Object Relational Graph with Teacher-Recommended Learning for Video Captioning: Supplementary Material

Ziqi Zhang[1,3]*, **Yaya Shi**[2]*, **Chunfeng Yuan**[1]†, **Bing Li**[1,6,7], **Peijin Wang**[3,5], **Weiming Hu**[1,3,4], **Zhengjun Zha**[2]
[1]National Laboratory of Pattern Recognition, CASIA
[2]University of Science and Technology of China [3]University of Chinese Academy of Sciences
[4]Center for Excellence in Brain Science and Intelligence Technology, CAS
[5]Aerospace Information Research Institute, CAS [6]PeopleAI, Inc.
[7]State Key Laboratory of Communication Content Cognition, People's Daily Online
{zhangziqi2017}@ia.ac.cn,{shiyaya,zhazj}@mail.ustc.edu.cn,{cfyuan,bli,wmhu}@nlpr.ia.ac.cn

1. The effect of different top-k soft targets in TRL

In the main paper, we demonstrate that the probabilities of most soft targets are too small. These small value soft targets always presents semantically unrelated words, and may introduce noise to the caption model. Therefore, we only select top-k words as the soft targets for the teacher-recommended learning (TRL) method. The curve in the Fig.1 depicts the performance on CIDEr with different top-k soft targets on the MSR-VTT dataset. It can be found that the model gets sweat point at $k = 50$, and too large or too small k bring negative effects for the system.

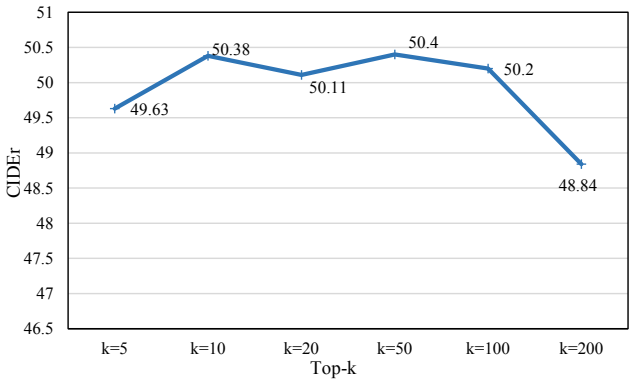

Figure 1. The CIDEr curve with different top-k soft targets on MSR-VTT dataset. The temperature is set to 2.0 and the KL-ratio is set to 0.3.

2. More fair comparisons on VATEX with the same visual features

In the Tab.1, the results on the top-block of the table are under the official provided I3D features. We utilize more powerful feature extractor C3D and IRV2 (InceptionResnetV2) to capture more discriminative appearance and motion representations, and the results reported in the main paper are shown on the bottom-block. For a fair comparison, we directly apply the I3D features provided by[2]. We get the superior results than Wang[2], and the ablation studies on the middle-block illustrate the effectiveness of our proposed ORG-TRL methods.

Methods	Features			VATEX			
	I3D	C3D	IRV2	B@4	M	R	C
Shared Enc[2]	✓	×	×	28.9	21.9	47.4	46.8
Shared Enc-Dec[2]	✓	×	×	28.7	21.9	47.2	45.6
Baseline(Ours)	✓	×	×	29.5	20.8	47.2	40.6
Baseline+ORG(Ours)	✓	×	×	30.7	21.5	47.9	44.9
Baseline+TRL(Ours)	✓	×	×	30.4	21.5	47.9	44.5
Baseline+ORG+TRL(Ours)	✓	×	×	**31.3**	**21.9**	**48.3**	**47.1**
Baseline(Ours)	×	✓	✓	30.2	21.3	47.9	44.6
Baseline+ORG(Ours)	×	✓	✓	31.5	21.9	48.7	48.8
Baseline+TRL(Ours)	×	✓	✓	31.5	22.1	48.7	49.3
Baseline+ORG+TRL(Ours)	×	✓	✓	**32.1**	**22.2**	**48.9**	**49.7**

Table 1. The performance on VATEX dataset. Wang's methods (top-block); our methods with official I3D features (middle-block); our methods with more powerful C3D and InceptionResnetV2 features (bottom-block).

3. The output regularization effect of TRL

The regularization of output distribution is first proposed in the work[1], where two output regularizers: a maximum entropy based confidence penalty and label smoothing are proposed. In our work, the caption model is under the

*Equal contribution.
†Corresponding author.

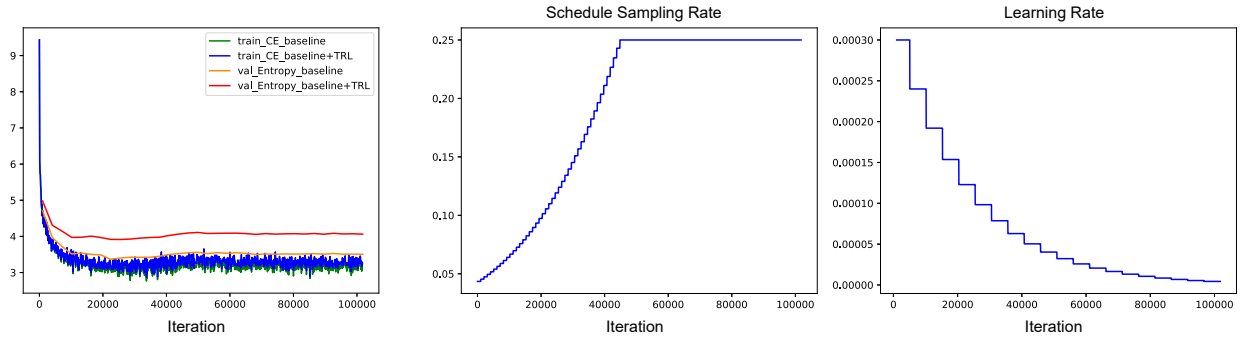

Figure 2. Regularization effect of TRL on output distribution(Left). The schedule sampling rate at the training phase(middle). Learning rate curve(Right).

co-guidance of teacher-enforced learning using hard target and teacher-recommended learning using soft targets. Formally, the loss function is the summation of the hard target's **Cross-Entropy** (CE) and the soft targets' KL-Divergence. This process pulls the output distribution to the language model's and guides the caption model to learn more semantically related words. Therefore, the TRL can be treated as an knowledge based output regularizer. In the Fig.6 at the main paper, we show two instances to demonstrate the effect of TRL in increasing the probabilities of content-specific words. Specifically, in order to verify the regularization effect of the whole output distribution, we conduct two experiments: the baseline model and the additional of TRL method. To measure the degree of regularization, we calculate the **Entropy** of the whole output distribution in validation:

$$\mathcal{H} = -\sum_{d \in D} p_d log(p_d) \tag{1}$$

where D is the vocabulary size, p_d is the probability of the word d. The higher entropy value, the smoother the output distribution, and the higher the degree of regularization.

As shown in Fig.2, the training Cross-Entropy of both baseline and baseline+TRL are similar. However, the validation Entropy curve of baseline+TRL is above the baseline's, which illustrates the output distribution under TRL is more smooth, and verifies the regularization effect to the output distribution.

4. Examples of generate captions on three datasets

In Fig.3 - Fig.5 show more examples of generating captions on the MSVD, MSR-VTT and VATEX datasets. Compared with the red words in the generations via baseline model and the green words in the generations via our ORG-TRL methods, it can be found that our ORG-TRL based model can capture more relational and detailed information in the video.

References

[1] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

[2] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

GT: a cat is playing in a box
Baseline: a cat is playing
ORG-TRL: a cat is jumping into a box

GT: two boys are ridding on skateboards
Baseline: a boy is riding a bicycle
ORG-TRL: a boy is riding a skateboard

GT: a man is playing bowling
Baseline: a man is playing
ORG-TRL: a man is playing a ball

GT: a lady is putting make up on her eyebrows
Baseline: a woman is applying makeup
ORG-TRL: a woman is applying mascara

Figure 3. Some examples on the MSVD dataset.

GT: a person showing the process being done on the plant
Baseline: there is a woman is doing some experiment
ORG-TRL: a woman is showing how to use a science experiment

GT: someone is stirring the dish that being cooked in a pan on the stove
Baseline: in a kitchen someone is preparing a dish in the kitchen
ORG-TRL: a person is cooking something in a pot on the stove

GT: a group having fun outside
Baseline: a man and a woman are talking
ORG-TRL: a group of kids are playing in the forest

GT: a person pours some liquid into a bottle
Baseline: a person is using a machine
ORG-TRL: a man pours a liquid into a container

Figure 4. Some examples on the MSR-VTT dataset.

GT: Unknown
Baseline: a woman is demonstrating how to tie a knot.
ORG-TRL: a person is demonstrating how to tie a knot with a rope.

GT: Unknown
Baseline: a woman is braiding her hair with a hair dryer.
ORG-TRL: a woman is having her hair braided by another woman.

GT: Unknown
Baseline: a young boy is sitting on the floor and playing a game of cards.
ORG-TRL: a man is holding cards in his hand and he is trying to cut it.

GT: Unknown
Baseline: a man and a woman are standing in front of a microphone.
ORG-TRL: two men are standing in front of a television and talking to each other.

Figure 5. Some examples on the VATEX dataset. The ground-truth is unknown, because the caption model is tested on the online testing system.