Prophet Attention: Predicting Attention with Future Attention

Fenglin Liu[1], Xuancheng Ren[2], Xian Wu[3], Shen Ge[3], Wei Fan[3], Yuexian Zou[1], Xu Sun[2]

[1] ADSPLAB, School of ECE, Peking University,

[2] MOE Key Laboratory of Computational Linguistics, School of EECS, Peking University, [3] Tencent

Introduction

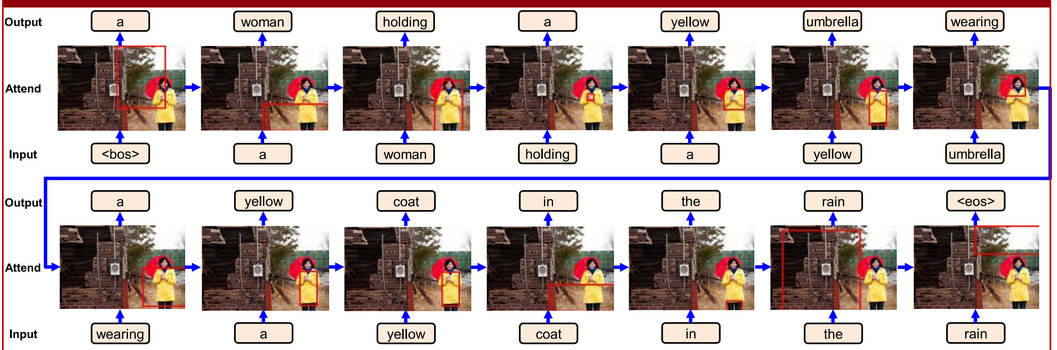

Figure 1. Illustration of the sequence of the attended image regions from a state-of-the-art system [1] in generating each word for a complete image description. At each time step, only the top-1 attended image region is shown. As we can see, the attended image regions are grounded more on the input words than the output words, such as the timesteps that input yellow and umbrella, demonstrating poor grounding accuracies of the current attention model.

Limitation & Challenge:

- Current attention models for image captioning have a "deviated focus" problem that they calculate the attention weights based on current hidden state, which contains the information of past generated words, instead of the one to be generated.
- These attention models has to predict attention weights without knowing the word it should ground.
- Figure 1 shows that the attended image regions are more grounded on current input word than the output one.

Those problems impair the captioning performance and ruin the model interpretability.

Solution:

- We propose the Prophet Attention to ground the image regions with proper generated words. In the training stage, our method utilizes the words that will be generated in the future to calculate the "ideal" attention weights towards image regions. These "ideal" attention weights are further used to regularize the "deviated" attention, which based on the input words that have already been generated.

Approach

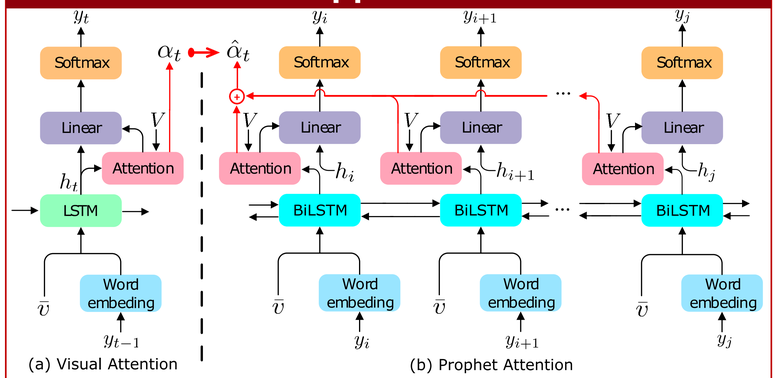

Figure 2. Our approach (right) calculates "ideal" attention weights $\hat{\alpha}_t$ based on future generated words $y_{i:j}(j \geq t)$ as a target for the conventional attention model (left) based on previous generated words $y_{1:t-1}$.

Conventional Attention Model [2]:

For each step t, the output h_t of the LSTM is used to attend to the relevant visual feature V and generates the attention weights α_t:

$$\alpha_t = f_{\text{Att}}(h_t, V) = \text{softmax}\left(w_\alpha \tanh\left(W_h h_t \oplus W_V V\right)\right)$$

Prophet Attention:

We first use the conventional image captioning model to generate the whole sentence $y_{1:T}$. Then, we employ a Bidirectional LSTM (BiLSTM) to encode the $y_{1:T}$. The information of $y_{i:j}$ is converted to $h_{i:j}$, and then the "ideal" attention weights $\hat{\alpha}_t$ are calculated by:

$$\hat{\alpha}_t = f_{\text{Prophet}}(h'_{i:j}, V) = \frac{1}{j-i+1}\sum_{k=i}^{j} f_{\text{Att}}(h'_k, V).$$

By minimizing the L1 loss, the model converges the "deviated" attention weights α_t calculated on previous words $y_{1:t-1}$ towards "ideal" attention weights $\hat{\alpha}_t$ calculated on future words $y_{i:j}(j \geq t)$.

$$\mathcal{L}_{\text{Att}}(\theta) = \sum_{t=1}^{T} \|\alpha_t - \hat{\alpha}_t\|_1$$

Approach

- Now, we introduce two variants of Prophet Attention.

Constant Prophet Attention: Since the attention weight is mainly determined by the single word that is to be generated at the current time step t, the intuition is to set $i = j = t$:

$$\hat{\alpha}_t = f_{\text{Prophet}}(h'_{i:j}, V) = f_{\text{Att}}(h'_t, V)$$

Dynamic Prophet Attention: If y_t belongs to a noun phrase (NP), we use all the words in the NP to calculate $\hat{\alpha}_t$. Then, when y_t is a non-visual (NV) word, we will remove our method:

$$\hat{\alpha}_t = f_{\text{Prophet}}(h'_{i:j}, V) = \begin{cases} \frac{1}{n-m+1}\sum_{k=m}^{n} f_{\text{Att}}(h'_k, V) & \text{if } y_t \in \text{NP: } y_{m:n} \\ \text{MASK} & \text{if } y_t \in \text{NV: } \{y_{\text{NV}}\} \\ f_{\text{Att}}(h'_t, V) & \text{otherwise} \end{cases}$$

Experiments

- We evaluate our method on two image captioning datasets.

Methods	Flickr30k Entities						Methods	MSCOCO				
	Fl_{all}	Fl_{loc}	B-4	M	C	S		B-4	M	R-L	C	S
NBT [33]	-	-	27.1	21.7	57.5	15.6	Up-Down [2]	36.3	27.7	56.9	120.1	21.4
Up-Down [2]	4.53	13.0	27.3	21.7	56.6	16.0	ORT [17]	38.6	28.7	58.4	128.3	22.6
GVD [59]	3.88	11.7	26.9	22.1	60.1	16.1	AoANet [20]	38.9	29.2	58.8	129.8	22.4
Cyclical [35]‡	4.98	13.53	27.4	22.3	61.4	16.6	X-Trans. [38]‡	39.7	29.5	59.1	132.8	23.4
Up-Down*	4.19	12.1	26.4	21.5	57.0	15.6	Up-Down*	36.7	27.9	57.1	123.5	21.3
w/ DPA	**5.45**†	**15.3**†	**27.2**†	**22.3**†	**60.8**†	**16.3**†	w/ DPA	**38.6**†	**29.1**†	**58.3**†	**129.0**†	**22.2**†
GVD*	3.97	11.8	26.6	22.1	59.9	16.3	AoANet*	38.8	29.0	58.7	129.6	22.6
w/ DPA	**4.79**†	**15.5**†	**27.6**†	**22.6**†	**62.7**†	**16.7**†	w/ DPA	**40.5**†	**29.6**†	**59.2**†	**133.4**†	**23.3**†

Table 1. Performance of offline evaluations on the Flickr30k Entities and the MSCOCO image captioning datasets. DPA represents the Dynamic Prophet Attention. B-4, M, R-L, C and S are short for BLEU-4, METEOR, ROUGE-L, CIDEr and SPICE, respectively.

- As we can see, the proposed method exhibits compelling effectiveness in boosting the baseline systems.

References

[1] Attention on attention for image captioning. *In ICCV, 2019.*

[2] Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *In CVPR, 2017.*

Contact Us

- {fenglinliu98, renxc, zouyx, xusun}@pku.edu.cn
- {kevinxwu, shenge, Davidwfan}@tencent.com