Learning to Disambiguate by Asking Discriminative Questions
Supplementary Material

Yining Li\textsuperscript{1}  Chen Huang\textsuperscript{2}  Xiaou Tang\textsuperscript{1}  Chen Change Loy\textsuperscript{1}
\textsuperscript{1}Department of Information Engineering, The Chinese University of Hong Kong
\textsuperscript{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 \cite{krishna2017visual} 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 \cite{jia2014caffe} and optimize the model parameters using Adam \cite{kingma2015adam} 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 \cite{russakovsky2015imagenet}, and fixed during finetuning for efficiency.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
Part of speech & Top attributes \\
\hline
<NN> & man, woman, table, shirt, person \\
<JJ> & white, black, blue, brown, green \\
<VB> & wear, stand, hold, sit, look \\
<CD> & one, more_than_one \textsuperscript{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 \\
\hline
\end{tabular}
\caption{Part-of-speech constraint on \(n\)-gram expressions to extract attributes}
\end{table}

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

\begin{thebibliography}{9}
\end{thebibliography}

\textsuperscript{1}We merge all numbers that are greater than one into one label “more_than_one”.}

Figure 1: Object category distribution of VDQG dataset.

Figure 2: $N$-gram sequence distribution of VDQG dataset (a) and Visual Genome dataset (b).
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 attribute-conditioned LSTM. The last two rows show some failure cases.