Beyond a Pre-Trained Object Detector: Cross-Modal Textual and Visual Context for Image Captioning

Anonymous CVPR submission

Paper ID 4855

1. Proposed Graphical Model

In this section, we derive the graphical model with a newly introduced node T shown in Figure 1.

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{x}) &= \prod_i p(y_i|\boldsymbol{x}, y_{1:i-1}) \\
&= \prod_i \sum_{O,T} p(\boldsymbol{o}, \boldsymbol{t}|\boldsymbol{x}, y_{1:i-1}) p(y_i|\boldsymbol{x}, \boldsymbol{o}, \boldsymbol{t}, y_{1:i-1}) \\
&= \prod_i \sum_{O,T} p(\boldsymbol{o}, \boldsymbol{t}|\boldsymbol{x}) p(y_i|\boldsymbol{x}, \boldsymbol{o}, \boldsymbol{t}, y_{1:i-1}) \quad (1) \\
&= \prod_i \sum_{O,T} p(\boldsymbol{o}|\boldsymbol{x}) p(\boldsymbol{t}|\boldsymbol{x}) p(y_i|\boldsymbol{x}, \boldsymbol{o}, \boldsymbol{t}, y_{1:i-1}) \quad (2) \\
&= \prod_i \sum_{O,T} p(\boldsymbol{o}|\boldsymbol{x}) p(\boldsymbol{t}|\boldsymbol{x}) p(y_i|\boldsymbol{o}, \boldsymbol{t}, y_{1:i-1}) \quad (3) \\
&\simeq \prod_i \sum_{T} p(\boldsymbol{t}|\boldsymbol{x}) p(y_i|\boldsymbol{o}, \boldsymbol{t}, y_{1:i-1}) \quad (4) \\
&\simeq \prod_i p(y_i|\boldsymbol{o}, \boldsymbol{t}, y_{1:i-1}) \quad (5)
\end{aligned}
$$

Between Equation 1 and Equation 2, given the input image $\boldsymbol{x}$, $\boldsymbol{o}$ and $\boldsymbol{t}$ are conditionally independent given the structure of the graphical model shown in Figure 1. Between Equation 2 and Equation 3, assume $\boldsymbol{o}$ and $\boldsymbol{t}$ completely encode all necessary information of $\boldsymbol{x}$, y_i is conditionally independent of $\boldsymbol{x}$. Between Equation 3 and Equation 4, researchers typically take argmax and threshold to select a fixed set of detected objects from the object detector. Between Equation 4 and Equation 5, we propose to retrieve a fixed set of top-k most relevant text descriptions for each image crop. Eventually, we arrive at the same result as Equation 4 in the main paper.

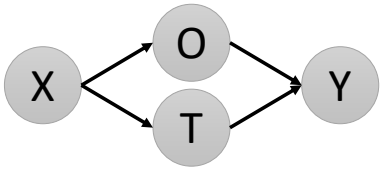
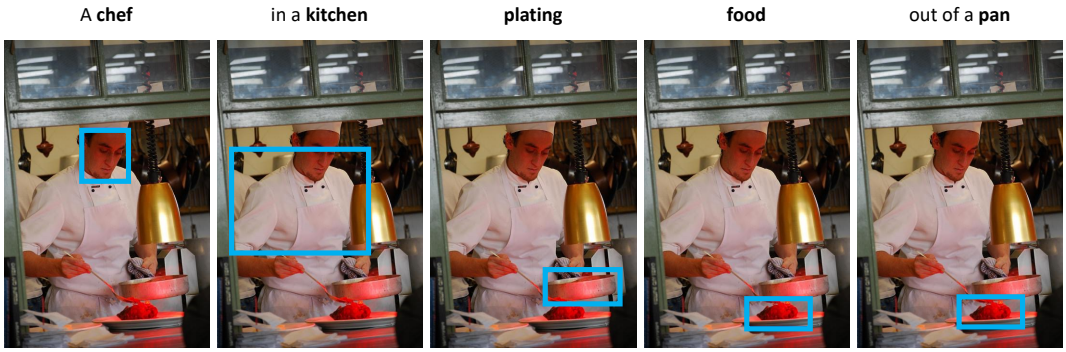


Figure 1. Graphical model with a newly introduced node T, which represents a set of retrieved text descriptions of image sub-regions.

2. Additional Qualitative Results

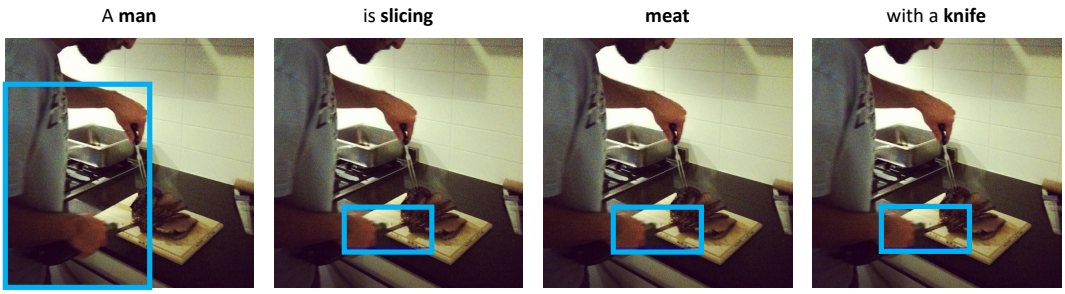
In Section 4.3 of the main paper, we quantitatively show that image conditioning helps refine the object features to aid with grounding on the Flickr30K dataset. In this section, we show qualitative examples in Figure 2-6 to further support this claim.



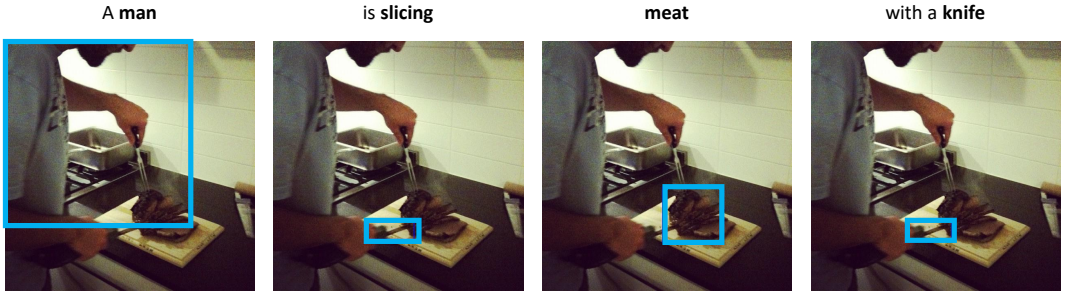
(a) Most attended object (in blue box) for generating each word using objects from a pre-trained object detector.

(b) Most attended object (in blue box) for generating each word using objects refined by our image conditioning module.

Figure 2. Example of: *A chef in a kitchen plating food out of a pan.*

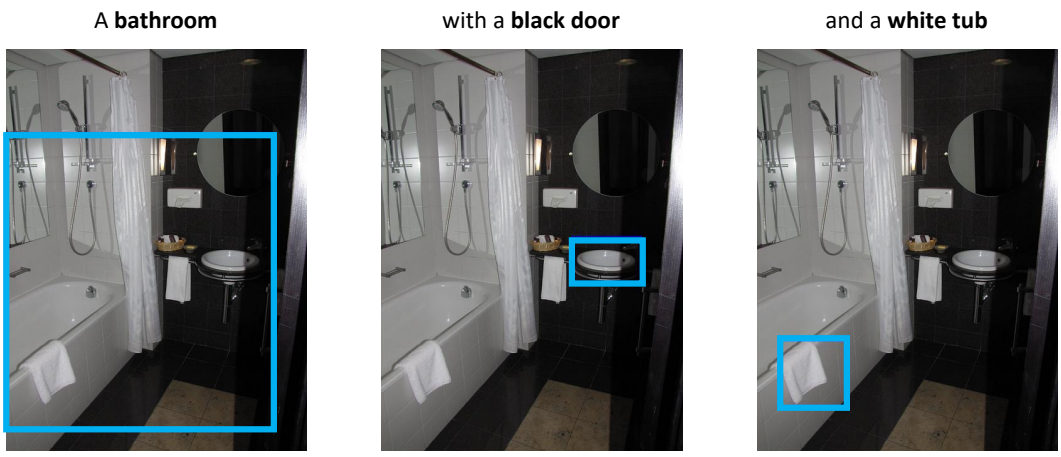


(a) Most attended object (in blue box) for generating each word using objects from a pre-trained object detector.



(b) Most attended object (in blue box) for generating each word using objects refined by our image conditioning module.

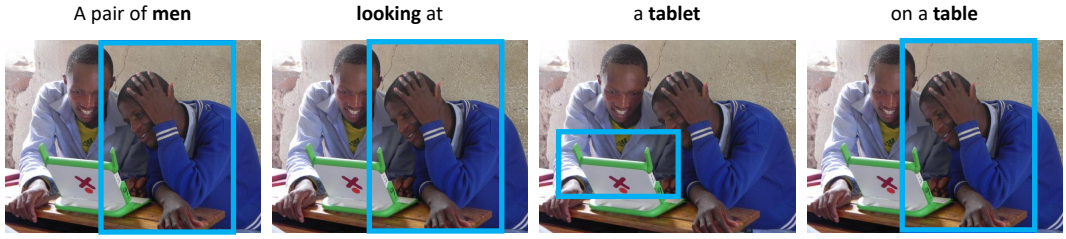
Figure 3. Example of: *A man is slicing meat with a knife.*



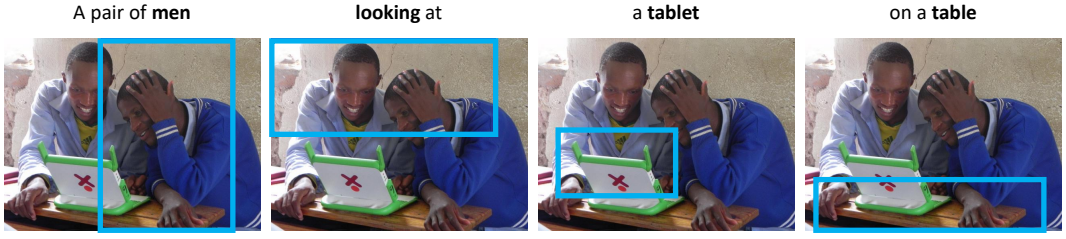
(a) Most attended object (in blue box) for generating each word using objects from a pre-trained object detector.

(b) Most attended object (in blue box) for generating each word using objects refined by our image conditioning module.

Figure 4. Example of: *A bathroom with a black door and a white tub.*

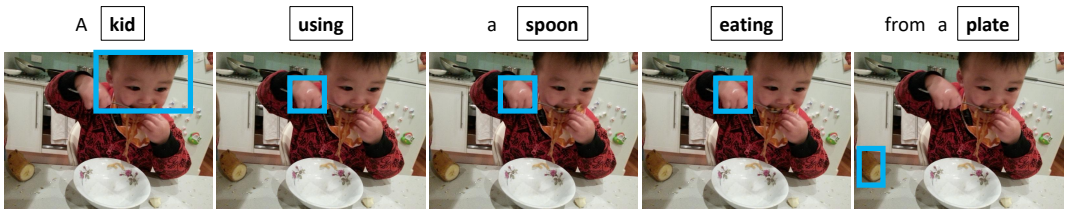


(a) Most attended object (in blue box) for generating each word using objects from a pre-trained object detector.

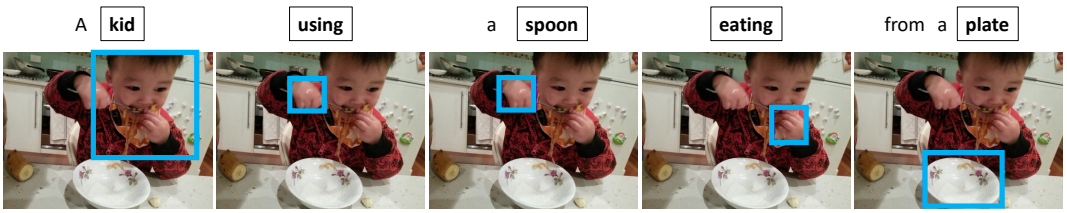


(b) Most attended object (in blue box) for generating each word using objects refined by our image conditioning module.

Figure 5. Example of: *A pair of men looking at a tablet on a table.*



(a) Most attended object (in blue box) for generating each word using objects from a pre-trained object detector.



(b) Most attended object (in blue box) for generating each word using objects refined by our image conditioning module.

Figure 6. Example of: *A pair of men looking at a tablet on a table.*