Why Does a Visual Question Have Different Answers?

Anonymous Author(s)

ABSTRACT

Visual Question Answering (VQA) is a popular task of returning the answer for a question about an image. A key problem for this task is that different people may return different answers. Moreover, little is known why such answers differ. We propose a taxonomy of nine plausible reasons explaining why, and asked crowd workers to annotate which of these reasons led to answer disagreements for roughly 35,000 visual questions asked by blind and sighted people. Our results highlight why disagreements arise in practice, as well as which reasons are unique to different domains. We also propose a problem of predicting directly from a visual question (plus optionally answers) which reasons will lead the answers to differ, and present two implementations of a machine learning model for this purpose. We demonstrate that these systems can predict such reasons with a precision as high as 94%.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Information systems** → *Crowdsourcing*;

KEYWORDS

crowdsourcing, visual question answering

ACM Reference format:
Anonymous Author(s). . Why Does a Visual Question Have Different Answers?. In *Proceedings of* , , , 14 pages.
https://doi.org/

1 INTRODUCTION

An important task is to answer questions about images [5, 10]. However, a challenge is that multiple people can return a diversity of answers for the same visual question (VQ) [32]. A critical step towards returning a desired answer is understanding why different people provide different answers.

Previous research has proposed several reasons why answer differences may arise from a crowd. Reasons that are commonly discussed include ambiguity [59], spam [27], and subjectivity [67]. Moreover, complementary works examine how to resolve these differences and arrive at a true answer [22, 23, 36]. Yet, prior works address a single reason rather than bringing all such reasons of differences (henceforth referred to as 'reasons for answer disagreement' or 'disagreement-sources') under a single umbrella.

Our work fills this gap in the literature in order to support the design of machines that can automatically select for themselves the appropriate actions needed to resolve answer differences. First, we propose a taxonomy of nine plausible reasons of why people disagree when answering a VQ. These reasons are illustrated in Figure 1. Generally, we established that disagreements can occur because there are issues with the question-image (QI) pair (first and second column), or there are issues with the ten answers (last column). We asked crowd-workers to annotate which of these reasons led to answer disagreement for roughly 35,000 VQs asked by blind and sighted users. Results show that three prominent sources cause answer disagreement in over 90% of the VQs. We also propose the novel problem of predicting these disagreement-sources directly from a VQ (plus optionally answers), and present two machine learning systems for automatically addressing this problem. Experimental result demonstrate promising performance from one of the systems in anticipating why answers will differ.

2 RELATED WORK

Understanding Why a Crowd Disagrees

A precursor to deciding how to respond when different people provide different responses for a task is understanding why people return different responses. A variety of reasons have been explored as possible causes of crowd disagreement including difficulty [71], ambiguity [38, 42], subjectivity [53], and spam or malicious answers [27]. However, to the best of our knowledge, no work has yet enumerated a comprehensive list of possible reasons for disagreements and no study has examined their prevalence in practice. Accordingly, motivated from the domain of visual question answering, we propose a taxonomy of reasons that could lead to disagreements from a crowd and conduct large-scale analysis to uncover the significance of each reason in practice.

Resolving Crowd Disagreement

Previous efforts have studied various causes of disagreement separately (usually in domains outside VQA), and posed solutions for resolving crowd disagreements. Commonly, disagreement is considered a measure of poor quality in the annotation task, for example because the task is poorly defined or because the annotators' lack of knowledge. Numerous works try to employ disagreement as a valuable signal for

, ,
. ACM ISBN …$15.00
https://doi.org/

Figure 1: Examples of VQs asked by blind and sighted users, and corresponding answers from 10 different people. As illustrated, the 10 answers can differ for a variety of reasons, including reasons arising from the question-image (QI) pair (first and second column), or from the answers (third column). We propose a system which can take a QI pair (and optionally, the 10 answers) as input, and automatically predict the reason(s) for which the answers may differ, if they do.

uncovering a true answer [3, 6, 7, 22, 24, 34–36, 59, 63, 67]. Some works identify which among multiple responses to trust [60, 71]. Others embrace context to resolve ambiguity [2]. Each work embeds assumptions regarding why answers differ such as because of ambiguity, subjectivity, difficulty, etc in order to pose a solution for recovering a true answer. Yet, a challenge is knowing which assumptions are valid when and so which methods to use when. Our work most closely relates to the CrowdVerge system which anticipates whether a crowd will disagree [32]. Our work goes a step further and offers a solution to automatically uncover *why* a crowd will disagree, critical information for deciding which method is best-suited to resolve disagreements.

Visual Question Answering (VQA) Datasets

Motivated by the recent excitement about the VQA problem, many datasets have been introduced to foster active research in developing artificial intelligence systems that automatically answer VQs [4, 5, 29, 31, 37, 40, 41, 43, 45, 50, 57, 69, 70, 72, 73]. Yet, a limitation of prior work is they assume the goal is to return a single answer despite the fact that VQs often lead to multiple answers from different people [5, 32]. Our work enriches prior work by introducing meta-data revealing why crowds disagree when their answers differ for two existing popular VQA datasets: VizWiz [33] and VQA [5]. Our extension of these datasets provides a critical foundation for the development of machine learning algorithms that can automatically identify why answers will differ for VQs coming from a diversity of users including people who are both blind (i.e., VizWiz) and sighted (i.e., VQA dataset).

Visual Dialog

Several works have attempted to support a continuous dialog to enable the remote humans supplying the answers to return a single desired answer. For example, Be My Eyes provides a direct connection between the asker and answerer [1]. Chorus:View simulates a direction connection between a asker and answerer by embracing a back-end crowd to engage as a single conversational partner [44]. Visual Dialog proposes an algorithm to automatically engage in a conversation with a person [20]. Our work can offer an alternative way of creating a dialog, by (a) identifying whether the answering crowd will agree on the answer to the given VQ, along with a reason for disagreement, if any, and (b) providing automatic feedback on how to best resolve the disagreement, thereby helping users to ask VQs that will achieve answer convergence more quickly and cheaply.

3 METHODOLOGY

We now describe the datasets and crowdsourcing system we designed to collect disagreement-source labels.

Datasets

We employ two popular VQA datasets that reflect a diversity of VQs coming from blind and sighted users. We describe these datasets below.

VizWiz: The *VizWiz* dataset [33] originates from blind users [10], who snapped photos using a smart-phone and recorded spoken questions about the photos. These VQs 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 "has the milk expired?" [11]. The VQs represent a real-world use-case scenario where a person is interactively exploring and learning about his/her surrounding physical world. Since blind people cannot see and verify the quality of the pictures they take, many images are ill-framed, lack proper illumination, or are out-of-focus. Each VQ comprises an image and a transcription of the spoken question (the QI pair), and ten answers crowdsourced from Amazon Mechanical Turk (AMT) workers. For our initial analysis, we used the entire *VizWiz* dataset, excluding the VQs where all answers are identical using exact string matching (i.e. no answer disagreement). This resulted in 29,974 VQs from the *VizWiz* dataset.

VQA_2.0: We also examine VQs asked by sighted users from the VQA 2.0 Balanced Real Images dataset [31]. Unlike the VQs from the *VizWiz* dataset, the images and questions in this datasets were created separately. The images were taken from the MS-COCO dataset [46], and the questions came from crowd-workers, who were instructed to ask such a question about the image that can 'stump' a 'smart robot' [5]. Most of the images have high photographic quality. Like the *VizWiz* dataset, each VQ comprises a QI pair and ten crowdsourced answers. For our experiment, we have randomly selected 5,032 QI pairs from the training set of the Balanced Real Images dataset, for which the ten crowdsourced answers were not identical (using exact string matching).

Taxonomy Design

Informed by existing literature, and an initial inspection of a subset of the VQs, we propose a taxonomy of nine plausible reasons as to why the answers to a visual question can differ.

We classify our nine reasons into two groups, based on whether they originate due to issues or problems with the QI pair, or due to issues with the ten answers. This grouping helps to localize the source of answer-disagreement to either the QI pair, or the 10 answers. We hypothesize that disagreement-resolution strategies for issues with QI pair will be different from those for issues with answers. Since the long-term goal is to develop automated systems that can identify and predict answer-disagreement in the crowd, localizing the source of crowd-disagreement will serve as first-steps for choosing disagreement-resolution strategies.

For VQs with **issues with the QI pair**, we propose the following six sources of answer-disagreement:

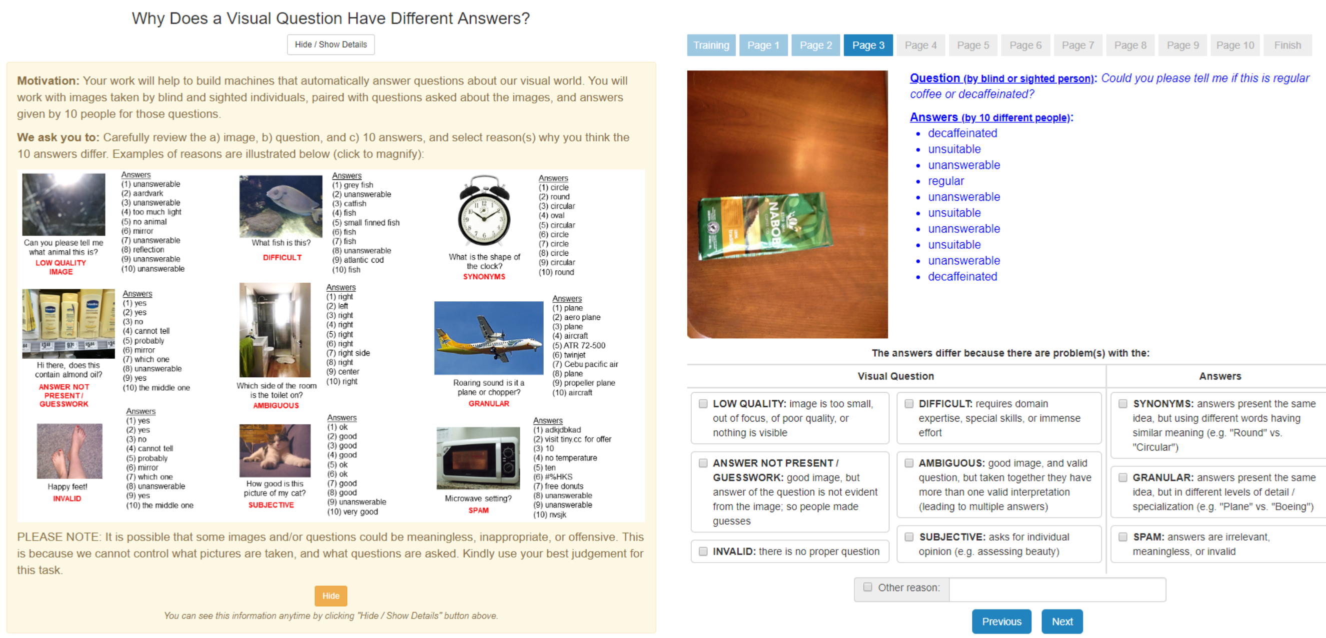
- ***Low Quality Image*** **(LQI)**: image is too small, out of focus, having poor quality, or nothing is visible.
- ***Answer Not Present / Guesswork*** **(IVE)**: good image, but answer to the question is not present in the image (Insufficient Visual Evidence), so some answers reflect guesses.
- ***Invalid*** **(INV)**: a proper or semantically correct question is absent [51].
- ***Difficult*** **(DFF)**: questions that require domain expertise (e.g. identifying if a skin-rash is due to bug bite), special skills, or too much effort (e.g. counting the number of sheep in a field full of sheep) [71].
- ***Ambiguous*** **(AMB)**: good image and valid question, but taken together they have more than one valid interpretation, leading to multiple answers [38, 42, 67].
- ***Subjective*** **(SBJ)**: opinion-driven questions, such as assessing beauty, fashion sense, emotions [15, 53, 67].

For VQs with **issues in the crowdsourced answers**, we propose the following three sources of answer-disagreement:

- ***Synonyms*** **(SYN)**: answers present the same idea, but using different words having similar meaning (e.g. 'round' versus 'circular') [51].
- ***Granular*** **(GRN)**: answers present the same idea, but in different levels of detail / specialization (e.g. 'plane' versus 'Boeing')
- ***Spam*** **(SPM)**: a person inadequately answers a simple, straight-forward visual question [25, 27, 65, 66].

Though our taxonomy can cover a wide range of disagreements, we kept a reason called ***Other*** **(OTH)**, linked to a free-entry text-box. Workers who felt none of the above reasons are well qualified, could enter what they thought was the relevant reason.

To develop this taxonomy we employed a three step process: (1) we examined causes cited in existing crowdsourcing literature, and identified six of the nine labels – INV, DFF, AMB, SBJ, SYN, and SPM; (2) we inspected VQs from the two datasets and introduced three labels that we identified were missing – LQI, IVE, and GRN; (3) finally, we used a pilot



(a) (b)

Figure 2: (a) Task instructions with examples to train crowd workers about all the disagreement-sources. (b) The user interface crowd workers used for choosing why different answers are observed for a given QI pair, and the 10 corresponding answers.

crowdsourcing task with 100 VQs which allowed crowdworkers to highlight missing labels, by either selecting the "OTH" category, or leaving feedback comments. We did not find any major labels appearing in the crowdworkers' OTH answers or feedback comments of the actual crowdsourcing experiment, that we missed in the pilot study.

Crowdsourcing System

We used Amazon Mechanical Turk (AMT) platform to crowdsource our disagreement labels. Our system worked as follows: on accepting a HIT (Human Intelligence Task) hosted by us on the AMT platform, the user was presented with the task instructions (Figure 2a) and a training task. The task instructions showed examples of each of the disagreement-source labels. The layout of the training and actual tasks are shown in Figure 2b. It contains a QI pair, the ten (crowdsourced) answers, and a list of checkboxes for selecting the labels. Checkboxes support selecting multiple labels. We included the definition of the label in the click-area for quick reference. The labels are grouped into the two classes (issues with QI pair, and issues with answers, as discussed in Section 3) illustrating which disagreement-sources fall into which category, and also guiding the crowd worker in deciding which label(s) to select.

For the training task, the correct labels were pre-determined by us, and the worker had to select the correct labels to proceed to the actual task. The worker was shown the correct labels is (s)he had chosen a wrong label and clicked "Next".

After the training task, the worker was presented with ten VQs for annotation. The worker was made to select at least one label in the current VQ before proceeding to the next. There was an optional feedback form in the end. We presented each HIT to five crowd-workers, and thus collected five sets of labels for each VQ.

4 DESCRIPTIVE ANALYSIS

Our first aim is to understand why people disagree when answering visual questions. We analyze the 175,040 crowdsourced labels collected, to learn: (1) what are the most common reasons for answer disagreement? and (2) how many unique reasons typically provoke answer disagreement for a single VQ?

Common Sources of Answer Disagreement

We first quantify how often differing answers from numerous people can be explained by our nine proposed disagreement-sources. We tally how many of the 35,008 VQs are labelled with each disagreement-source label. To account for different

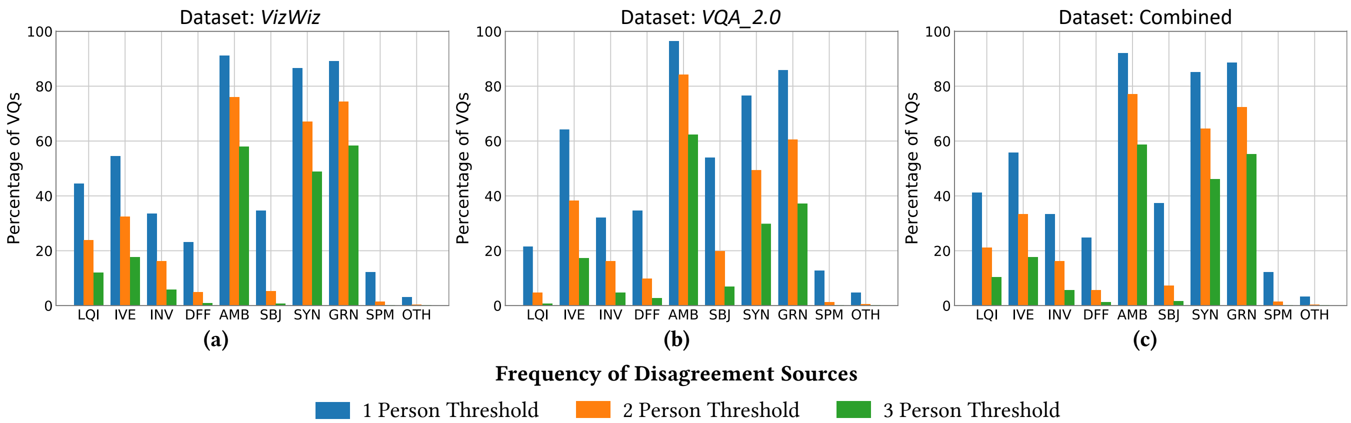


Figure 3: (a) - (c): Summary of why the ten crowdsourced answers of a VQ are different. The histograms show the frequency of each disagreement-source label (Sec. 3) for (a) 29,974 VQs asked by blind people, (b) 5,034 VQs asked by sighted people, and (c) combination of the previous two. The plots are computed based on increasing thresholds of inter-worker agreement required to make a disagreement-source label valid: only one (out of five) worker has to select a label, at least two workers must agree on a label, and at least three workers must agree on the label. Our findings show that ambiguous questions (AMB), synonymous answers (SYN), and varying levels of answer granularity (GRN), are the three most popular reasons of answer disagreement.

levels of trust in the crowd workers, we report results based on increasing thresholds of inter-worker agreement:

- *Trust All*: only one worker has to select a disagreement-source label (1-person validity threshold)
- *Trust Any Pair*: at least two workers must agree on a label (2-person threshold)
- *Trust Majority*: at least three workers must agree on a label (3-person threshold)

Results are shown in Figure 3.

'Ambiguous', 'Synonyms', and 'Granular': In both *VizWiz* and *VQA_2.0* datasets, ambiguous questions (AMB), synonymous answers (SYN), and varying levels of answer granularity (GRN), are the three **most common disagreement-sources** (Figure 3). For example, AMB is the top reason for answer difference, with at least two people selecting it as the reason for 76% of the *VizWiz* VQs, and 84% of the *VQA_2.0* VQs (Figures 3a, b; 2-person threshold). GRN is the closely-following second choice, with it being the reason for 74% of *VizWiz* and 61% of *VQA_2.0* (Figures 3a, b; 2-person threshold). Synonymous answers (SYN) is the third most common reason, occuring for 67% of *VizWiz* and 49% of *VQA_2.0* questions (Figures 3a, b; 2-person threshold). Therefore, a promising way to resolve a large portion of the answer differences in VQs is to establish techniques that handle ambiguity, synonyms, and granularity. Previous works that trained systems to ask non-ambiguous, discriminating questions [45], improve task clarity [28], and model ambiguity [67] may be effectively applied in such scenarios.

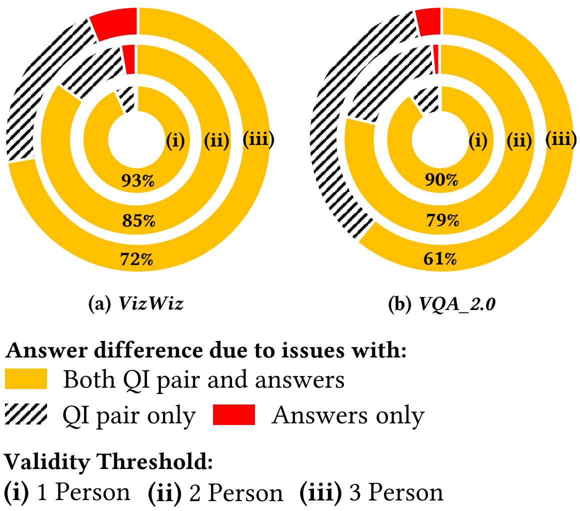
We visually inspected the VQs to identify plausible reasons why these disagreement-sources arise.

In the *VizWiz* dataset, most ambiguous (AMB) examples are of the form "What (object) is this . . . ?", and AMB is selected because the images show multiple objects (e.g. 'store', 'shopping area', 'shopping cart'). AMB also occurs because users were engaged in a dialog with the VizWiz mobile application [10], which resulted in some questions having the form 'Okay, how about now?' or 'Okay, is this correct?', which are apparent continuations of previously asked questions.

Synonym (SYN) occurs when the answerers used different words or phrases to present the same idea ('man', 'guy', 'male person').

Granularity (GRN) is most observed for questions trying to elicit **colour-related information** (colours of clothing, make-up, or everyday objects), with answerers providing varying levels of detail ('green', 'green-yellow', 'green, yellow and blue rims').

In the *VQA_2.0* dataset, AMB is often chosen when the question is lengthy (e.g. 'What weather related event can be seen under the clouds in the horizon?'). We hypothesize that overly long questions can be confusing [16], and therefore people produce a diversity of answers based on their individual understanding of the question. AMB also occurs when questions are intentionally ambiguous (e.g. 'Q: Where are the baby elephants? Ans 1: right, Ans 2: on the grass, Ans 3: next to mom and dad, etc.). These intentionally ambiguous questions are present because the *VQA_2.0* questions were created with the aim of stumping a smart robot [5].



(a) *VizWiz* (b) *VQA_2.0*

Figure 4: Proportions of VQs where disagreement occurs due to issues with the QI pair only (red), issues with the 10 answers only (striped), or issues with both (yellow, with percentage), for both *VizWiz* (a) and *VQA_2.0* (b) datasets, across the three validity thresholds. Most VQs have issues with both the QI pair *and* the ten answers.

Synonyms (SYN) in *VQA_2.0* occur in the same context as in the *VizWiz* dataset, i.e. using different words having similar meaning ('suitcase' versus 'luggage').

Granularity (GRN) is typically chosen when the description on an item is asked (e.g. 'What is the person using / holding / carrying ...?'). The answers, as we hypothesized (Section 3), provide varying levels details for the item (e.g. 'ball', 'tennis ball', 'green tennis ball').

Other disagreement-sources:
Overall, across both datasets, we found that **spam (SPM)** was the **rarest disagreement-source**. It affected approximately 1% of VQs in both *VizWiz* and *VQA_2.0* (Figures 3a, b; 2-person threshold). This is interesting because the issue of spam has received a lot of attention in the crowdsourcing literature (e.g. [25, 27, 65, 66] to name a few). While detecting spam remains important, our findings suggest this line of work will have considerably less impact than approaches addressing the other disagreement-sources.

Since the collection of the *VizWiz* and the *VQA_2.0* datasets are very different – with *VizWiz* arising from daily visual challenges of blind users, and *VQA_2.0* containing questions which are hard for machines to answer – we expected that **reasons for answer difference across the two datasets** would be wildly different. However, it is interesting to note that the top four reasons (AMB, GRN, SYN, IVE) are identical for both datasets, across all the validity thresholds (Figures 3a, b). This suggests that there can be a wide variety of topics that humans disagree about, but only a finite number of core reasons why people disagree.

For example, while people agree that answers differ due to **Insufficient Visual Evidence (IVE)** in both datasets, the reasons to choose that label are largely distinct in the two datasets. As images in the *VizWiz* dataset are often poorly framed, they do not contain the answer to questions (e.g. 'What is in the can?' when no can is visible), resulting in IVE. Whereas in the *VQA_2.0* dataset, IVE occurs because many VQs require deductive or speculative information, which are not immediately evident from the image (e.g. 'Could the smaller giraffe reach the hay mounted on the wall?', or 'Is there likely a shower in the area with the toilet and sink?').

More generally, our disagreement-source labels also highlight **how often answer differences arise because of issues with the QI pair (LQI, IVE, INV, DFF, AMB, and SBJ), versus issues with the answers themselves (SYN, GRN, and SPM)**, across both datasets. Figure 4 shows the proportions of VQs having one or both of these issues. We initially expected that the VQs would have only one of the two issues (not both). However, our results suggest otherwise. Only a handful of VQs strictly have one issue (3% with answer-issues only, and 13% with question-issues only, for the 2-person validity threshold). The majority of VQs (85%, in the 2-person threshold) have answer-disagreement due to issues with both the QI pair *and* the ten answers. This indicates that trying localize the source of disagreement to either the QI pair or the ten answers will not be very useful, and disagreement-resolution strategies for VQA systems need to consider the entire visual question along with its answers holistically.

Number of Unique Disagreement-Sources

Next, we quantify the number of reasons leading to answer-disagreements for each example, again employing the three levels of trust in crowd workers: 1-person, 2-person, and 3-person thresholds. Results are shown in Figure 5.

Overall, we find that there are typically multiple reasons for answer differences across both datasets (Figures 5: a, b). Most commonly, i.e., for more than 55% of the *VizWiz* examples, and for almost 50% of the *VQA_2.0* dataset examples, there are three unique reasons (Figures 5: a, b; 2-person threshold). From inspection, we find that these three labels are commonly AMB, SYN and GRN, the three most common disagreement-sources. Two and four reasons are also common for both datasets (Figures 5: a, b).

This leads us to examine (1) how often two disagreement-source labels co-occur, and, (2) how often a label occurs on its own, without other labels (label *clarity*).

We measured co-occurrence of two disagreement-source labels d_i and d_j using an adaptation of causal power [17] as follows:

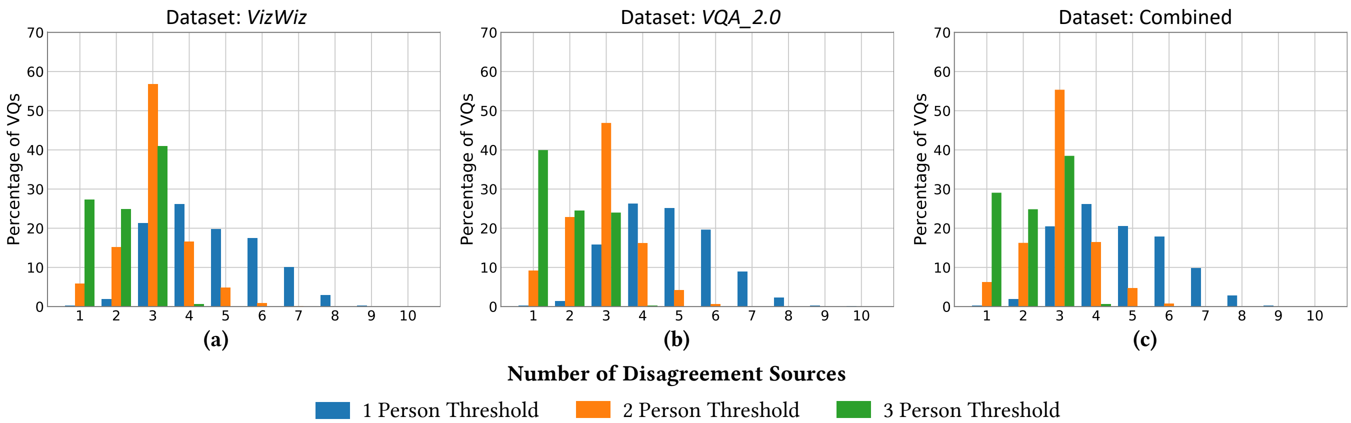


Figure 5: (a) - (c): Summary of how many unique reasons are identified as the sources of answer disagreement, when five crowd workers identify why the ten previously crowdsourced answers are different, for (a) 29,974 VQs asked by blind people, (b) 5,034 VQs asked by sighted people, and (c) combination of (a) and (b). Across both datasets, most commonly there are three unique reasons for answer-disagreement. Visual inspections show that these are the three most popular reasons: 'ambiguous', 'synonyms', and 'granularity'.

Disagreement Source Label	Co-occurs with: (%)										Label Clarity (%)
	LQI	IVE	INV	DFF	AMB	SBJ	SYN	GRN	SPM	OTH	
SYN	0	0	0	0	89	5	0	93	0	0	7
GRN	0	0	0	0	93	13	91	0	0	0	7
INV	52	91	0	30	0	10	0	0	0	0	9
AMB	0	0	0	0	0	28	85	91	0	0	9
SBJ	0	1	10	1	75	0	13	32	0	0	25
DFF	32	67	43	0	0	2	0	0	1	0	33
LQI	0	66	39	17	0	0	0	0	8	0	34
IVE	54	0	56	28	0	1	0	0	3	0	44
SPM	28	14	1	1	0	0	0	0	0	4	72
OTH	0	0	0	0	0	0	0	0	17	0	83

(a) *VizWiz*

Disagreement Source Label	Co-occurs with: (%)										Label Clarity (%)
	LQI	IVE	INV	DFF	AMB	SBJ	SYN	GRN	SPM	OTH	
INV	2	94	0	14	21	27	0	0	0	0	6
SYN	0	0	0	0	83	0	0	92	0	0	8
GRN	0	0	0	0	87	0	82	0	0	0	13
AMB	0	0	7	0	0	5	66	77	0	0	23
LQI	0	58	30	20	0	0	0	0	3	3	42
IVE	19	0	47	24	0	22	0	0	2	0	53
DFF	12	44	13	0	0	0	0	0	3	1	56
SBJ	0	26	16	0	9	0	0	0	0	0	74
OTH	12	0	0	1	0	0	0	0	13	0	87
SPM	5	12	0	7	0	0	0	0	0	5	88

(b) *VQA_2.0*

Figure 6: Co-occurrences of the disagreement-source labels for (a) *VizWiz* and (b) *VQA_2.0* datasets (1-person threshold). Label clarity denotes how often a label occurs alone. The most frequently occurring labels – AMB, SYN, and GRN – also co-occur with each other. These labels have the lowest clarity, i.e. they rarely occur alone.

*If the **co-occurrence** between labels d_i and d_j is x, then in all the VQs where d_i is chosen, d_j is chosen in x% of them.*

We chose this metric (instead of, e.g., Pearson's correlation coefficient) because this helps to correct for self-correlations, as well as for cases where d_j is chosen in the absence of d_i, and vice-versa. Further mathematical explanations are presented in [17, 49].

We also measured the *clarity* of a label d as follows:

*If the **clarity** of a label d is x, then in all the VQs where d is chosen, no other label is chosen in x% of them.*

In other words, disagreement-source d occurs alone (or, is 'clearly expressed') in only x% of the VQs it is selected. In the rest $(100 - x)$% of the VQs, d co-occurs with at least one other label. For brevity, we discuss these metrics for 1-person validity threshold only. Results are shown in **Figure 6**, for all possible pairs of labels, separately for the *VizWiz* and the *VQA_2.0* datasets.

In both *VizWiz* and *VQA_2.0*, the labels SYN (synonyms) and GRN (granular) have some of the highest co-occurrences with other labels. For example, in *VizWiz*, for all the VQs where SYN was chosen, GRN co-occurs for 93% of those questions, followed by AMB for 89%. Likewise, in all the question where GRN occurs, AMB occurs in 93% of them, and SYN occurs in 91% of them. Thus, the labels SYN and GRN were commonly chosen together by the workers. While synonyms meant workers found words having similar meaning (e.g., 'round' vs. 'circular'), granularity meant workers found answers explaining the same thing in greater detail (e.g., 'mostly red' vs. 'red, black and blue'). On inspection we found that these labels commonly occur together because the when 10 answers do provide varying levels of detail, they often do so using synonyms (e.g. ‘money’, ‘currency’, ‘10 dollar bill’).

In the *VQA_2.0* dataset, questions with INV (invalid question) label has IVE (insufficient visual evidence, i.e. answer not present in image) occurring in 94% of cases. Since many questions are intentionally designed to outwit machines, the answers to such questions may not be immediately evident from the image, and so require some deductions. Hence, crowd workers may consider questions that were invalid (INV) as also having insufficient visual evidence (IVE).

5 PREDICTING WHY ANSWERS DIFFER

Having seen that answer disagreements can arise for different reasons, we next explore a novel problem of predicting why the answers will differ, given the QI pair and optionally the answers. We introduce two machine learning models, and describe the experiments to assess their accuracy in making predictions.

Machine Learning Setup

Problem Definition and Evaluation: We pose the task of predicting answer-disagreement-source(s) as a multi-label binary classification problem. The input is a QI pair and optionally the crowdsourced answers. The output is a binary value for each of the 10 disagreement-source labels, indicating whether that label is the reason for answer-disagreement of the VQ. In other words, we consider each disagreement label as a distinct binary classification problem. We evaluate each classifier using a precision-recall curve, and the average precision score.

Ground Truth: We compute binary ground truth labels for all the 10 disagreement-source labels for each VQ. Specifically, we examined the five crowdsourced labels per VQ, and considered a disagreement-source label as ‘1’ (i.e present), if at least two people selected that label. The 2-person threshold is a reasonable choice when five answers are crowdsourced, as indicated by [12, 47].

Train/Validation/Test Split: We used the whole *VizWiz* dataset, including the QI pairs where all answers were identical (i.e., 3% of the total VQs) to grow the size of the training set. Employing the train/validation/test split from [33], we have 20,000 training (65%), 3,173 validation (10%), and 7,988 test (25%) samples. For the 5,031 VQs from the *VQA_2.0* dataset, we introduced a 65/10/25 split which resulted in 3,230 training, 513 validation, and 1,291 test examples.

Baseline: To the best of our knowledge, no prior work has tried to predict the reason(s) why a VQ will have different answers. So the best option available today is to randomly guess the reasons. Thus, we compare our system against a *Status Quo* predictor, which randomly assigns a binary value for each of the ten disagreement-labels, to simulate random guessing.

Machine Learning Models

Random Forest: We proposed a random forest [13] model. We chose to extract features that describe the image, question, and 10 answers.

As *image features*, we use the Computer Vision API[1] from Microsoft Cognitive Services to extract: (a) number of category labels assigned to the image (‘outdoor’, ‘abstract’, ‘food’, etc), (b) number of tags assigned to the image (‘pizza’, ‘sign’, ‘water’, ‘television’, etc.), (c) number of distinct colours detected in the image, and (d) number of faces detected in the image. Intuitively, the number of categories and tags associated with an image informs two things: (1) the number of different ways an image can be interpreted (e.g. ‘sitting at a table’ versus ‘eating’), and (2) the number of objects in an image competing for an person’s attention. When an image is assigned multiple tags like ‘sitting at a table’ and ‘eating’, then a question of the form ‘What is the person doing?’ will be considered ambiguous (AMB), as the answer could either be ‘sitting’ or ‘eating’ or both. Also, if an image is associated with a number of categories and tags, it indicates there are multiple salient objects in the image competing for the viewer’s attention. Therefore, answers to ‘What is this?’ questions will result in a variety of answers, depending on which object attracts the viewer’s attention. This will give rise to synonyms (SYN) and varying granularity (GRN) in the answers.

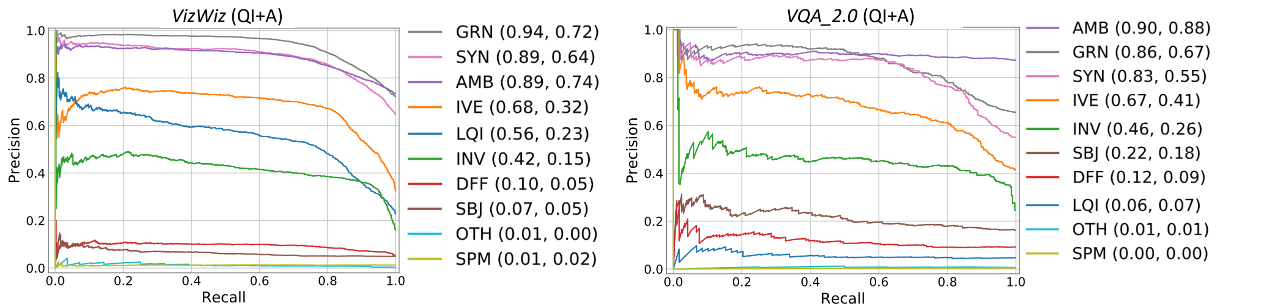
For *question features*, we considered the following: (a) number of words in the question, as from Section 4 we saw that lengthy questions tended to be ambiguous, (especially for *VQA_2.0*), (b) whether the word ‘colo(u)r’ is present in the question, as a binary label, and (c) the most common answer-type from the 10 crowdsourced answers [32, 33], namely numeric, yes/no, other, unanswerable. Intuitively, the most common answer-type can indicate whether answer disagreement can occur. For instance, a generic ‘other’ question has more chance to produce a wide variety of answers, than a ‘yes/no’ question.

For *answer features*, we counted the number of words in each of the ten answers, as difference in the number of words indicates a difference in the answer text.

We used the random forest implementation of Scikit-Learn [55], with 1,000 trees, ‘balanced’ class-weights (so that all output labels get equal priority, despite class imbalance in training data), and maximum tree-depth of 20.

Deep Learning: Given the many successes of deep learning systems, we also developed a deep learning model for

[1] https://azure.microsoft.com/services/cognitive-services/computer-vision/



(a) **(b)**

Figure 7: Performance curves of our random forest model, for (a) *VizWiz* and (b) *VQA_2.0* datasets. The legends show average precision scores of our model (left), and *Status Quo* baseline (right), for each label.

our prediction problem. We adapted our architecture from the hybrid neural network proposed in [5]. It takes as input the raw image, question, and (optionally) the 10 answers. The text inputs are converted to numeric form using GloVE (Global Vectors for Word Representation) pre-trained 100-dimensional word embedding [56], which was trained on the entire text corpus of Wikipedia 2014. Then they are passed through a 256-dimensional Long Short Term Memory (LSTM) [30] model. The image is encoded using the popular VGG16 [62] pre-trained vision model, which takes a 224×224 colour image as input, and outputs a 4096-dimensional vector. The image and text encodings are then combined and passed through two fully connected layers with ReLu (Rectified Linear Unit) activation functions, and are finally output via a 10-node sigmoid activated output layer, corresponding to 10 probabilities for each disagreement-source label.

Performance Analysis

Overall Performance: We first examine the performance of the models to predict the disagreement-source directly from the VQ and answers. Figure 7 shows the precision-recall curves for the proposed random forest model as well as the average precision scores for it and the *Status Quo* approach. As observed, the random forest model outperforms *Status Quo* for most disagreement-causes.

For the *VizWiz* dataset, Ambiguity (AMB), Synonyms (SYN) and Granularity (GRN) are predicted with the highest average precision (Figure 7a). The model success appears to correlate with most frequent disagreement-sources in the dataset, probably because there are more training examples.

The model performs worst for detecting spam (SPM) for the *VizWiz* dataset. Intuitively, this makes sense since a worker's choice to return bogus results is probably independent of the task at hand. Hence, we would not expect that information about the QI pair or the answers will help in deciding when a worker will submit spam.

While the model does demonstrate some predictive power for detecting difficulty (DFF) and subjectivity (SBJ), it is a less strong predictor compared to AMB, SYN and GRN. This is may be because detecting whether a question is difficult or subjective requires understanding the meaning of the question, which is possible through more sophisticated natural language processing and semantic analysis, rather than simple statistical features such as word count or presence of certain words like 'colour'. Another reason for non-performance is possibly the lack of sufficient training examples (e.g. less than 10% in *VizWiz*).

For the *VQA_2.0* dataset, performance is similar to *VizWiz* with respect to labels AMB, SYN, GRN, IVE, DFF, SPM and OTH. Significant differences are observed for LQI and SBJ labels. We hypothesize the diffence in LQI performance is due to the nature of the images themselves. While images in *VizWiz* are typically more blurred or ill-formed when they are LQI and so easier to detect, images from MS-COCO [46] typically share with other images in the dataset that they are high photographic quality even when they are LQI.

For completeness, we include the precision-recall curves for the deep learning model in the Supplementary Materials. However, we exclude it from the main paper as the model was unable to perform better than *Status Quo*. We attribute the poor performance to the limited amount of training data and huge class imbalance. Specifically, the relatively low number of training samples (20,000 as opposed to millions for large-scale systems) makes it difficult to effectively learn the weights necessary for this end-to-end machine-learning task. This issue is compounded by the huge class imbalance of AMB, SBJ, and GRN, where a standard 'yes predictor' for those three labels would already yield promising performance; i.e. predict AMB, SYN and GRN for all samples.

	Disagreement Source Labels																			
	LQI		IVE		INV		DFF		AMB		SBJ		SYN		GRN		SPM		OTH	
	VizWiz	*VQA*	*VizWiz*	*VQA*	*VizWiz*	*VQA*	*VizWiz*	*VQA*	*VizWiz*	*VQA*	*VizWiz*	*VQA*	*VizWiz*	*VQA*	*VizWiz*	*VQA*	*VizWiz*	*VQA*	*VizWiz*	*VQA*
QI+A	0.56	0.06	0.68	0.67	0.42	0.46	0.10	0.12	0.89	0.90	0.07	0.22	0.89	0.83	0.94	0.86	*0.01**	*0.00**	0.01	0.01
QI	0.43	0.06	0.61	0.65	0.34	0.44	0.08	0.11	0.82	0.90	0.06	0.21	0.81	0.80	0.87	0.84	*0.01**	*0.00**	0.01	0.01
Q	0.46	0.07	0.62	0.66	0.34	0.49	0.09	0.11	0.78	0.90	0.06	*0.17**	0.80	0.81	0.85	0.84	0.02	0.01	0.01	0.01
I	0.24	*0.05**	0.32	0.42	0.16	*0.23**	0.05	0.11	0.74	0.88	0.05	*0.16**	0.66	0.56	0.74	*0.66**	*0.01**	0.01	0.01	0.01
Status Quo	0.23	0.07	0.32	0.41	0.15	0.26	0.05	0.09	0.74	0.88	0.05	0.18	0.64	0.55	0.72	0.67	0.02	0.01	0.00	0.00

(a) ***VQA*** **=** ***VQA_2.0*** **dataset**

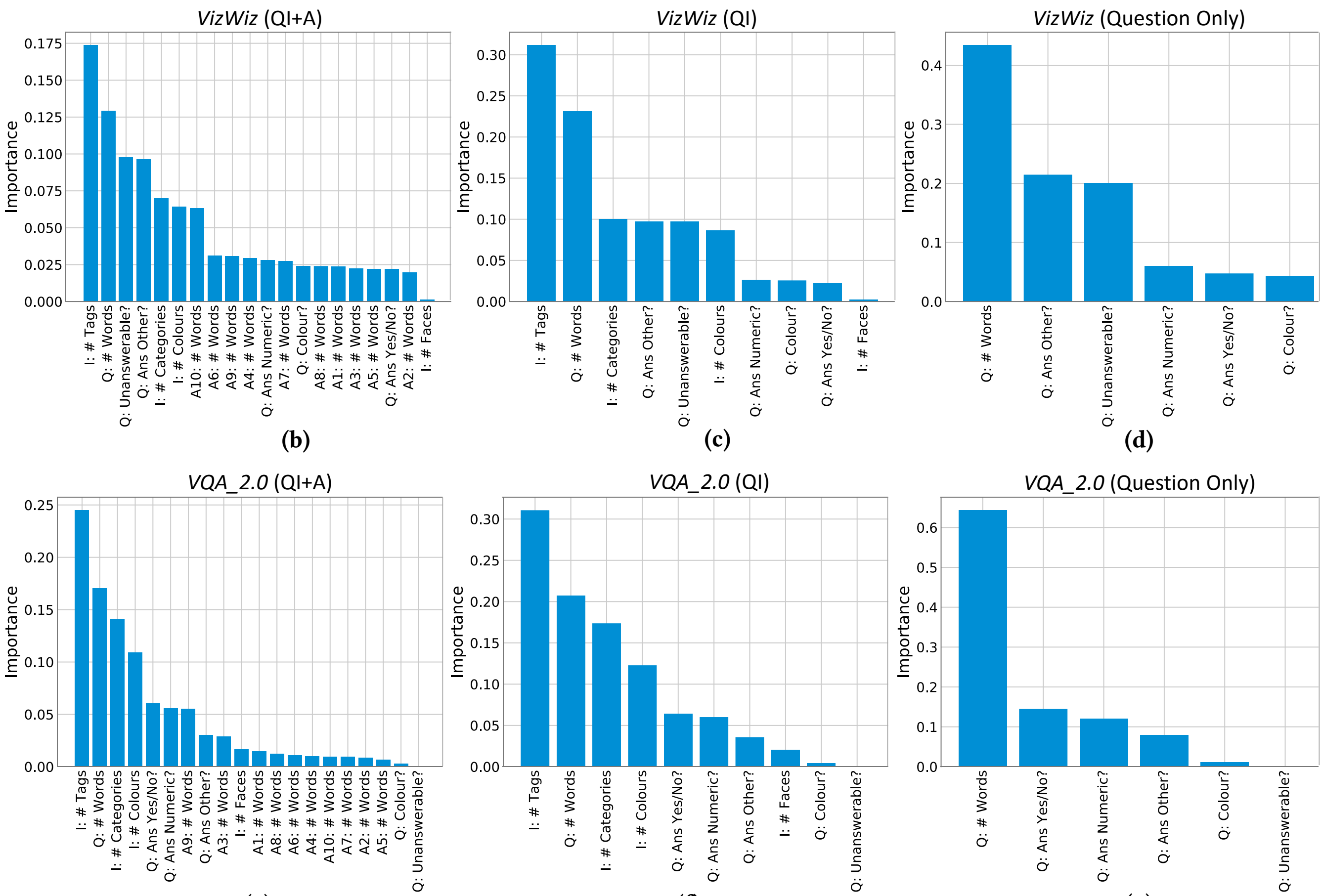


Figure 8: (a): Average precision scores of our random forest model, against the *Status Quo* (random) baseline, for *VizWiz* and *VQA_2.0* datasets. The figures are for four ablations: training and testing on question, image and answer features ('QI+A'), question and image features only ('QI'), question features alone ('Q'), and image features alone ('I'). Italicized values with asterisk (*) indicate instances where our model performed worse than *Status Quo*. (b) - (g): Importance of our handcrafted features for predicting disagreement-sources, as returned by our random forest model, trained on 'QI+A', 'QI' and 'Q', for the *VizWiz* (b)-(d) and *VQA_2.0* (e)-(g) datasets. Question, Image and Answer feature-names start with Q:, I:, and A#: respectively.

Predictive Cues: We also conducted ablation studies to investigate what cues are most predictive of the disagreement-source, using the top-performing random forest model. Specifically, the four ablations we trained on are (a) question, image and answer features ('QI+A'), (b) question and image features only ('QI'), (c) question features alone ('Q'), and (d) answer features alone ('A'). We also report the importance of individual features in Figures 8(b)-(c). Specifically, it reveals the learned importance of each feature for three ablations of the model, across both datasets. The importance values are obtained from the Scikit-Learn implementation of the random forest classifier. Specifically, we used the "gini impurity" criteria [14]

Figure 8(a) shows the average precision results for these four ablations of our random forest model, and the *Status Quo* baseline across both *VizWiz* and *VQA_2.0* datasets. Overall, reducing the number of input features causes a performance deterioration. Still, except for 'I', all variations perform better than *Status Quo* for all the labels across both datasets.

As noted above, for the *VizWiz* dataset, Ambiguity (AMB), Synonyms (SYN) and Granularity (GRN) are predicted with the highest precision. This can be partially due to the high frequency of these labels in the dataset. However, except for 'I', all other variants perform significantly better than *Status Quo* baseline. It is interesting that even without the answer features, our model is able to predict answer related issues like SYN and GRN from 'QI' and 'Q' features only. This can be attributed to the features: number of image tags, number of image categories, and number of words in the question, as seen from Figures 8(c) - (d). As we hypothesized, count of image tags and categories inform about the number of different salient objects in the image, and more objects lead to answers with synonyms and varying granularity.

Low Quality Image (LQI) and Insufficient Visual Evidence (IVE) are predicted fairly well by 'QI+A', 'QI' and 'Q'. We believe that the combination of the features: number of image tags, and count of words in the question, play a significant role here. A question with low word count (e.g. 'What is this?') is probably not invalid by itself. However, an image with low tag count is possibly blurred, or does not contain enough identifiable entities, resulting in LQI. A combination of low tag count in image, and high word count in question suggests that IVE is about to occur. Since the 'Q' ablation performs better for these labels than 'I', we believe that some other question features (like a combination of the four answer-type binary variables) may also play a role. We hypothesize that the performance for LQI and IVE drops significantly for 'I', since these additional question based features are not available.

We hypothesized in Section 4 that a VQ seeking colour related information leads to particular disagreement-sources. So we included the two colour related features: count of distinct salient and accent colours present in the image, and whether the word 'colo(u)r' is present in the question. Interestingly for both datasets, presence of 'colo(u)r' is not as important as the number of colours present in the image (Figure 8b,c,e,f). This indicates that answer disagreement occurs with higher probability if many different colours are visible in the image. For answering such VQs, people will use different names for the visible colours, or will probably list the colours in varying order, even if the question does not explicitly mention the word 'colo(u)r'.

6 EXISTING SOLUTIONS

Various solutions exist to resolve crowd disagreements that arise due to different reasons. Yet, currently a system designer has no way of knowing which solution(s) to apply, without first reviewing the VQ with answers, and then identifying the reason(s) for which the disagreement occurred. Using our proposed taxonomy, a trained VQA system will be able to detect which specific reason(s) will cause disagreement(s) to occur, and thereby recruit appropriate disagreement resolution solution(s). We discuss this mapping between our taxonomy of disagreement causes, and some of the existing solutions below.

Low Quality Images (LQI) occur due to poor resolution, camera framing error, or lack of focus. Solutions like blur detection and correction [48, 61], image-sharpening [52, 54, 58], and tools supporting blind photography [39] (esp. for *VizWiz*) can be applied in this scenario. For invalid questions (INV), solutions include question-text processing [64], followed by automated techniques for grammatical error detection [21] and correction [19]. Difficult visual questions (DFF) can be tackled by combining methods for assessing difficulty of textual questions [8] and difficulty of image annotation tasks [71]. Ambiguity (AMB) can be handled using solutions proposed for measuring image specificity (i.e. whether an image elicits a converging textual description from the crowd) [38], and for determining the different shades of meaning present in textual product label attributes [42]. Subjectivity (SBJ) can be modelled and resolved by techniques proposed by [53, 68]. Synonymous answers (SYN) due to using different words having same meaning, or due to spelling errors, can be detected and corrected using methods described by [9, 18, 26]. Lastly, spam answers and malicious behaviour of crowdworkers (SPM) are discussed at length by [27, 66], while solutions for spam prevention and resolution are proposed by [25, 65].

7 CONCLUSION

We proposed a taxonomy of nine reasons why answers to VQs vary and a novel machine learning problem of automatically predicting directly from a VQ (plus optionally answers) why answers will differ. We crowdsourced "disagreement-source" labels for VQs asked by blind and sighted people and found ambiguity in the question, synonyms in the answers, and varying granularity in the answers are the primary reasons answers differ. Our experiments with two machine learning models demonstrate it is possible to predict why answers will differ. We will publicly share our new dataset and all code to facilitate future extensions of this work.

ACKNOWLEDGEMENTS

We gratefully acknowledge funding from Anonymous. We also thank the crowd workers for their valuable time and effort to provide the annotations.

REFERENCES

[1] 2015. Be My Eyes. https://www.bemyeyes.com/. (2015). [Online; accessed 21-September-2018].

[2] Ehsan Amid and Antti Ukkonen. 2015. Multiview Triplet Embedding: Learning Attributes in Multiple Maps. In *International Conference on Machine Learning*. 1472–1480.

[3] Hossein Amirkhani and Mohammad Rahmati. 2014. Agreement/Disagreement Based Crowd Labeling. *Applied intelligence* 41, 1 (2014), 212–222.

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural Module Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 39–48.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[6] Lora Aroyo and Chris Welty. 2013. Crowd Truth: Harnessing Disagreement in Crowdsourcing a Relation Extraction Gold Standard. *WebSci2013. ACM* 2013 (2013).

[7] Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Journal of Human Computation* 1 (2014), 31–34.

[8] Yoram Bachrach, Thore Graepel, Tom Minka, and John Guiver. 2012. How to grade a test without knowing the answers—a Bayesian graphical model for adaptive crowdsourcing and aptitude testing. *arXiv preprint arXiv:1206.6386* (2012).

[9] Steven D Baker and John O Lamping. 2011. Identifying a synonym with n-gram agreement for a query phrase. (April 12 2011). US Patent 7,925,498.

[10] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and others. 2010. VizWiz: Nearly Real-Time Answers to Visual Questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*. ACM, 333–342.

[11] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual Challenges in the Everyday Lives of Blind People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2117–2126. https://doi.org/10.1145/2470654.2481291

[12] Jonathan Bragg, Daniel S Weld, et al. 2013. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI conference on human computation and crowdsourcing*.

[13] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[14] Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. 1984. *Classification and Regression Trees (Wadsworth Statistics/Probability)*. Chapman and Hall/CRC.

[15] Michele A Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P Bigham, and Amy Hurst. 2012. Crowdsourcing subjective fashion advice using VizWiz: challenges and opportunities. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 135–142.

[16] Fabio Calefato, Filippo Lanubile, and Nicole Novielli. 2018. How to ask for technical help? Evidence-based guidelines for writing questions on Stack Overflow. *Information and Software Technology* 94 (2018), 186–207.

[17] Patricia W Cheng. 1997. From covariation to causation: a causal power theory. *Psychological review* 104, 2 (1997), 367.

[18] Maarten Clements, Arjen P De Vries, and Marcel JT Reinders. 2008. Detecting synonyms in social tagging systems to improve content retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 739–740.

[19] Daniel Dahlmeier and Hwee Tou Ng. 2011. Grammatical error correction with alternating structure optimization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 915–923.

[20] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2.

[21] Rachele De Felice and Stephen G Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 169–176.

[22] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2017. Crowdsourcing Ground Truth for Medical Relation Extraction. *arXiv preprint arXiv:1701.02185* (2017).

[23] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. (2018). https://arxiv.org/abs/1808.06080

[24] ANCA DUMITRACHE, OANA INEL, BENJAMIN TIMMERMANS, and LORA AROYO. 2017. Crowdsourcing Ambiguity-Aware Ground Truth. *Collective Intelligence* (2017).

[25] Carsten Eickhoff and Arjen P de Vries. 2013. Increasing Cheat Robustness of Crowdsourcing Tasks. *Information retrieval* 16, 2 (2013), 121–137.

[26] Yoko Futagi, Paul Deane, Martin Chodorow, and Joel Tetreault. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning* 21, 4 (2008), 353–367.

[27] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1631–1640.

[28] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity Is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, 5–14.

[29] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question. In *Advances in Neural Information Processing Systems*. 2296–2304.

[30] Felix A. Gers, JÃijrgen Schmidhuber, and Fred Cummins. 2000. Learning to Forget: Continual Prediction with LSTM. *Neural Computation* 12, 10 (2000), 2451–2471. https://doi.org/10.1162/089976600300015015

[31] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*, Vol. 1. 9.

[32] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3511–3522.

[33] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions From Blind People. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[34] Oana Inel and Lora Aroyo. 2017. Harnessing Diversity in Crowds and Machines for Better Ner Performance. In *European Semantic Web Conference*. Springer, 289–304.

[35] Oana Inel, Lora Aroyo, Chris Welty, and Robert-Jan Sips. 2013. Domain-Independent Quality Measures for Crowd Truth Disagreement. *Detection, Representation, and Exploitation of Events in the Semantic Web* (2013), 2.

[36] Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. Crowdtruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data. In *International Semantic Web Conference*. Springer, 486–504.

[37] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 1359–1367.

[38] Mainak Jas and Devi Parikh. 2015. Image Specificity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2727–2736.

[39] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P Bigham. 2011. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility.* ACM, 203–210.

[40] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference On.* IEEE, 1988–1997.

[41] Kushal Kafle and Christopher Kanan. 2017. An Analysis of Visual Question Answering Algorithms. In *2017 IEEE International Conference on Computer Vision (ICCV).* IEEE, 1983–1991.

[42] Adriana Kovashka and Kristen Grauman. 2015. Discovering Attribute Shades of Meaning with the Crowd. *International Journal of Computer Vision* 114, 1 (2015), 56–73.

[43] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and others. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[44] Walter S Lasecki, Phyo Thiha, Yu Zhong, Erin Brady, and Jeffrey P Bigham. 2013. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility.* ACM, 18.

[45] Y. Li, C. Huang, X. Tang, and C. C. Loy. 2017. Learning to Disambiguate by Asking Discriminative Questions. In *2017 IEEE International Conference on Computer Vision (ICCV).* 3439–3448. ISSN:.

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision.* Springer, 740–755.

[47] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2009. Turkit: tools for iterative tasks on mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation.* ACM, 29–30.

[48] Renting Liu, Zhaorong Li, and Jiaya Jia. 2008. Image partial blur detection and classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE, 1–8.

[49] Christian C Luhmann and Woo-kyoung Ahn. 2005. The meaning and computation of causal power: Comment on Cheng (1997) and Novick and Cheng (2004). (2005).

[50] Mateusz Malinowski and Mario Fritz. 2014. A Multi-World Approach to Question Answering about Real-World Scenes Based on Uncertain Input. In *Advances in Neural Information Processing Systems.* 1682–1690.

[51] Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2017. Crowdsourcing question-answer meaning representations. *arXiv preprint arXiv:1711.05885* (2017).

[52] Richard A Muller and Andrew Buffington. 1974. Real-time correction of atmospherically degraded telescope images through image sharpening. *JOSA* 64, 9 (1974), 1200–1210.

[53] An Thanh Nguyen, Matthew Halpern, Byron C Wallace, and Matthew Lease. 2016. Probabilistic Modeling for Crowdsourcing Partially-Subjective Ratings. In *Fourth AAAI Conference on Human Computation and Crowdsourcing.*

[54] Kenneth A Parulski and Michael S Axman. 1997. Automatic image sharpening in an electronic imaging system. (Dec. 9 1997). US Patent 5,696,850.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[56] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).* 1532–1543.

[57] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring Models and Data for Image Question Answering. In *Advances in Neural Information Processing Systems.* 2953–2961.

[58] John GM Schavemaker, Marcel JT Reinders, Jan J Gerbrands, and Eric Backer. 2000. Image sharpening by morphological filtering. *Pattern Recognition* 33, 6 (2000), 997–1012.

[59] V. Sharmanska, D. Hernández-Lobato, J. M. Hernández-Lobato, and N. Quadrianto. 2016. Ambiguity Helps: Classification with Disagreements in Crowdsourced Annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2194–2202. ISSN:.

[60] Aashish Sheshadri and Matthew Lease. 2013. Square: A Benchmark for Research on Computing Crowd Consensus. In *First AAAI Conference on Human Computation and Crowdsourcing.*

[61] Till Sieberth, Rene Wackrow, and Jim H Chandler. 2016. Automatic detection of blurred images in UAV image sets. *ISPRS Journal of Photogrammetry and Remote Sensing* 122 (2016), 1–16.

[62] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[63] Guillermo Soberón, Lora Aroyo, Chris Welty, Oana Inel, Hui Lin, and Manfred Overmeen. 2013. Measuring Crowd Truth: Disagreement Metrics Combined with Worker Behavior Filters. In *CrowdSem 2013 Workshop.*

[64] Tomek Strzalkowski and Sanda Harabagiu. 2006. *Advances in open domain question answering.* Vol. 32. Springer Science & Business Media.

[65] Jeroen Vuurens, Arjen P de Vries, and Carsten Eickhoff. 2011. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11).* 21–26.

[66] Jeroen BP Vuurens and Arjen P De Vries. 2012. Obtaining high-quality relevance judgments using crowdsourcing. *IEEE Internet Computing* 16, 5 (2012), 20–27.

[67] M. Wan and J. McAuley. 2016. Modeling Ambiguity, Subjectivity, and Diverging Viewpoints in Opinion Question Answering Systems.

In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 489–498. ISSN:.

[68] Mengting Wan and Julian McAuley. 2016. Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 489–498.

[69] P. Wang, Q. Wu, C. Shen, A. Dick, and A. v. d. Hengel. 2018. FVQA: Fact-Based Visual Question Answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP, 99 (2018), 1–1.

[70] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit Knowledge-Based Reasoning for Visual Question Answering. *arXiv preprint arXiv:1511.02570* (2015).

[71] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. 2010. The Multidimensional Wisdom of Crowds. In *Advances in Neural Information Processing Systems*. 2424–2432.

[72] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. 2015. Visual Madlibs: Fill in the Blank Description Generation and Question Answering. In *Computer Vision (ICCV), 2015 IEEE International Conference On*. IEEE, 2461–2469.

[73] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded Question Answering in Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4995–5004.