Object-Centric Unsupervised Image Captioning: Supplements

Zihang Meng[1], David Yang[2], Xuefei Cao[2], Ashish Shah[2], and Ser-Nam Lim[2]

[1] University of Wisconsin-Madison [2] Meta AI
zihangm@cs.wisc.edu; xuefeicao01@gmail.com;
dzyang,ashishbshah,sernamlim@fb.com

Fig. 1: Qualitative results of the model trained on COCO images and GCC sentences. Images are from the COCO test split.

1 Comparison between COCO caption and LN-COCO

Table 1 (borrowed from [4]) shows the statistics of COCO captions [1] and Localized Narratives [4]. We can see that Localized Narratives contain more comprehensive and semantically meaningful sentences. This is why we choose LN-COCO as the ground truth for evaluation. We refer readers to the website of [4] for more qualitative examples of the captions.

In our experiments, during test time, we set the maximum prediction length of the model to 20 for GCC and SS (the same as [3,2]), and 100 for Localized Narratives.

Dataset	Words	Nouns	Pronous	Adjectives	Adpositions	Verbs
COCO Captions [1]	10.5	3.6	0.2	0.8	1.7	0.9
Localized Narratives [4]	36.5	10.8	3.6	1.6	4.7	4.2

Table 1: Comparison between COCO captions and Localized Narratives. The number shows the amount of certain type of words per sentence.

2 Details of how we change the object ratio in Table 5

In Table 5 of the main paper, we take the code of [3] change the object ratio to different values for ablation. In [3], they mine pseudo image-sentence pairs from existing images and sentences according to the number of overlapping objects in the image and the sentence. If the number of overlapping objects is larger than a certain threshold, then this image-sentence pair is added to the training set. We change the threshold to $2, 1, 0$ respectively which correspond to the three rows in Table 5. Each row corresponds to a certain (average) object ratio. (In the original implementation of [3], part of the training pairs have at least 1 overlapping object, and another part have at least 2 overlapping objects)

3 Qualitative results on GCC

In the main paper we have qualitative results for models trained on SS, LN-OpenImages and here we give some qualitative examples for the model trained on GCC. The results are in Fig. 1. Overall we can see that our model generates reasonable captions for the images which cover most of the main objects, their actions, and sometimes the relationship between the main objects and the background objects. The last one is a failure example where the generated caption says "vase on the floor" while the correct one should be "vase on the table". This type of error could be caused by the fact that the our training image region–sentence pairs do not always obey the correct spatial relationship, thus the model sometimes give wrong predictions about the spatial relationship. Another type of error which is hard for our model to give correct predictions is some finegrained information. For example, in the third example in Fig. 1, the man is looking at somewhere above while our generated caption says "looking at the laptop". This kind of finegrained information like the gaze direction is completely missed by the pre-trained object detector and can be possibly addressed by utilizing a specialized pre-trained detector (e.g., gaze predictor) which can extract this kind of information from the training data.

References

1. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
2. Feng, Y., Ma, L., Liu, W., Luo, J.: Unsupervised image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4125–4134 (2019)
3. Honda, U., Ushiku, Y., Hashimoto, A., Watanabe, T., Matsumoto, Y.: Removing word-level spurious alignment between images and pseudo-captions in unsupervised image captioning. arXiv preprint arXiv:2104.13872 (2021)
4. Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: European Conference on Computer Vision. pp. 647–664. Springer (2020)