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

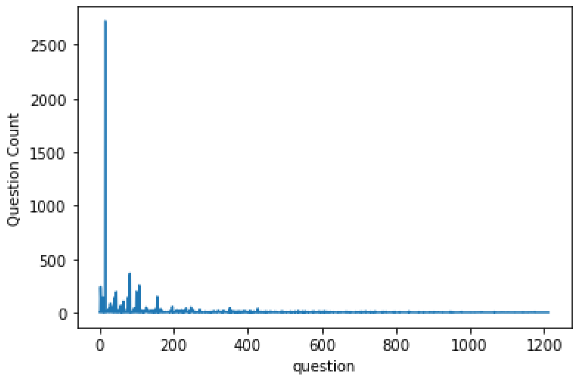

Figure 1: Question Frequency Distribution

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”.

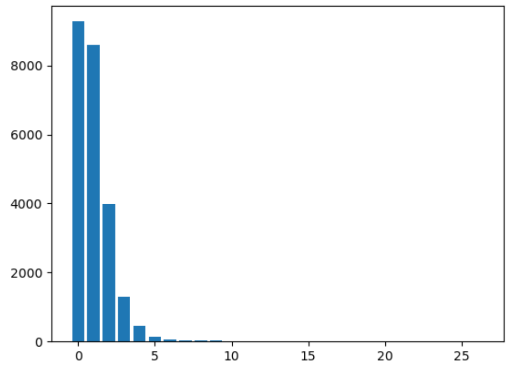

Figure 3: Number of bounding boxes

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

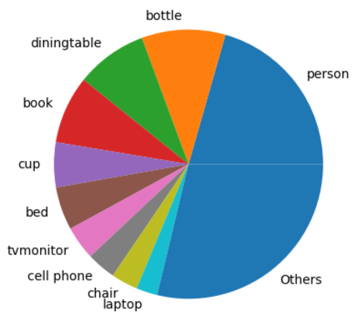

Figure 4: Predicted Classes Distribution

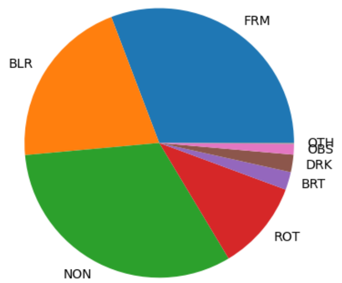

Figure 5: Images with Flaws Distribution

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.

3 Related Work and Background

3.1 Related Tasks

Over the years, there has been rapid progress in bridging vision and language in the research community. Various models have been developed and applied to a wide range of vision-and-language tasks. We'll briefly describe the problem and training objective of five well-established tasks and include relevant papers.

Visual Question Answering (VQA) has received a lot of attention in recent years. A natural language question regarding an image is presented and the model will output the correct answer to it. Various models are proposed for this specific task (Agrawal et al., 2016; Tan and Bansal, 2019; Ben-younes et al., 2017; Fukui et al., 2016).

Visual commonsense reasoning (VCR) is similar to VQA, but besides predicting the answer, the model is also expected to select the correct answer justification among multiple choices (Lu et al., 2019; Li et al., 2019; Tan and Bansal, 2019).

Visual grounding (VG) aims to localize an image region or object given a natural language question. Common approaches to this task including bounding box proposals and reranking a set of image region candidates (Lu et al., 2019; Li et al., 2019; Fukui et al., 2016; Yu et al., 2018).

Image captioning recognizes the context of an image and adds a descriptive sentence to it. Models targeting this task are mainly attention-based neural networks that could extract deep features (Anderson et al., 2018; Donahue et al., 2016; Xu et al., 2016).

Image-text matching aims to measure the visual-semantic similarity between a text and an image. It has been widely applied to other applications such as image search for a given query (Kim et al., 2018; Nam et al., 2017).

Besides these five tasks, there are also many other popular vision-and-language transfer tasks such as caption-based image retrieval (Lu et al., 2019) and emotion recognition and sentiment analysis (Delbrouck et al., 2020).

3.2 Related Techniques

In this section, we briefly review the techniques used in previous related work with a focus on the feature representation, attention mechanisms and fusion methods.

3.2.1 Feature representation

Generating feature representation has been a crucial step in transforming raw data to meaningful features. There has been substantial past works in developing single modality models separately for vision and language representations. For image representations, common backbone models include Faster R-CNN (Ren et al., 2016), R2+1D-(152)

(Tran et al., 2015), VGG (Simonyan and Zisserman, 2015), ResNet (He et al., 2015) and fishNet (Sun et al., 2019). In terms of language representations, BERT (Devlin et al., 2019), a transformer language model pretrained on the BookCorpus (Zhu et al., 2015) and English Wikipedia is a popular choice nowadays.

Language embeddings and image embeddings are then combined through summation or multiplication (Lu et al., 2019; Delbrouck et al., 2020) to form a joint representation for fine tuning at a later stage. Another common approach to generate joint representation involves early fusion of image and text. Models are pretrained with multi-modal objectives such as masked language modeling with image and sentence-image prediction (Lu et al., 2019; Li et al., 2019; Tan and Bansal, 2019) to directly generate joint representations.

3.2.2 Attention Mechanism

Attention mechanism has been proven effective in many tasks including VQA. Previous attention approaches commonly used in VQA or related tasks can be classified into following four categories:

- Self attention that aggregate information inside each modality by query-key-value attention mechanism (Vaswani et al., 2017).
- Question-guided visual attention (Xu et al., 2016; Xu and Saenko, 2016) that compute attention on image region.
- Co-attention that jointly reason about both question and visual attention to interact across the two modalities. It uses image representation to guide the question attention and question representation to guide image attention (Lu et al., 2017, 2019; Yu et al., 2019). There also exist many variants of co-attention such as bilinear attention (Kim et al., 2018) and dense co-attention network (DCN) (Nguyen and Okatani, 2018) that considers interactions between every pair of question words and image regions.
- The intra- & inter-modal attention (DFAF) (Peng et al., 2019) that consider both inter-modality attention and intra-modality attention, where attention for intra-modality feature aggregation is dynamically modulated by the other modality using the pooled features.

3.2.3 Fusion Methods

The common multi-modal fusion approach is that visual and language features are extracted from the image and question independently as the first step, and then they are fused to compute the final results.

In previous studies, many works employed simple fusion methods that use element-wise product or summation of the visual and language features as final fused feature, and fed it to fully connected layers to predict results. (Lu et al., 2019; Delbrouck et al., 2020)

More complex fusion techniques have also been explored. For instance, Multimodal Compact Bilinear pooling (MCB) (Fukui et al., 2016) uses bilinear pooling method to compute the outer product between visual and language vectors as their fusion. In contrast to element-wise product approach, MCB allows a multiplicative interaction between all elements of both vectors. MUTAN fusion model (Ben-younes et al., 2017) is able to represent full bilinear interactions between visual and language modalities using Tucker decomposition of the correlation tensors, while maintaining the size of the model tractable. Multi-modality Latent Interaction Network (MLIN) (Gao et al., 2019) first encodes question and visual features into latent summarization vectors and then it go through the process of interaction, propagation and aggregation to achieve multi-modality reasoning.

4 Baselines

We would run the following 4 models as baselines.

- LXMERT (Tan and Bansal, 2019) constructs separate representations for images and texts, use (Anderson et al., 2018) for images. Have two encoders each for one modality with self-attention, and then a cross-modality encoder with cross-attention.
- UNITER (Chen et al., 2020) has two embedders and three ”uniters” for image/text representation. Uses different pretraining techniques and fewer sentence-image pairs then LXMERT does. Views VQA as a multi-label classification problem in implementation. Assigns soft target scores to most frequent answers based on human responses.
- Pythia (Singh et al., 2019) proposes Look, Read, Reason, Answer mechanism, use spatial attention, lstm and OCR in the model.

Method	Binary (yes/no)				Number				Overall	
	Framing	Blur	None	**Accu**	Framing	Blur	None	**Accu**	None	**Accu**
UNITER	-	-	-	-	-	-	-	-	-	-
Pythia	-	-	-	-	-	-	-	-	-	54.72
Gail-VisQA-Ultra	-	-	-	68.12	-	-	-	**28.81**	-	53.68
LXMERT	-	-	-	74.00	-	-	-	24.76	-	**55.40**
LXMERT-PIR	-	-	-	-	-	-	-	-	-	-

Table 1: Test-set results. last row: LXMERT with Multi-stage Progressive Image Restoration

- Decoupled Box (Changpinyo et al., 2019) avoids labeling bounding boxes by decoupling box proposal and featurization on downstream tasks.

Since their pre-trained models are all available and we are interested in how their different attention / representation / fusion methods would effect the performance. Their overall results on VizWiz dataset are listed in table 1. All metrics follow VizWiz official setting, i.e.

$$accu = min(1, \frac{\textit{\# humans that provided that answer}}{3})$$

As table 1 suggested, we are particularly interested in each model's performance with flawed images, and we mainly focus on binary (yes or no) questions and questions related to numbers. From VizWiz challenge leader board we observe that current SoTA model achieves 56.33% accuracy, but most models have relatively weaker performance in answering questions related to "numbers," with only 27.1% accuracy. We first aim to explore this discrepancy.

According to our data analysis, many input images are flawed. Blurring and framing issues are the most common types of flaw. Therefore, we expect to produce model results for each sub-category of image quality issues shown in table. 1

Aiming to boost performance with flawed images, we would test different fusing methods that inject de-noising technique, such as applying multi-stage progressive image restoration (Zamir et al., 2021) to our baseline models. Additionally, We would evaluate alternative representation / encoding / fusion methods following VizWiz official evaluation metric.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering.

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.

Hedi Ben-younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for 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.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*.

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

Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. 2016. Long-term recurrent convolutional networks for visual recognition and description.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding.

Peng Gao, Haoxuan You, Zhanpeng Zhang, Xiaogang Wang, and Hongsheng Li. 2019. Multi-modality latent interaction network for visual question answering.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

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.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2017. Hierarchical question-image co-attention for visual question answering.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching.

Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering.

Gao Peng, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic fusion with intra- and inter- modality attention flow for visual question answering.

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

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read.

Shuyang Sun, Jiangmiao Pang, Jianping Shi, Shuai Yi, and Wanli Ouyang. 2019. Fishnet: A versatile backbone for image, region, and pixel level prediction.

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.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, attend and tell: Neural image caption generation with visual attention.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering.

Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. 2018. Rethinking diversified and discriminative proposal generation for visual grounding.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. 2021. Multi-stage progressive image restoration.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.