Learning to Disambiguate by Asking Discriminative Questions Supplementary Material

Yining Li[1] Chen Huang[2] Xiaoou Tang[1] Chen Change Loy[1]
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]Robotics Institute, Carnegie Mellon University
{ly015, xtang, ccloy}@ie.cuhk.edu.hk, chenh2@andrew.cmu.edu

1. VDQG Dataset

Object Category. We selected 87 object categories from the annotation of Visual Genome datasets [3] to construct the VDQG dataset. Figure 1 shows the list of object category and the number of samples belonging to each object category.

Question Type. Figure 2 visualizes the most frequent n-gram ($n \leq 4$) sequences of questions in the VDQG dataset as well as the Visual Genome dataset. We observe that the question type distributions of these two datasets are similar to each other. A significant difference is that there is almost no "why" type question in VDQG dataset, which is reasonable because this type of question is hardly used to distinguish similar objects.

Examples. We show some examples of the VDQG dataset in Fig. 3.

2. Implementation Details

Attributes. We built an attribute set by extracting the commonly used n-gram expressions ($n \leq 3$) from region descriptions available in the Visual Genome dataset. And the part-of-speech constraint has been taken into consideration to select for discriminative expressions. Table 1 shows the part-of-speech constraints we use and the most frequent attributes.

Model Optimization. We implement our model using Caffe [1] and optimize the model parameters using Adam [2] algorithm. For the attribute recognition model, we use a batchsize of 50 and train for 100 epochs. For the attribute-conditioned LSTM model, we use a batchsize of 50 and train for 30 epochs, where gradient clipping is applied for stability. The parameters of CNN network has been pre-trained on ImageNet [4], and fixed during finetuning for efficiency.

[1]We merge all numbers that are greater than one into one label "more_than_one".

Table 1: Part-of-speech constraint on n-gram expressions to extract attributes

Part of speech	Top attributes
<NN>	man, woman, table, shirt, person
<JJ>	white, black, blue, brown, green
<VB>	wear, stand, hold, sit, look
<CD>	one, more_than_one [1]
<JJ,NN>	white plate, teddy bear, young man
<VB,NN>	play tennis, hit ball, eat grass
<IN,NN>	on table, in front, on top, in background
<NN,NN>	tennis player, stop sign, tennis court
<VB,NN,NN>	play video game
<IN,NN,NN>	on tennis court, on train track

3. Qualitative Results

Figure 4 shows some examples of discriminative question generated using our approach. The experimental result shows that our model is capable of capturing distinguishing attributes and generate discriminative questions based on the attributes. Some failure cases are shown at the last two rows in Figure 4. We observe that the failure cases are caused by different reasons. Specifically, the first two failure examples are caused by incorrect attribute recognition; the following two failure examples are caused by pairing attributes of different type of objects (*e.g.* pairing "green" of the grass and "white" of the people's clothes); and the last two failure examples are caused by incorrect language generation.

References

[1] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 1

[2] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 1

[3] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 1

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1

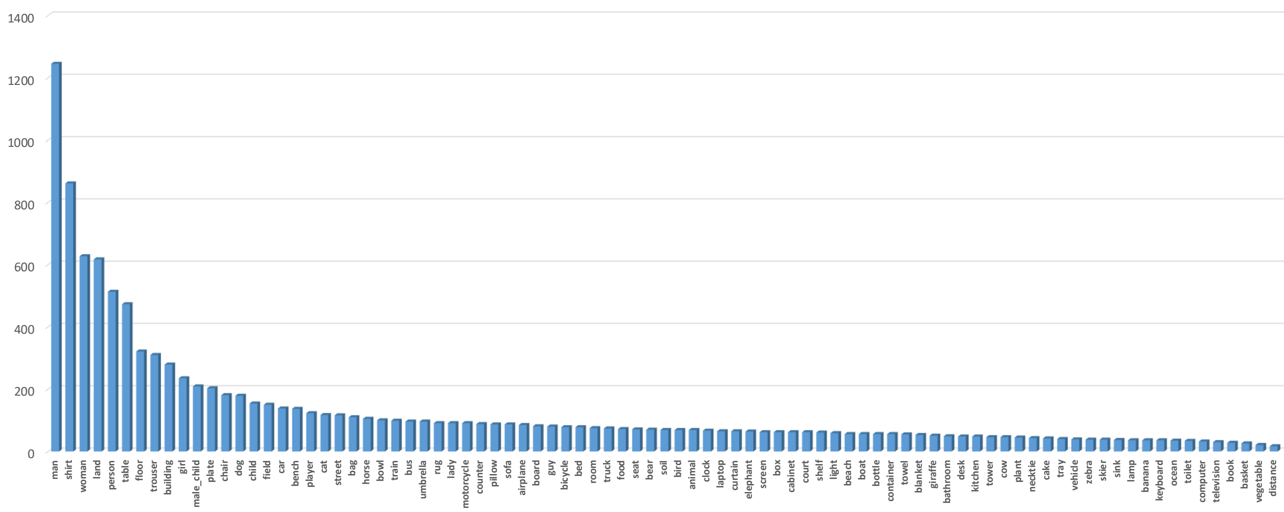
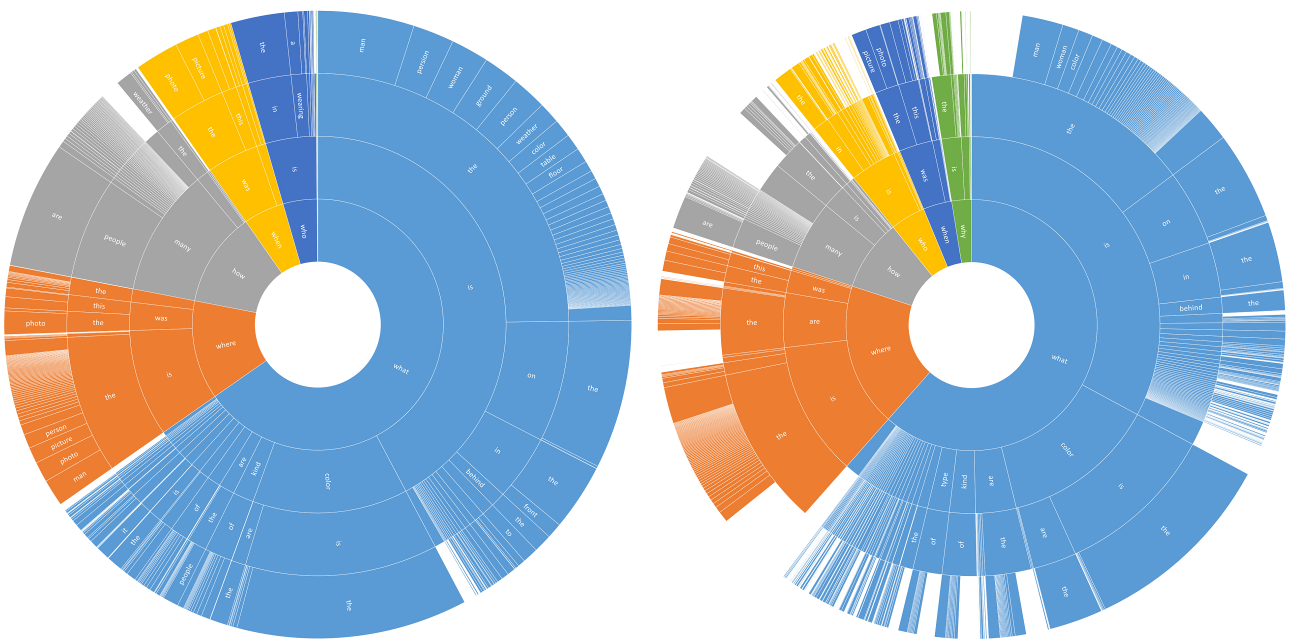
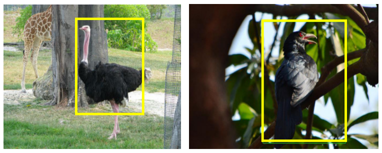


Figure 1: Object category distribution of VDQG dataset.



(a) VDQG (b) Visual Genome

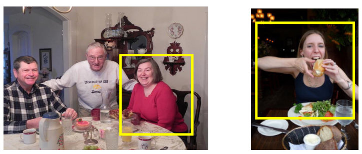
Figure 2: N-gram sequence distribution of VDQG dataset (a) and Visual Genome dataset (b).



+ What kind of bird is this?
+ What is the bird standing on?
+ What color is the bird’s feet?
- How many birds are there?
- What is the bird doing?
- When was the photo taken?

+ What is the man doing?
+ What is the man wearing on his head?
+ What is the man holding?
- What gender is the person in the shirt?
- What is the man wearing?
- Who is wearing the shirt?

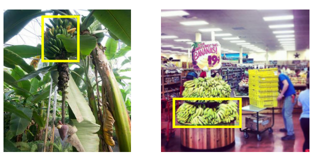
+ What is in the dog’s mouth?
+ Where is the dog?
+ What is the dog doing?
- What kind of animal is in the picture?
- How many dogs are there?
- When was this picture taken?



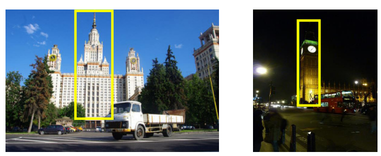
+ How many people are there?
+ What color is the woman’s top?
+ What color is the table?
- What is on the table?
- Where is the woman?
- What is on the woman’s head?

+ What color is the man’s pants?
+ What is the man wearing on the head?
+ What is in the background?
- Who is on the skateboard?
- What is the person doing?
- What is the man wearing?

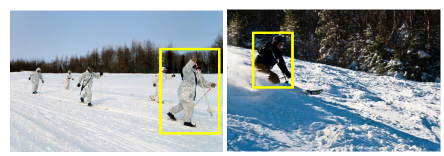
+ What color is the woman’s umbrella?
+ What is the woman wearing?
+ What color is the woman’s hair?
- What is the woman holding?
- Who is holding an umbrella?
- What is in front of the woman?



+ Where are the bananas?
+ What is beside the bananas?
+ How many bananas are there?
- What colors are the bananas?
- What fruit is shown?
- What is green?

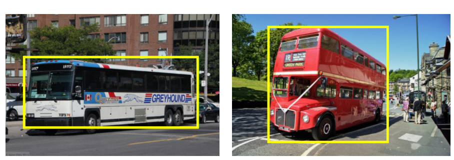


+ What color is the building?
+ When is the picture taken?
+ What time of day is it?
- Where is the clock?
- Hat is on the top of the tower?
- Where is the weather like?



+ What color is the skier’s jacket?
+ How many people are in the picture?
+ How many skiers are there?
- What is on the ground?
- What is the skier doing?
- What is the person wearing?

+ What is on the bench?
+ What color is the bench?
+ What color is the wall?
- What is the bench made of?
- Where is the bench?
- How many people are there sitting on the bench?



+ What color is the bus?
+ How many decks does the bus have?
+ What is on the side of the bus?
- Where is the bus?
- How many buses are there in the photo?
- What is behind the bus?

+ What is on the ground?
+ How many people are there?
+ How is the weather?
- What is the color of the ground?
- What is in the distance?
- Where was the photo taken?

Figure 3: Example of image pairs and the associated positive and negative question annotations in the proposed VDQG dataset. Positive and negative questions are written in blue and red, respectively.

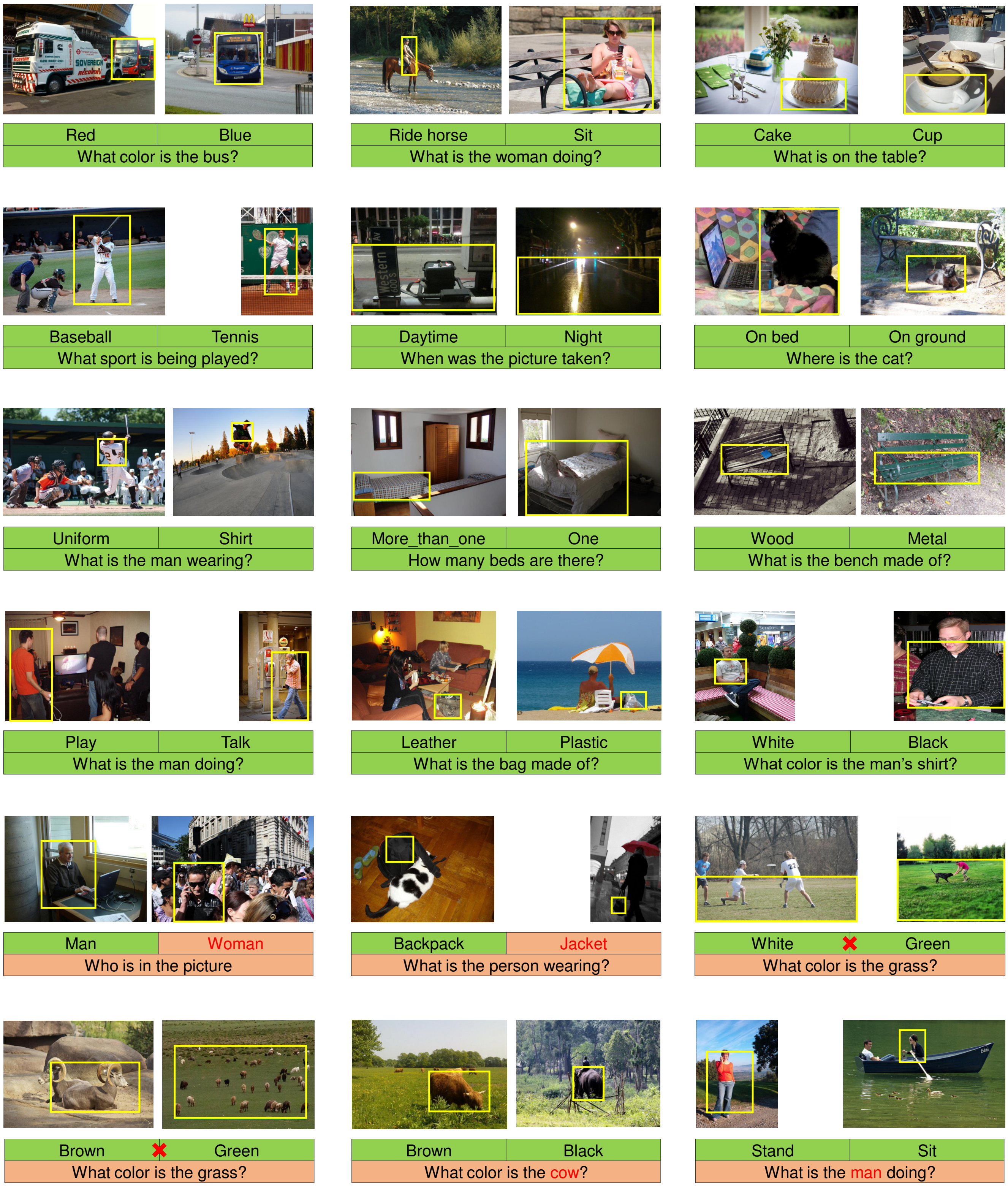


Figure 4: Discriminative questions generated by our approach. Under each ambiguous pair, the first line shows the distinguishing attribute pair selected by the attribute model, and the second line shows the questions generated by the attibute-conditioned LSTM. The last two rows show some failure cases.