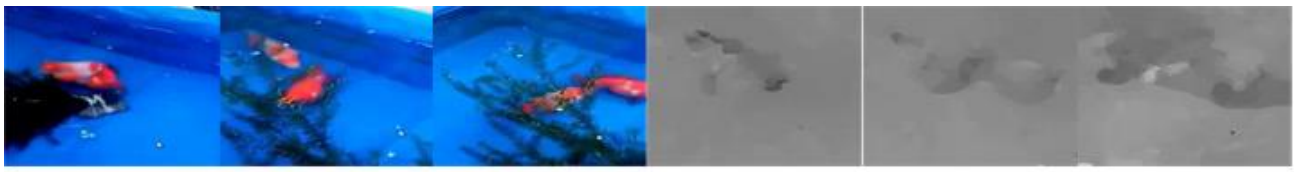
Controlling Description: generate 'THERE BE'
[POS]: ADV VERB ART NOUN VERB VERB NOUN.
Ours: There is a slideshow of pictures of people is shown.

Original Description:
[POS]: ART NOUN PREP ART ADJ NOUN VERB VERB AUX ART NOUN.
Ours: A man in a suit is talking to a woman.

Controlling Description: generate 'THERE BE'
[POS]: ADV VERB ART NOUN VERB VERB AUX ART NOUN.
Ours: There is a man is talking about the latest news.

Original Description:
[POS]: ART NOUN VERB VERB AUX ART NOUN.
Ours: A man is talking about a car.

Controlling Description: generate 'THERE BE'
[POS]: ADV VERB ART NOUN VERB VERB PREP ART NOUN.
Ours: There is a car is shown in a video.



Original Description:
[POS]: ART NOUN VERB VERB NOUN.
Ours: Some fish are swimming.

Controlling Description: change quantity
[POS]: NUM NOUN VERB VERB PREP ART NOUN.
Ours: Two fish are swimming in the aquarium.

Original Description:
[POS]: ART NOUN PREP ADJ NOUN VERB NOUN.
Ours: A video of disney characters dance.

Controlling Description: change quantity
[POS]: NUM ADJ NOUN VERB VERB NOUN.
Ours: Several disney characters are dancing.

Original Description:
[POS]: ART NOUN VERB VERB PREP ART NOUN.
Ours: A wrestling match is being played.

Controlling Description: change quantity
[POS]: NUM NOUN VERB VERB PREP ART NOUN.
Ours: Two men are wrestling in a gym.

Original Description:
[POS]: ART ADJ NOUN VERB VERB PREP ART NOUN.
Ours: A small dog is playing with a kid.

Controlling Description: change quantity
[POS]: NUM NOUN VERB VERB AUX NOUN PREP ART NOUN.
Ours: Two children are playing with a dog in the grass.

References

[1] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 1, 2

[2] R. Pasunuru and M. Bansal. Reinforced video captioning with entailment rewards. *arXiv preprint arXiv:1708.02300*, 2017. 1

[3] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 290–298, 2017. 1

[4] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017. 1, 2

[5] R. S. Sutton, A. G. Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998. 1, 2

[6] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1

[7] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Yang Wang. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4213–4222, 2018. 1

[8] W. Zaremba and I. Sutskever. Reinforcement learning neural turing machines-revised. *arXiv preprint arXiv:1505.00521*, 2015. 2