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.

$$p(y|x) = \prod_{i} p(y_i|x, y_{1:i-1})$$

$$= \prod_{i} \sum_{o, t} p(o, t|x, y_{1:i-1})p(y_i|x, o, t, y_{1:i-1})$$

$$= \prod_{i} \sum_{o, t} p(o|x)p(t|x)p(y_i|x, o, t, y_{1:i-1})$$

$$= \prod_{i} \sum_{o} p(o|x)p(t|x)p(y_i|x, o, t, y_{1:i-1})$$

$$\approx \prod_{i} \sum_{T} p(t|x)p(y_i|o, t, y_{1:i-1})$$

$$\approx \prod_{i} p(y_i|o, t, y_{1:i-1})$$

Between Equation 1 and Equation 2, given the input image $x$, $o$ and $t$ are conditionally independent given the structure of the graphical model shown in Figure 1. Between Equation 2 and Equation 3, assume $o$ and $t$ completely encode all necessary information of $x$, $y_i$ is conditionally independent of $x$. 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.

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

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

A bathroom with a black door and a white tub

(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 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.

A pair of men looking at a tablet on a table

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