Supplementary Material: Modeling Relationships in Referential Expressions with Compositional Modular Networks

Ronghang Hu[1] Marcus Rohrbach[1] Jacob Andreas[1] Trevor Darrell[1] Kate Saenko[2]
[1]University of California, Berkeley [2]Boston University
{ronghang,rohrbach,jda,trevor}@eecs.berkeley.edu, saenko@bu.edu

Abstract

This document accompanies the paper "Modeling Relationships in Referential Expressions with Compositional Modular Networks". In this document, we describe some implementation details of our model, and show additional visualization results on each dataset.

1. Implementation details

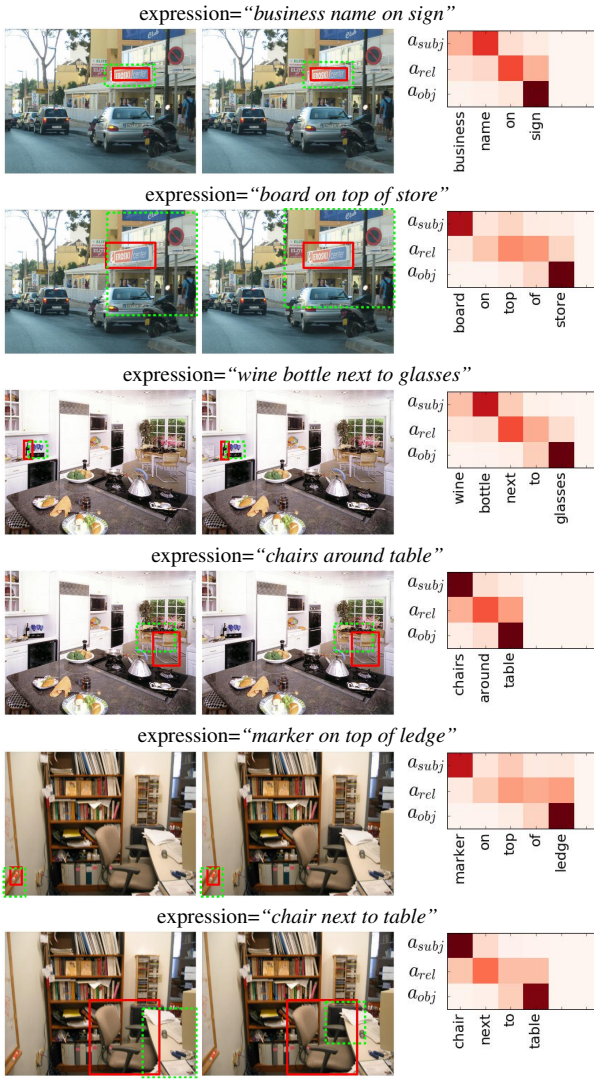
The language representation in our model is based on a 2-layer bi-directional LSTM network. In our implementation, both the forward and the backward LSTM in each layer have 1000-dimensional hidden states, so $h_t^{(1,fw)}$, $h_t^{(1,bw)}$, $h_t^{(2,fw)}$ and $h_t^{(2,bw)}$ are all 1000-dimensional vectors, and the final h_t is 4000-dimensional. During training, dropout is added on top of h_t as regularization.

In our experiments in Sec. 4.2, 4.3 and 4.4 in the main paper, we use the Faster-RCNN VGG-16 network architecture [2] for visual feature extraction. The input images are first forwarded through the convolutional layers of the network, and the features of each image region are extracted by ROI-pooling over the convolutional feature map, followed by subsequent fully connected layers. Parameters in the localization module, the relationship module and the language representation in our model are initialized randomly with Xavier initializer [1]. During training, each batch contains one image with all referential expressions annotated over that image. We train for 300000 iterations, with 0.95 momentum and an initial learning rate of 0.005, multiplied by 0.1 after every 120000 iterations.

2. Visualized results

We show more visualized results in Figure 1 for localizing the subject entity and the object entity from the relationship expressions in the Visual Genome dataset.

Figure 2 and 3 contain additional prediction examples in the Google-Ref dataset and the Visual-7W dataset, respectively.



(a) ground-truth (b) our prediction (c) attention weights

Figure 1. Grounded relationship expressions in the Visual Genome dataset, trained with weak supervision (subject-GT). (a, b) ground-truth region pairs and our predicted region pairs respectively (subject in red solid box and object in green dashed box). (c) attention weights for subject, relationship and object (darker is higher).

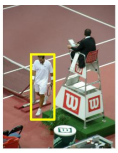
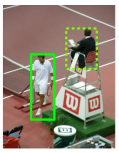
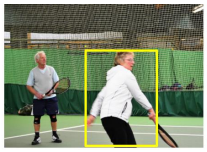
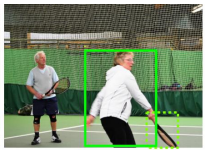
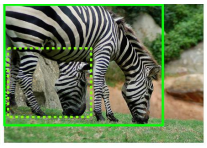
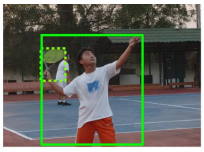
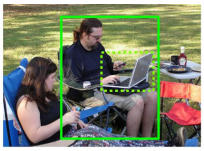
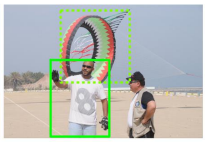
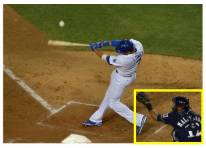
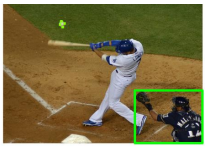
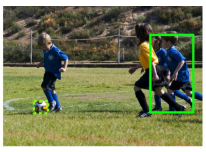
ground-truth	our prediction	ground-truth	our prediction	ground-truth	our prediction
expression="*the man nearest the surfboard*"		expression="*a tennis player walks beside a tower seat*"		expression="*a man wearing a red sweater and holding a glass of wine*"	
					
correct		correct		correct	
expression="*an old woman playing tennis*"		expression="*zebra on the right of other zebra*"		expression="*car in front of the motor cycle*"	
					
correct		correct		correct	
expression="*a baby eating some cake*"		expression="*a green coloured bus is standing near another bus*"		expression="*a man wearing a white T shirt and orange pants hits a ball with a racket*"	
					
correct		correct		correct	
expression="*a sugar donut to the far right of three other donuts*"		expression="*a green and light green striped couch a man is sitting on*"		expression="*a man working on his computer*"	
					
correct		incorrect		correct	
expression="*a man with gloves near a kite*"		expression="*the catcher reaching to catch the ball*"		expression="*the black umbrella held by the woman on the right*"	
					
correct		correct		correct	
expression="*a person with white T-shirt is trying to get the ball*"		expression="*a man reading his laptop in the waiting room*"		expression="*a young boy in a blue uniform is kicking a soccer ball*"	
					
incorrect		correct		incorrect	

Figure 2. Examples of referential expressions in the Google-Ref dataset. The left column shows the ground-truth region and the right column shows the grounded subject region (our prediction) in solid box and the grounded object region in dashed box. A prediction is labeled as correct if the predicted subject region matches the ground-truth region.

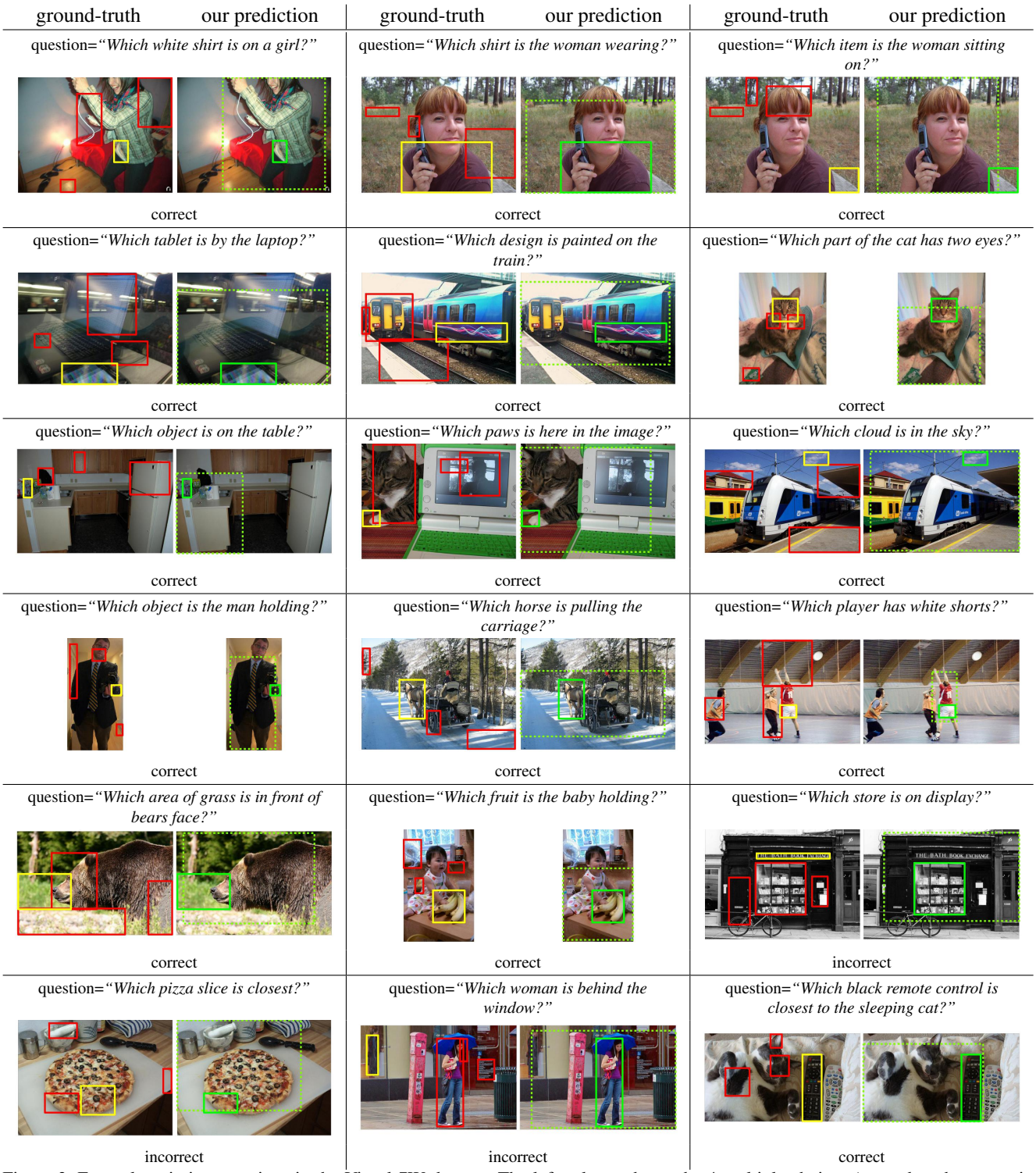


Figure 3. Example pointing questions in the Visual-7W dataset. The left column shows the 4 multiple choices (ground-truth answer in yellow) and the right column shows the grounded subject region (predicted answer) in solid box and the grounded object region in dashed box. A prediction is labeled as correct if the predicted subject region matches the ground-truth region.

References

[1] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010. 1

[2] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 1