Visual Question Answering for Blind People

Senyu Tong* **Tianyi Lin** * **Yutian Zhao***
{senyut, tianyil2, yutianzh}@andrew.cmu.edu

1 Task Definition

In this project, we will use VizWiz dataset (Gurari et al., 2018) to address the visual question answering for blind people. The high level task of our project is to predict an accurate answer to a visual question given an image and question about it.

1.1 Subtasks

We will break down the task into the following three specific subtasks. Given the questions are collected from blind users of a mobile phone application, most of them are conversational and may not be answerable. Therefore, our first subtask would be to identify whether a question is valid and answerable according to the given image. The second task would be to output the correct answer if the question is answerable. Since every question in the VizWiz dataset has 10 answers, it is also important to determine how to select or combine these answers during training, which will be our third subtask.

1.2 Input-output Representation

Limited by computing resources, we would not try to modify any pre-training mechanism. Therefore, we would seek to incorporate task-specific knowledge to augment current large pre-trained model to boost performance. Specifically, according to our data analysis, we would like to address image framing issue particularly for images with text detected. Further discussion will be put into later sections.

Following (Tan and Bansal, 2019) and (Chen et al., 2020)'s work, we would like to use both joint representation and coordinated representations. Our tasks would be to generate two embeddings separately for images and texts, and a cross-modality output for their joint representation.

*Everyone Contributed Equally – Alphabetical order

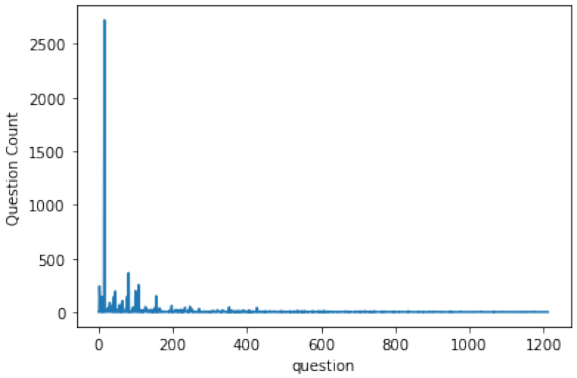

Figure 1: Question Frequency Distribution

For image representation, we would follow (Anderson et al., 2018). The input image embedding would be the features (positions and Region of Interest) of detected objects in it.

For language representation, the input would be word-level sentence embedding tokenized by WordPiece as in BERT (Devlin et al., 2019).

2 Data Analysis

2.1 Analysis of Questions

We examine the questions in both sentence and word level. We first calculate the distribution of the questions and observe that the percentage of questions that appear more than once is 54.4%. "What is this?" is the most common question that occupies more than 13% of all questions while the second most common question "What color is this?" is only about 1.7%. The frequency distribution is shown in Figure 1. We also analyze question diversity by computing statistics on sentence lengths. The mean question length is 6.76 words and the 25th and 75th percentile lengths are four and eight words respectively. The longest question has 62 words and the shortest one has only two words.

We also analyze the words in each sentence. We

Figure 2: Answer popularity word cloud excluding "unanswerable" and "unsuitable image" answers

consider the words that appear only once in all questions as rare words, which are about 2.64% of all words. The percentage of questions having at least one rare word is 12.09%. The most common first word of each question is "What", which is consistent with the observation of most common questions in the sentence-level analysis. We also observe that the questions in this dataset often begin with a rare first word. The percentage of questions starting with a first word that occurs for less than 5% and 10% of all questions is 32.68% and 38.24% respectively.

2.2 Analysis of Answers

We first analyze the percentage of answerable questions. Since VizWiz (Gurari et al., 2018) images are collected by blind people, a large portion of images have low quality and hence are not answerable. Based on our analysis, only 73.04% of visual questions are tagged as answerable.

We also analyze the diversity of answers by calculating statistics on answer length. The mean and median answer lengths are 1.66 and 1.00 words respectively, and the max answer length is 21 words. The result indicates that the visual questions in dataset tend to have short answers. We also generate a word cloud computed on all answers in the dataset excluding "unanswerable" and "unsuitable image", which is shown in Figure 2.

Lastly, we analyze the answer confidence level and agreement level. There is a total of 20523 questions in the dataset. Each question has 10 answers, and each answer has a confidence level ranging from "yes", "maybe" to "no". Among all questions, there are 18.82% questions have all 10 answers with a confidence level of "yes", 57.02% questions have at least 8 answers with a confidence level of "yes" and 94.85% questions have at least half of the answers with a confidence level of "yes".

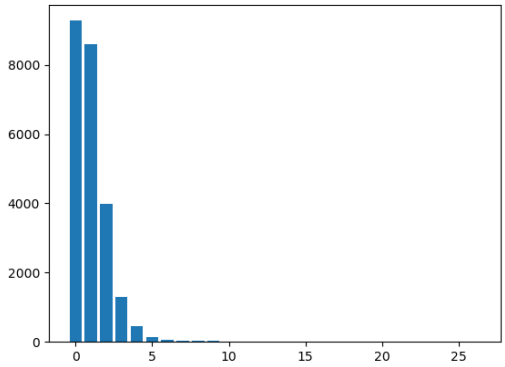

Figure 3: Number of bounding boxes

The dataset also has a pretty high human agreement level. Although we use a very strict agreement measure (exact string matching), we still observe that 40.54% of visual questions have more than 5 people agreed on the most popular answer, and 73.20% of visual questions have more than 3 people agreed on the most popular answer.

2.3 Analysis of Images

We run pre-trained YOLO (You Only Look Once) v3 model (Redmon et al., 2016) as a naive, coarse-grained, first-step baseline to analyze the training images. The weights of the model have been obtained by training on COCO dataset (Lin et al., 2015), and we hereby adapt the setting, trying to detect objects from 80 classes.

We count the number of bounding boxes on each image after we run objectness score thresholding and Non-Maximum Suppression to avoid overlapping. We observe that, we get 24,480 bounding boxes in total, and most images have either 0 (9,280 out of 23,953) or 1 (8,609 images) bounding box. The average number of bounding boxes per image is 1.02, and the third-quartile count is 2.0. The bar graph is shown in Figure 3.

After investigating the generated bounding boxes with labels, we observe that YOLO detects "Person" and "Bottle" objects most often. We also find that the coarse-grained model classify human hands as "Person" in most cases, and the 80 classes in COCO dataset are obviously insufficient.

From the experiment of running YOLO we conclude that, though most images may only contain one major object, it is difficult to successfully detect it. Ultrafine-grained semantic labels might be needed. Further study could be conducted when

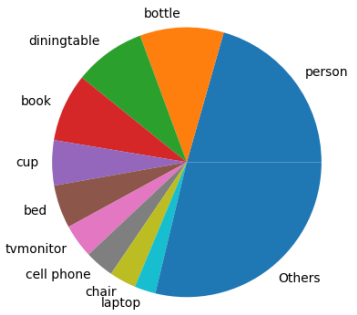

Figure 4: Predicted Classes Distribution

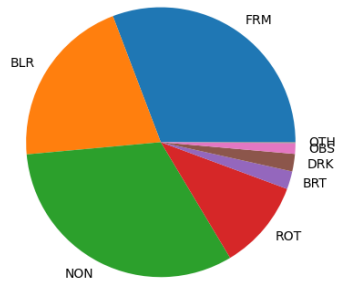

Figure 5: Images with Flaws Distribution

we incorporate the labeled image captions to our current predictions, to check where most errors are from. And we will adopt more advanced method such as Faster R-CNN (Changpinyo et al., 2019) to further explore the image data.

Apart from object detection bounding boxes analysis, we also analyze image quality. We observe that, a large number of pictures are covered by hands and many pictures are blurred. We then only focus on images that are tagged "answerable," and we find among 19,873 images that score 1 or below for category "recognizable," 52% of them have the "frame" flaw. Figure 5 shows the pie chart of flaws.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering.

Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. 2019. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.