Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

SUPPLEMENTARY MATERIALS

1. Implementation Details

1.1. Bottom-Up Attention Model

Our bottom-up attention Faster R-CNN implementation uses an IoU threshold of 0.7 for region proposal suppression, and 0.3 for object class suppression. To select salient image regions, a class detection confidence threshold of 0.2 is used, allowing the number of regions per image k to vary with the complexity of the image, up to a maximum of 100. However, in initial experiments we find that simply selecting the top 36 features in each image works almost as well in both downstream tasks. Since Visual Genome [1] contains a relatively large number of annotations per image, the model is relatively intensive to train. Using 8 Nvidia M40 GPUs, we take around 5 days to complete 380K training iterations, although we suspect that faster training regimes could also be effective.

1.2. Captioning Model

In the captioning model, we set the number of hidden units M in each LSTM to 1,000, the number of hidden units H in the attention layer to 512, and the size of the input word embedding E to 1,000. In training, we use a simple learning rate schedule, beginning with a learning rate of 0.01 which is reduced to zero on a straight-line basis over 60K iterations using a batch size of 100 and a momentum parameter of 0.9. Training using two Nvidia Titan X GPUs takes around 9 hours (including less than one hour for CIDEr optimization). During optimization and decoding we use a beam size of 5. When decoding we also enforce the constraint that a single word cannot be predicted twice in a row. Note that in both our captioning and VQA models, image features are fixed and not finetuned.

1.3. VQA Model

In the VQA model, we use 300 dimension word embeddings, initialized with pretrained GloVe vectors [2], and we use hidden states of dimension 512. We train the VQA model using AdaDelta [4] and regularize with early stopping. The training of the model takes in the order of 12–18 hours on a single Nvidia K40 GPU. Refer to Teney et al. [3] for further details of the VQA model implementation.

2. Additional Examples

In Figure 1 we qualitatively compare attention methodologies for image caption generation, by illustrating attention weights for the ResNet baseline and our full Up-Down model on the same image. The baseline ResNet model hallucinates a toilet and therefore generates a poor quality caption. In contrast, our Up-Down model correctly identifies the couch, despite the novel scene composition. Additional examples of generated captions can be found in Figures 2 and 3. Additional visual question answering examples can be found in Figures 4 and 5.

References

[1] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 1

[2] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, 2014. 1

[3] D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *CVPR*, 2018. 1

[4] M. D. Zeiler. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 1

Figure 1. Qualitative differences between attention methodologies in caption generation. For each generated word, we visualize the attended image region, outlining the region with the maximum attention weight in red. The selected image is unusual because it depicts a bathroom containing a couch but no toilet. Nevertheless, our baseline ResNet model (top) hallucinates a toilet, presumably from language priors, and therefore generates a poor quality caption. In contrast, our Up-Down model (bottom) clearly identifies the out-of-context couch, generating a correct caption while also providing more interpretable attention weights.

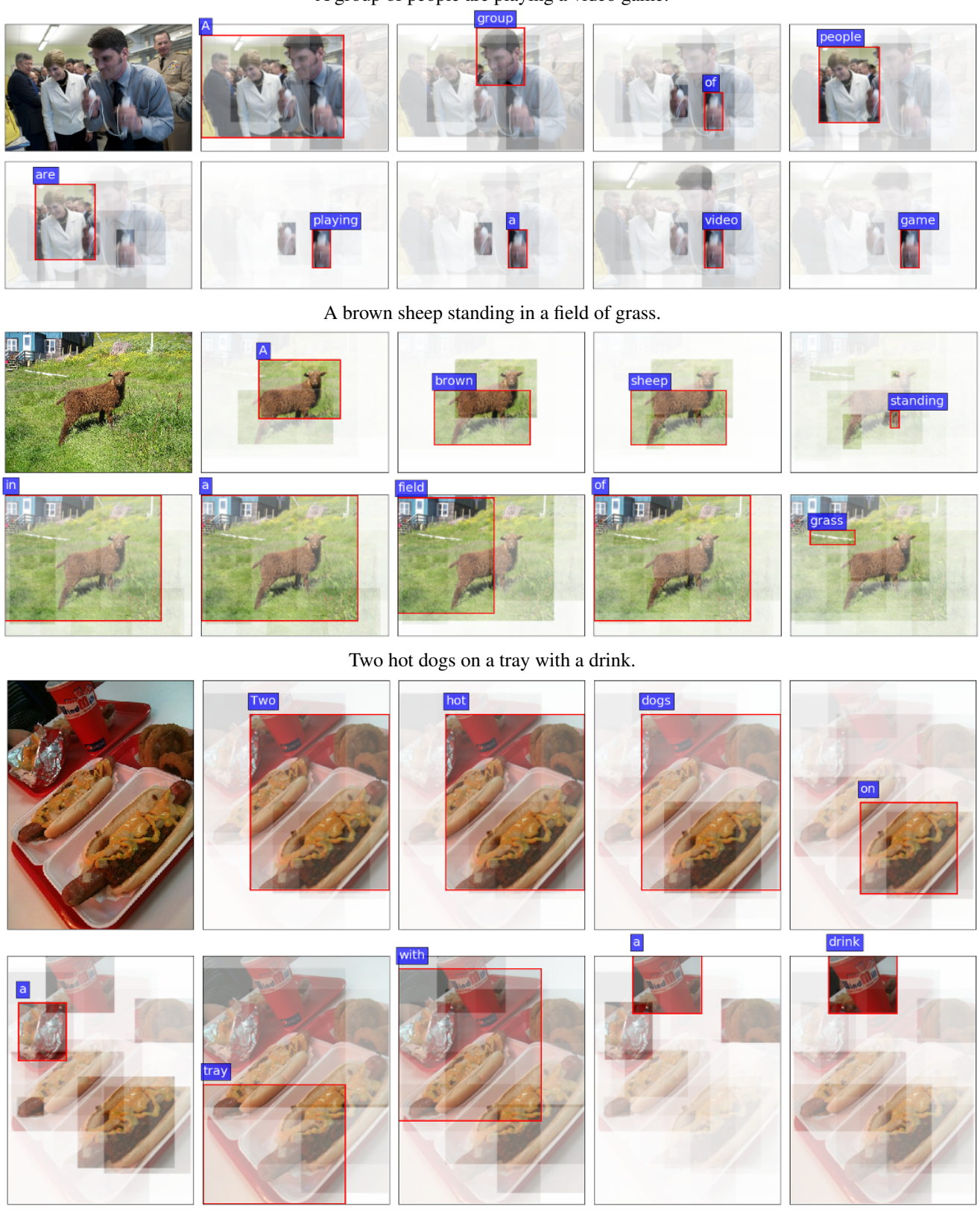


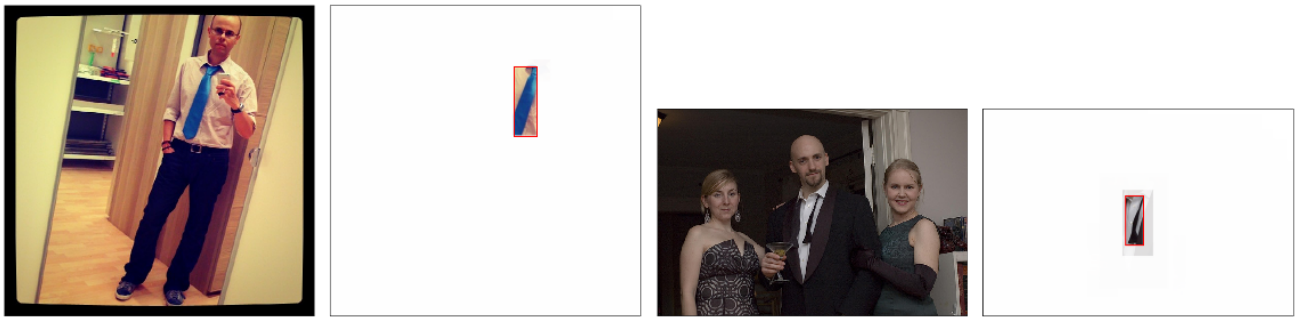
Figure 2. Examples of generated captions showing attended image regions. Attention is given to fine details, such as: (1) the man's hands holding the game controllers in the top image, and (2) the sheep's legs when generating the word 'standing' in the middle image. Our approach can avoid the trade-off between coarse and fine levels of detail.

Figure 3. Further examples of generated captions showing attended image regions. The first example suggests an understanding of spatial relationships when generating the word 'together'. The middle image demonstrates the successful captioning of a compositionally novel scene. The bottom example is a failure case. The dog's pose is mistaken for laying, rather than jumping – possibly due to poor salient region cropping that misses the dog's head and feet.

Question: What color is illuminated on the traffic light? Answer left: green. Answer right: red.

Question: What is the man holding? Answer left: phone. Answer right: controller.

Question: What color is his tie? Answer left: blue. Answer right: black.



Question: What sport is shown? Answer left: frisbee. Answer right: skateboarding.

Question: Is this the handlebar of a motorcycle? Answer left: yes. Answer right: no.

Figure 4. Further examples of successful visual question answering results, showing attended image regions.

Question: What is the name of the realty company? Answer left: none. Answer right: none.

Question: What is the bus number? Answer left: 2. Answer right: 23.

Question: How many cones have reflective tape? Answer left: 2. Answer right: 1.

Question: How many oranges are on pedestals? Answer left: 2. Answer right: 2.

Figure 5. Examples of visual question answering (VQA) failure cases. Although our simple VQA model has limited reading and counting capabilities, the attention maps are often correctly focused.