CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question

Danna Gurari and Kristen Grauman
The University of Texas at Austin
2317 Speedway, Stop D9500, Austin, TX 78712
{dgurari,grauman}@cs.utexas.edu

ABSTRACT

Visual question answering systems empower users to ask any question about any image and receive a valid answer. However, existing systems do not yet account for the fact that a visual question can lead to a single answer or multiple different answers. While a crowd often agrees, disagreements do arise for many reasons including that visual questions are ambiguous, subjective, or difficult. We propose a model, *CrowdVerge*, for automatically predicting from a visual question whether a crowd would agree on one answer. We then propose how to exploit these predictions in a novel application to efficiently collect all valid answers to visual questions. Specifically, we solicit fewer human responses when answer agreement is expected and more human responses otherwise. Experiments on 121,811 visual questions asked by sighted and blind people show that, compared to existing crowdsourcing systems, our system captures the same answer diversity with typically 14-23% less crowd involvement.

Author Keywords

Visual Question Answering; Machine Learning; Crowdsourcing

ACM Classification Keywords

H.5.m. Info. Interfaces and Presentation (e.g. HCI): Misc

INTRODUCTION

What would be possible if a person had access to a system that could answer any question about the visual world? Blind users could quickly figure out the denomination of their currency and so whether they spent the appropriate amount for a product. Hikers could learn about their bug bites to help decide whether to seek out professional medical care. Factory managers could identify how many defective products are on an assembly line and so monitor a factory's efficiency. These examples reflect the tip of the iceberg for the vast range of benefits blind and sighted users could derive from a *visual question answering (VQA)* system (e.g., **Figure 1**).

CHI 2017, May 06 - 11, 2017, Denver, CO, USA

ISBN 978-1-4503-4655-9/17/05…$15.00

DOI: http://dx.doi.org/10.1145/3025453.3025781

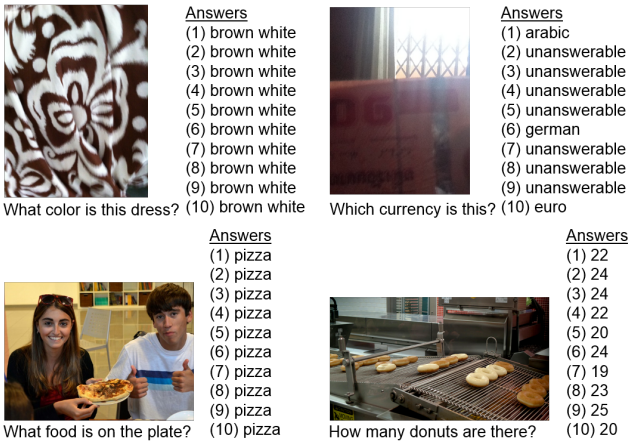

Figure 1. Examples of visual questions and corresponding answers from 10 different people. The examples include visual questions asked by both blind users (top row) and sighted users (bottom row). As observed, the crowd sometimes all agree on a single answer (first column) and at other times offer different answers (last column). We propose a *CrowdVerge* system to automatically predict whether multiple people will give the same answer when given an image and question about it.

While progress has been made in building VQA systems to accurately answer visual questions [4, 5, 6, 7, 24], existing systems do not yet account for the fact that some visual questions lead multiple people to provide the same answer while other visual questions lead multiple people to provide different answers. Yet, our analysis of over 450,000 visual questions asked by blind and sighted users reveals that these two outcomes arise in approximately equal proportions. We find humans disagree for a variety of reasons including because visual questions are ambiguous, subjective, or difficult (e.g., as observed in **Figure 1**, counting how many donuts or recognizing a foreign currency from a poor quality image).

Our goal is to *account for whether different people would agree on a single answer to a visual question to improve upon today's VQA systems*. We propose multiple *CrowdVerge* prediction systems to automatically decide if a visual question will lead to human agreement and demonstrate the value of their predictions for a new task of capturing the diversity of all plausible answers with less human effort.

Our work is partially inspired to improve how to employ crowds as the computing power at run-time. Towards satisfying existing users, gaining new users, and supporting a wide range of applications, a crowd-powered VQA system

should be low cost, have fast response times, and yield high quality answers. Yet, today's status quo is to assume a fixed number of human responses per visual question [6, 10]. In other words, users currently often either incur extra costs and delays by collecting extra answers when they are redundant or compromise on quality by not collecting all valid answers.

Our work is also inspired to improve how to employ crowds at design-time to produce the information needed to design automated VQA methods. Specifically, researchers in fields as diverse as computer vision [5], computational linguistics [4], and machine learning [24] rely on large datasets, which include visual questions and human-supplied answers, to train and evaluate VQA algorithms. In general, "bigger" data is better. Current methods to create these datasets assume a fixed number of human answers per visual question [5, 35], thereby either compromising on quality by not collecting all plausible answers or compromising on efficiency by collecting additional answers when they are redundant.

To our knowledge, this work is the first to propose a crowdsourcing system which dynamically solicits the number of human responses based on each visual question. Our goal is to actively solicit extra answers only for visual questions likely to have multiple answers. We show in our experiments that employing predictions from our proposed system to answer more than 100,000 visual questions can eliminate over 11 40-hour work weeks and save $1800 with no loss to captured answer diversity, compared to today's status quo [5, 6].

While the application of our prediction system can offer cost and time savings when answering visual questions, it is also useful to understand what aspects of a visual question are informative for predicting answer (dis)agreement. To address this, we examine the influence of the question alone, image alone, and combination of both sources for predicting answer (dis)agreement for a visual question.

The key contributions of our work are as follows:

- Analysis demonstrating the prevalence and reasons for human answer disagreements in three VQA datasets.
- A new problem and system for predicting whether a crowd will (dis)agree when answering a visual question.
- A novel application for efficient answer collection which solicits additional answers from additional members of a crowd only when disagreement is anticipated.

RELATED WORK

Visual Question Answering Services

A commonality across communities as diverse as human computer interaction, machine learning, computational linguistics, and computer vision is they adopt a one-size-fits-all approach when deciding the number of answers their systems should return per visual question [4, 5, 6, 24] . For example, crowd-powered systems aim to supply a fixed number of answers [6] and automated systems return a single answer [4, 5, 24]. By analyzing 10 answers per visual question for nearly half a million visual questions from sighted and blind people, we were inspired to rethink generally held assumptions about how to design VQA systems. We instead propose to predict whether a crowd will agree when answering a visual question. Our experiments demonstrate that knowing whether a crowd will agree is an important factor to consider to design *faster*, *cheaper*, and *more accurate* systems that spend just enough to collect all valid answers for each visual question.

Analyses of Crowd Disagreement

More broadly, our work relates to methods that account for crowd disagreement [3, 17, 28, 31, 34]. For example, researchers have suggested ways to resolve disagreement due to task difficulty [34] and ambiguity/specificity [3, 17]. Some methods decide which workers to trust most when aggregating multiple responses into a final, single response [31, 34]. Other methods leverage context to automatically disambiguate which of multiple outcomes is the desired outcome [3]. Unlike prior work, we focus on the task of visual question answering. Moreover, rather than try to resolve specific sources of disagreement (e.g., task difficulty, ambiguity), we instead aim to automatically predict whether disagreement will arise for *any* reason. Such information empowers system designers to create systems that *only* collect multiple answers and decide how to resolve disagreement when a disagreement is expected. We demonstrate the advantage of our system to predict disagreement over relying on the uncertainty of a VQA algorithm in its predicted answer [5].

Answer Collection from a Crowd

Our work relates to methods that propose how to employ crowd workers to answer questions about images. Such approaches aim to collect a pre-specified, fixed number of answers per visual question [5, 6, 10, 35]. For those systems that treat response time as a first priority, a variable number of answers may arise but this is due to varying crowdsourcing conditions such as the available supply of workers [6, 10]. Unlike prior work, our goal is to collect answers in a way that is both economical *and* complete in capturing the diversity of plausible answers for all visual questions. To our knowledge, our work is the first to predict the number of answers to collect for a visual question. Experiments demonstrate that our (dis)agreement predictions are useful to significantly reduce human effort for capturing the diversity of valid answers.

Continuous Dialogue with the Crowd

Be My Eyes [1] and Chorus:View [20] empower users to engage in a continuous communication channel with a crowd worker when asking a visual question. The aim is to expedite arriving at desired answers by, for example, permitting the crowd worker to clarify ambiguous questions. Our work offers an alternative by demonstrating how a crowdsourcing service might instead solicit multiple answers for a one time back-and-forth rather than enacting a more costly, continuous communication channel with a single voice, whether from a single person [1] or the consensus of a crowd [20].

High Quality Work with Fixed Human Budget

Our work relates to methods that actively allocate a limited human budget to where it will best contribute to improve the quality of results. For example, one method distributes a budget between three different levels of human effort when deciding how to segment images [16]. Another method predicts

when to employ algorithms versus crowd workers to segment images [14]. Another method spends a budget between less costly crowd workers and more costly experts for biomedical citation screening [26]. To our knowledge, our work is the first towards deciding how to spend a budget for the task of visual question answering, which is distinct from prior work that focused on spending a budget for image analysis or language analysis alone. Furthermore, our aim is to spend a budget to capture the *diversity* of all valid results for every task rather than to collect a *single* result for every task.

Minimizing Human Labeling

Our aim to actively decide how to allocate human effort to improve results is also somewhat related to active learning [29]. Specifically, active learners try to use as little human effort as possible to train accurate prediction models. Some methods iteratively supplement a training dataset with the most informative images for training a classifier [27, 33]. Other methods solicit redundant labels to prevent incorrect/noisy labels [21, 30]. While active learners aim to minimize human input to improve the accuracy of a prediction model, our method aims to minimize human input while still exhaustively capturing all plausible answers to all visual questions.

Scalable Annotation Collection

Our work relates to crowdsourcing systems designed to produce "big data". The aim is to collect human annotations both to teach machine learning algorithms to behave like a human (e.g., recognize a cat) and evaluate how similarly algorithms behave to humans (e.g., did the algorithm accurately indicate the image shows a cat?) [2, 5, 12, 18, 22]. Towards scaling up, one system reduces the number of tasks by intelligently deciding what questions to ask each crowd worker in what order [12]. Another system instructs crowd workers that annotation errors are okay in order to speed up their productivity [18]. However, a common inefficiency for such approaches is they collect a fixed number of redundant responses per task to establish trusted high quality annotations. While prior work modifies the crowdsourcing task itself (i.e., asked questions [12], instructions [18]), we instead propose to only modify how many redundant annotations to collect per task to more efficiently create human-annotated datasets.

Crowdsourcing and Computer Vision

Finally, our work relates to systems which mix crowdsourcing with computer vision. While our hybrid system design reinforces existing work by also demonstrating advantages in combining crowd and algorithm efforts, it differs by addressing the VQA task rather than the object detection [15], event detection [19], and image description [13, 37] tasks.

PAPER OVERVIEW

The remainder of the paper is organized into four sections. We first examine how often visual questions lead to answer agreement, why disagreement arises, and how many different answers typically arise. Next, we explore: 1) For a novel visual question, can a machine correctly predict whether multiple independent members of a crowd would supply the same answer? and 2) If so, what insights does our machine-learned system reveal regarding what humans are most likely to agree about? Then, we propose a novel resource allocation system for efficiently capturing the diversity of all answers for a set of visual questions. Finally, we end with concluding remarks.

DATASETS AND ANSWER (DIS)AGREEMENTS

Our first aim is to better understand the information gained for the visual question answering task by collecting multiple answers from different people. Towards this aim, we investigate (1) how often do multiple answers lead to answer agreement (redundant information)?, (2) what does answer disagreement tell us about a visual question?, and (3) given that multiple people can disagree on the answer, how many different answers typically capture all the perspectives from the crowd? Our findings validate that, in practice, visual questions regularly lead to both a single answer and multiple valid answers. Our findings also enrich our understanding for why different answers arise and the typical amount of answer diversity.

Visual Questions

We compiled 461,360 visual questions that come from three publicly-available datasets. We chose to analyze visual questions coming from datasets that are widely studied [5, 6] in order to offer practical guidance to the many system designers already working on the VQA problem. We also chose to analyze visual questions asked by blind and sighted people in order to address the interests of a diversity of users. Altogether, the three datasets represent three different types of images asked about by two different types of users.

VizWiz [6]: We include 1,499 visual questions asked by blind users. Each visual question was created by a person using a mobile phone who took a picture and recorded his/her spoken question. These visual questions often address accessibility issues for daily tasks, with a focus on asking for objective information; e.g., "what type of beverage is in this bottle?" or "what color is this shirt?" [8]. Accordingly, images often show familiar, everyday objects such as food, beverages, computer screens, clothing, and household items. Yet, because blind people cannot see and verify the quality of the pictures they take, many images are poorly framed, poorly lit, or blurry.

VQA Real Images [5]: We also examine 369,861 visual questions asked by sighted users about images collected from the photo-sharing website Flickr. In particular, three open-ended questions were collected about each of 153,287 images by instructing three Amazon Mechanical Turk (AMT) crowd workers to look at the given image and generate a text-based question about it that would "stump a smart robot" [5]. The images show complex scenes that include at least one from 91 categories of objects that would be easily recognizable by a four year old; e.g., dog, chair, person [22]. Consequently, questions are often grounded in images and task-independent while images are often high quality.

VQA Abstract Scenes [5]: The remaining 90,000 visual questions are asked by sighted users about 30,000 abstract scenes. The abstract scenes were created by crowd workers who were instructed to add objects from 100 clipart options to create scenes in artificial indoor and outdoor environments [5]. As with the previous dataset, three crowd workers were recruited to provide three questions per image that would "stump a

	VizWiz	VQA - Real Images			VQA - Abstract Scenes		
Answer Type:	*All*	*Yes/No*	*Number*	*Other*	*Yes/No*	*Number*	*Other*
# VQAs (%):	*1,499*	*140,777 (38%)*	*45,822 (12%)*	*183,262 (50%)*	*36,717 (41%)*	*12,956 (14%)*	*40,327 (45%)*
At Most One Disagreement	20%	74%	49%	35%	74%	79%	36%
- Unanimous Agreement	9%	54%	35%	22%	57%	65%	22%
- Exactly One Disagreement	11%	20%	14%	13%	17%	14%	14%

Table 1. Frequency of answer agreement for visual questions asked by blind (VizWiz) and sighted (VQA) people. Shown for each dataset (or answer type in a dataset) is the percentage of visual questions that lead to at most one disagreement (row 1), unanimous agreement (row 2), and exactly one disagreement (row 3) from 10 crowdsourced answers. Both crowd agreement and disagreement regularly arise for a variety of visual questions that elicit a range of answer types. On average, across all three datasets, the crowd agrees on the answer for nearly half (i.e., 53%) of all visual questions.

smart robot". While questions are often grounded in images and task-independent, the images are semi-realistic.

Answers

We compiled 4,613,600 answers to analyze answer (dis)agreement trends. In particular, every visual question is paired with 10 answers collected from 10 AMT crowd workers. Answers were collected following the excellent crowdsourcing protocol discussed in [5], which shows a worker an image with associated question and asks the worker to respond with "a brief phrase and not a complete sentence" [5][1]. We leveraged the answers already included with the datasets for the VQA Real Images and Abstract Scenes. For the VizWiz visual questions, we crowdsourced the collection of 14,990 answers[2] and share them at `http://vision.cs.utexas.edu/CrowdVerge/`. To do so, we slightly modified the aforementioned crowdsourcing system by adding instructions telling crowd workers to state a visual question is unanswerable if a question cannot be answered from the image. We added this instruction because, unlike the other two datasets, there was reasonable doubt that questions would be grounded in images since blind people cannot validate an image captures the content of their question.

Each answer was subsequently post-processed. Following prior work [5], we converted all letters to lower case, converted numbers to digits, and removed punctuation and articles (i.e., "a", "an", "the"). We further processed our collected answers for the VizWiz visual questions by fixing spelling mistakes, removing filler phrases (e.g., "it is"), and resolving differences among answers where one description is subsumed in another description (e.g., "dr pepper" = "dr pepper soda").

Answer (Dis)Agreements

We next quantify the tendency for real-world visual questions to lead multiple people to offer the same answer. We tally the number of different answers observed among 10 crowdsourced answers using exact string matching in order to establish a lower bound of expected answer agreement (i.e., more sophisticated natural language processing would reveal greater agreement by, for example, resolving synonyms).

To enrich our analysis, we employ labels included with the VQA datasets that indicate for each visual question which of the following answer types is elicited: "yes/no", "number", or "other". Each label represents the most popular option from the 10 labels assigned to the 10 answers per visual question.

We report results for both when crowds unanimously agree as well as when nine of the 10 people agree on an answer for all three datasets (**Table 1**). These results capture when at most one untrusted result is permitted from the crowd when inferring whether a crowd agrees. Overall, we observe at most one disagreement for 53% of all visual questions across the three datasets. In absolute terms, this means that approximately $11,719 and 230 40-hour work weeks were spent to collect the redundant answers for the 244,159 visual questions that led to agreement[3]. This finding supports our hypothesis that great savings can be achieved with no loss to captured answer diversity *if* a crowdsourcing system could know whether a given visual question would lead to answer agreement from a crowd. Moreover, savings are possible whether collecting visual questions from blind or sighted users.

Our findings also highlight that both outcomes, crowd agreement and disagreement, regularly arise for a variety of answer types (**Table 1**). For example, we observe high agreement for "yes/no" visual questions for both VQA datasets. Still, such visual questions do evoke disagreements, such as when they are seeking subjective information (e.g., "Does this picture look scary?", **Figure 2g**). In addition, we observe moderate to high levels of agreement for "number" visual questions. We hypothesize the greater likelihood for disagreement for the real images than for the abstract scenes arises because larger counting problems (i.e., many objects), which are more difficult to get correct, occur more often in real scenes. Finally, we observe moderate agreement levels for "other" visual questions. Yet, disagreements do arise for these visual questions, often because of a greater diversity of opinions regarding the true answer (e.g., **Figure 2b**) as well as ways to express the same concept (e.g., **Figure 2h**). Given that the tendency for agreement differs for *yes/no*, *counting*, and *other* visual questions, we hypothesize that the question wordings that lead to these different types of answers (e.g., "How many... ?" versus "Is the... ?" versus "Why is... ?") will be informative of whether a

[1] See [5] for a screen shot of the user interface.

[2] We collected our own answers because the existing dataset includes a variable number of answers per visual question that was insufficient for our analysis; i.e. typically zero to five answers.

[3] Our cost and time estimates are based on the crowdsourcing methodology established by Antol et al. [5]. They paid $0.006 per answer. In addition, when we used their system to collect 15,000 crowdsourced answers for the VizWiz visual questions, crowd workers took on average 17 seconds to submit an answer. Our estimation is also based on assuming eight redundant answers, one correct answer, and one untrusted (e.g., spam) answer.

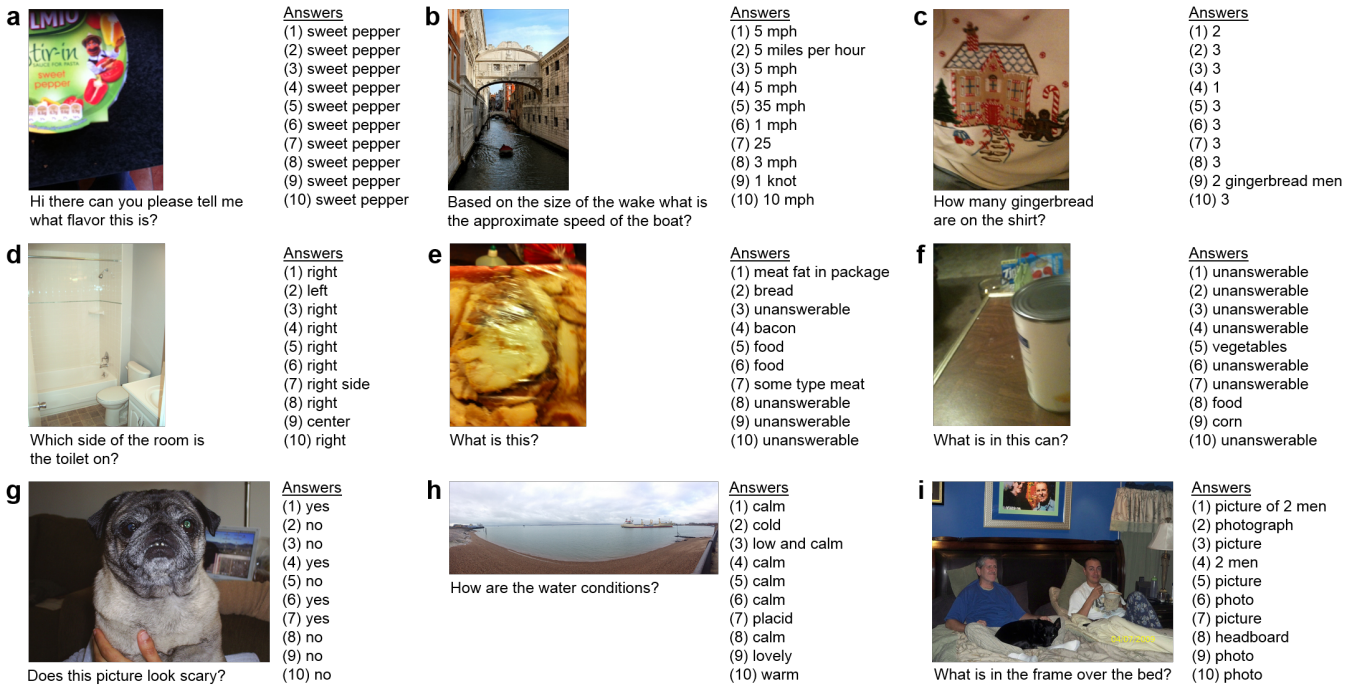

Figure 2. Examples of visual questions asked by blind and sighted users with corresponding answers collected from 10 crowd workers. As observed, (a) unanimous answer agreement arises when questions are visually grounded and capture commonsense knowledge. (b-i) Answer disagreement arises for a variety of reasons: (b) expert skill needed, (c) human mistakes, (d) ambiguous question, (e) ambiguous visual content, (f) insufficient visual evidence, (g) subjective question, (h) answer synonyms, and (i) varying answer granularity. (Best viewed on pdf.)

crowd will agree. We capitalize on this observation in the next section to design prediction systems that automatically decide whether a visual question will lead to agreement.

From visual inspection of hundreds of visual questions that lead to different amounts of answer disagreement, we identified eight reasons why people disagree. Each reason is exemplified in **Figure 2**. Disagreements can arise due to crowd worker skill, both because a difficult task necessitates domain expertise and because a crowd worker may inadequately answer a seemingly simple question (**Figure 2b,c**). Crowds disagree also because of ambiguity in the question and visual content (**Figure 2d,e**). Further reasons for disagreement include insufficient visual evidence to answer the question, subjective questions, synonymous answers, and varying levels of answer granularity (**Figure 2f–i**). We observed each of these reasons for answer disagreement in each of the three studied datasets. Our findings highlight that answer disagreement can reflect numerous possible aspects of a visual question including its quality, specificity, and difficulty.

Answer Diversity

We next examine how many unique answers are observed in practice for visual questions. We measure answer agreement by counting how many different answers are observed in 10 crowdsourced answers per visual question using exact string matching. Although this approach does not fully resolve all conceptually equivalent responses, it does reveal an upper bound of expected disagreement. Specifically, more lenient agreement schemes that employ more sophisticated natural language processing methods (e.g., inferring agreement for a synonymous answer) would lead to greater answer agreement.

Across all 461,360 visual questions, we tally how many visual questions yield k unique answers where $k = \{1, 2, ..., 10\}$. We examine the influence of different levels of trust in the crowd as well as the influence of different datasets. Specifically, we tally the number of valid answers observed when requiring a minimum of $m = 1$, $m = 2$, or $m = 3$ members of the crowd to offer the same answer for the answer to be valid. As a point of reference, prior work deems answers as valid using blind trust (i.e., $m = 1$ person) [25] as well as more conservative schemes (i.e., $m = 3$ people) [5]. We conduct our analysis on each of the three datasets independently.

Our findings were surprising in that only 1% of visual questions from sighted and blind people led to no answer agreement from a crowd for both real image datasets (i.e., VizWiz and VQA). In other words, at least two people agree upon an answer for 99% of the visual questions. This highlights that multiple, independent people asked to answer the same visual question typically converge on at least one answer, despite the open-ended nature of how answers are collected.

We found that a visual question typically leads to at most three different answers for all datasets (**Figure 3**). This gives an upper bound of expected answer diversity. As discussed above, we anticipate that visual questions will lead to fewer valid answers with less stringent answer agreement schemes. While our findings suggest that the visual questions asked by sighted and blind people are predominantly answerable, accurately reflecting the wisdom of the crowd may require responding with multiple answers.

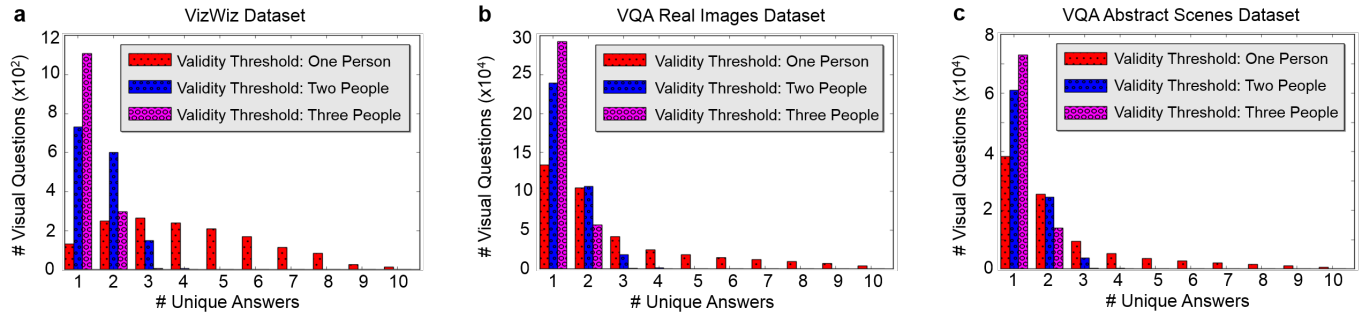

Figure 3. Summary of answer diversity outcomes showing how frequently different numbers of unique answers arise when asking 10 crowd workers to answer a visual question for (a) 1,499 visual questions asked by blind people, (b) 369,861 visual questions asked by sighted people about real images and (c) 90,000 visual questions asked by sighted people about clipart abstract scenes. Results are shown based on different degrees of answer agreement required to make an answer valid: only one person has to offer the answer, at least two people must agree on the answer, and at least three people must agree on the answer. Our findings demonstrate that a large diversity of open-ended visual questions are answerable (i.e., lead to answer agreement from a crowd), which motivates the question of how to efficiently collect all valid answers from a crowd (i.e., typically one answer to at most three answers).

We observe the same trend for the amount of answer diversity for most agreement thresholds for all three datasets (**Figure 3**). Most commonly there is one unique answer, followed by two and three answers respectively. As expected, moving from requiring no answer agreement to a more conservative agreement between three people shifts the distribution to more sharply peak at less overall diversity (i.e., 1 unique answer).

CROWDVERGE: PREDICTING (DIS)AGREEMENT

As observed in the previous section, a variety of visual questions regularly lead to both a single answer and multiple different answers from a crowd. Yet, currently, a person who asks a visual question cannot know which outcome will arise unless (s)he actually collects answers from a crowd. In this section, we introduce a model which we call *CrowdVerge* to address this problem. In particular, a user can learn from a *CrowdVerge* system whether a given visual question will lead the crowd to converge on a single answer or diverge and offer multiple different answers. We evaluate two implementations of *CrowdVerge* and investigate what these systems reveal are predictive cues for answer (dis)agreement.

CrowdVerge Model and Implementations

We pose the prediction task as a binary classification problem. Specifically, a CrowdVerge model takes as input an image and associated question and outputs a binary label indicating whether a crowd will agree on the same answer. The goal is to detect which visual questions to assign a disagreement label, regardless of the disagreement cause (e.g., subjectivity, ambiguity, difficulty). We consider random forest and deep learning implementations of *CrowdVerge*.

Answer (Dis)Agreement Labels

Each visual question is assigned either an answer *agreement* or *disagreement* label. We employ the 10 crowdsourced answers per visual question to assign labels. A visual question is assigned an answer *agreement* label when there is an exact string match for at least 9 of the answers (after answer pre-preprocessing, as discussed in the previous section) and an answer *disagreement* label otherwise. Our rationale is to permit up to one careless/spam answer per visual question.

Random Forest System

For our first system, we use domain knowledge to guide the learning process. We compile a set of features that we hypothesize inform whether a crowd will arrive at an undisputed, single answer. Then we apply a machine learning tool to reveal the significance of each feature. We propose features based on the observation that answer agreement often arises when 1) a lay person's attention can be easily concentrated to a few salient regions in an image and 2) a lay person would find the requested task easy to address.

As *image-based* features, we represent the estimated number of prominent objects in the image, regardless of the object category. To extract these features, we employ a state-of-the-art salient object subitizing [36] (SOS) method. It produces five probabilities that indicate whether an image contains 0, 1, 2, 3, or 4+ salient objects. Intuitively, the number of salient objects shows how many regions in an image are competing for an observer's attention, and so may correlate with the ease in identifying a region of interest. Moreover, we hypothesize this feature will capture our observation that counting problems typically lead to disagreement for images showing many objects, and agreement otherwise.

We also employ *question-based features*. One feature is the number of words in the question. Intuitively, a longer question offers more information and we hypothesize additional information makes a question more precise. The remaining features capture the first two words in the question. We encode them as two one-hot vectors. Each one-hot vector is created using the learned vocabularies that define all possible words at the first and second word location of a question respectively (using training data, as described in the next section). Intuitively, early words in a question inform the type of answers that might be possible and, in turn, possible reasons/frequency for answer disagreement. For example, we expect "why is" to regularly elicit many opinions and so disagreement. This intuition about the beginning words of a question is also supported by our analysis in the previous section which shows that different answer types yield different biases of eliciting answer agreement versus disagreement.

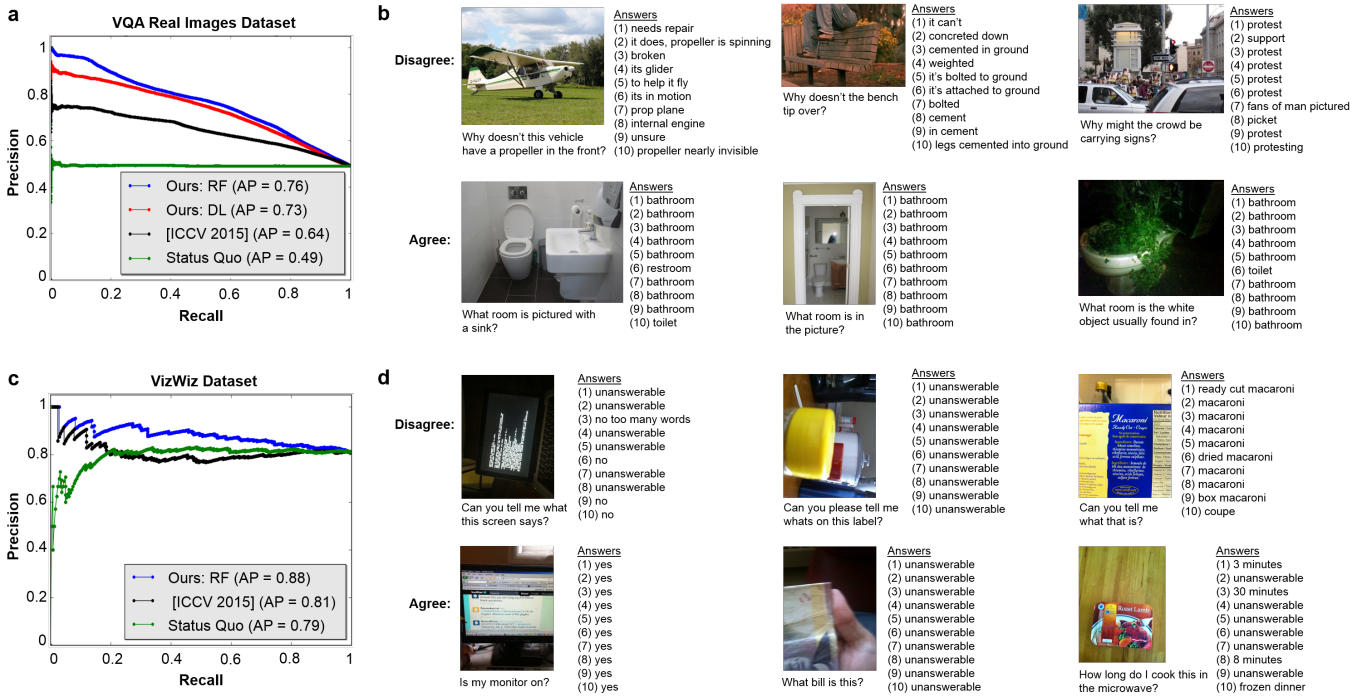

Figure 4. Precision-recall curves and average precision (AP) scores for all benchmarked systems on the (a) VQA and (c) VizWiz datasets. Our random forest (RF) and deep learning (DL) *CrowdVerge* systems outperform a related automated VQA baseline, showing the importance in modeling human disagreement as opposed to system uncertainty. Also shown are examples of prediction results from our top-performing RF classifier for the (b) VQA and (d) VizWiz datasets. Included are the top three visual questions with crowdsourced answers for the most confidently predicted instances that lead to answer disagreement and agreement. These examples illustrate a strong language prior for making predictions. (Best viewed on pdf.)

We employ a random forest *classification model* [9] to predict an answer (dis)agreement label for a visual question. This model consists of an ensemble of decision tree classifiers. We train the system to learn the unique weighted combinations of the aforementioned image-based and question-based features that each decision tree applies to make a prediction. At test time, given a novel visual question, the trained system converts a feature descriptor of the visual question into a final prediction that reflects the majority vote prediction from the ensemble of decision trees. The system returns the final prediction with a probability indicating the system's confidence in that prediction. We employ the Matlab implementation of random forests, using 25 trees and the default parameters.

Deep Learning System

As an alternative to random forests, we next consider a deep learning approach for our classifier. We adapt the deep learning architecture used in [5]. The question is encoded with a 1024-dimensional Long Short Term Memory (LSTM) model that takes in a one-hot descriptor of each word in the question. The image is described with the 4096-dimensional output from the last fully connected layer of the Convolutional Neural Network, VGG16 [32]. The system performs an element-wise multiplication of the image and question features, after linearly transforming the image descriptor to 1024 dimensions. The final layer of the architecture is a softmax layer.

We train the system to predict (dis)agreement labels with training examples, where each example includes an image, question, and label. At test time, given a novel visual question, the system outputs an unnormalized log probability indicating its confidence in the disagreement label, which we normalize to produce probabilities in [0,1]. Larger values reflect greater likelihood for crowd disagreement.

Analysis of Prediction System

We now describe our studies to assess the predictive power of the *CrowdVerge* systems to decide whether visual questions will lead to answer (dis)agreement from a crowd.

We evaluate our methods on two datasets that represent visual questions asked by blind and sighted users. We chose to focus only on visual questions about real images to align our analysis with real practical challenges (fewer image-based challenges arise with simple abstract scenes). One dataset is the VQA Real Images [5]. From the 369,861 visual questions about real images, 248,349 are kept for for training and the remaining 121,512 are employed for testing (i.e., Training and Validation 2015 v1.0 datasets). The other dataset is the VizWiz [6] dataset. We apply a random 80/20 train/test split to the 1,499 visual questions for which we collected answers, resulting in 1,200 training images and 299 test images.

To our knowledge, no prior work addresses predicting answer (dis)agreement for visual questions. Thus, the best a user can achieve today is to randomly decide if a visual question will lead to disagreement. For this reason, we compare our systems to a `Status Quo` predictor which returns a random value in its confidence in disagreement. We also compare our systems to a related VQA algorithm [23, 5] which produces for a given visual question an answer with a confidence score. This system parallels the deep learning architecture we adapt. However, it

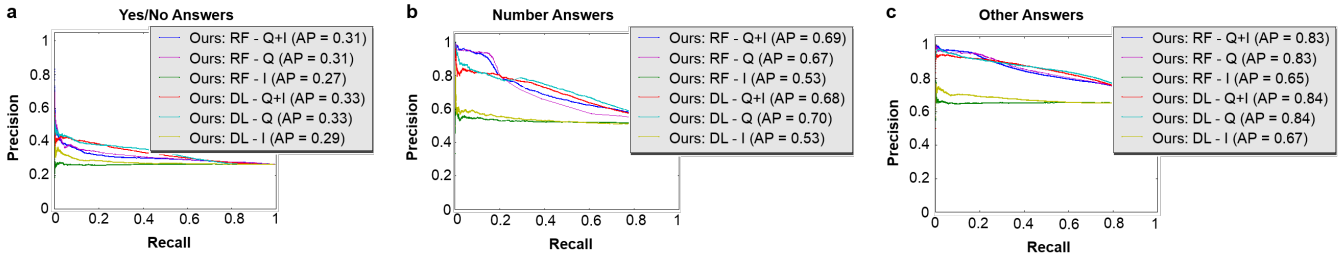

Figure 5. Precision-recall curves and average precision (AP) scores for our random forest (RF) and deep learning (DL) classifiers with different features (Question Only - Q; Image Only - I; Q +I) for visual questions that lead to (a-c) three answer types. (Best viewed on pdf.)

predicts *the system's uncertainty in its own answer*, whereas we are interested in *the collective disagreement from a crowd* on the answer. Still, it is a useful baseline to see if an existing algorithm could serve our purpose.

Classification Performance

We evaluate the predictive power of the *CrowdVerge* systems on the two datasets separately. We first show performance of the baseline and our two *CrowdVerge* systems using *precision-recall curves*. We also report the *average precision* (AP), which indicates the area under a precision-recall curve. Precision, recall, and AP values range from 0 to 1 with better-performing prediction systems having larger values.

Figures 4a,c show precision-recall curves for all prediction systems on both datasets[4]. Our proposed *CrowdVerge* systems outperform the `Status Quo` and `ICCV 2015` [5] baselines on both datasets; e.g., for the VQA dataset, `Ours: RF` yields a 27 percentage point improvement with respect to AP over `Status Quo` and a 12 percentage point improvement with respect to AP over `ICCV 2015` [5]. Our findings demonstrate there is value in learning the (dis)agreement task specifically, rather than employing an algorithm's confidence in its answers. More generally, our results demonstrate it is possible to predict whether a crowd will agree on a single answer from a given image and associated question, even for the poor quality images and more free-form natural language questions often observed from blind people (i.e., VizWiz dataset). Despite the significant variety of questions and image content and despite the variety of reasons for which the crowd can disagree, our learned model is able to produce quite accurate results.

We observe our Random Forest classifier outperforms our deep learning classifier; e.g., `Ours: RF` yields a three percentage point improvement with respect to AP while consistently yielding improved precision-recall values over `Ours: DL` (**Figure 4a**). In general, deep learning systems hold promise to replace handcrafted features to pick out the discriminative features. Interestingly, however, we find that handcrafted features (the SOS [36] results, etc.) actually do have an advantage over standard VQA deep learning architectures for our task. We hypothesize this is due to having inadequate training data for training the higher-capacity deep learning model.

We show examples of prediction results where our top-performing `RF` classifiers make their most confident predictions (**Figures 4b,d**). In these examples, for the VQA dataset, the predictor expects human agreement for "what room... ?" visual questions and disagreement for "why... ?" visual questions. For the VizWiz dataset, the predictor expects human disagreement for "can you... ?" visual questions. These examples highlight that the *CrowdVerge* systems may have a strong language prior towards making predictions, as we will discuss in the next section.

[4]We do not show results from the deep learning implementation on the VizWiz dataset because we had an insufficient number of training examples to successfully train such a system.

Predictive Cues for Answer (Dis)Agreement

We now explore what makes a visual question lead to crowd answer agreement versus disagreement. Here we focus on the larger VQA Real Images dataset.

We analyze the predictive power of our random forest (`RF`) and deep learning (`DL`) classification systems for visual questions that lead to the three types of answers ("yes/no", "number", "other") independently. Moreover, we enrich our analysis by also examining the predictive performance of both *CrowdVerge* systems when they are trained and tested exclusively with image and question features respectively. **Figure 5** shows precision-recall curves for both *CrowdVerge* systems with question features alone (`Q`), image features alone (`I`), and both question and image features together (`Q+I`).

When comparing AP scores (**Figure 5**), we observe our `Q+I` predictors yield the greatest predictive performance for visual questions that lead to "other" answers, followed by "number" answers, and finally "yes/no" answers. Interestingly, this resembles the trend we saw in **Table 1**. Accordingly, we hypothesize that the question wordings that lead to the different types of answers yield different predictive strength.

We observe that question-based features offer greater predictive power than image-based features for all answer types, when comparing AP scores for `Q` and `I` classifiers (**Figure 5**). Still, image features contribute to performance improvements for our random forest classifier for visual questions that lead to "number" answers, as illustrated by comparing AP scores for `Our RF: Q+I` and `Our RF: Q` (**Figure 5b**). Our overall finding that most of the predictive power stems from language-based features parallels feature analysis findings in the automated VQA literature [5, 25]. Further work improving visual content cues for VQA agreement is warranted.

Our findings suggest that the Random Forest classifier's overall advantage over the deep learning system arises because of "number" visual questions, as indicated by higher AP scores (**Figure 5**). For example, the advantage of the initial higher

precision (**Figure 4a**; `Ours: RF` vs `Ours: DL`) is also observed for "number" visual questions (**Figure 5b**; `Ours: RF - Q+I` vs `Ours: DL - Q+I`). We hypothesize this advantage arises due to the strength of the Random Forest classifier in pairing the question prior ("How many?") with the image-based SOS features that indicates the number of objects in an image. Specifically, we expect "how many" to lead to agreement only for small counting problems.

CAPTURING ANSWER DIVERSITY WITH LESS EFFORT

We next present a novel resource allocation system for efficiently capturing the *diversity of valid answers* for a batch of visual questions. Today's status quo is to either uniformly collect N answers [5] or let external crowdsourcing conditions determine the number of answers [6] per visual question. Our system instead spends a human budget by predicting how many answers to collect per visual question based on whether multiple answers are predicted to be redundant.

Answer Collection System

Suppose we have a budget B which we can allocate to collect extra answers for a subset of visual questions. Our system automatically decides to which visual questions to allocate the "extra" answers in order to maximize captured answer diversity for all visual questions.

The aim of our system is to accrue additional costs and delays from collecting extra answers only when extra responses will provide more information. Our system involves three steps to collect answers for all N visual questions (**Figure 6**). First, the system applies our *CrowdVerge* system to every visual question in the batch. We employ the random forest classifier, our top-performing option. Then, the system ranks the N visual questions based on predicted scores from the classifier, from visual questions most confidently predicted to lead to answer "agreement" to those most confidently predicted to lead to answer "disagreement" from a crowd. Finally, the system solicits more (R) human answers for the B visual questions predicted to reflect the greatest likelihood for crowd disagreement and fewer (S) human answers for the remaining visual questions. More details below.

Analysis of Answer Collection System

We now describe our studies to assess the benefit of our allocation system to reduce human effort to capture the diversity of all answers to visual questions.

Experimental Design

We evaluate the impact of actively allocating extra human effort to answer visual questions as a function of the available budget of human effort. Specifically, for a range of budget levels, we compute the total measured answer diversity (as defined below) resulting for the batch of visual questions. The goal is to capture a large amount of answer diversity with little human effort. This is beneficial both when using a system like VizWiz to adequately answer a blind person's visual question as well as to economically create a dataset for the development of automated VQA systems.

We conduct our studies on the 121,811 test visual questions from the VQA Real Images and VizWiz datasets. For each

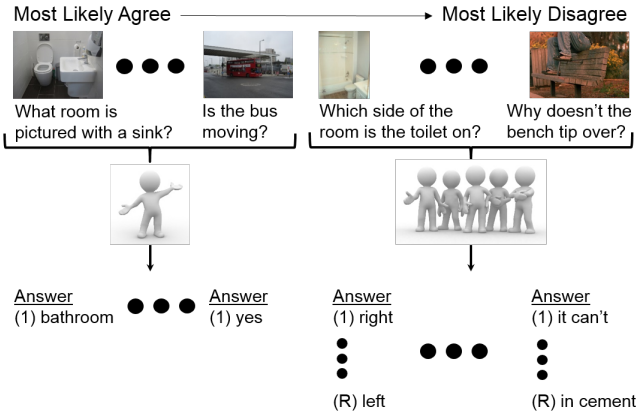

Figure 6. We propose a novel application of predicting the number of redundant answers to collect from the crowd per visual question to efficiently capture the diversity of all answers for all visual questions. For a batch of visual questions, our system first produces a relative ordering using the predicted confidence in whether a crowd would agree on an answer (upper half). Then, the system allocates a minimum number of annotations to all visual questions (bottom, left half). Finally, the extra available human budget is allocated to visual questions most confidently predicted to lead to crowd disagreement (bottom, right half).

visual question, we establish the set of true answers as all unique answers which are observed at least twice in the 10 crowdsourced answers per visual question. We require agreement by two workers to avoid the possibility that careless or spam answers are treated as ground truth.

System Implementation

We collect either the minimum of $S = 1$ answer per visual question or the maximum of $R = 5$ answers per visual question. Our number of answers roughly aligns with existing crowd-powered VQA systems, for example with the VizWiz application, "On average, participants received 3.3 (SD=1.8) answers for each question" [6]. Our maximum number of answers also supports the possibility of capturing the maximum of three unique, valid answers typically observed in practice (recall study above). While more elaborate schemes for distributing responses may be possible, we will show this approach already proves quite effective in our experiments. We simulate answer collection by randomly selecting answers from the 10 crowd answers per visual question.

Baselines

We compare our approach to two baselines. We leverage the `ICCV 2015` [5] predictor's output confidence score from the publicly-shared model [5, 23] to rank the order of priority for visual questions to receive redundancy. We also leverage a `Status Quo` system, which randomly prioritizes which images receive redundancy. This system illustrates the best a user can achieve today with crowd-powered systems [6, 10] or with current dataset collection methods [5, 35].

Evaluation Methodology

We measure total diversity of answers captured by a resource allocation system for a batch of visual questions Q as follows:

$$D(Q) = \sum_{i=1}^{|B|} |r_i \cap q_i| + \sum_{j=1}^{|Q \setminus B|} |s_j \cap q_j| \quad (1)$$

where q_i represents the set of all true answers for the i-th visual question, r_i represents the set of unique answers captured in the R answers collected for the i-th visual question, and s_j represents the set of unique answers captured in the S answers collected for the j-th visual question. Total diversity comes from the first term when the *maximum* extra human budget (B) is available and total diversity comes from the second term when *no* extra human budget is available. Given a partial extra human budget (B), the aim is to have perfect predictions such that the minimum number of answers (S) are allocated only for visual questions with one true answer in order for all diverse answers to be safely captured.

We measure diversity per visual question as the number of all true answers collected per visual question (e.g., $|r_i \cap q_i|$). Larger values reflect greater captured diversity. The motivation for this measure is to only give total credit to visual questions when all valid, unique human answers are collected.

Results

Our system consistently offers significant gains over today's `Status Quo` approach for visual questions from sighted (**Figure 7a**) and blind (**Figure 7b**) users. For example, for the VQA dataset, our system accelerates the collection of 70% of the diversity by 21% over the `Status Quo` baseline. Our system also accelerates the collection of the diversity one would observe with the VizWiz system (i.e., average of 3.3 answers per visual question) by 23% for the VQA dataset and 14% for the VizWiz dataset. We hypothesize the greater performance gains on the VQA dataset than the VizWiz dataset arise due to the greater amount of training data; i.e., the difference in order of magnitude is over 100. In absolute terms for the VQA dataset, our system eliminates the collection of 92,180 answers with no loss to captured answer diversity. This translates to eliminating over 11 40-hour work weeks and saving $1800, assuming workers are paid $0.02 per answer and take 18 seconds to answer a visual question[5]. Our approach fills an important gap in the crowdsourcing answer collection literature for targeting the allocation of extra answers only to visual questions where a diversity of answers is expected.

Figure 7 also illustrates the advantage of our system over a related VQA algorithm [5] for our novel application of cost-sensitive answer collection from a crowd. As observed, relying on an algorithm's confidence in its answer offers a valuable indicator over today's status quo of passively budgeting. While we acknowledge this method is not intended for our task specifically, it serves as an important baseline to ensure VQA system uncertainty is insufficient to gauge crowd convergence. We attribute the further performance gains of our prediction system to it directly predicting whether *humans* will disagree rather than predicting a property of a specific *algorithm* (i.e., confidence of the VQA algorithm [5] in its answer prediction).

A valuable area for future work to achieve further cost-savings, time-savings, and user satisfaction for answer collection is to employ fine-grained predictions of the exact number of answers to expect per visual question. This information would guide how many answers to collect. One possible challenge would be successfully training a fine-grained prediction system, since this depends on employing a sufficiently large dataset that represents each possible number of valid answers per visual question (e.g., 3 vs 5 answers) with a balanced, large number of examples. Another possible direction is to develop online approaches, such as has been done in prior work [11], by dynamically deciding when enough answers are collected from the crowd to capture all valid answers. Additionally, establishing how to pair such a system with low-latency response mechanisms, such as quikTurkit [6], would be a valuable step towards making such systems amenable for real-time use.

[5]To make a fair comparison to the VizWiz system [6], we use the timing and price information reported in that paper.

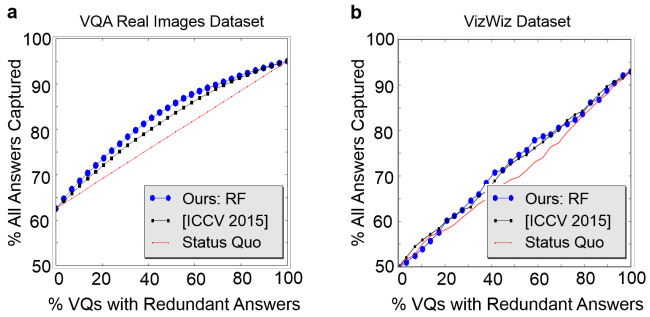

Figure 7. We show results for our system, a related VQA algorithm, and the status quo (which lacks any active prioritization) for (a) 121,512 visual questions from the VQA dataset and (b) 299 visual questions from the VizWiz dataset. Boundary conditions are one answer (leftmost) and five answers (rightmost) for all visual questions. Our approach typically accelerates capturing answer diversity compared to `Status Quo` selection by over 20% for the VQA dataset and 14% for the VizWiz dataset.

CONCLUSIONS

We proposed a new problem of predicting whether different people would give the same answer to a visual question. Towards motivating the practical implications for this problem, we analyzed nearly half a million visual questions and demonstrated there is nearly a 50/50 split between visual questions that lead to answer agreement versus disagreement. We observed that crowd disagreement arose for various types of answers (yes/no, counting, other) for many different reasons. We next proposed a system that automatically predicts whether a visual question will lead to a single versus multiple answers from a crowd. Our method outperforms a strong existing VQA system limited to estimating *system* uncertainty rather than *crowd* disagreement. Finally, we demonstrated how to employ the prediction system to accelerate the collection of diverse answers from a crowd by typically at least 14%-23% over today's status quo of fixed redundancy allocation for visual questions asked by blind and sighted users.

ACKNOWLEDGMENTS

We gratefully acknowledge funding from the National Science Foundation (IIS-1065390) and Office of Naval Research (YIP N00014-12-1-0754); thank Dinesh Jayaraman, Aron Yu, Yu-Chuan Su, Suyog Jain, and Chao-Yeh Chen for their assistance with setting up experiments; and thank Aishwarya Agrawal for sharing her crowdsourcing answer collection code.

REFERENCES

1. Be My Eyes. http://www.bemyeyes.org/.
2. L. Von Ahn and L. Dabbish. 2004. Labeling Images With a Computer Game. In *ACM Conference on Human Factors in Computing Systems (CHI)*. 319–326.
3. E. Amid and A. Ukkonen. 2015. Multiview Triplet Embedding: Learning Attributes in Multiple Maps. In *International Conference on Machine Learning (ICML)*. 1472–1480.
4. J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. 2016. Learning to Compose Neural Networks for Question Answering. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*. 1545—1554.
5. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. 2015. VQA: Visual Question Answering. In *IEEE International Conference on Computer Vision (ICCV)*. 2425–2433.
6. 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. 2010. VizWiz: Nearly Real-time Answers to Visual Questions. In *ACM symposium on User interface software and technology (UIST)*. 333–342.
7. E. Brady, M. R. Morris, and J. P. Bigham. 2015. Gauging Receptiveness to Social Microvolunteering. In *ACM Conference on Human Factors in Computing Systems (CHI)*. 1055–1064.
8. E. Brady, M. R. Morris, Y. Zhong, S. White, and J. P. Bigham. 2013. Visual Challenges in the Everyday Lives of Blind People. In *ACM Conference on Human Factors in Computing Systems (CHI)*. 2117–2126.
9. L. Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.
10. M. A. Burton, E. Brady, R. Brewer, C. Neylan, J. P. Bigham, and A. Hurst. 2012. Crowdsourcing Subjective Fashion Advice Using VizWiz: Challenges and Opportunities. In *ACM SIGACCESS conference on Computers and accessibility (ASSETS)*. 135–142.
11. P. Dai, Mausam, and D. S. Weld. 2010. Decision-Theoretic Control of Crowd-Sourced Workflows. In *AAAI Conference on Artificial Intelligence*. 1168–1174.
12. J. Deng, O. Russakovsky, J. Krause, M. Bernstein, A. Berg, and L. Fei-Fei. 2014. Scalable Multi-label Annotation. In *ACM Conference on Human Factors in Computing Systems (CHI)*. 3099–3102.
13. A. Guo, X. Chen, H. Qi, S. White, S. Ghosh, C. Asakawa, and J. P. Bigham. 2016. VizLens: A Robust and Interactive Screen Reader for Interfaces in the Real World. In *ACM symposium on User interface software and technology (UIST)*. 651–664.
14. D. Gurari, S. D. Jain, M. Betke, and K. Grauman. 2016. Pull the Plug? Predicting If Computers or Humans Should Segment Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 382–391.
15. K. Hara, J. Sun, R. Moore, D. Jacobs, and J. Froehlich. 2014. Tohme: Detecting Curb Ramps in Google Street View Using Crowdsourcing, Computer Vision, and Machine Learning. In *ACM symposium on User interface software and technology (UIST)*. 189–204.
16. S. D. Jain and K. Grauman. 2013. Predicting Sufficient Annotation Strength for Interactive Foreground Segmentation. In *IEEE International Conference on Computer Vision (ICCV)*. 1313–1320.
17. M. Jas and D. Parikh. 2015. Image Specificity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2727–2736.
18. R. Krishna, K. Hata, S. Chen, J. Kravitz, D. A. Shamma, L. Fei-Fei, and M. S. Bernstein. 2016. Embracing Error to Enable Rapid Crowdsourcing. In *ACM Conference on Human Factors in Computing Systems (CHI)*. 3167–3179.
19. G. Laput, W. S. Lasecki, J. Wiese, R. Xiao, J. P. Bigham, and C. Harrison. 2015. Zensors: Adaptive, Rapidly Deployable, Human-Intelligent Sensor Feeds. In *ACM Conference on Human Factors in Computing Systems (CHI)*. 1935–1944.
20. W. S. Lasecki, P. Thiha, Y. Zhong, E. Brady, and J. P. Bigham. 2013. Answering Visual Questions with Conversational Crowd Assistants. In *ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*. 1–8.
21. C. H. Lin, Mausam, and D. S. Weld. 2014. To Re(label), or Not To Re(label). In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. 151–158.
22. T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *IEEE European Conference on Computer Vision (ECCV)*. 740–755.
23. J. Lu, X. Lin, D. Batra, and D. Parikh. 2015. Deeper LSTM and normalized CNN Visual Question Answering model. `https://github.com/VT-vision-lab/VQA_LSTM_CNN`.
24. M. Malinowski and M. Fritz. 2014. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Advances in Neural Information Processing Systems (NIPS)*. 1682–1690.
25. M. Malinowski, M. Rohrbach, and M. Fritz. 2015. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In *IEEE European Conference on Computer Vision (ECCV)*. 1–9.
26. A. T. Nguyen, B. C. Wallace, and M. Lease. 2015. Combining Crowd and Expert Labels Using Decision Theoretic Active Learning. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. 120–129.

27. G. Patterson, G. V. Horn, S. Belongie, P. Perona, and J. Hays. 2015. Tropel: Crowdsourcing Detectors with Minimal Training. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. 150–159.

28. A. Sarkar, C. Morrison, J. F. Dorn, R. Bedi, S. Steinheimer, J. Boisvert, and L. Walsh. 2016. Setwise Comparison: Consistent, Scalable, Continuum Labels for Computer Vision. In *ACM Conference on Human Factors in Computing Systems (CHI)*. 261–271.

29. B. Settles. 2010. *Active Learning Literature Survey*. Technical Report. University of Wisconsin, Madison. 65 pages.

30. V. S. Sheng, F. Provost, and P. G. Ipeirotis. 2008. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. 614–622.

31. A. Sheshadri and M. Lease. 2013. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. 156–164.

32. K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

33. S. Vijayanarasimhan and K. Grauman. 2011. Cost-sensitive Active Visual Category Learning. In *International Journal of Computer Vision (IJCV)*, Vol. 91. 24–44.

34. P. Welinder, S. Branson, S. Belongie, and P. Perona. 2010. The Multidimensional Wisdom of Crowds. In *Advances in Neural Information Processing Systems (NIPS)*. 2424–2432.

35. L. Yu, E. Park, A. C. Berg, and T. L. Berg. 2015. Visual Madlibs: Fill in the blank Image Generation and Question Answering. In *IEEE International Conference on Computer Vision (ICCV)*. 2461–2469.

36. J. Zhang, S. Ma, M. Sameki, S. Sclaroff, M. Betke, Z. Lin, X. Shen, B. Price, and R. Mech. 2015. Salient Object Subitizing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4045–4054.

37. Y. Zhong, W. S. Lasecki, E. Brady, and J. P. Bigham. 2015. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *ACM Conference on Human Factors in Computing Systems (CHI)*. 2353–2362.