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. Social media platforms, such as Facebook and Twitter, help people 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 [27, 42]. 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 visual question answering (VQA) problem, which aims to accurately answer any question about any image.

Over the past three years, many VQA datasets have emerged in the vision community to catalyze research on the VQA problem [7, 8, 17, 18, 20, 21, 24, 29, 33, 40, 41, 44, 46]. Historically, progress in the research community on a computer vision problem is typically preceded by a large-scale, publicly-shared dataset [13, 26, 31, 34, 43]. 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, 9] and asking questions about the pictures they take [9, 12, 25]. 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 [9] which established a mobile phone application that supported blind people to ask over 70,000 visual questions [11] by taking a photo and asking a question about it. We begin our work by implementing a rigorous filtering process to remove visual questions that could compromise the safety or privacy of any individuals associated with them; e.g., 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 existing VQA datasets [7, 8, 17, 18, 20, 21, 24, 29, 33, 40, 41, 44, 46]. We finally evaluate numerous algorithms for predicting answers [18, 23] and predicting if a visual question can be answered [28]. 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 for modern algorithms. Our findings suggest the reasons stem from the fact VizWiz is the first vision dataset to in-

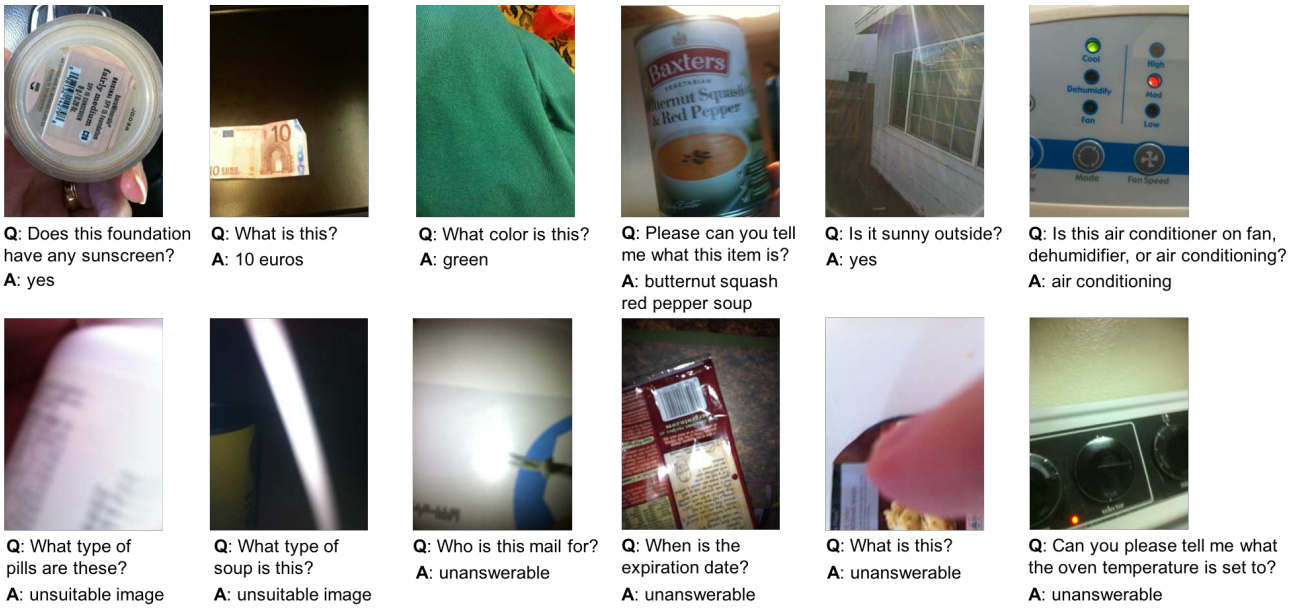

Figure 1. Examples of visual questions asked by blind people and corresponding answers agreed upon by crowd workers. The examples include questions that both can be answered from the image (top row) and cannot be answered from the image (bottom row).

troduce 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 [9, 12, 25]; 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).

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, 9, 25]. 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 [9, 25] or volunteers [1]. Such VQA systems have been shown to be valuable for many daily tasks including grocery shopping [9], locating a specific object in a complex scene [10], and choosing clothes to wear [12]. 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 [9], disrupting the natural flow of a blind user's life. 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., [13, 26, 31, 34, 43]) or created artificially (e.g., [7, 8, 20]). 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 [7, 20] or crowd workers are employed to add pre-defined clipart objects to pre-defined indoor and outdoor scenes [8]. 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 experiments show these 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 [7, 8, 17, 18, 20, 21, 24, 29, 33, 40, 41, 44, 46]. 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 [26]) or artificially constructed (e.g., Abstract Scenes [8], computer graphics [7, 20]). In addition, questions were generated either automatically [7, 20, 21, 29, 33, 44], from crowd workers [8, 17, 18, 21, 24, 46], or from in-house participants [21, 41]. 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.

Answerability Visual Questions. The prevailing assumption when collecting answers to visual questions is that the questions are answerable from the given images [7, 8, 17, 18, 20, 24, 29, 33, 41, 40, 44, 46]. The differences when constructing VQA datasets thus often lies in whether to collect answers from anonymous crowd workers [7, 8, 17, 21, 24], automated methods [20, 29], or in-house annotators [29, 40, 41]. 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 assist blind photographers to improve the image focus [3], lighting [9], or composition [19, 38, 45].

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 [21, 28, 32, 37]. 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 [28] 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 who were seeking answers to their daily visual questions [9, 11]. It is built off of previous work [9] which accrued 72,205 visual questions over four years using the VizWiz application, which is available for iPhone and Android mobile phone platforms. A person asked a visual question by taking a picture and then recording a spoken question. The 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 approach for asking visual questions in VizWiz with approaches employed for 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 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. VizWiz is also unique because, in contrast to all other VQA datasets, the people asking the

Dataset	Which Images?	Who Asked?	How Asked?
DAQUAR [29]	NYU Depth V2 [35]	In-house participants, Automatically generated (templates)	———
VQA v1.0: Abstract [8]	Abstract Scenes	Crowd workers (AMT)	Typed
VQA v1.0: Real [8]	MSCOCO [26]	Crowd workers (AMT)	Typed
Visual Madlibs [44]	MSCOCO [26]	Automatically generated (templates)	———
FM-IQA [17]	MSCOCO [26]	Crowd workers (Baidu)	Typed
KB-VQA [41]	MSCOCO [26]	In-house participants	Typed
COCO-QA [33]	MSCOCO [26]	Automatically generated (captions)	———
VQA v2.0: Real [18]	MSCOCO [26]	Crowd workers (AMT)	Typed
Visual7W [46]	MSCOCO [26]	Crowd workers (AMT)	Typed
CLEVR [20]	Synthetic Shapes	Automatically generated (templates)	———
SHAPES [7]	Synthetic Shapes	Automatically generated (templates)	———
Visual Genome [24]	MSCOCO [26] & YFCC100M [36]	Crowd workers (AMT)	Typed
FVQA [40]	MSCOCO [26] & ImageNet [15]	In-house participants	Typed
TDIUC [21]	MSCOCO [26] & YFCC100M [36]	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.

questions could not "see" the images. Consequently, questions could be unrelated to the images for a variety of reasons that are exemplified in **Figure 1**.

3.2. Anonymizing and Filtering Visual Questions

We faced many challenges with preparing the dataset for public use because our visual questions were collected "in the wild" from real 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" content (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 spell-checker 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 the producers (e.g., askers) or consumers (e.g., research community) of the dataset vulnerable. Accordingly, we obtained from two committees that decide whether proposed research is ethical – the Collaborative Institutional Training Initiative board and Institutional Review Board – approval to publicly release the filtered dataset.

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". 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, includ-

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).

ing 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 panes, etc. We do not expect crowd workers to understand such nuances without extensive instructions and training.

In total, $\sim$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 identifies visual questions that would be good to generate artificially so datasets could address all needs of blind people without requiring them to release personal information.

3.3. Collecting Answers

We next collected answers for a final set of 31,173 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 [8]. 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". While both additions to the annotation protocol indicate a visual question is unanswerable, this annotation approach enables more fine-grained understanding for why a visual question is unanswerable. The final set of answers should represent common sense from sighted people.

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 27.88% for VizWiz versus 13.4% for VQA 2.0 [8] (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 many visual questions do a poor job in narrowing the scope of plausible answers. In contrast, initial wordings such as "How many..." and "Is..." often narrow plausible answers to numbers and "yes/no" respectively.

We also analyze question diversity by computing statistics summarizing the number of words in each question. The median and mean question lengths are five and 6.68 words respectively and 25th and 75th percentile lengths are four and seven words respectively. This resembles the statistics found in the existing artificially constructed VQA datasets, nicely summarized in [14] and [21]. We also observe three words regularly suffice for a question: "What is this?". As observed in **Figure 2**, this short object recognition question is the most common question. Longer and multi-sentence questions also occasionally arise, typically

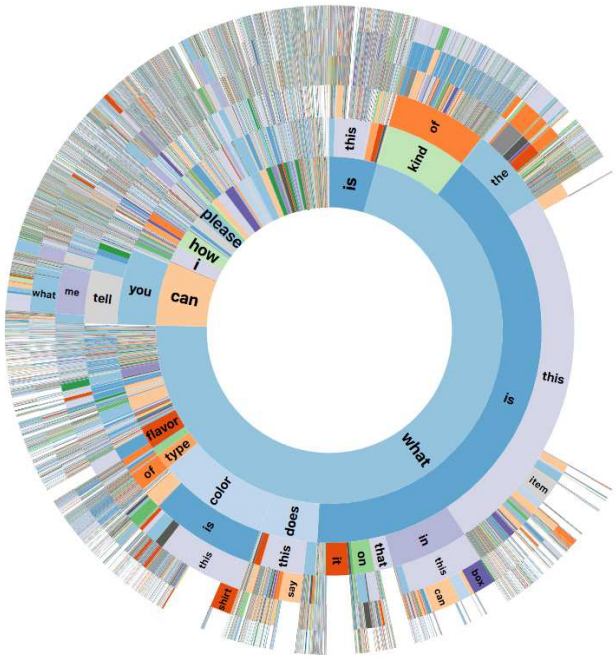

Figure 2. Distribution of the first six words for all questions in VizWiz. The innermost ring represents the first word and each subsequent ring represents a subsequent word. The arc size is proportional to the number of questions with that word/phrase.

because people offer auxiliary information to disambiguate 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 when the audio recording device captures too much content or 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. We first address a concern that our dataset has high quality images showing a single, iconic object, which is a possibility since our filtering process erred on removing "suspicious" scene-based and blurry images and the remaining visual questions contain many object recognition questions. Following prior work [15], we 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.

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) which excludes the

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

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

answers "Unanswerable" and "Unsuitable Image". This visually highlights the fact that there are a large number of unique answers; i.e., ∼58,789. While in absolute terms this number is an order of magnitude smaller than existing larger-scale datasets such as VQA 2.0 [8], 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 [8]. 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 28.63% of visual questions are not answerable. This finding validates the practical importance of the recent efforts [21, 28, 32, 37] 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 content (i.e., "Unanswerable").

We also analyze answer diversity by computing statistics for the number of words in each answer. The median and mean answer lengths are 1.0 and 1.66 words respectively. These statistics resemble what is observed for numerous artificially constructed VQA datasets, as summarized in [14] and [21]. We also compute the percentage of answers with different answer lengths: 67.32% have one word, 20.74% have two words, 8.24% have three words, 3.52% have four words, and the remaining 0.01% have more than four words. Interestingly, our answers are longer on average than observed by Antol et al. [8], who used a similar crowdsourcing system. We attribute this discrepancy in part to many

VizWiz visual questions asking to read multi-word text.

We finally compute the level of human agreement on answers, using exact string matching. Despite that humans provided open-ended text as answers, we observe agreement from independent people on the answer for most visual questions (i.e., 97.7%). More than three people agreed on the most popular answer for 72.83% of visual questions, exactly three people agreed for 15.5% of visual questions, and exactly two people agreed for 9.67% of visual questions. This agreement level is the lower bound since less stringent agreement measures (e.g., that resolve synonyms) may lead to greater agreement.

5. VizWiz Benchmarking

We now investigate the difficulty of the VizWiz dataset for existing algorithms. We divide the final 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). All results below are reported for the test dataset.

5.1. Visual Question Answering

We assess the difficulty of the VizWiz dataset for modern VQA algorithms and evaluate how well models trained on VizWiz generalize (more details in Supp. Materials).

Baselines. We benchmark nine methods. Included are three top-performing VQA methods [6, 18, 23], which we refer to as `Q+I` [23], `Q+I+A` [18], and `Q+I+BUA` [6]. These baselines are trained on the VQA V2.0 dataset [18] to predict the 3,000 most frequent answers in the training dataset. [18] relies on image and question information alone, [23] adds an attention mechanism to specify image regions to focus on, and [6] combines bottom-up and top-down attention mechanisms to focus on objects and other salient image regions. We introduce three fine-tuned classifiers built on the three networks, which we refer to as `FT` [18], `FT` [23], and `FT` [6]. We also train the three networks from scratch using the VizWiz data alone, and we refer to these as `VizWiz` [18], `VizWiz` [23], and `VizWiz` [6].

Evaluation Metrics. We evaluate with respect to four metrics: Accuracy [8], CIDEr [39], BLEU4 [30], and METEOR [16]. Accuracy [8] was introduced as a good metric when most answers are one word. Since nearly half the answers in VizWiz exceed one word, we also use image description metrics provided by [13] which are designed for evaluating longer phrases and/or sentences.

Results. We first analyze how existing prediction models [6, 18, 23] perform on the VizWiz test set. As observed in the first three rows of **Table 3**, these models perform poorly, as indicated by low values for all metrics; e.g., $\sim$0.14 accuracy for all algorithms. We attribute the poor generalization of these algorithms largely to their inability to predict answers observed in the VizWiz dataset; i.e., only 824 out of the top 3,000 answers in VizWiz are included in the dataset (i.e., VQA 2.0 [18]) used to train the models.

We observe in **Table 3** that fine-tuning (i.e., rows 4–6) and training from scratch (i.e., rows 7–9) yield significant performance improvements over relying on the three prediction models [6, 18, 23] as is. We find little performance difference between fine-tuning and training from scratch for the three models. While the number of training examples in VizWiz is relatively small, we hypothesize the size is sufficient for teaching the models to retain knowledge about answer categories that are applicable in this setting. Despite the improvements, further work is still needed to achieve human performance (i.e., 0.75 accuracy)[1].

We next analyze what predictive cues may lead to algorithm success/failure. We observe models that add the attention mechanism [6, 23] consistently outperform relying on image and question information alone [18]. Still, the improvements are relatively small compared to improvements typically observed on VQA datasets. We hypothesize this improvement is small in part because many images in VizWiz include few objects and so do not need to attend to specific image regions. We also suspect attention models perform poorly on images coming from blind photographers since such models were not trained on such images.

We further enrich our analysis by evaluating the nine algorithms for visual questions that lead to different answer types (their frequencies in VizWiz are shown in parentheses): "yes/no" (4.80%), "number" (1.69%), "other" (58.91%), and "unanswerable" (34.6%). Results are shown in **Table 4**. Overall, we observe performance gains by fine-tuning algorithms (rows 4–6) and training from scratch (rows 7–9), with the greatest gains for "unanswerable" vi-

[1]Performance is measured by partitioning the dataset into 10 sets of one answer per visual question and then evaluating one answer set against the remaining nine answer sets for all 10 partitions using the accuracy metric.

Method	Acc	CIDEr	BLEU	METEOR
Q+I [18]	0.137	0.224	0.000	0.078
Q+I+A [23]	0.145	0.237	0.000	0.082
Q+I+BUA [6]	0.134	0.226	0.000	0.077
FT [18]	0.466	0.675	0.314	0.297
FT [23]	0.469	0.691	0.351	0.299
FT [6]	**0.475**	**0.713**	0.359	**0.309**
VizWiz [18]	0.465	0.654	0.353	0.298
VizWiz [23]	0.469	0.661	0.356	0.302
VizWiz [6]	0.469	0.675	**0.396**	0.306

Table 3. Performance of VQA methods on the VizWiz test data with respect to four metrics. Results are shown for three variants of three methods [6, 18, 23]: use models as is, fine-tuned (FT), and trained on only VizWiz data (VizWiz). The methods use different combinations of image (I), question (Q), and attention (A) models.

	Yes/No	Number	Unans	Other
Q+I [18]	0.598	0.045	0.070	0.142
Q+I+A [23]	0.605	0.068	0.071	0.155
Q+I+BUA [6]	0.582	0.071	0.060	0.143
FT [18]	0.675	0.220	0.781	0.275
FT [23]	**0.681**	0.213	0.770	0.287
FT [6]	0.669	0.220	0.776	**0.294**
VizWiz [18]	0.597	**0.262**	**0.805**	0.264
VizWiz [23]	0.608	0.218	0.802	0.274
VizWiz [6]	0.596	0.210	**0.805**	0.273

Table 4. Accuracy of nine VQA algorithms for visual questions that lead to different answer types.

sual questions and smallest gains for "number" and "other" visual questions. Exemplar failures include when asking for text to be read (e.g., captchas, cooking directions) and things to be described (e.g., clothes).

Finally, we evaluate how well algorithms trained on VizWiz predict answers for the VQA 2.0 test dataset [18]. The six models that are fine-tuned and trained from scratch for the three models [6, 18, 23] do not generalize well; i.e., accuracy scores range from 0.218 to 0.318. This result suggests that VizWiz provides a domain shift to a different, difficult VQA environment compared to existing datasets.

5.2. Visual Question Answerability

We next turn to the question of how accurately an algorithm can classify a visual question as answerable.

Baselines. We benchmark eight methods. We use the only publicly-available method for predicting when a question is not relevant for an image [28]. This method uses NeuralTalk2 [22] pre-trained on the MSCOCO captions dataset [26] 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 [28]. We use the model as is (i.e., `Q+C` [28]), fine-tuned to the VizWiz data (i.e., `FT` [28]), and trained from scratch on the VizWiz data only (i.e., `VizWiz` [28]). We also employ our top-performing VQA algorithm by using its output probability that the predicted answer is "unanswerable" (`VQA` [18]). We enrich our analysis by further investigating the influence of different features on the predictions: question alone (i.e., `Q`), caption alone (i.e., `C`), image alone using ResNet-152 CNN features (i.e., `I`), and the question with 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 precision-recall curves. As observed, all methods outperform the status quo approach

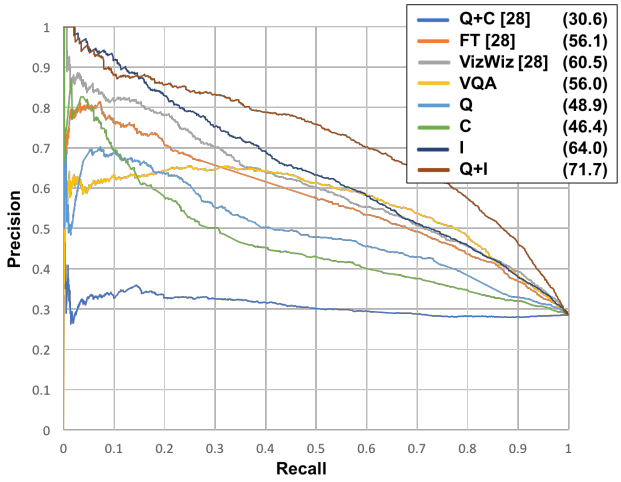

Figure 5. Precision-recall curves and average precision scores for the answerability models tested on the VizWiz test dataset.

by 25% to 41%; i.e., AP score of 30.6 for [28] 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 adding 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 demonstrates this dataset is difficult for modern algorithms. Improving algorithms on VizWiz can simultaneously educate 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. We share the dataset and code to facilitate future work (`http://vizwiz.org/data/`).

Acknowledgements: We thank Erin Brady, Samuel White, and Yu Zhong for supporting the VizWiz deployment, the anonymous users of the VizWiz application for sharing their visual questions, the authors of [8] for sharing their answer collection code, and the anonymous crowd workers for providing the annotations. This work is supported by NSF awards to D.G. (IIS-1755593), K.G. (IIS-1514118), and J.L. (IIS-1704309); AWS Machine Learning Research Award to K.G.; Google Research Award to J.P.B.; Microsoft Research support to J.P.B.; and New York State support from the Goergen Institute for Data Science to J.L.

References

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

[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, 4

[6] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *arXiv preprint arXiv:1707.07998*, 2017. 7, 8

[7] 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

[8] 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

[9] 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

[10] 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. 2

[11] 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, 3

[12] 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

[13] 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, 2, 7

[14] 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. 5, 6

[15] 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

[16] 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. 7

[17] 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, 3, 4

[18] 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, 3, 4, 7, 8

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

[20] 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

[21] K. Kafle and C. Kanan. An analysis of visual question answering algorithms. *arXiv preprint arXiv:1703.09684*, 2017. 1, 3, 4, 5, 6

[22] 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

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

[24] 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, 3, 4

[25] 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

[26] 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, 2, 3, 4, 8

[27] 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

[28] 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, 3, 6, 8

[29] 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, 3, 4

[30] 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

[31] 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, 2

[32] 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, 6

[33] 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, 3, 4

[34] 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, 2

[35] 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

[36] 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

[37] 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, 6

[38] 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

[39] 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

[40] 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, 3, 4

[41] 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, 3, 4

[42] 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

[43] 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, 2

[44] 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, 3, 4

[45] 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

[46] 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, 3, 4