

Let's Agree to Disagree: A Meta-Analysis of disagreement among crowdworkers during Visual Question Answering

Samridhi Ojha and Anuparna Banerjee

Abstract

The realm of **Visual question answering** has evolved itself by utilizing the principles of *Artificial Intelligence* and that has enabled people to receive natural free-form answers to any form of question asked about an image. Answering a question on the basis of an image could be difficult since it requires understanding not only the relevant details from the image, but also the various constructs of the associated natural language question. Previous studies show that subjective, opinion or difficult visual questions may have multiple responses. Our work is focused on identifying the reasons behind this disagreement of crowd workers in arriving at a unanimous answer to a visual question. In this paper, we propose two different user-interfaces to collect the responses from crowdworkers. We analysed the responses gathered from crowd-workers to suggest the reasons for disagreement among answers to visual questions across three different datasets which could be further utilized to develop intelligent computer vision algorithms.

Introduction

The domain of **Visual Question Answering (VQA)** lies in the intersection of Natural Language Processing (NLP) and Computer Vision and provides natural responses to any visual question (Antol et al. 2015). Blind users can use this technique to learn more about the environment around them and improve upon their involvement in the community (Lasecki et al. 2013; Antol et al. 2015). Visual Questions can have multiple answers depending upon the interpretation of the question asked and the relevant aspects of an image. Most Visual Question Answering (VQA) systems developed to-date cannot predict which images and their associated questions could have multiple answers. A model has been proposed in (Gurari and Grauman 2017) which identifies questions with multiple responses by leveraging the power of *crowd*. These results could be further enhanced if we know the reasons behind different answers for the same visual question. The reasons could be attributed to different subjective interpretations of the question, varied background knowledge of the task, different visual perception of the images etc.

Past research works have discussed about the disagreement of responses gathered from crowdworkers for visual questions. (Gurari and Grauman 2017) describes some plausible reasons behind these disagreements like ambiguity in questions and images, granularity, lack of required worker skills, etc. Our work *accounts for the different reasons of disagreement among crowdworkers in answering a visual question*. Instead of developing our own theories explaining these disagreements, we plan to involve the crowd to help us understand the discrepancies in the crowd's answer choices. To our knowledge, no previous study has been done in this area.

Visual Question Answering systems could be improved if we have a gold-standard dataset of images and their associated question-answer pairs. However, due to worker disagreements to arrive at a unanimous answer to a question, formulation of this kind of gold-standard dataset is hard to achieve (Kairam and Heer 2016). Our aim is to categorize visual questions on the basis of the different reasons of this disagreement which could be helpful in future research to create such a standardized dataset for visual question-answers.

Previous research studies have indicated that user-interface design choices have impacted the quality of results obtained from a crowdsourcing system (Rahmanian and Davis 2014). Using this idea, we plan to develop two different user interfaces to capture the crowdsourced reasoning behind different answers to visual questions.

Related Work

Our work could be split into three major areas: Crowd Disagreement, Visual Question Categorization and User Interface Design.

Crowd Disagreement

Previous research studies in the domain of Visual Question Answering have expressed concerns on the difficulties that arise due to visual questions having multiple valid answers. In (Kafle and Kanan 2016), the authors describe using evaluation metrics like Wu-Palmer Similarity (WUPS) index to calculate the similarity between two valid answers. However, this metric is useful for natural language expressions and does not account for questions arising from the

images. The authors also discuss about a strategy to address the problem by employing multiple crowd workers to identify questions leading to different answers and then, placing a lower emphasis on these questions in order to utilize the existing evaluation tools. This presents a potential problem in the quality of predictions of VQA systems.

The work done in (Inel et al. 2014) is another attempt to understand the differences in the crowd opinion. They perform their experiments on medical text and newspaper extracts.

Another approach to this problem is to collect responses from the experts in the domain and then compare them with the crowd answers to determine the disagreement among them. This was done in (Aroyo and Welty 2013). Our experiment works on datasets from blind people who do not require an expert opinion on their visual questions. These visual questions are based on their day-to-day experiences which a crowd can answer in general. Unavailability of answers to some of their questions is an area VQA systems could improve. One of the reasons of unavailability of answers could be unavailability of a unanimous answer. We explore this latent area of disagreement in answers to a visual question.

Our work delves into understanding the rationale behind multiple answers to visual questions. We plan to categorize such visual questions into groups with similar reasoning. This could help future works to develop efficient evaluation metrics for each of these categories. This could help create accurate VQA systems, even for questions with multiple answers.

Visual Question Categorization

Researchers have proposed various machine learning approaches that classify open-ended questions to fine-grained domains (Li and Roth 2006), (Mishra, Mishra, and Sharma 2012), (Suzuki et al. 2003). These classification systems allow a user to get a valid answer to the given question, however, it does not tell if the question will have a single or multiple responses. In computer vision, crowd disagreement may arise due to multiple reasons. (Gurari and Grauman 2017) proposed a system to identify if the question will have one or multiple responses, but it mostly focuses on the binary categorization of question (Gurari and Grauman 2017). Although previous studies have shown that a question may have multiple responses, they did not discuss the cause for the disagreement in the responses. This study provides deeper understanding of response disagreement.

User Interface Design

User-interface impacts the performance of the crowdworkers (Khanna et al. 2010), (Rahmanian and Davis 2014), (Finnerty et al. 2013). (Khanna et al. 2010) noticed the improvement in the performance of crowdworkers when simple yet effective user-friendly interface was administered in a crowdsourcing system. Likewise, (Rahmanian and Davis 2014) used three different user-interfaces to collect the responses from the crowd. They noticed that crowdworkers preferred ‘rate user-interfaces’ as the cognitive load on the

users is low. These findings are consistent with the experiment conducted by (Finnerty et al. 2013) where ‘simple user interface and simple tasks’ generated better results. Our research is different from the other related works because instead of evaluating crowdworkers’ performance in two different user-interfaces, we want to validate responses collected from two user interfaces.

Methods

We follow the methodology discussed in Crowdsourcing Annotations for Visual Object Detection (Su, Deng, and Fei-Fei 2012). We explain our methods and designs in detail in this section.

User Interface Designs

In our research experiment, we designed and deployed two different user interfaces Figure 1 and 2 to ensure the accuracy and consistency of the responses collected from our crowd workers.

User interface impacts the accuracy of the results, errors are minimized if designs and crowdsourcing tasks are simple (Finnerty et al. 2013). Both our interfaces have simple designs and neatly organized annotation tasks. Workers select single or multiple options from the list of non-overlapping reasonings. In order to assist crowd worker with their selection, we added definitions in the collapsible panels next to each reasoning. In a single Human Intelligent Task (HIT), there are 4 images. Question about the image are displayed on top and the responses collected from crowds appears on the left side of the image. The possible reasons for disagreement among crowd workers are below the images. A worker completes all four tasks before submitting his/her responses. We decided to use four images for our interfaces because it allows us to gather more data at a reasonable rate without compromising the data quality. We also added a ‘Comment’ box where workers can give us their feedback and share their experiences.

Both interfaces display same 144 images, questions, and responses. The web interfaces for both crowdsourcing platforms are similar in all ways except for the option selection part. The first interface has a radio button that restricts workers to select only one best option for each VQA, whereas, the second interface has checkboxes where a worker can select multiple best options for each VQA. We use these two designs as we want to evaluate how the responses will differ when we restrict crowd to submit one versus multiple responses. Furthermore, checkboxes give us richer data which can be manipulated to extract useful information.

Reasoning Categories

We use disagreement reasons covered by (Gurari and Grauman 2017). Instead of using all eight reasons, we finalized following six non-overlapping reasonings as our reasoning categories.

- *Task Difficulty*: Questions that require domain expertise to answer will fall in this category. This category also contains questions that are difficult to answer such as counting lamp posts in a low-resolution image.





Visual Question Answers			
<p>Question: What color is the man's jacket?</p> <p>Answer Choices:</p> <ol style="list-style-type: none"> 1. beige 2. tan 3. beige 4. tan 5. khaki 6. tan 7. tan 8. tan 9. tan 10. tan <p> <input type="radio"/> Task Difficulty <input checked="" type="radio"/> Ambiguity <input type="radio"/> Granularity <input type="radio"/> Subjectivity <input type="radio"/> Insufficient Visual Evidence <input type="radio"/> Synonym </p>		<p>Question: How many people are wearing hats?</p> <p>Answer Choices:</p> <ol style="list-style-type: none"> 1. 4 2. 4 3. many 4. 4 5. 4 6. 4 7. 4 8. 4 9. 3 10. 4 <p> <input type="radio"/> Task Difficulty <input type="radio"/> Ambiguity <input type="radio"/> Granularity <input type="radio"/> Subjectivity <input type="radio"/> Insufficient Visual Evidence <input type="radio"/> Synonym </p>	
<p>Question: What color is the board?</p> <p>Answer Choices:</p> <ol style="list-style-type: none"> 1. red 2. red and white 3. red 4. red 5. red 6. red and white 7. red 8. red 9. red 10. red <p> <input type="radio"/> Task Difficulty <input type="radio"/> Ambiguity <input type="radio"/> Granularity <input type="radio"/> Subjectivity <input type="radio"/> Insufficient Visual Evidence <input type="radio"/> Synonym </p>		<p>Question: Are the bikes safe?</p> <p>Answer Choices:</p> <ol style="list-style-type: none"> 1. yes 2. yes 3. yes 4. no 5. yes 6. yes 7. yes 8. no 9. yes 10. yes <p> <input type="radio"/> Task Difficulty <input type="radio"/> Ambiguity <input type="radio"/> Granularity <input type="radio"/> Subjectivity <input type="radio"/> Insufficient Visual Evidence <input type="radio"/> Synonym </p>	

Figure 1: Crowdsourcing Interface with Radio Button Option



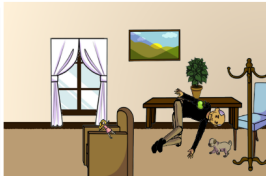

<p>Question: Who looks happier?</p> <p>Answer Choices:</p> <ol style="list-style-type: none"> 1. old person 2. man 3. man 4. man 5. old man 6. man 7. man 8. man 9. man 10. grandpa <p> <input type="checkbox"/> Task Difficulty <input type="checkbox"/> Ambiguity <input type="checkbox"/> Granularity <input type="checkbox"/> Subjectivity <input type="checkbox"/> Insufficient Visual Evidence <input type="checkbox"/> Synonym </p>		<p>Question: Is the mouse under the chair?</p> <p>Answer Choices:</p> <ol style="list-style-type: none"> 1. yes 2. yes 3. no 4. no 5. yes 6. no 7. yes 8. yes 9. yes 10. yes <p> <input type="checkbox"/> Task Difficulty <input type="checkbox"/> Ambiguity <input type="checkbox"/> Granularity <input type="checkbox"/> Subjectivity <input type="checkbox"/> Insufficient Visual Evidence <input type="checkbox"/> Synonym </p>	
<p>Question: What color is the dog?</p> <p>Answer Choices:</p> <ol style="list-style-type: none"> 1. gray 2. gray 3. brown 4. brown 5. gray 6. gray 7. gray 8. brown 9. brown 10. tan <p> <input type="checkbox"/> Task Difficulty <input type="checkbox"/> Ambiguity <input type="checkbox"/> Granularity <input type="checkbox"/> Subjectivity <input type="checkbox"/> Insufficient Visual Evidence <input type="checkbox"/> Synonym </p>		<p>Question: Is the man dancing with two turtles?</p> <p>Answer Choices:</p> <ol style="list-style-type: none"> 1. yes 2. yes 3. no 4. yes 5. yes 6. yes 7. yes 8. no 9. yes 10. yes <p> <input type="checkbox"/> Task Difficulty <input type="checkbox"/> Ambiguity <input type="checkbox"/> Granularity <input type="checkbox"/> Subjectivity <input type="checkbox"/> Insufficient Visual Evidence <input type="checkbox"/> Synonym </p>	

Figure 2: Crowdsourcing Interface with Checkboxes Options

- **Ambiguity:** Questions and images that lack clarity and may be interpreted differently by different people will fall in this category.
- **Granularity:** If responses are different due to varying granularity, then such questions fit in this category.
- **Subjectivity:** Opinion related questions and responses such as defining beauty, ugly, etc, fits in this category. Different crowd workers may have different opinions on subjective visual questions.
- **Insufficient Visual Evidence (IVE):** Often times, the object may not appear or is not visible in an image that is asked in the question. Such questions fall in this category.
- **Synonym:** This category contains questions where the crowd workers agree but use synonym to capture the same

connotation.

Project Work-flow

We recruited 4 crowd workers to work on a single HIT in each web interface. In total, for each image, we got responses from 8 different workers. In radio button interface, we got four responses for a HIT whereas, in checkboxes, we got four to sixteen responses for a single HIT. Once all HITs were completed, we aggregated our results and used various evaluation metrics to analyze them. We used 'Majority Voting' metric to evaluate and compare reasonings from two different user interfaces. We also used 'Accuracy' metric to identify all reasoning options majority of the voters agreed upon. In addition, we displayed reasonings that frequently appeared together in a Histogram.

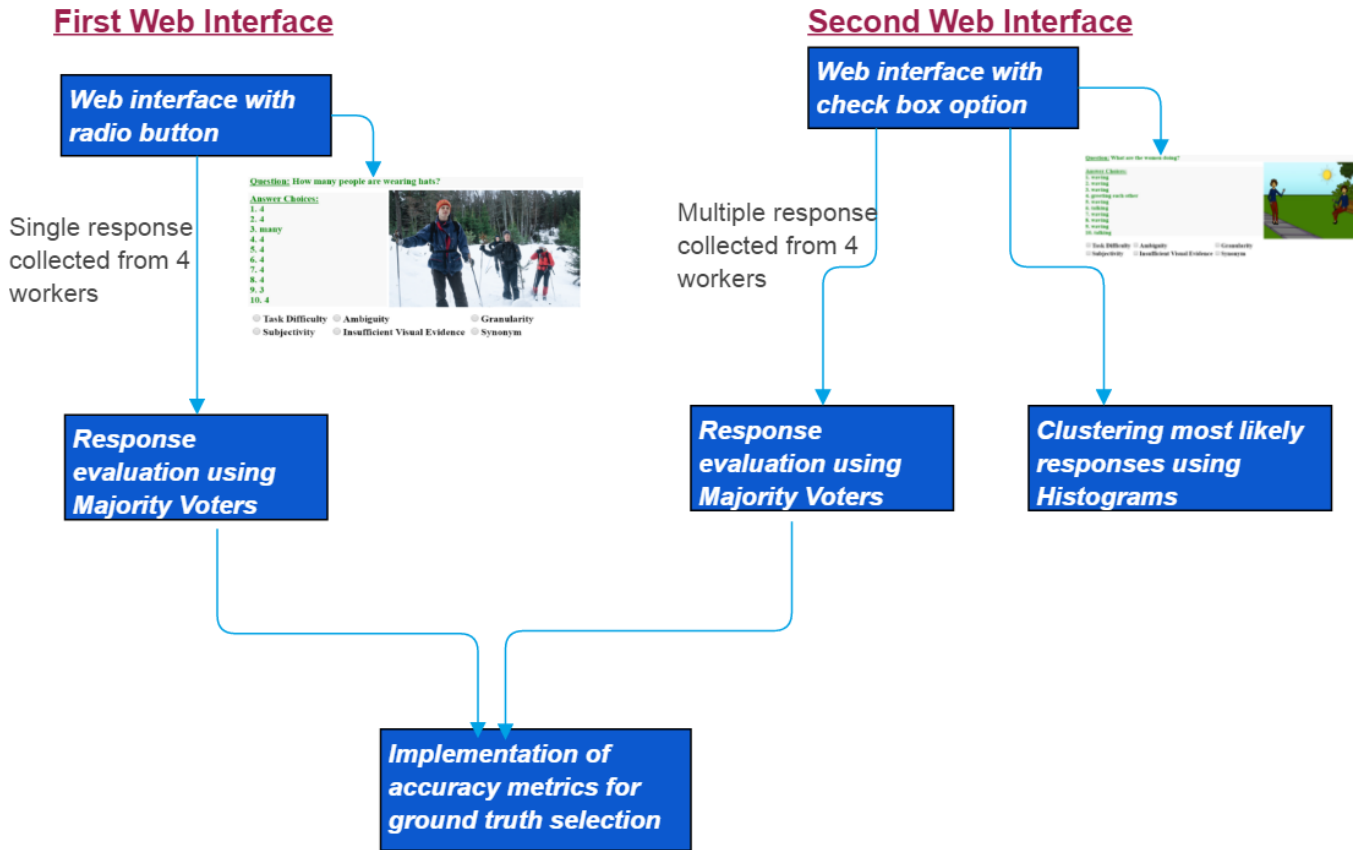


Figure 3: Workflow of our Crowdsourcing System. We have included interfaces and our evaluation metrics in the workflow

Experimental Design

Experiments

Our design identifies the reasons behind the disagreement among crowd workers in answering a visual question. We designed two different interfaces (crowd tasks) to capture the cause of this disagreement. Our interfaces use deterministic user interface elements like radio buttons and check-boxes to aid the normalization and aggregation of the data collected.

Our first interface asks the crowd workers to pick the most relevant cause of disagreement from the exclusive set of causes. Our hypothesis is based on our notion that disagreement among answers to visual questions could have multiple causes. Our task to choose the most relevant one coerces the crowd workers to weigh the reasons amongst one another and pick the best possible explanation.

Our second interface allows the crowd-workers to pick multiple reasons to describe the cause of disagreement among crowd-workers. This interface is based on the premise that crowd-workers would mark similar reasons of disagreement together thus forming a cluster of reasons that appear together.

There have been several experiments conducted in the field of Natural Language Processing (NLP) (Inel et al. 2014) to identify the disagreements in crowd responses. To

our knowledge, our work is first to consider the disagreements between crowd workers in the realm of VQA. Our two interfaces aim to capture heightened user experience which we assume would provide us with responses of good quality.

Datasets

We use the datasets worked upon by (Gurari and Grauman 2017). These datasets comprise of images and visual questions from the following sources:

- *VQA Dataset*: The VQA dataset (Antol et al. 2015) consists of images from the *real-world* and *abstract scenes*. This dataset is the most comprehensive collection of images available today. The most important characteristic of this dataset is its "free-form" and "open-ended" nature which we presume would lead to subjectivity in answer choices, a factor of disagreement we are interested to consider.
- *VizWiz Dataset*: This dataset (Bigham et al. 2010) is unique in its nature since it represents the concerns raised by the visually impaired. The images in this dataset are taken by blind users. The corresponding questions are the ones asked by the blind people. Our work would help the visually impaired since it would assist in understanding the differences in answers to questions raised by them and lead to development of better VQA systems.

Our experiment collects results from 144 images across the above-mentioned datasets. We use 48 images each from the three datasets - *real images from VQA*, *abstract scenes from VQA* and *images from VizWiz*.

We computed the randomness of the answers obtained from past VQA systems in the form of a measure called "entropy". The expression for entropy is given as below:

$$E = \sum_{i=1}^N -p_i \log p_i \quad (1)$$

where p_i represents the fraction of N answers that match the i -th most popular answer."

We find that higher values of entropy indicate greater disagreement among crowd-workers in answering a visual question. We classify our image datasets into 3 categories based on the generated entropy score - '0.5 to 1.0', '1.0 to 1.5' and 'more than 1.5'. Each group contributes 33% in the overall 48 images used for our experiment from a single dataset. Together, we have 144 images with a 33% representation of each of the entropy group across the datasets - VQA and VizWiz (Refer Figure 4). For each group, images are drawn at random to generate high quality results.

To our knowledge, our works reflects represents the best possible explanation to disagreement among crowd-workers from a small sample of 144 images, till date.

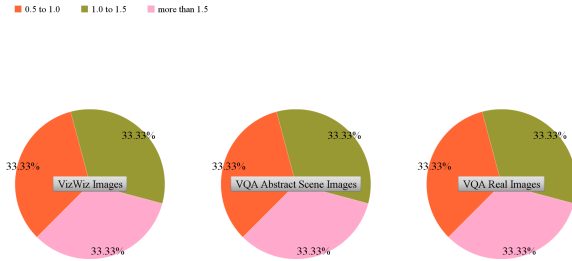


Figure 4: Percentage distribution of entropy groups across datasets

Crowd

We decided to use Amazon Mechanical Turk (AMT) platform to collect responses from crowd workers. AMT is one of the most popular crowdsourcing platforms where workers belonging to diverse demography participate (Ross et al. 2010). We do not have any restrictions on the qualifications for any crowd-worker to participate in our experiment. We recruited 4 unique workers for each HIT for both our user interfaces.

Evaluation Metrics

We evaluate our experiments using some disagreement-aware metrics described in (Soberón et al. 2013). The authors have utilized these metrics to compare worker performances in the domain of NLP for relation extraction. We use the similar metrics to evaluate the disagreements between workers:

1. **Majority Vote:** This is one of the most popular methods utilised to arrive at a conclusive decision for the task in hand. For every visual question, the crowd worker selects the most relevant reasoning behind the disagreement in the answers. In this metric evaluation, we prefer the reasoning R_i over R_j ($R_i \succ R_j$; $R_i, R_j \subseteq S_r$, the set of all possible reasons defined in page 2 under section "Methods") if majority of the crowdworkers say so.

One of the disadvantages of using this metric is that an equal weightage is placed to all crowdworkers and hence, we do not consider the possibility that some workers could have a better domain knowledge and expertise performing the task than others.

2. **Accuracy:**

We also use the metric defined by (Antol et al. 2015) to compute the accuracy of the results we obtain from the crowdsourcing platform. We believe that a reasoning should be given a greater preference if at least 3 crowd workers judged it as the best explanation of the disagreement. Unlike *Majority Vote*, here, we consider other causes of disagreement too, albeit with a reduced weightage.

$$\text{Accuracy} = \min\left(\frac{C}{3}, 1\right) \text{ where,}$$

C = The number of crowd workers who completed the task successfully.

Usability Metrics

We took the inspiration from (Mifsud 2015) to develop the following metrics. Since we have two interfaces, we can compare the performance of the crowd and the crowd task effectively using the following usability metrics:

1. **Time-based task efficiency:** We can calculate the time-based task efficiency of the two interfaces to determine which interface (crowd-task) is more efficient in collecting answers.

$$\text{Efficiency} = \frac{\sum_{j=1}^C \sum_{i=1}^N \frac{n_{ij}}{t_{ij}}}{NC}$$

where N = The total number of crowd tasks performed on the given interface,

C = The number of crowd-workers,

n_{ij} = The result of crowd-task, i by crowd-worker j ; if the crowd-worker successfully completes the task, then $n_{ij} = 1$, if not, then $n_{ij} = 0$.

t_{ij} = The time spent by crowd-worker j to complete task i .

Experimental Results

We present the results obtained from our experiments here.

	0.5 to 1.0	1.0 to 1.5	more than 1.5
VQA Real Images	Subjectivity	Subjectivity	Subjectivity
VQA Abstract Images	Subjectivity	Subjectivity	Subjectivity
VizBiz Images	IVE	Synonym; IVE	Synonym

Table 1: Majority Vote Results on Radio Button Interface

Radio Button Interface

This interface was implemented using radio buttons. The crowd-workers marked the most relevant cause of disagreement among answers to visual questions. We computed the metric *Majority Vote* across all datasets within every entropy group. The results are shown in Table 1.

Checkbox Interface

We used checkboxes in this interface to record all the possible reasons for disagreement. We computed the reasons for disagreement using *Majority Vote*. The results are displayed in Table 2. We also computed the number of reasonings selected by crowd workers in each dataset. We gathered a total of 192 responses (48 images * 4 crowd-workers) and the split of these responses has been provided in Table 3. We also constructed *Histograms* (Refer Figure 5) to display reasonings that often appeared together across all three datasets.

	0.5 to 1.0	1.0 to 1.5	more than 1.5
VQA Real Images	Subjectivity	Subjectivity	Subjectivity
VQA Abstract Images	Subjectivity	Subjectivity	Subjectivity
VizBiz Images	IVE	IVE	Synonym

Table 2: Majority Vote Results on Checkbox Interface

	1 response	2 responses	3 or more
VQA Real Images	151	35	6
VQA Abstract Images	145	42	5
VizBiz Images	156	31	5

Table 3: Responses count for three datasets on Checkbox Interface

Crowd Responses from the two Interfaces

We used *Accuracy* metrics to evaluate responses agreed by most crowd workers collected from both the interfaces. The distribution of responses for both interfaces are displayed in Table 4. Distribution of number of common responses collected from two interfaces are different. From table 2, 3, and

	0.5 to 1.0	1.0 to 1.5	more than 1.5
VQA Real Images	Subjectivity	Subjectivity	Subjectivity
VQA Abstract Images	Subjectivity	Subjectivity	Subjectivity
VizBiz Images	IVE	IVE	Synonym

Table 4: Reasonings for disagreement using Accuracy Metrics

4, we can see that findings are consistent across different platforms and metrics.

Analysis of Usability Metrics

We used Time-based task Efficiency metrics to calculate the performance of two Interfaces. Results for two interfaces are displayed in Table 5. We calculated the average time taken to complete the tasks by workers in two different interfaces. Workers took longer time to complete the tasks on Radio Button Interface for all the three datasets as displayed in Figure 6.

	Radio Interface	Checkbox Interface
Number of HITs/sec	0.0015	0.167

Table 5: Average Time spent on HITs by crowdworkers

We force our workers to make selections from our six reasonings and do not ask the cause behind that. There may be instances where workers will not agree with our reasonings. Our interface does not allow workers to insert their own reasoning, furthermore, it does not record the rationale behind it. The future work may allow user to add their own responses with detail explanation. This can be utilised to identify other potential reasons for disagreement and the rationale behind selecting a reason for disagreement.

Conclusion

In this paper, we explored the idea that the differences in human interpretation may lead to differences in approaching a single ground truth in the field of visual question/answering. We performed crowdsourcing tasks to identify and evaluate the various causes of disagreement among crowd workers in answering a visual question. We utilized different evaluation metrics to support our claim. We learnt that there are various factors which affect the human judgement – the ambiguity in the visual question, clarity of the associated image, design of the crowdsourcing task, to name a few. We realized a proper understanding of the crowd disagreement could help us achieve a faster and a cost-effective solution to obtain ground truth answers for visual questions.

As future work, it would be interesting to note how harnessing the disagreements among crowd workers could be utilized to train, evaluate, and improve the existing Visual Question Answering systems.

Acknowledgments

We would like to thank the *crowd workers* for participating in our experiments and *Danna Gurari* for her suggestions.

References

- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.
- Aroyo, L., and Welty, C. 2013. Measuring crowd truth for medical relation extraction. In *2013 AAAI Fall Symposium Series*.
- Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; et al.

DISTRIBUTION OF REASONING THAT APPEARED TOGETHER

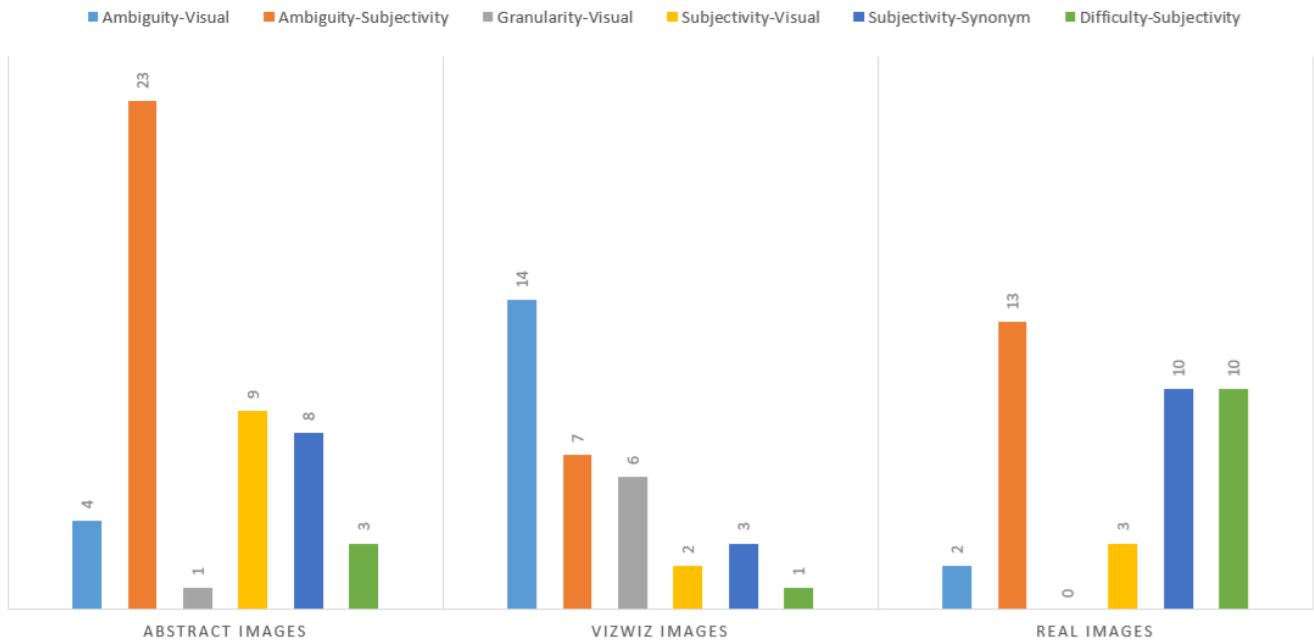


Figure 5: Graphical Representation of Responses Appearing Together on Checkbox Interface

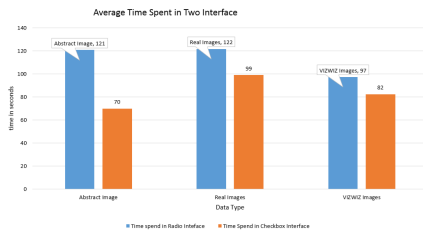


Figure 6: Box Plot for Time Taken in three datasets for Radio Button Interface

2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, 333–342. ACM.

Finnerty, A.; Kucherbaev, P.; Tranquillini, S.; and Convertino, G. 2013. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, 14. ACM.

Gurari, D., and Grauman, K. 2017. Crowdverge: Predicting if people will agree on the answer to a visual question.

Inel, O.; Khamkham, K.; Cristea, T.; Dumitrache, A.; Rutjes, A.; van der Ploeg, J.; Romaszko, L.; Aroyo, L.; and Sips, R.-J. 2014. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *International Semantic Web Conference*, 486–504. Springer.

Kafle, K., and Kanan, C. 2016. Answer-type prediction

for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4976–4984.

Kairam, S., and Heer, J. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1637–1648. ACM.

Khanna, S.; Ratan, A.; Davis, J.; and Thies, W. 2010. Evaluating and improving the usability of mechanical turk for low-income workers in india. In *Proceedings of the first ACM symposium on computing for development*, 12. ACM.

Lasecki, W. S.; Thiha, P.; Zhong, Y.; Brady, E.; and Bigham, J. P. 2013. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, 18. ACM.

Li, X., and Roth, D. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering* 12(03):229–249.

Mifsud, J. 2015. Usability metrics – a guide to quantify the usability of any system.

Mishra, M.; Mishra, V. K.; and Sharma, H. 2012. Question classification: Semantic feature. *Research Journal of Engineering and Technology* 3(4):III.

Rahmanian, B., and Davis, J. G. 2014. User interface design for crowdsourcing systems. In *Proceedings of the 2014*

International Working Conference on Advanced Visual Interfaces, 405–408. ACM.

Ross, J.; Irani, L.; Silberman, M.; Zaldivar, A.; and Tomlinson, B. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, 2863–2872. ACM.

Soberón, G.; Aroyo, L.; Welty, C.; Inel, O.; Lin, H.; and Overmeen, M. 2013. Measuring crowd truth: Disagreement metrics combined with worker behavior filters. In *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web-Volume 1030*, 45–58. CEUR-WS. org.

Su, H.; Deng, J.; and Fei-Fei, L. 2012. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, volume 1.

Suzuki, J.; Taira, H.; Sasaki, Y.; and Maeda, E. 2003. Question classification using hdag kernel. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, 61–68. Association for Computational Linguistics.