Hybrid Information of Transformer for Image Captioning

Yuchen Ren[13], Ziqiang Chen[12], Jinyu Hu[13], Lei Chen[3]*
[1]University of Science and Technology of China,[2]School of Software Engineering
[3]Institute of Intelligent Machines,Chinese Academy of Sciences,China
{ycren,sa517034,Jinyuhu}@mail.ustc.edu.cn, chenlei@iim.ac.cn

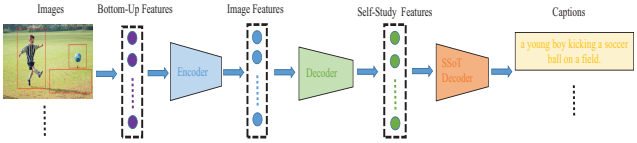

Figure 1. Brief conceptual diagram of our proposed SSoTNet, SSoT Decoder mainly includes SSoT module and Transformer decoder in the encoder-decoder framework.

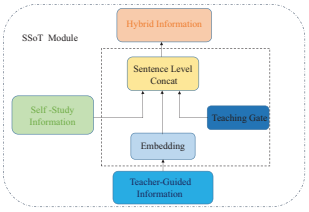

Figure 2. "Scheduled Sampling of Transformer" (SSoT).

In the current image captioning challenge, Transformer has emerged and has replaced the RNN with its efficient parallel mechanism and more powerful encoder-decoder framework to show encouraging results in some methods. In order to maintain its parallelism, most current methods adopt the teacher forcing method in the decoding stage of the training process, which encourages the model to predict all previous ground truth tokens in the sequence as the next ground truth token. However, during the process of test, the model cannot use ground truth tokens and must instead condition on its own prior predictions, resulting in train-test discrepancy:exposure bias. There are also a small number of methods that use Scheduled Sampling instead of Teacher-forcing for high performance, unfortunately, this destroys the parallelism of Transformer. In order to effectively deal with the balance between exposure bias and parallelism for image captioning, we propose to use the Scheduled Sampling of Transformer (SSoT) module to help Transformer generate captions for images.

As shown in Fig.1, first, we use the encoder to encode the image features, and then the encoded feature vectors enter the decoder, after generating the first decoding of the prediction "self-study information", we apply SSoT module to our image captioning model, each group of sentences and ground truth "teacher-guide information" are mixed with a certain probability as the guide feature vector "hybrid information" of the second decoding of the decoder.

As shown in Fig.2, given self-study information $S = (S_1, S_2, \cdots, S_n)$ from the first decoding, teacher-guided information $t = (T_1, T_2, \cdots, T_n)$, the teaching gate $g \in [0, 1]$ represents the possibility to use the token in teacher-guided information as final decoder input and is scheduled by the following piece-wise linear function:

$$g(step) = \min \quad N_{ic} * \frac{step - N_{st}}{N_{it}}\Big), g_{\max}\Big(\tag{1}$$

where $g_{\max}$ is the maximum teaching gate, and step means the current training step. N_{st} represents the Scheduled Sampling starting step, and N_{it} means how many steps as an interval to increase g. Finally, we acquire the whole token series "hybrid information" $H = (H_1, H_2, \cdots, H_n)$ as the guide of second decoding to achieve parallelism. Unlike scheduled sampling for RNN or LSTM, we replace the ground truth at a certain moment that needs to be changed with the word generated at the last moment, that is, change at the word level. In scheduled sampling for Transformer, we choose to concatenate the information at the sentence level , replacing the token of the teacher-guided information sentence determined by the teaching gate in a batch with the sentence token corresponding to the self-study information.For the sentence level concatenation, hybrid information $H_x \in (H_1, H_2, \cdots, H_n)$ can be formulated as:

$$H_x = \begin{cases} T_x & \text{with probability } g(step) \\ S_x & \text{with probability } 1 - g(step) \end{cases} \tag{2}$$

where x is the number of certain sentence in the current batch, for sentence level concatenation, the T_x or S_x is selected for the input according to the teaching gate $g(step)$. In summary, the function of SSoT can be formulated as:

$$H = SSoT(T, S, g(step)) \tag{3}$$

*Corresponding author.

Table 1. CIDEr-D Score Optimization of VizWiz-Captions dataset.

Method	P_{max}	B@1	B@4	M	R	C	S	Steps/Sec
Transformer	-	65.52	22.96	19.96	46.66	59.24	15.16	4.2
AoANet	0.50	65.86	**23.52**	**20.21**	46.97	**61.42**	15.23	2.6
SSoTNet	0.25	**65.96**	23.49	20.17	46.96	60.86	**15.30**	3.7
SSoTNet	0.50	65.84	23.35	20.12	**46.98**	60.74	15.21	3.7
SSoTNet	0.75	65.91	23.44	20.15	46.89	60.94	15.27	3.7

Table 2. Performance of our model and other state-of-the-art methods on MS-COCO "Karpathy" test split, where B@N,M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr-D and SPICE scores.

Model		Cross-Entropy Loss						CIDEr-D Score Optimization					
Approach	Parallelism	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
SCST[6]	×	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
LSTM-A[10]	×	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
Up-Down[1]	×	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
RFNet[4]	×	76.4	35.8	27.4	56.8	112.5	20.5	79.1	36.5	27.7	57.3	121.9	21.2
GCN-LSTM[9]	×	77.3	36.8	27.9	57.0	116.3	20.9	80.5	38.2	28.5	58.3	127.6	21.2
SGAE[8]	×	-	-	-	-	-	-	**80.8**	38.4	28.4	58.6	127.8	22.0
AoANet[3]	×	**77.4**	**37.2**	**28.4**	**57.5**	**119.8**	**21.3**	80.2	38.9	**29.2**	**58.8**	**129.8**	22.1
Transformer[7]	√	74.0	32.7	26.2	54.5	104.0	19.2	79.9	38.7	27.9	57.9	121.2	21.9
SSoTNet	√	76.0	35.1	27.9	56.1	113.4	21.0	**80.8**	**39.2**	28.7	**58.8**	125.8	**22.5**

We conduct experiments on the VizWiz-Captions dataset[2] and test online. Table 1 shows both the basic Transformer and SSoT are faster to train than the Scheduled Sampling model (AoANet) without parallelism, and the SSoT effect is better when the maximum probability of the teaching gate is 0.25. However, most scores of AoANet are higher because of its stronger attention on the basic self-attention.

Table 2 shows the performance comparisons between the state-of-the-art models and our proposed SSoT on the MS-COCO dataset[5]. For the cross entropy loss training stage, the METEOR and SPICE scores surpass most methods that do not have parallelism except AoANet, and SSoTNet has better performance than the basic Transformer.As for the CIDEr-D score optimization stage, the result shows that our model achieves superior performance in addition to maintaining parallelism, which reaches a new state-of-the-art of 39.2 BLEU-4 and 22.5 SPICE scores.

We introduce a simple technique to mitigate the discrepancy between train and test time, when we use Transformer framework for image captioning with parallelism. More remarkably, we obtain new state-of-theart performances of the Transformer through SSoT.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. *arXiv preprint arXiv:2002.08565*, 2020.

[3] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019.

[4] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *ECCV*, 2018.

[5] Tsung Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[6] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

[7] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[8] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019.

[9] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[10] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ECCV*, 2017.