Supplementary Material: Show, Conceive and Tell: Image Captioning with Prospective Linguistic Information

Anonymous ACCV 2020 submission

Paper ID 808

1 Impact of Coarse Caption Quality

We verify how the quality of the coarse caption influences the performance of the Pro-LSTM model. Different checkpoints of the AoA* model, which is the coarse caption generator, are exploited to generate coarse captions. The quality of the coarse caption is assessed using its CIDEr-D score on the Karpathy's test split. We leverage the aforementioned model with *baseline + AAD* (*in Table 4 of the paper*) as a new baseline model to evaluate the impact of different coarse captions.

Table 1 demonstrates that the performance of our proposed Pro-LSTM model generally improves as the CIDEr-D of the coarse caption increases. However, it can be noticed that too poor quality coarse captions harm the performance of the Pro-LSTM model, as the prospective information contained in these captions may be erroneous. The erroneous prospective information would mislead the Pro-LSTM model and result in poor performance. When the quality of the coarse captions improves, the Pro-LSTM model outperforms the baseline model thanks to the semantically correct prospective information. We notice that when the CIDEr-D score of the coarse caption reaches 106.9, the Pro-LSTM model achieves comprehensive improvement than the *baseline + AAD* model in terms of CIDEr-D. Considering that 106.9 is much lower than 118.3, which is the CIDEr-D score of the *baseline + AAD* model, our proposed Prospective information guided Attention (ProA) mechanism does enhance the captioning model even when the coarse caption is not that satisfying.

Table 1. The performance of using different coarse captioner with different quality (CIDEr-D score) in the XE training phase.

CIDEr-D of coarse caption	Bleu-1	Bleu-4	Meteor	Rouge-L	CIDEr-D	SPICE
baseline+AAD	76.8	36.5	28.0	56.9	118.3	21.3
59.9	76.2	36.3	27.7	56.5	117.0	20.9
78.2	76.4	36.5	27.7	56.6	117.3	21.0
92.2	76.5	36.5	27.9	56.8	117.6	21.1
101.3	76.7	36.7	27.8	57.0	118.1	21.3
106.9	76.8	36.6	28.1	57.2	118.4	21.3
112.7	77.1	36.9	28.1	57.1	119.0	21.4
115.8	77.5	**37.2**	**28.2**	57.2	119.7	21.4
118.4	**77.8**	37.1	**28.2**	**57.3**	**120.2**	**21.5**

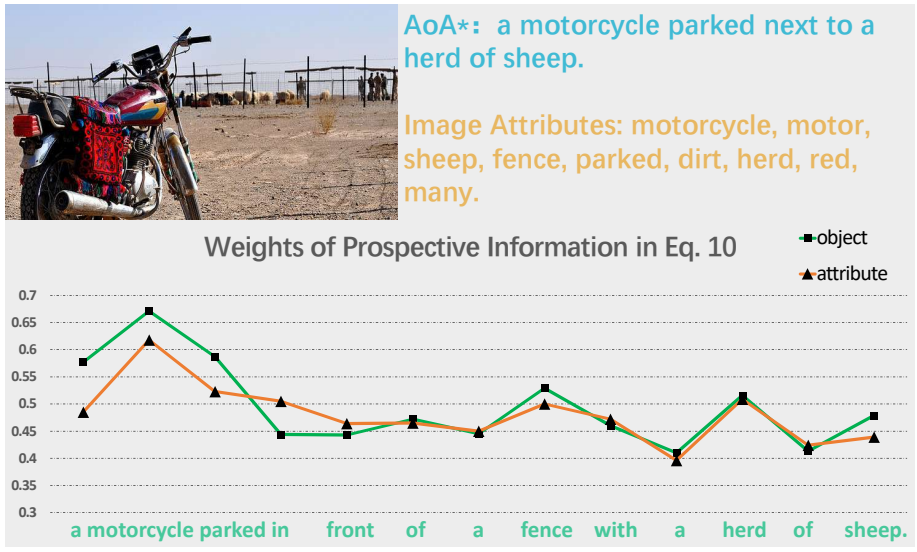
2 More Qualitative Results

More qualitative results are shown in Figures 1 - 2 in which the weights of prospective feature in the ProA module (*in Eq.10 of the paper*) and the visual attention in AAD are plotted. Considering that the weights are $1024d$ vectors in practice, we show the mean value of $\bar{\boldsymbol{\alpha}}_t$ in both object attention and attribute attention for simplicity.

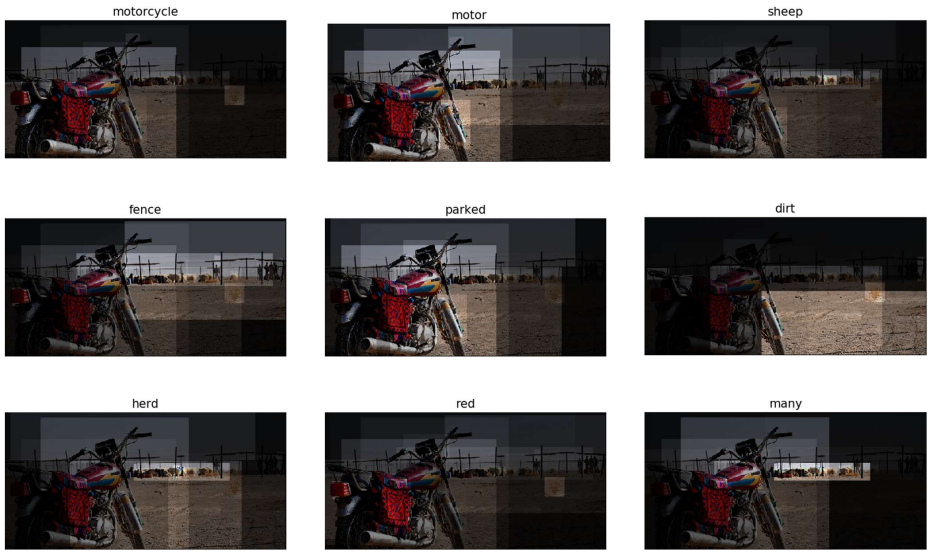
For example, Figure 1 shows the corresponding results for an image selected the MSCOCO test set. The coarse caption and the top-ranked predicted image attributes are shown in the top right of the figure. Although the coarse caption is roughly a semantically correct description for this image, it fails to point out the key object *'fence'*, which may help to describe the scene more reasonably. Nevertheless, the caption generated by the Pro-LSTM model, which is shown in green at the bottom of Figure 1(a), successfully describes the *fence* with the help of ProA and AAD. We first visualize how ProA influences our model by showing the weights of prospective information in the modified object attention and attribute attention sub-modules in the ProA module. Generally, these two weights are close in all the time steps, suggesting that the influence of prospective information is similar in the two modalities. It can be noticed our model assigns more weights to the prospective information to generate the nouns like *'motor'*, *'herd'* and *'sheep'* since these nouns are already predicted by the coarse caption. More importantly, the prospective information is successfully exploited to generate the new key word *'fence'*. This is because the *'sheep'*, which has been generated in the coarse caption, frequently appears with the *'fence'* in the countryside landscape images. Thus, the prospective information guides our model to attend to the features that are correlated to the *'fence'* in both modalities by forming relatively global linguistic contexts, which further leads to the generation of the new instance *'fence'*. We also visualize the attended areas of some image attributes in the AAD to show how we integrate the object features. Fig. 1(b) shows the attended areas for the top-ranked image attributes, where the lighter the area is the larger its corresponding weight is. The attribute *'fence'* is accurately detected since the AAD precisely attend to the most related areas inside the image. Consequently, the *'fence'* is successfully depicted by the Pro-LSTM model thanks to the plausible image attributes.

Similarly, in Fig. 2, the AoA* model fails to point out the *'clock tower'*. However, with the help of AAD, although the linguistic information in the coarse caption is thoroughly utilized after the generation of *'building'*, our model still delineates the *'clock tower'* as these two attributes are predicted with relatively high confidence by the AAD.

To conclude, we can witness that the collaboration of ProA and AAD enables our model to spot the new instances in the image so as to polish the coarse caption.

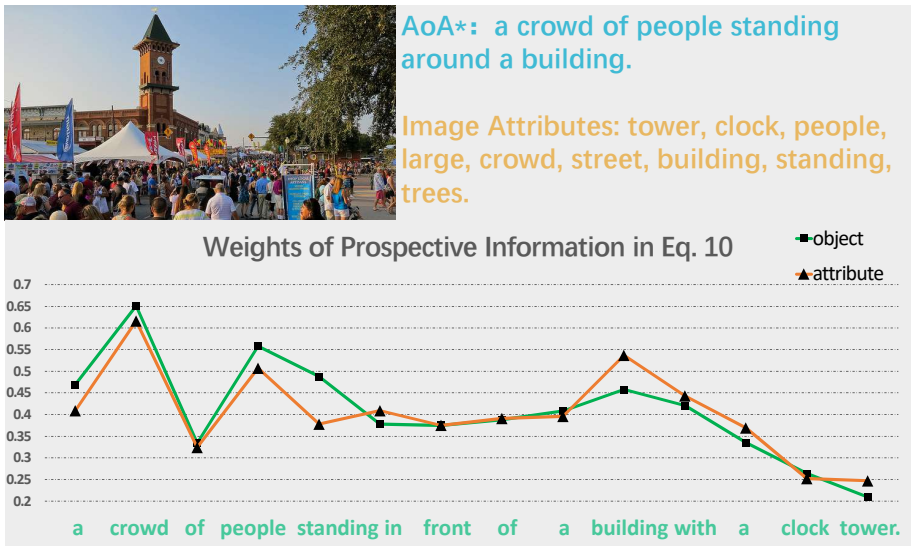


(a)

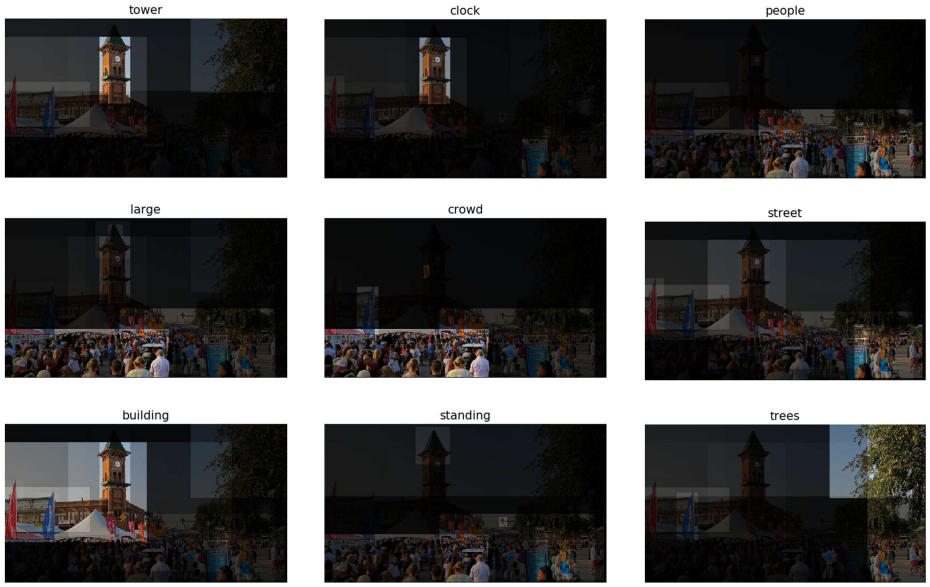


(b)

Fig. 1. The visualization of how ProA and AAD refine the coarse caption.



(a)



(b)

Fig. 2. The visualization of how ProA and AAD refine the coarse caption.

3 Implementation Details

3.1 Attentive Attribute Detector

We leverage the focal loss [1] to train AAD (*in section 4.2 of the paper*) as shown in Eq. 1, where l_i denotes whether the i^{th} attribute is in the ground truth captions or not, and δ and γ are empirically set to 0.95 and 2. The AAD model was trained 10 epochs in the MSCOCO dataset. The Adam Optimizer [2] was used with a batch size of 10. The learning rate was initially set to 5e-4 and decayed by a factor of 0.8 every 3 epochs.

$$\begin{aligned} loss_i^{fl} = & - l_i\delta(1-p_i)^{\gamma}log(p_i) \\ & - (1-l_i)(1-\delta)p_i^{\gamma}log(1-p_i) \end{aligned} \tag{1}$$

The performance of AAD is evaluated by the average F1 score (*in Table 1 of the paper*), here we explain how to compute it in detail. The precision and recall of attribute detection for each image are defined in Eq. 2, in which $\mathbf{I}_{gt}$ is the set of ground truth attributes of an image, $\mathbf{I}_d$ is the set of top-L detected attributes, and $|\cdot|$ represents the cardinality of a set. The F1 score is defined as the harmonic mean of precision and recall.

$$\text{precision} = \frac{|\mathbf{I}_{gt} \bigcap \mathbf{I}_d|}{|\mathbf{I}_{gt}|}, \quad \text{recall} = \frac{|\mathbf{I}_{gt} \bigcap \mathbf{I}_d|}{|\mathbf{I}_d|}, \quad \text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

3.2 Prospective attention guided LSTM

The Pro-LSTM model (*in section 4.1 of the paper*) adopted the bottom-up 36×2048 object feature provided in [3]. The dual encoder was formed with 6 attention blocks, where the sizes of hidden layer and the feed-forward network were set to 1024 and 2048 respectively. The Adam optimizer was also adopted to train the Pro-LSTM model. In the cross-entropy (XE) training phase, we trained our model for 15 epochs with a batch size of 40. The learning rate was initially set as 2e-4 and then reduced by a factor 0.8 every 3 epochs. We optimized the CIDEr-D score with SCST [4] for another 10 epoch with a batch size of 40. The learning rate was initially set as 2e-5 and then reduced by a factor 0.5 when the CIDEr-D on the validation set does not improve for 4500 iterations. The gradients were clipped by the maximum absolute value of 0.1 in both training phases. We used the beam search strategy with the beam size of 2 to generate the image captions. Note that our model was trained for much fewer epochs than other compared methods with the aid of prospective information in the coarse caption. In practice, after only 1 epoch XE training, the CIDEr-D score on the validation set is over 110.

References

1. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)

2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 3. (2018) 6
4. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1179–1195