Visual Question Answering on Image Sets: Supplementary Material

Ankan Bansal[1*], Yuting Zhang[2], and Rama Chellappa[1]

[1] University of Maryland, College Park, {ankan,rama}@umd.edu
[2] Amazon Web Services (AWS), yutingzh@amazon.com

1 Supplementary Material

1.1 Predictions from VQA Models

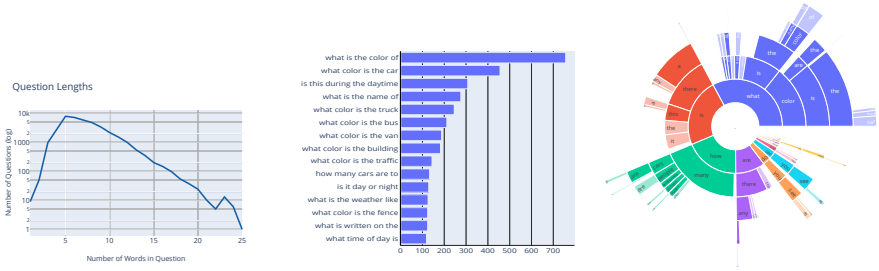
We show that current models fail to predict "Do not know" when the question cannot be answered, even for easy cases. Figure 1 shows some predictions from the recent Pythia model [1]. Directly using such a model for ISVQA is not optimal because ISVQA requires rejecting some of the images from a given set. Solving ISVQA requires either the ability to predict "not possible" to individual images or development of algorithms which can relate information across images.

Fig. 1: For the given image, a recent state-of-the-art model [1], gives the following answers for the following questions:
1.) What is the color of the painting above the computer? → blue (98%).
2.) Which bike is the woman riding? → zebra (52%), 0 (15%).
3.) What is written on the sign board? → nothing (99%), unknown (0.05%), can't tell (0.01%), not possible (0.01%).
4.) Which animal is the man petting? → zebra (86%), giraffe (13%).
Existing methods fail to answer "do not know" or "not possible" for simple questions. The ISVQA dataset will require this ability.

* This work was done when Ankan Bansal was an intern at AWS.

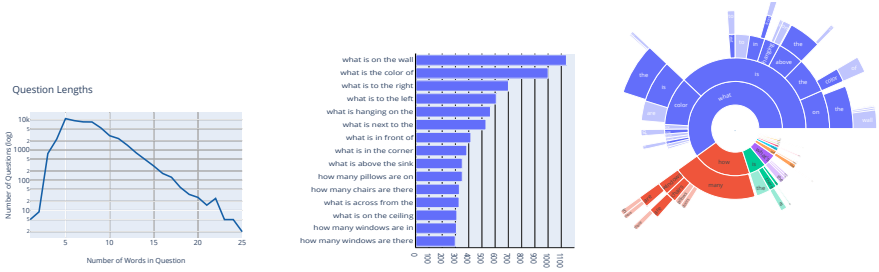
1.2 Dataset Analysis

In this section we provide an analysis of the nuScenes, Gibson-Room, and Gibson-Building datasets separately for a clearer understanding of their statistics. Figure 2 presents the statistics of the nuScenes datasets. Similarly, figures 3 and 4 show the respective statistics for Gibson-Room and Gibson-Building. Note that we had shown the statistics of the complete dataset in Fig. 4 in the paper.



(a) Question lengths (b) Frequent questions (c) Types of questions

Fig. 2: Statistics of the nuScenes dataset. (Left) The distribution of questions over numbers of words. (Middle) The most frequent types of questions in the dataset. (Right) The first five words of the questions. (Best viewed digitally)

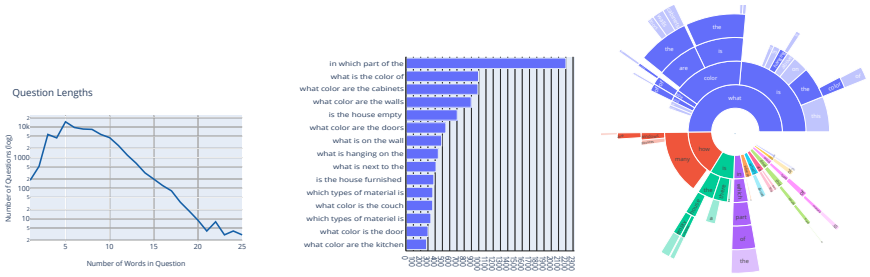


(a) Question lengths (b) Frequent questions (c) Types of questions

Fig. 3: Statistics of the Gibson-Room dataset. (Left) The distribution of questions over numbers of words. (Middle) The most frequent types of questions in the dataset. (Right) The first five words of the questions. (Best viewed digitally)

1.3 Annotation Screen

Figure 5 shows an example view shown to the annotators for collecting question-answer annotations. We provide clear, simple instructions and show the image set in a way to provide the maximum information about the scene to the annotators. We ask the annotators to ask a question about the scene and provide the corresponding answer in the given text input regions.



(a) Question lengths (b) Frequent questions (c) Types of questions

Fig. 4: Statistics of the Gibson-Building dataset. (Left) The distribution of questions over numbers of words. (Middle) The most frequent types of questions in the dataset. (Right) The first five words of the questions. (Best viewed digitally)

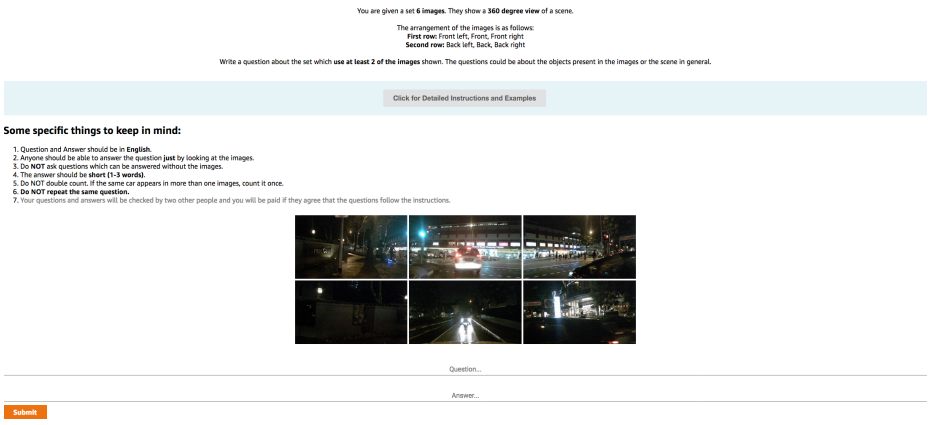


Fig. 5: Annotation screen for Amazon Mechanical Turk

1.4 Video Samples

We are attaching a few sample videos generated for Gibson using the method discussed in section 3.1 in the main paper. The videos shown in the attachments are taken from different locations and contain varied scene elements.

References

1. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Conference on Computer Vision and Pattern Recognition (2019)