VizWiz Grand Challenge: Answering Visual Questions from Blind People

Danna Gurari[1], Qing Li[2], Abigale J. Stangl[3], Anhong Guo[4], Chi Lin[1],
Kristen Grauman[1], Jiebo Luo[5], and Jeffrey P. Bigham[4]
[1] University of Texas at Austin, [2] University of Science and Technology of China,
[3] University of Colorado Boulder, [4] Carnegie Mellon University [5] University of Rochester

Abstract

The study of algorithms to automatically answer visual questions currently is motivated by visual question answering (VQA) datasets constructed in artificial VQA settings. We propose VizWiz, the first goal-oriented VQA dataset arising from a natural VQA setting. VizWiz consists of over 31,000 visual questions originating from blind people who each took a picture using a mobile phone and recorded a spoken question about it, together with 10 crowdsourced answers per visual question. VizWiz differs from the many existing VQA datasets because (1) images are captured by blind photographers and so are often poor quality, (2) questions are spoken and so are more conversational, and (3) often visual questions cannot be answered. Evaluation of modern algorithms for answering visual questions and deciding if a visual question is answerable reveals that VizWiz is a challenging dataset. We introduce this dataset to encourage a larger community to develop more generalized algorithms that can assist blind people.

1. Introduction

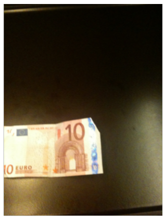
A natural application of computer vision is to assist blind people, whether that may be to overcome their daily visual challenges or break down their social accessibility barriers. For example, modern object recognition tools from private companies, such as TapTapSee [3] and CamFind [2], already empower people to snap a picture of an object and recognize what it is as well as where it can be purchased. In addition, social media platforms, such as Facebook and Twitter, provide people a way to maintain connections with friends by enabling them to identify and tag friends in posted images as well as respond to images automatically described to them [26, 41]. A desirable next step for vision applications is to empower a blind person to directly request in a natural manner what (s)he would like to know about the surrounding physical world. This idea relates to the recent explosion of interest in the generic visual question answering (VQA) problem, which aims to accurately answer any question about any image.

Over the past three years, 14 VQA datasets have emerged in the vision community to catalyze research on the VQA problem [6, 7, 16, 17, 19, 20, 23, 28, 32, 39, 40, 43, 45]. Historically, progress in the research community on a given computer vision problem is typically preceded by a large-scale, publicly-shared dataset [12, 25, 30, 33, 42]. However, a limitation of available VQA datasets is that all come from artificially created VQA settings. Moreover, none are "goal oriented" towards the images and questions that come from blind people. Yet, blind people arguably have been producing the big data desired to train algorithms. For nearly a decade, blind people have been both taking pictures [4, 8] and asking questions about the pictures they take [8, 11, 24]. Moreover, blind people often are early adopters of computer vision tools to support their *real* daily needs.

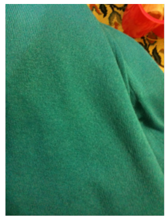
We introduce the first publicly-available vision dataset originating from blind people, which we call "VizWiz", in order to encourage the development of more generalized algorithms that also address the interests of blind people. Our work builds off previous work [8] which established a mobile phone application that supported blind people to ask over 70,000 visual questions [10] by taking a photo and asking a question about it. We begin our work by implementing a rigorous filtering process to remove all visual questions that could risk compromising the safety and/or privacy of any individuals associated with them, since blind people often willingly share personal information with strangers to overcome personal obstacles [5]. We then crowdsource answers to support algorithm training and evaluation. We next conduct experiments to characterize the images, questions, and answers and uncover unique aspects differentiating VizWiz from the many existing VQA datasets [6, 7, 16, 17, 19, 20, 23, 28, 32, 39, 40, 43, 45]. We finally evaluate numerous algorithms for predicting answers [17, 22] and predicting if a visual question can be answered [27]. Our findings highlight VizWiz is a difficult dataset for modern vision algorithms and offer new perspectives about the VQA problem.

It is also useful to understand why VizWiz is challenging

Q: Does this foundation have any sunscreen?
A: yes

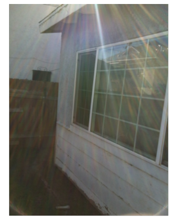


Q: What is this?
A: 10 euros

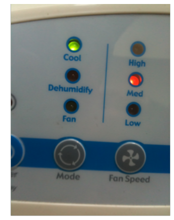


Q: What color is this?
A: green

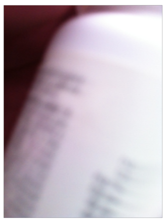
Q: Please can you tell me what this item is?
A: butternut squash red pepper soup



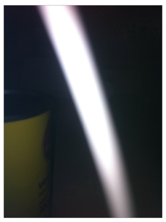
Q: Is it sunny outside?
A: yes



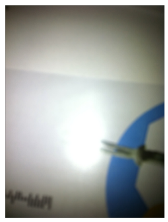
Q: Is this air conditioner on fan, dehumidifier, or air conditioning?
A: air conditioning



Q: What type of pills are these?
A: unsuitable image

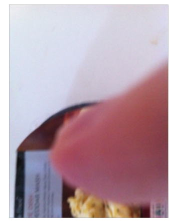


Q: What type of soup is this?
A: unsuitable image



Q: Who is this mail for?
A: unanswerable

Q: When is the expiration date?
A: unanswerable



Q: What is this?
A: unanswerable

Q: Can you please tell me what the oven temperature is set to?
A: unanswerable

Figure 1. Examples of visual questions asked by blind people and corresponding answers agreed upon by crowd workers. The examples include questions that both can be answered from the image (top row) and cannot be answered from the image (bottom row).

for modern algorithms. Our findings suggest the reasons stem from the fact VizWiz is the first vision dataset to introduce images and questions from blind people as well as questions that originally were spoken. Unlike existing vision datasets, images are often poor quality, including due to poor lighting, focus, and framing of the content of interest. Unlike existing VQA datasets, the questions can be more conversational or suffer from audio recording imperfections such as clipping a question at either end or catching background audio content. Finally, there is no assurance that questions can be answered since blind people cannot verify their images capture the visual content they are asking about for a plethora of reasons; e.g., blur, inadequate lighting, finger covering the lens, etc. Several of the aforementioned issues are exemplified in **Figure 1**.

More broadly, VizWiz is the first goal-driven VQA dataset to capture real-world interests of real users of a VQA system. Furthermore, it is the first VQA dataset to reflect a use case where a person asks questions about the physical world around himself/herself. This approach is critical for empowering blind people to overcome their daily visual-based challenges. Success in developing automated methods would mitigate concerns about the many undesired consequences from today's status quo for blind people of relying on humans to answer visual questions [8, 11, 24]; e.g., humans often must be paid (i.e., potentially expensive), can take minutes to provide an answer (i.e., slow), are not always available (i.e., potentially not scalable), and pose privacy issues (e.g., when credit card information is shared).

We introduce the first publicly-available vision dataset originating from blind people, which we call "VizWiz", in order to encourage the development of more generalized algorithms that also address the interests of blind people. Our work builds off previous work [8] which established a mobile phone application that supported blind people to ask over 70,000 visual questions [10] by taking a photo and asking a question about it. We begin our work by implementing a rigorous filtering process to remove all visual questions that could risk compromising the safety and/or privacy of any individuals associated with them, since blind people often willingly share personal information with strangers to overcome personal obstacles [5]. We then crowdsource answers to support algorithm training and evaluation. We next conduct experiments to characterize the images, questions, and answers and uncover unique aspects differentiating VizWiz from the many existing VQA datasets [6, 7, 16, 17, 19, 20, 23, 28, 32, 39, 40, 43, 45]. We finally evaluate numerous algorithms for predicting answers [17, 22] and predicting if a visual question can be answered [27]. Our findings highlight VizWiz is a difficult dataset for modern vision algorithms and offer new perspectives about the VQA problem.

2. Related Works

VQA for Blind Users. For nearly a decade, human-powered VQA systems have enabled blind people to overcome their daily visual challenges quickly [1, 8, 24]. With such systems, users employ a mobile phone application to capture a photo (or video), ask a question about it, and then receive an answer from remotely located paid crowd workers [8, 24] or volunteers [1]. Such VQA systems have been shown to be valuable for many daily tasks including grocery shopping [8], locating a specific object in a complex scene [9], and choosing clothes to wear [11]. Yet, these systems are limited because they rely on humans to provide answers. An automated solution would be preferred for reasons such as cost, latency, scalability, and enhanced privacy. For example, the latency between sending out an image and getting the answer back may take minutes [8], disrupting the natural flow of a blind user's life and career. Our work describes the unique challenges for creating public datasets with data captured in natural settings from real-world users and, in particular, blind users. Our work also offers the first dataset for enabling algorithm development on images and questions coming from blind people, which in turn yields new vision-based and language-based challenges.

Images in Vision Datasets. When constructing vision datasets, prior work typically used images gathered from the web (e.g., [12, 25, 30, 33, 42]) or created artificially (e.g., [6, 7, 19]). Such images are typically high quality and safe for public consumption. For example, images curated from the web intrinsically pass a human quality assessment of "worthy to upload to the internet" and typically are internally reviewed by companies hosting the images (e.g., Google, Facebook) to ensure the content is appropriate. Alternatively, artificially constructed images come from controlled settings where either computer graphics is employed to synthesize images with known objects and scenes [6, 19] or crowd workers are employed to add any pre-defined clipart objects to pre-defined indoor and outdoor scenes [7]. In contrast, images collected "in the wild" can contain inappropriate or private content, necessitating the need for a review process before releasing the data for public consumption. Moreover, images from blind photographers regularly are poor quality, since blind people cannot validate the quality of the pictures they take. Our experimental findings show that such poor quality images pose new challenges for modern vision algorithms.

VQA Datasets. Over the past three years, a plethora of VQA datasets have been publicly shared to encourage a larger community to collaborate on developing algorithms that answer visual questions [6, 7, 16, 17, 19, 20, 23, 28, 32, 39, 40, 43, 45]. While a variety of approaches have been proposed to assemble VQA datasets, in all cases the visual questions were contrived. For example, all images were either taken from an existing vision dataset (e.g., MSCOCO [25]) or artificially constructed (e.g., Abstract Scenes [7], computer graphics [6, 19]). In addition, questions were generated either automatically [6, 19, 20, 28, 32, 43], from crowd workers [7, 16, 17, 20, 23, 45], or from in-house participants [20, 40]. We introduce the first VQA dataset which reflects visual questions asked by people who were authentically trying to learn about the visual world. This enables us to uncover the statistical composition of visual questions that arises in a real-world situation. Moreover, our dataset is the first to reflect how questions appear when they are spoken (rather than automatically generated or typed) and when each image and question in a visual question is created by the same person. These differences reflect a distinct use case scenario where a person interactively explores and learns about his/her surrounding physical world. Our experiments show the value of VizWiz as a difficult dataset for modern VQA algorithms, motivating future directions for further algorithm improvements.

Answering Visual Questions. The prevailing assumption when collecting answers to visual questions is that the questions are answerable from the given images [6, 7, 16, 17, 19, 23, 28, 32, 40, 39, 43, 45]. The differences when constructing VQA datasets thus often lies in whether to collect answers from anonymous crowd workers [6, 7, 16, 20, 23], automated methods [19, 28], or in-house annotators [28, 39, 40]. Yet, in practice, blind people cannot know whether their questions can be answered from their images. A question may be unanswerable because an image suffers from poor focus and lighting or is missing the content of interest. In VizWiz, $\sim$28% of visual questions are deemed unanswerable by crowd workers, despite the availability of several automated systems designed to guide blind photographers to improve the image focus [3], lighting [8], or composition [18, 37, 44].

We propose the first VQA dataset which naturally promotes the problem of predicting whether a visual question is answerable. We construct our dataset by explicitly asking crowd workers whether a visual question is answerable when collecting answers to our visual questions. Our work relates to recent "relevance" datasets which were artificially constructed to include irrelevant visual questions by injecting questions that are unrelated to the contents of high quality images [20, 27, 31, 36]. Unlike these "relevance" datasets, our dataset also includes questions that are unrelated because images are too poor in quality (e.g., blur, over/under-saturation). Experiments demonstrate VizWiz is a difficult dataset for the only freely-shared algorithm [27] designed to predict whether a visual question is relevant, and so motivates the design of improved algorithms.

3. VizWiz: Dataset Creation

We introduce a VQA dataset we call "VizWiz", which consists of visual questions asked by blind people in a real world setting where the people were seeking answers to their daily visual questions. This dataset is built off of previous work [8] which accrued 72,205 visual questions over

Dataset	Which Images?	Who Asked?	How Asked?
DAQUAR [28]	NYU Depth V2 [34]	Automatically generated (templates)	———
VQA v1.0: Abstract [7]	Abstract Scenes	Crowd workers (AMT)	Typed
VQA v1.0: Real [7]	MSCOCO [25]	Crowd workers (AMT)	Typed
Visual Madlibs [43]	MSCOCO [25]	Automatically generated (templates)	———
FM-IQA [16]	MSCOCO [25]	Crowd workers (Baidu)	Typed
KB-VQA [40]	MSCOCO [25]	In-house participants	Typed
COCO-QA [32]	MSCOCO [25]	Automatically generated (captions)	———
VQA v2.0: Real [17]	MSCOCO [25]	Crowd workers (AMT)	Typed
Visual7W [45]	MSCOCO [25]	Crowd workers (AMT)	Typed
CLEVR [19]	Synthetic Shapes	Automatically generated (templates)	———
SHAPES [6]	Synthetic Shapes	Automatically generated (templates)	———
Visual Genome [23]	MSCOCO [25] & YFCC100M [35]	Crowd workers (AMT)	Typed
FVQA [39]	MSCOCO [25] & ImageNet [14]	In-house participants	Typed
TDIUC [20]	MSCOCO [25] & YFCC100M [35]	Crowd workers (AMT), In-house participants, Automatically generated	Typed
Ours - VizWiz	Blind people use mobile phones to take a picture and ask question		Spoken

Table 1. Comparison of visual questions from 14 existing VQA datasets and our new dataset called VizWiz.

four years using the VizWiz mobile application[1], an application available for iPhone and Android mobile phone platforms. Blind people used this application to ask about their daily visual accessibility challenges [8, 10]. A person asked a visual question by taking a picture and then recording a spoken question. The VizWiz application was released May 2011, and used by 11,045 users. 48,169 of the collected visual questions were asked by users who agreed to have their visual questions anonymously shared. These visual questions serve as the starting point for the development of our dataset. We begin this section by comparing the approaches for asking visual questions in VizWiz with those employed to construct the many existing VQA datasets. We then describe how we created the dataset.

3.1. Visual Question Collection Analysis

We summarize in **Table 1** how the process of collecting visual questions for VizWiz is unlike the processes employed for 14 existing VQA datasets. A clear distinction is that VizWiz contains images from blind photographers. The quality of such images offer challenges not typically observed in existing datasets, such as significant amounts of image blur, poor lighting, and poor framing of image content. Another distinction is that questions are spoken. Speaking to technology is increasingly becoming a standard interaction approach for people with their technology (e.g., Apple's Siri, Google Now, Amazon's Alexa) and VizWiz yields new challenges stemming from this question-asking modality, such as more conversational language and audio recording errors. A further distinction is VizWiz is the first dataset where a person both takes the picture and then asks a question about it. This reflects a novel use-case scenario in which visual questions reflect people's daily interests about their physical surroundings explored with their mobile devices. VizWiz is also unique because, in contrast to all other VQA datasets, the people asking the questions could not "see" the images. Consequently, questions could often be unrelated to the images for a variety of reasons that are exemplified in **Figure 1**.

[1]http://www.vizwiz.org

3.2. Anonymizing and Filtering Visual Questions

We faced many challenges with preparing the dataset for use by the research community because our visual questions were collected "in the wild" from real world users of a VQA system. The challenges related to protecting the privacy and safety of the many individuals involved with the dataset. This is especially important for visually impaired people, because they often make the tradeoff to reveal personal information to a stranger in exchange for assistance [5]; e.g., credit card numbers and personal mail. This is also important for those reviewing the dataset since visual questions can contain "adult-like" (e.g., nudity), and so potentially offensive content. Our key steps to finalize our dataset for public use involved anonymizing and filtering candidate visual questions.

Anonymization. Our aim was to eliminate clues that could reveal who asked the visual question. Accordingly, we removed the person's voice from the question by employing crowd workers from Amazon Mechanical Turk to transcribe the audio recorded questions. We applied a spellchecker to the transcribed sentences to fix misspellings. We also re-saved all images using lossless compression in order to remove any possible meta-data attached to the original image, such as the person's location.

Filtering. Our aim also was to remove visual questions that could make either the producers (e.g., askers) or consumers (e.g., research community) of the dataset vulnerable. As part of this task, we obtained from two committees that review proposed research to ensure it is ethical – the Collaborative Institutional Training Initiative (CITI) board and Institutional Review Board (IRB) – approval to publicly release the filtered dataset. Our team for this task brought extensive experience conducting human subject studies, including with blind people.

We initiated this work by developing a taxonomy of vulnerabilities (see Supplementary Materials for details). We identified the following categories that came from erring on the safe side to protect all people involved with the dataset:

1. Personally-Identifying Information (PII); e.g., any part of a person's face, financial statements, prescriptions.
2. Location; e.g., addressed mail, business locations.
3. Indecent Content; e.g., nudity, profanity.
4. Suspicious Complex Scenes: the reviewer suspects PII may be located in the scene but could not locate it.
5. Suspicious Low Quality Images: the reviewer suspects image processing to enhance images could reveal PII.

We next performed two rounds of filtering. We first instructed AMT crowd workers to identify all images showing PII, as reflected by "any part of a person's face, anyone's full name, anyone's address, a credit card or bank account number, or anything else that you think would identify who the person who took the photo is".[2] Then, two of the in-house domain experts who established the vulnerability taxonomy jointly reviewed all remaining visual questions and marked any instances for removal with one of the five vulnerability categories or "Other". This phase also included removing all instances with a missing question (i.e., 7,477 visual questions with less than two words in the question).

Table 2 shows the resulting number of visual questions tagged for removal in each round of human review, including a breakdown by vulnerability issue. We attribute the extra thousands of flagged visual questions from domain experts to their better training on the potential vulnerabilities. For example, location information, such as zip codes and menus from local restaurants, when augmented with additional information (e.g., local libraries have lists of blind members in the community) could risk exposing a person's identity. Also, blurry and/or bright images, when post-processed, could reveal PII. Additionally, people's faces can appear in reflections on monitor screens, window pains, etc. We do not expect crowd workers to understand such nuances without extensive instructions and training.

[2]An important distinction when crowdsourcing visual questions gathered "in the wild" is that it is critical to ensure the crowd is comfortable reviewing adult content (e.g., apply an "Adult Qualification" in AMT). The alternative, which we experienced, is that the crowdsourcing company may cancel your jobs in response to complaints about indecent content.

Filter	# of VQs
Crowd Workers	4,626
In-House Experts	2,693
- PII	895
- Location	377
- Indecent Content	55
- Suspicious Complex Scene	725
- Suspicious Low Quality Image	578
- Other	63

Table 2. We report the number of visual questions filtered in our iterative review process by crowd workers and then in-house domain experts (including with respect to each vulnerability category).

In total, ∼31% of visual questions (i.e., 14,796) were filtered from the original 48,169 candidate visual questions. While our taxonomy of vulnerabilities helps guide what visual questions to filter from real-world VQA datasets, it also reveals valuable visual questions to artificially generate to augment VQA datasets so that such datasets still address the needs of blind people, yet without relying on their data.

3.3. Collecting Answers

We next collected answers for our final set of 33,373 visual questions. The original VizWiz application prioritized providing a person near real-time access to answers and permitted the person to receive answers from crowd workers, IQ Engines, Facebook, Twitter, or email. Since our aim is to enable the training and evaluation of algorithms, we collected new answers to all visual questions for this purpose.

To collect answers, we modified the excellent protocol used for creating VQA 1.0 [7]. As done before, we collected 10 answers per visual question from AMT crowd workers located in the US by showing crowd workers a question and associated image and instructing them to return "a brief phrase and not a complete sentence". We augmented this user interface to state that "you will work with images taken by blind people paired with questions they asked about the images". We also added instructions to answer "Unsuitable Image" if "an image is too poor in quality to answer the question (i.e., all white, all black, or too blurry)" or "Unanswerable" if "the question cannot be answered from the image". Consequently, answers capture basic common sense from lay people who can see and are fluent in English. Moreover, answers reveal whether a visual question is answerable. We then post-processed and spell-checked[3] each answer (details in Supplementary Materials).

4. VizWiz: Dataset Analysis

Our aim in this section is to characterize the visual questions and answers in VizWiz. We analyze (1) What is the diversity of natural language questions?, (2) What is the diversity of images?, (3) What is the diversity of answers?, and (4) How often are visual questions unanswerable? A valuable outcome of this analysis is it enriches our understanding of the interests of blind users in a real VQA set-up.

4.1. Analysis of Questions

We first examine the diversity of questions asked by visualizing the frequency that questions begin with different words/phrases. Results are shown in a sunburst diagram in **Figure 2**. While many existing VQA datasets include a small set of common initial words (e.g., “What”, “When”, “Why”, “Is”, “Do”), we observe from the upper left quadrant of **Figure 2** that VizWiz often begins with a rare first word. In fact, the percentage of questions starting with a first word that occurs for less than 5% of all questions is 24.2% for VizWiz versus 13.4% for VQA 2.0 [7] (based on random subset of 40,000 VQs). We attribute this finding partially to the use of more conversational language when speaking a question; e.g., “Hi”, “Okay”, and “Please”. We also attribute this finding to the recording of the question starting after the person has begun speaking the question; e.g., “Sell by or use by date of this carton of milk” or “oven set to thanks?”. Despite such questions being incomplete, it is still reasonable the intended question can be inferred and so answered; e.g., “What is the oven set to?” We also observe in **Figure 2** that most questions begin with “What”. This suggests the initial words in VizWiz often do a poor job in narrowing the scope of plausible answers (and so meaningful approaches for answering). In contrast, initial wordings such as “How many...” and “Is...” conveniently narrow plausible answers to numbers and “yes/no” respectively. Yet, “yes/no” and “number” answers constitute only 2.02% and 1.65% of visual questions in VizWiz.

We also analyze question diversity by computing statistics summarizing the number of words in each question. The median and mean question lengths are six words and eight words respectively and 25th and 75th percentile lengths are four and eight words respectively. Our findings show the statistics found in the existing artificially constructed VQA datasets, nicely summarized in [13] and [20], match statistics observed in practice. We also observe three words can often suffice for a question: “What is this?”.

[3]Note that manual review of automatically detected spelling errors is particularly important in this dataset since tasks, such as reading captchas, often include invalid words.

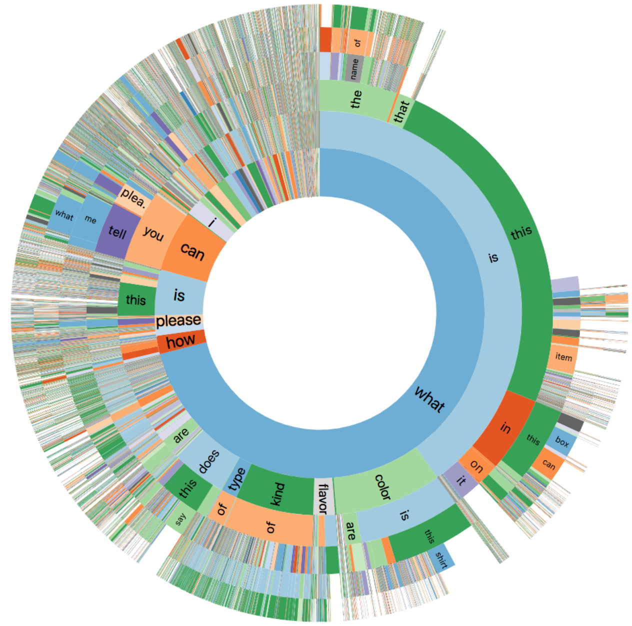


Figure 2. Distribution of the first six words for all questions in VizWiz. The innermost ring represents the first word and each subsequent ring represents a subsequent word. The arc size is proportional to the number of questions with that initial word/phrase.

As observed in **Figure 2**, this short object recognition question is the most common question. Longer questions and multi-sentence questions also occasionally arise, typically because people offer auxiliary information to disambiguate what is the desired response; e.g., “Which one of these two bags would be appropriate for a gift? The small one or the tall one? Thank you.” Longer questions also can arise if the audio recording device captures too much content or captures background audio content; e.g., “I want to know what this is. I’m have trouble stopping the recordings.”

4.2. Analysis of Images

We next investigate the diversity of images. In particular, a valid concern is our dataset has high quality images showing a single, iconic object since our filtering process erred on removing “suspicious” scene-based and blurry images from our dataset paired with containing many object recognition questions (see **Figure 2**). Following prior work [14], we first computed the average image from all images in VizWiz. **Figure 3** shows the result. As desired from a diverse dataset, the resulting gray image confirms our dataset does not conform to a particular structure across all the images. We also tallied how many images had at least two crowd workers give the answer “unsuitable image”. We found 28% of images were labelled as such.

Figure 3. The average image created using all images in VizWiz.

Figure 4. Popularity of all answers in VizWiz, with the text size proportional to the number of times the answer occurs.

4.3. Analysis of Answers

We next analyze the diversity of the answers. We first visualize the popularity of different answers in **Figure 4** using a word map (cropped to fit in the paper), excluding the answers "Unanswerable" and "Unsuitable Image". This visually highlights the fact that there are a large number of unique answers; i.e., ∼54,253.[4] While in absolute terms this number is an order of magnitude smaller than existing larger-scale datasets such as VQA 2.0 [7], we find the answer overlap with existing datasets can be low. For example, only 824 out of the top 3,000 answers in VizWiz are included in the top 3,000 answers in VQA 2.0 [7]. This observation is used in the next section to explain why existing prediction systems perform poorly on the VizWiz dataset.

We also tally how often a visual question is unanswerable, as indicated by at least half the crowdsourced answers for a visual question stating the answer is "unanswerable" or "unsuitable image". We find 27.9% of visual questions are not answerable. This finding validates the practical importance of the recent efforts [20, 27, 31, 36] to augment VQA datasets with irrelevant visual questions. Moreover, our dataset offers more fine-grained annotations that enable research to automatically identify whether the answerability issue is due to inadequate image quality (e.g., "Unsuitable Image") or image composition (i.e., "Unanswerable").

We also analyze answer diversity by computing statistics summarizing the number of words in each answer. The median and mean answer lengths are 1.0 words and 1.8 words respectively. This shows the answer statistics observed for numerous artificially constructed VQA datasets, as nicely summarized in [13] and [20], resemble what is observed in practice. We also compute the percentage of answers with different answer lengths: 52.1% have one word, 32.9% have two words, 8% have three words, 3.4% have four words, and the remaining 3.6% have more than four words. Interestingly, our answers are longer on average than that observed by Antol et al. [7], who used a similar crowdsourcing approach. We attribute this discrepancy in part to the observation that VizWiz visual questions often request someone to read multi-word text.

We finally compute the level of human agreement on answers. Despite the fact that crowd workers provide open-ended text as answers, we observe agreement from independent people occurs for nearly all visual questions (i.e., 97.7%). More than three people agreed on the most popular answer for 72.4% of visual questions, exactly three people agreed for 15.5% of visual questions, and exactly two people agreed for 9.8% of visual questions.

5. VizWiz Benchmarking

We now investigate the difficulty of the VizWiz dataset for existing algorithms. We divide the dataset into training, validation, and test sets of 20,000, 3,173, and 8,000 visual questions, respectively (i.e, approximately a 65/10/25 split).[5] All results below are reported for the test dataset.

5.1. Visual Question Answering

We first evaluate the performance of modern VQA algorithms and their variants for predicting answers to the visual questions in the VizWiz dataset.

Baselines. We benchmark six methods. Included are two modern, top-performing VQA methods [17, 22], which we refer to as `Q+I` [22] and `Q+I+A` [17]. Both of these baselines are trained on the VQA V2.0 dataset [17] to predict the 3,000 most frequent answers in the training dataset. The former method [17] relies on image and question information alone to make predictions whereas the latter method [22] augments an attention mechanism to specify which regions in an image should be the primary focus. We introduce two fine-tuned classifiers built upon on the two aforementioned networks (see Supplementary Materials for fine-tuning details), which we refer to as `FT` [17] and `FT` [22]. We also train the two aforementioned networks [17, 22] from scratch using VizWiz data alone, and we refer to these as `VizWiz` [17] and `VizWiz` [22].

Evaluation Metrics. We employ five metrics commonly used for evaluating VQA and image description algorithms: Accuracy [7], CIDEr [38], BLEU4 [29], and ME-

[4] See Supplementary Materials for a plot showing the cumulative coverage of all answers by the most frequent answers.

[5] Practical issues led to a dataset with ∼2,000 fewer visual questions.

TEOR [15]. The accuracy method [7] was introduced based on the observation that most answers in the VQA 1.0 dataset were one word in length. However, since nearly half of the answers in VizWiz exceed one word in length, we augment the traditional image description evaluation metrics which are designed to evaluate longer phrases and/or sentences.

Results. We first analyze how existing prediction models [17, 22] perform on the VizWiz test set. As observed in the first two rows of **Table 3**, these models perform poorly, as indicated by low values for all metrics; e.g., ~0.14 accuracy for both algorithms. We attribute the poor generalization of these algorithms largely to their inability to predict the answers observed in the VizWiz dataset; i.e., only 824 out of the top 3000 answers in VizWiz are included in the dataset (i.e., VQA 2.0 [17]) used to train both models.

We observe in **Table 3** that fine-tuning (i.e., rows 3 and 4) and training from scratch (i.e., rows 5 and 6) yield significant performance improvements over relying on the two prediction models [17, 22] as is. Interestingly, training from scratch outperforms fine-tuning, despite the relatively small number of training examples in VizWiz. We hypothesize training from scratch has an advantage because it is not retaining knowledge about answer categories that are not applicable in this setting.

In all three variants of the two algorithms, we found the model that augments the attention mechanism [22] consistently is a better predictor than the the model that relies on the image and question information alone [17]. However, the improvements are relatively small on the VizWiz dataset compared to improvements typically observed on VQA datasets. We attribute this distinction in part to our observation that images in VizWiz usually include fewer objects than the images in MSCOCO and so less commonly need to attend to a fine-grained image region. We also suspect the attention mechanism struggles on VizWiz images because such it was not intentionally designed for the low quality images that are typical from blind photographers.

Method	Acc	CIDEr	BLEU	METEOR
Q+I [17]	0.137	0.224	0.000	0.078
Q+I+A [22]	0.145	0.237	0.000	0.082
FT [17]	0.466	0.675	0.314	0.297
FT [22]	**0.469**	**0.691**	0.351	0.299
VizWiz [17]	0.465	0.654	0.353	0.298
VizWiz [22]	**0.469**	0.661	**0.356**	**0.302**

Table 3. Performance of six methods with respect to four evaluation metrics for the open-ended task of answering visual questions on the VizWiz test dataset. We show results for three variants of two VQA methods [17, 22]: use models as is, fine-tune them (FT), or re-train them using only VizWiz data (VizWiz). The two VQA methods are based on different combinations of information coming from the image (I), question (Q), and attention models (A).

5.2. Visual Question Answerability

We next turn to the question of how accurately an algorithm can classify a visual question as answerable or not using only the image and question.

Baselines. We benchmark eight methods. We include three variants of the only publicly-available method for predicting when a question is not relevant for an image [27]. This method uses NeuralTalk2 [21] pre-trained on the MSCOCO captions dataset [25] to generate a caption for each image. The algorithm then measures the similarity between the proposed caption and the question to predict a relevance score. The model is trained on the QRPE dataset [27]. As done above, we use the model as is (i.e., `Q+C` [27]), introduce a variant of the method that is fine-tuned to the VizWiz data (i.e., `FT` [27]), and introduce a variant trained from scratch on the VizWiz data only (i.e., `VizWiz` [27]). We also employ our top-performing VQA algorithm by using its output probability that the predicted answer is "unanswerable" (`VQA` [17]). We enrich our analysis by further investigating in the influence of different features on the predictions: question alone (i.e., `Q`), caption alone (i.e., `C`), image alone using the ResNet-152 CNN features (i.e., `I`), and the question with the image (i.e., `Q+I`).

Evaluation Metrics. We report the performance of each method to predict if a visual question is not answerable using a precision-recall curve. We also report the average precision (AP); i.e., area under a precision-recall curve.

Results. **Figure 5** shows the resulting precision-recall curves. As observed, all our proposed methods outperform

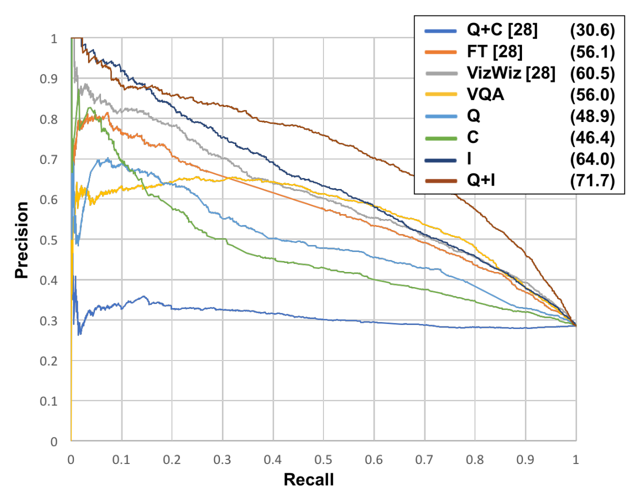


Figure 5. Precision-recall curves and average precision scores for the answerability models tested on the VizWiz test dataset.

today's status quo approach by at least 25% and as much as 41%; i.e., AP score of 30.6 for [27] versus 71.7 for `Q+I`. We hypothesize this large discrepancy arises because the irrelevance between a question and image arises for more reasons in VizWiz than for QRPE; e.g., low quality images and fingers blocking the camera view. When comparing the predictive features, we find the image provides the greatest predictive power (i.e., AP = 64) and is solidly improved by augmenting the question information (i.e., AP = 71.7). Again, we attribute this finding to low quality images often leading visual questions to be unanswerable.

6. Conclusions

We introduced VizWiz, a VQA dataset which originates from a natural use case where blind people took images and then asked questions about them. Our analysis of the images, questions, and answers in this dataset reveal new image-based and language-based challenges. Moreover, our analysis of existing modern VQA and answerability algorithms demonstrate this dataset is difficult for modern algorithms. Importantly, improving algorithms on VizWiz can simultaneously educate more people about the technological needs of blind people while providing an exciting new opportunity for researchers to develop assistive technologies that eliminate accessibility barriers for blind people. Accordingly, we will freely share our dataset with ground truth and all code to facilitate future extensions on this work (`http://vizwiz.org/data/`).

Acknowledgments

We thank Aishwarya Agrawal for sharing her crowdsourcing answer collection code, the anonymous crowd workers for their countless hours in answering questions, and the anonymous VizWiz users for sharing their visual questions and so making this dataset possible.

References

[1] Be my eyes. http://www.bemyeyes.org/. 2, 3

[2] http://camfindapp.com/. 1

[3] http://www.taptapseeapp.com/. 1, 3

[4] D. Adams, L. Morales, and S. Kurniawan. A qualitative study to support a blind photography mobile application. In *International Conference on PErvasive Technologies Related to Assistive Environments*, page 25. ACM, 2013. 1

[5] T. Ahmed, R. Hoyle, K. Connelly, D. Crandall, and A. Kapadia. Privacy concerns and behaviors of people with visual impairments. In *ACM Conference on Human Factors in Computing Systems*, pages 3523–3532. ACM Conference on Human Factors in Computing Systems (CHI), 2015. 1, 2, 4

[6] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, 2016. 1, 2, 3, 4

[7] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 1, 2, 3, 4, 5, 6, 7, 8

[8] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. Vizwiz: Nearly real-time answers to visual questions. In *ACM symposium on User interface software and technology (UIST)*, pages 333–342, 2010. 1, 2, 3, 4

[9] J. P. Bigham, C. Jayant, A. Miller, B. White, and T. Yeh. Vizwiz:: Locateit-enabling blind people to locate objects in their environment. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 65–72. IEEE, 2010. 3

[10] E. Brady, M. R. Morris, Y. Zhong, S. White, and J. P. Bigham. Visual challenges in the everyday lives of blind people. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 2117–2126, 2013. 1, 2, 4

[11] M. A. Burton, E. Brady, R. Brewer, C. Neylan, J. P. Bigham, and A. Hurst. Crowdsourcing subjective fashion advice using VizWiz: Challenges and opportunities. In *ACM SIGACCESS conference on Computers and accessibility (ASSETS)*, pages 135–142, 2012. 1, 2, 3

[12] X. Chen, H. Fang, T. Lin, R. Vedantam, S. K. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 1, 3

[13] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra. Visual dialog. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6, 7

[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009. 4, 6

[15] D. Elliott and F. Keller. Image description using visual dependency representations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1292–1302, 2013. 8

[16] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *arXiv preprint arXiv:1505.05612*, 2015. 1, 2, 3, 4

[17] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *arXiv preprint arXiv:1612.00837*, 2016. 1, 2, 3, 4, 7, 8

[18] C. Jayant, H. Ji, S. White, and J. P. Bigham. Supporting blind photography. In *ASSETS*, 2011. 3

[19] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*, 2016. 1, 2, 3, 4

[20] K. Kafle and C. Kanan. An analysis of visual question answering algorithms. *arXiv preprint arXiv:1703.09684*, 2017. 1, 2, 3, 4, 6, 7

[21] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015. 8

[22] V. Kazemi and A. Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017. 1, 2, 7, 8

[23] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 1, 2, 3, 4

[24] W. S. Lasecki, P. Thiha, Y. Zhong, E. Brady, and J. P. Bigham. Answering visual questions with conversational crowd assistants. In *ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, number 18, pages 1– 8, 2013. 1, 2, 3

[25] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *IEEE European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 1, 3, 4, 8

[26] H. MacLeod, C. L. Bennett, M. R. Morris, and E. Cutrell. Understanding blind people's experiences with computer-generated captions of social media images. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 5988–5999. ACM, 2017. 1

[27] A. Mahendru, V. Prabhu, A. Mohapatra, D. Batra, and S. Lee. The promise of premise: Harnessing question premises in visual question answering. *arXiv preprint arXiv:1705.00601*, 2017. 1, 2, 3, 7, 8, 9

[28] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1682–1690, 2014. 1, 2, 3, 4

[29] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318. Association for Computational Linguistics, 2002. 7

[30] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2751–2758. IEEE, 2012. 1, 3

[31] A. Ray, G. Christie, M. Bansal, D. Batra, and D. Parikh. Question relevance in VQA: Identifying non-visual and false-premise questions. *arXiv preprint arXiv:1606.06622*, 2016. 3, 7

[32] M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2935–2943, 2015. 1, 2, 3, 4

[33] 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. *International Journal on Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 3

[34] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. *Computer Vision–ECCV 2012*, pages 746–760, 2012. 4

[35] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 4

[36] A. S. Toor, H. Wechsler, and M. Nappi. Question part relevance and editing for cooperative and context-aware vqa (c2vqa). In *International Workshop on Content-Based Multimedia Indexing*, page 4. ACM, 2017. 3, 7

[37] M. Vázquez and A. Steinfeld. An assisted photography framework to help visually impaired users properly aim a camera. In *ACM Transactions on Computer-Human Interaction (TOCHI)*, volume 21, page 25, 2014. 3

[38] R. Vedantam, L. C. Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 7

[39] P. Wang, Q. Wu, C. Shen, A. Dick, and A. Hengel. FVQA: fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1, 2, 3, 4

[40] P. Wang, Q. Wu, C. Shen, A. Hengel, and A. Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015. 1, 2, 3, 4

[41] S. Wu, J. Wieland, O. Farivar, and J. Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *CSCW*, pages 1180–1992, 2017. 1

[42] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492. IEEE, 2010. 1, 3

[43] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank image generation and question answering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2461–2469, 2015. 1, 2, 3, 4

[44] Y. Zhong, P. J. Garrigues, and J. P. Bigham. Real time object scanning using a mobile phone and cloud-based visual search engine. In *SIGACCESS Conference on Computers and Accessibility*, page 20, 2013. 3

[45] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded question answering in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004, 2016. 1, 2, 3, 4