Supplementary Material for Partially-Supervised Image Captioning

Peter Anderson

Macquarie University* Sydney, Australia p.anderson@mq.edu.au

Stephen Gould

Australian National University Canberra, Australia stephen.gould@anu.edu.au

Mark Johnson

Macquarie University Sydney, Australia mark.johnson@mq.edu.au

As supplementary material we provide additional caption examples for COCO novel object captioning in Figure 1, and for captions trained with Open Images in Figure 2. Further analysis of the impact of adding pre-trained word embeddings to the base model is included in Table 1.

Table 1: Analysis of the impact of adding fixed word embeddings (GloVe [1], dependency embeddings [2] or both) to the Up-Down [3] captioning model. Txofy indicates the model was decoded using constrained beam search [4] requiring the inclusion of at least x of the top y concepts randomly selected from the ground-truth image labels. Adding fixed embeddings has a slightly negative impact on the model when decoding without constraints (top panel). However, concatenating both embeddings (capturing both semantic and functional information) helps to preserve fluency during constrained decoding (bottom two panels).

	Out-of-Domain Val Scores				In-D	In-Domain Val Scores		
Model	SPICE	METEOR	CIDEr	F1	SPICE	METEOR	CIDEr	
Up-Down	14.4	22.1	69.5	0.0	19.9	26.5	108.6	
Up-Down-GloVe	14.0	21.6	66.4	0.0	19.5	26.2	104.1	
Up-Down-Dep	14.3	21.9	67.9	0.0	19.4	26.0	105.0	
Up-Down-Both	14.0	21.8	66.7	0.0	19.5	26.1	104.0	
Up-Down-GloVe + T2of3	18.0	24.4	80.2	28.3	22.2	27.9	109.0	
Up-Down-Dep + T2of3	17.8	24.4	79.5	23.8	21.8	27.5	107.3	
Up-Down-Both + T2of3	18.3	24.9	84.1	31.3	22.3	27.8	109.4	
Up-Down-GloVe + T3of3	19.0	24.6	80.1	45.2	23.0	27.4	101.4	
Up-Down-Dep + T3of3	19.0	24.5	79.0	42.2	22.3	26.9	98.4	
Up-Down-Both + T3of3	19.6	25.1	82.2	45.8	23.0	27.5	102.2	

^{*}Now at Georgia Tech (peter.anderson@gatech.edu)

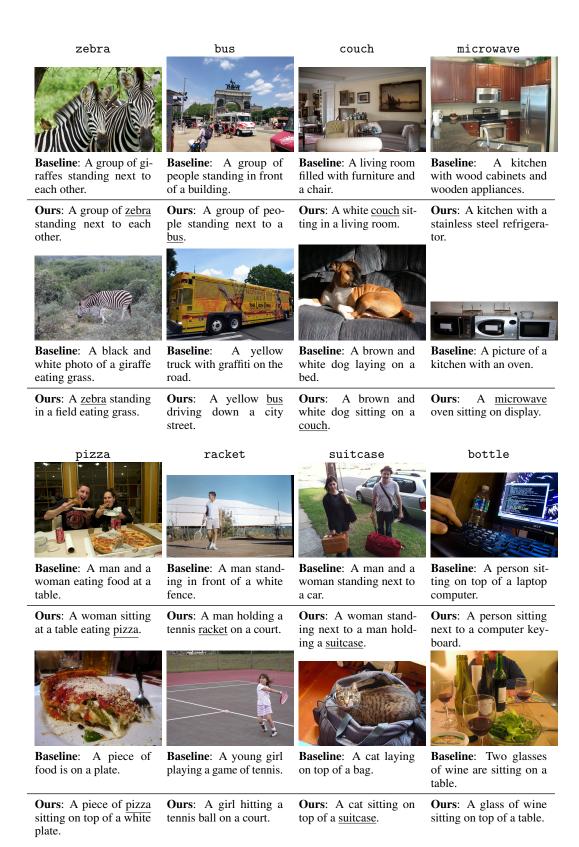


Figure 1: Further examples of captions generated by the Up-Down captioning model (top) and the same model trained with additional image labels using PS3 (bottom). All images shown contain held-out objects.



Figure 2: Further examples of captions generated by the Up-Down captioning model trained on COCO (top) and the same model trained with COCO and image labels from an additional 25 Open Images animal classes using PS3 (bottom). Several examples are failure cases (but no worse than the baseline).

References

- [1] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *EMNLP*, 2014.
- [2] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In ACL, 2014.
- [3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
- [4] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, 2017.