Reflective Decoding Network for Image Captioning

Lei Ke[1], Wenjie Pei[2], Ruiyu Li[2], Xiaoyong Shen[2], Yu-Wing Tai[2]

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Tencent YouTu Lab

ICCV 2019 Seoul, Korea

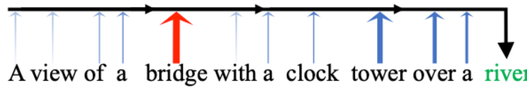
Introduction:

Existing problems:

- Traditional LSTM model tends to **focus more on the relatively closer vocabulary while neglecting the farther one**. For example, in the following figure, the word 'bridge' has an important hint on predicting the word 'river' , but the two words are separated by 6 words.
- Current mainstream caption decoder is weak in handling **long-term dependency** in sequential sentence, especially **when the visual content of an image is complex and hard to describe**.
- Although existing image captioning methods achieve high performance on standard caption dataset, we find that for **hard image captioning cases**, their performance **drops dramatically**.

Basis decoder: A black and white photo of a clock tower in the background.

Ours: A view of a bridge with a clock tower over a river.



Proposed solution:

- **Reflective Attention Module**: Modeling the dependencies between pairs of words at different time steps explicitly, taking into account the corresponding hidden states, instead of only memorizing the historical sequence information by balancing the overall relevance of all time steps like LSTM.
- **Reflective Position Module**: Modeling the relative position information individually in a supervised way, which equips our model with a strong perception of relative position for each word in the caption.

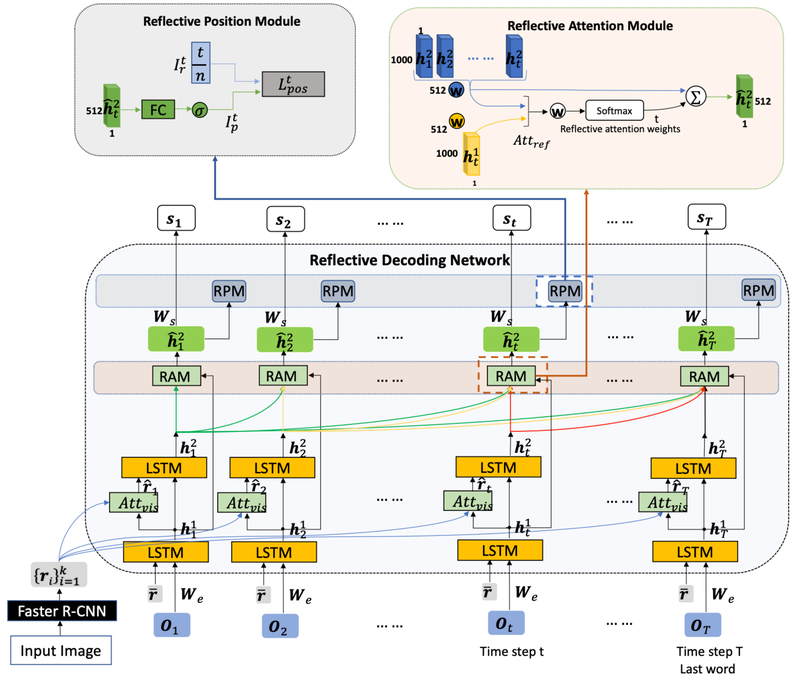
Summary:

- Enhance the **long sequential modeling ability** of the traditional caption decoder
- Explore **the coherence between words**
- Perceive the **relative position** of each word in the whole caption
- **Visualize** the word decision making process in **text domain** by considering long-term textual attention

References:

[1] Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." In CVPR, 2018.
[2] Jiang, Wenhao, et al. "Recurrent fusion network for image captioning." In ECCV, 2018.
[3] Rennie, Steven J., et al. "Self-critical sequence training for image captioning." In CVPR, 2017.
[4] Lu, Jiasen, et al. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning." In CVPR, 2017.
[5] Yao, Ting, et al. "Boosting image captioning with attributes." In ICCV, 2017.

Framework:



Quantitative Results:

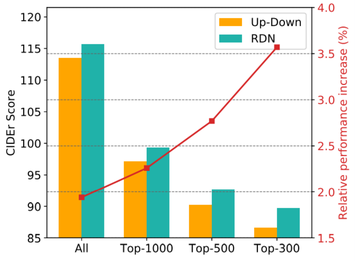
Ablation study on COCO 'Karpathy' test split

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Baseline	77.0	61.3	47.2	36.1	26.8	56.1	113.2	20.1
RDN$_{pos}$	77.4	61.6	47.5	36.3	27.0	56.5	114.3	20.4
RDN$_{ref}$	**77.6**	61.6	47.4	36.3	27.1	56.7	115.0	**20.5**
RDN	77.5	**61.8**	**47.9**	**36.8**	**27.2**	**56.8**	**115.3**	**20.5**

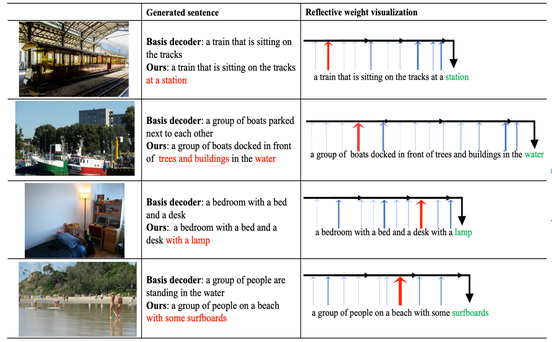
Comparison with other methods

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Review Net [47]	-	-	-	29.0	23.7	-	88.6	-
LSTM-A3 [49]	73.5	56.6	42.9	32.4	25.5	53.9	99.8	18.5
Att2in [34]	-	-	-	31.3	26.0	54.3	101.3	-
Adaptive [26]	74.2	58.0	43.9	33.2	26.6	-	108.5	-
Up-Down [2]	77.2	-	-	36.2	27.0	56.4	113.5	20.3
RFNet [17]	76.4	60.4	46.6	35.8	**27.4**	56.5	112.5	**20.5**
RDN	**77.5**	**61.8**	**47.9**	**36.8**	27.2	**56.8**	**115.3**	**20.5**

Evaluation on **hard image captioning**



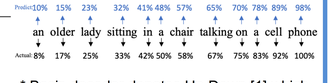
Sample Results:



Predicting **relative position in the sentence**

GT: an older woman is talking on a cell phone

RDN: an older lady sitting in a chair talking on a cell phone



* Basis decoder denotes Up-Down [1] which uses traditional two-layer LSTM

Attention weight distribution over the past generated hidden states for **multiple key words**