Learning to Recognize When and Why a Crowd Will Offer Different Answers to a Visual Question

Danna Gurari
University of Texas at Austin
School of Information
July 19, 2019
Task: Answer Visual Questions (VQs)

- Is my monitor on?
- Hi there can you please tell me what flavor this is?
- Does this picture look scary?
- Which side of the room is the toilet on?

Asked by sighted and blind people
Why Visual Question Answering Matters?

Human-powered systems:

Automated systems (in research)
Observation of Crowdsourced Answers

Is my monitor on?
- (1) yes
- (2) yes
- (3) yes
- (4) yes
- (5) yes
- (6) yes
- (7) yes
- (8) yes
- (9) yes
- (10) yes

Hi there can you please tell me what flavor this is?
- (1) sweet pepper
- (2) sweet pepper
- (3) sweet pepper
- (4) sweet pepper
- (5) sweet pepper
- (6) sweet pepper
- (7) sweet pepper
- (8) sweet pepper
- (9) sweet pepper
- (10) sweet pepper

Does this picture look scary?
- (1) yes
- (2) no
- (3) no
- (4) yes
- (5) no
- (6) yes
- (7) yes
- (8) no
- (9) no
- (10) no

Which side of the room is the toilet on?
- (1) right
- (2) left
- (3) right
- (4) right
- (5) right
- (6) right
- (7) right side
- (8) right
- (9) center
- (10) right
Key Observation #1: Crowd Can All Agree

Collecting redundant answers compromises efficiency!

(wastes time & money)
Key Observation #2: Crowd Answers Can Differ

Does this picture look scary?
(1) yes
(2) no
(3) no
(4) yes
(5) no
(6) yes
(7) yes
(8) no
(9) no
(10) no

Which side of the room is the toilet on?
(1) right
(2) left
(3) right
(4) right
(5) right
(6) right
(7) right side
(8) right
(9) center
(10) right

Unclear what to do.
Goal: Efficiently Collect All Plausible Answers

Is my monitor on?

Hi there can you please tell me what flavor this is?

Does this picture look scary?

Which side of the room is the toilet on?
This Talk

1: Learning to anticipate *when* a crowd will answer a VQ differently towards *efficiently capturing all plausible answers* for a batch of VQs

2: Learning to anticipate *why* a crowd will answer a VQ differently towards *supporting users to resolve answer differences*
This Talk

1: Learning to anticipate *when* a crowd will answer a VQ differently towards *efficiently capturing all plausible answers* for a batch of VQs

2: Learning to anticipate *why* a crowd will answer a VQ differently towards *supporting users to resolve answer differences*
Does this picture look scary?
(1) yes  (2) no  (3) no  (4) yes  (5) no  (6) yes  (7) yes  (8) no  (9) no  (10) no

Which side of the room is the toilet on?
(1) right  (2) left  (3) right  (4) right  (5) right  (6) right  (7) right side  (8) right  (9) center  (10) right

Is my monitor on?
(1) yes  (2) yes  (3) yes  (4) yes  (5) yes  (6) yes  (7) yes  (8) yes  (9) yes  (10) yes

Hi there can you please tell me what flavor this is?
(1) sweet pepper  (2) sweet pepper  (3) sweet pepper  (4) sweet pepper  (5) sweet pepper  (6) sweet pepper  (7) sweet pepper  (8) sweet pepper  (9) sweet pepper  (10) sweet pepper

Motivation: Crowd Answers Can Match or Differ
Redundancy Wasteful
Redundancy Valuable
Motivation: How Often Do Answers Match?

VizWiz (asked by blind people)  
[Bigham et al; UIST ’10]

VQA Real (asked by sighted people)  
[Antol et al; ICCV ‘15]

VQA Abstract Scenes (asked by sighted people)  
[Antol et al; ICCV ‘15]

Hi there can you please tell me what flavor this is?  
Why might the crowd be carrying signs?  
What is the girl sitting on?

(Checked for string match between 10 answers/VQ)
Motivation: How Often Do Answers Match?

Answers differ for 53% of nearly 500,000 visual questions!

<table>
<thead>
<tr>
<th># VQAs (%)</th>
<th>VizWiz</th>
<th>VQA - Real Images</th>
<th>VQA - Abstract Scenes</th>
</tr>
</thead>
<tbody>
<tr>
<td># VQAs (%)</td>
<td>1,499</td>
<td>369,861</td>
<td>90,000</td>
</tr>
<tr>
<td>At Most One Disagreement</td>
<td>20%</td>
<td>52%</td>
<td>58%</td>
</tr>
</tbody>
</table>
Idea: Predict If a Crowd’s Answers Will Differ

No (matches):
1. Is my monitor on?
2. Hi there can you please tell me what flavor this is?

Yes (differs):
1. Does this picture look scary?
2. Which side of the room is the toilet on?

[Gurari & Grauman; CHI 2017]
Method: Predict If a Crowd’s Answers Will Differ

**Given:**

- ![Image](image1.png)
  - Is my monitor on?
- ![Image](image2.png)
  - Hi there can you please tell me what flavor this is?

**Want:**

- Yes
- No

**Build Prediction System**

- ![Image](image3.png)
  - Does this picture look scary?
- ![Image](image4.png)
  - Which side of the room is the toilet on?

[ Gurari & Grauman; CHI 2017 ]
Method: Predict If a Crowd’s Answers Will Differ

Given:

<table>
<thead>
<tr>
<th>Is my monitor on?</th>
<th>Hi there can you please tell me what flavor this is?</th>
<th>Does this picture look scary?</th>
<th>Which side of the room is the toilet on?</th>
</tr>
</thead>
<tbody>
<tr>
<td>No (matches)</td>
<td></td>
<td>Yes (differs)</td>
<td></td>
</tr>
</tbody>
</table>

Create Training Data

<table>
<thead>
<tr>
<th>Given:</th>
<th>Want:</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
<td>Yes</td>
</tr>
<tr>
<td><img src="image2" alt="Image" /></td>
<td>No</td>
</tr>
<tr>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
</tbody>
</table>

Build Prediction System

[ Gurari & Grauman; CHI 2017 ]
Method: Predict If a Crowd’s Answers Will Differ

Build Prediction System

Train the Prediction System

Given:

Want:

Yes  No  Yes

1. Deep Learning (DL) algorithm
2. Random Forest, handcrafted features

[ Gurari & Grauman; CHI 2017]
Method: Predict If a Crowd’s Answers Will Differ

No (matches)
- Is my monitor on?
- Hi there can you please tell me what flavor this is?

Yes (differs)
- Does this picture look scary?
- Which side of the room is the toilet on?

Apply Prediction System

Answer:
- No
- Yes

Confidence:
- 60%
- 100%

[ Gurari & Grauman; CHI 2017 ]
Experiments

VQA Real Images
[Antol et al; ICCV ‘15]
248,349 training VQs, 121,512 testing VQs

VizWiz
[Bigham et al; UIST ‘10]
1,200 training VQs, 299 testing VQs

Status quo: randomly predict confidence in disagreement
Experiments

VQA Real Images
[Antol et al; ICCV ‘15]
248,349 training VQs, 121,512 testing VQs

Experiments

VizWiz
[Bigham et al; UIST ‘10]
1,200 training VQs, 299 testing VQs

VQA algorithm’s confidence in its predicted answer
Experiments

VQA Real Images
[Antol et al; ICCV ‘15]

248,349 training VQs, 121,512 testing VQs

Ours: RF (AP = 0.76)
Ours: DL (AP = 0.73)
[ICCV 2015] (AP = 0.64)
Status Quo (AP = 0.49)

VizWiz
[Bigham et al; UIST ‘10]

1,200 training VQs, 299 testing VQs

Ours: RF (AP = 0.88)
[ICCV 2015] (AP = 0.81)
Status Quo (AP = 0.79)

Our deep learning system outperforms all baselines!
Experiments

VQA Real Images
[Antol et al; ICCV ‘15]
248,349 training VQs, 121,512 testing VQs

VizWiz
[Bigham et al; UIST ‘10]
1,200 training VQs, 299 testing VQs

Our random forest classifier is the top-performing predictor!

Ours: RF (AP = 0.76)
Ours: DL (AP = 0.73)
[ICCV 2015] (AP = 0.64)
Status Quo (AP = 0.49)

Ours: RF (AP = 0.88)
[ICCV 2015] (AP = 0.81)
Status Quo (AP = 0.79)
### Experiments: What Predictive Cues Are Learned?

<table>
<thead>
<tr>
<th>Question</th>
<th>Answers</th>
<th>Predict Strong Agree</th>
<th>Predict Strong Disagree</th>
</tr>
</thead>
</table>
| Why doesn’t the bench tip over?   | (1) it can’t  
(2) concreted down  
(3) cemented in ground  
(4) weighted  
(5) it’s bolted to ground       | Strong Disagree      | Strong Agree            |
| What room is pictured with a sink?| (1) bathroom  
(2) bathroom  
(3) bathroom  
(4) bathroom  
(5) bathroom       | Strong Agree         | Strong Disagree         |
| Can you tell me what that is?     | (1) ready cut macaroni  
(2) macaroni  
(3) macaroni  
(4) macaroni  
(5) macaroni       | Strong Agree         | Strong Disagree         |
| Is my monitor on?                 | (1) yes  
(2) yes  
(3) yes  
(4) yes  
(5) yes       | Strong Agree         | Strong Disagree         |

Most of the predictive power comes from language-based features.
Experiments: What Predictive Cues Are Learned?

(quantitative analysis in paper)

Most of the predictive power comes from language-based features.
Idea #2: Efficiently Collect All Plausible Answers...

**No (matches)**
- Is my monitor on?
- Hi there can you please tell me what flavor this is?

**Yes (differs)**
- Does this picture look scary?
- Which side of the room is the toilet on?

(1) yes  (1) sweet pepper  (1) yes  (1) left
(R) no  (R) center
Given a Human Budget

Ranked:

Most Likely Match

1. Is my monitor on?
2. Hi there can you please tell me what flavor this is?

Most Likely Differ

1. Does this picture look scary?
2. Which side of the room is the toilet on?

(1) yes  (1) sweet pepper  (1) yes  (1) left
(1) no   (R) center
Efficiently Collect All Plausible Answers

VQA Real Images Dataset (121,512 VQs)

% All Answers Captured

% VQs with Redundant Answers

Ours: RF

[ICCV 2015]

Status Quo

(121,512 VQs)
Efficiently Collect All Plausible Answers

VQA Real Images Dataset
(121,512 VQs)

Accelerates capturing diversity by 23% compared to the VizWiz system.
Efficiently Collect All Plausible Answers

This translates to eliminating over 11 40-hour work weeks and saving $1800!
Efficiently Collect All Plausible Answers

Can obtain further performance gains by collecting just enough answers to capture all unique answers (instead of 1 or 5 answers only)
Efficiently Collect All Plausible Answers

Can obtain further performance gains by collecting just enough answers to capture all unique answers (instead of 1 or 5 answers only)

[Yang, Grauman, & Gurari; HCOMP 2018]
Efficiently Collect All Plausible Answers

**Is my monitor on?**
(1) yes  
(2) yes  
(3) yes  
(4) yes  
(5) yes  
(6) yes  
(7) yes  
(8) yes  
(9) yes  
(10) yes

**Does this picture look scary?**
(1) yes  
(2) no  
(3) no  
(4) yes  
(5) no  
(6) yes  
(7) yes  
(8) no  
(9) no  
(10) no

**What is in this can?**
(1) unanswerable  
(2) unanswerable  
(3) unanswerable  
(4) tuna  
(5) cat food  
(6) cat food  
(7) unanswerable  
(8) unanswerable  
(9) unanswerable  
(10) pet food

**What is this?**
(1) clothesline  
(2) game net  
(3) backyard  
(4) clothes line pole  
(5) unanswerable  
(6) tiki torch  
(7) garden  
(8) backyard  
(9) compound  
(10) backyard barbecue area
Is my monitor on? (1) yes

Does this picture look scary? (1) yes (2) no

What is in this can? (1) unanswerable (2) tuna (3) cat food (4) pet food

What is this? (1) clothesline (2) game net (3) backyard (4) clothes line pole (5) unanswerable (6) tiki torch (7) garden (8) backyard (9) compound (10) backyard barbecue area
This Talk

1: Learning to anticipate *when* a crowd will answer a VQ differently towards efficiently capturing all plausible answers for a batch of VQs

2: Learning to anticipate *why* a crowd will answer a VQ differently towards supporting users to resolve answer differences
Motivation: How to Resolve Answer Differences?

When does this expire?
(1) august 14 2013
(2) 08 14 2013
(3) 14 august 2013
(4) 14 aug 2013
(5) 14 aug 2013
(6) august 14th 2013
(7) 14aug2013
(8) 14 aug 2013
(9) 14 aug 2013
(10) aug 14 2013

Does this picture look scary?
(1) yes
(2) no
(3) no
(4) yes
(5) no
(6) yes
(7) yes
(8) no
(9) no
(10) no

Which side of the room is the toilet on?
(1) right
(2) left
(3) right
(4) right
(5) right
(6) right
(7) right side
(8) right
(9) center
(10) right

Synonymous Answers
Subjectivity
Ambiguity
# Dataset Creation: Why Answers May Differ

<table>
<thead>
<tr>
<th>Synonyms</th>
<th>Subjective</th>
<th>Ambiguous</th>
<th>Spam</th>
<th>Answer Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>How are the water conditions?</td>
<td>Could the floor use a mopping?</td>
<td>Which side of the room is the toilet on?</td>
<td>What kind of spice is this?</td>
<td>What is in this can?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Difficult</th>
<th>Low Quality Image</th>
<th>Invalid VQ</th>
<th>Granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many sheep are there?</td>
<td>Can you see the label and tell me what it is please?</td>
<td>I just wanted to say thank you for your assistance.</td>
<td>What book is this?</td>
</tr>
</tbody>
</table>
Dataset Creation: Source

Visual Questions (VQs) where 10 answers do not all match:

VizWiz: 29,974 VQs (asked by blind people)  
[ Gurari et al; CVPR ‘18]

VQA Real: 5,034 VQs (asked by sighted people)  
[ Goyal et al; CVPR ‘17]

Is my monitor on?

Does this picture look scary?

[Under Review]
Why Answers Differ?

~20% of answer differences arise because of the question-image pair alone

Dataset: VizWiz
Dataset: VQA_2.0
Dataset: Combined

Percentage of VQAs

Frequency of Disagreement Sources

1 Person Threshold  2 Person Threshold  3 Person Threshold

[Under Review]
Why Answers Differ?

Most common reasons are ambiguity, synonymous answers, & varying levels of answer granularity.

Frequency of Disagreement Sources

Dataset: VizWiz
Dataset: VQA_2.0
Dataset: Combined

Most common reasons are ambiguity, synonymous answers, & varying levels of answer granularity.

[Under Review]
Why Answers Differ?

Least common reason is spam (~1% across both datasets)
Methods: Predicting Why Answers Will Differ
Experiments: Predicting Why Answers Will Differ

VizWiz Dataset (7988 test examples)

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall</th>
<th>LQI</th>
<th>IVE</th>
<th>INV</th>
<th>DFF</th>
<th>AMB</th>
<th>SBJ</th>
<th>SYN</th>
<th>GRN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>30.15</td>
<td>23.59</td>
<td>33.69</td>
<td>18.15</td>
<td>5.70</td>
<td>74.70</td>
<td>5.14</td>
<td>66.61</td>
<td>71.94</td>
</tr>
<tr>
<td>QI-Relevance [31]</td>
<td>31.71</td>
<td>30.56</td>
<td>40.52</td>
<td>18.15</td>
<td>5.7</td>
<td>76.53</td>
<td>5.14</td>
<td>66.61</td>
<td>71.94</td>
</tr>
<tr>
<td>Unanswerable [21]</td>
<td>35.31</td>
<td>44.82</td>
<td>58.63</td>
<td>18.15</td>
<td>5.7</td>
<td>80.14</td>
<td>5.14</td>
<td>66.61</td>
<td>71.94</td>
</tr>
<tr>
<td>I</td>
<td>40.54</td>
<td>55.42</td>
<td>50.66</td>
<td>30.12</td>
<td>8.77</td>
<td>83.39</td>
<td>8.64</td>
<td>79.76</td>
<td>86.29</td>
</tr>
<tr>
<td>Q</td>
<td>40.5</td>
<td>35.87</td>
<td>54.66</td>
<td>39.24</td>
<td>12.32</td>
<td>84.41</td>
<td>11.00</td>
<td>79.46</td>
<td>85.10</td>
</tr>
<tr>
<td>Q+I</td>
<td>45.73</td>
<td>57.81</td>
<td>62.47</td>
<td>43.24</td>
<td>13.77</td>
<td>87.81</td>
<td>11.14</td>
<td>86.36</td>
<td>92.01</td>
</tr>
<tr>
<td>Q+I+A</td>
<td>50.02</td>
<td>65.58</td>
<td>77.42</td>
<td>56.54</td>
<td>10.49</td>
<td>89.70</td>
<td>11.26</td>
<td>90.42</td>
<td>95.44</td>
</tr>
<tr>
<td>Q+I+A_FT</td>
<td>50.01</td>
<td>64.93</td>
<td>77.40</td>
<td>56.78</td>
<td>10.10</td>
<td>89.48</td>
<td>13.16</td>
<td>90.52</td>
<td>95.50</td>
</tr>
</tbody>
</table>

Overall, outperforms related baselines by ~15%

[Under Review]
Experiments: Predicting Why Answers Will Differ

VizWiz Dataset (7988 test examples)

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall</th>
<th>LQI</th>
<th>IVE</th>
<th>INV</th>
<th>DFF</th>
<th>AMB</th>
<th>SBJ</th>
<th>SYN</th>
<th>GRN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>30.15</td>
<td>23.59</td>
<td>33.69</td>
<td>18.15</td>
<td>5.70</td>
<td>74.70</td>
<td>5.14</td>
<td>66.61</td>
<td>71.94</td>
</tr>
<tr>
<td>QI-Relevance [31]</td>
<td>31.71</td>
<td>30.56</td>
<td>40.52</td>
<td>18.15</td>
<td>5.7</td>
<td>76.53</td>
<td>5.14</td>
<td>66.61</td>
<td>71.94</td>
</tr>
<tr>
<td>Unanswerable [21]</td>
<td>35.31</td>
<td>44.82</td>
<td>58.63</td>
<td>18.15</td>
<td>5.7</td>
<td>80.14</td>
<td>5.14</td>
<td>66.61</td>
<td>71.94</td>
</tr>
<tr>
<td>I</td>
<td>40.54</td>
<td>55.42</td>
<td>50.66</td>
<td>30.12</td>
<td>8.77</td>
<td>83.39</td>
<td>8.64</td>
<td>79.76</td>
<td>86.29</td>
</tr>
<tr>
<td>Q</td>
<td>40.5</td>
<td>35.87</td>
<td>54.66</td>
<td>39.24</td>
<td>12.32</td>
<td>84.41</td>
<td>11.00</td>
<td>79.46</td>
<td>85.10</td>
</tr>
<tr>
<td>Q+I</td>
<td>45.73</td>
<td>57.81</td>
<td>62.47</td>
<td>43.24</td>
<td>13.77</td>
<td>87.81</td>
<td>11.14</td>
<td>86.36</td>
<td>92.01</td>
</tr>
<tr>
<td>Q+I+A</td>
<td>50.02</td>
<td>65.58</td>
<td>77.42</td>
<td>56.54</td>
<td>10.49</td>
<td>89.70</td>
<td>11.26</td>
<td>90.42</td>
<td>95.44</td>
</tr>
<tr>
<td>Q+I+A_FT</td>
<td>50.01</td>
<td>64.93</td>
<td>77.40</td>
<td>56.78</td>
<td>10.10</td>
<td>89.48</td>
<td>13.16</td>
<td>90.52</td>
<td>95.50</td>
</tr>
</tbody>
</table>

Promising results for informing people who are blind when their images are too low quality or has insufficient visual evidence.
Experiments: Predicting Why Answers Will Differ

VizWiz Dataset (7988 test examples)

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall</th>
<th>LQI</th>
<th>IVE</th>
<th>INV</th>
<th>DFF</th>
<th>AMB</th>
<th>SBJ</th>
<th>SYN</th>
<th>GRN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>30.15</td>
<td>23.59</td>
<td>33.69</td>
<td>18.15</td>
<td>5.70</td>
<td>74.70</td>
<td>5.14</td>
<td>66.61</td>
<td>71.94</td>
</tr>
<tr>
<td>QI-Relevance [31]</td>
<td>31.71</td>
<td>30.56</td>
<td>40.52</td>
<td>18.15</td>
<td>5.7</td>
<td>76.53</td>
<td>5.14</td>
<td>66.61</td>
<td>71.94</td>
</tr>
<tr>
<td>Unanswerable [21]</td>
<td>35.31</td>
<td>44.82</td>
<td>58.63</td>
<td>18.15</td>
<td>5.7</td>
<td>80.14</td>
<td>5.14</td>
<td>66.61</td>
<td>71.94</td>
</tr>
<tr>
<td>I</td>
<td>40.54</td>
<td>55.42</td>
<td>50.66</td>
<td>30.12</td>
<td>8.77</td>
<td>83.39</td>
<td>8.64</td>
<td>79.76</td>
<td>86.29</td>
</tr>
<tr>
<td>Q</td>
<td>40.5</td>
<td>35.87</td>
<td>54.66</td>
<td>39.24</td>
<td>12.32</td>
<td>84.41</td>
<td>11.00</td>
<td>79.46</td>
<td>85.10</td>
</tr>
<tr>
<td>Q+I</td>
<td>45.73</td>
<td>57.81</td>
<td>62.47</td>
<td>43.24</td>
<td>13.77</td>
<td>87.81</td>
<td>11.14</td>
<td>86.36</td>
<td>92.01</td>
</tr>
<tr>
<td>Q+I+A</td>
<td><strong>50.02</strong></td>
<td><strong>65.58</strong></td>
<td><strong>77.42</strong></td>
<td><strong>56.54</strong></td>
<td><strong>10.49</strong></td>
<td><strong>89.70</strong></td>
<td>11.26</td>
<td>90.42</td>
<td>95.44</td>
</tr>
<tr>
<td>Q+I+A_FT</td>
<td>50.01</td>
<td>64.93</td>
<td>77.40</td>
<td>56.78</td>
<td>10.10</td>
<td>89.48</td>
<td>13.16</td>
<td>90.52</td>
<td>95.50</td>
</tr>
</tbody>
</table>

Performs poorly for detecting difficult or subjective visual questions
Why Useful to Predict Reasons?

Synonyms
- How are the water conditions?

Subjective
- Could the floor use a mopping?

Ambiguous
- Which side of the room is the toilet on?

Spam
- What kind of spice is this?

Difficult
- How many sheep are there?

Low Quality Image
- Can you see the label and tell me what it is please?

Invalid VQ
- I just wanted to say thank you for your assistance.

Granularity
- What book is this?

Could modify visual question
- What is in this can?
Why Useful to Predict Reasons?

<table>
<thead>
<tr>
<th>Classification</th>
<th>Example Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonyms</td>
<td>How are the water conditions?</td>
</tr>
<tr>
<td>Subjective</td>
<td>Could the floor use a mopping?</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>Which side of the room is the toilet on?</td>
</tr>
<tr>
<td>Spam</td>
<td>What kind of spice is this?</td>
</tr>
<tr>
<td>Answer Missing</td>
<td>What is in this can?</td>
</tr>
<tr>
<td>Difficult</td>
<td>How many sheep are there?</td>
</tr>
<tr>
<td>Low Quality Image</td>
<td>Can you see the label and tell me what it is please?</td>
</tr>
<tr>
<td>Invalid VQ</td>
<td>I just wanted to say thank you for your assistance.</td>
</tr>
<tr>
<td>Granularity</td>
<td>What book is this?</td>
</tr>
<tr>
<td>Reason</td>
<td>Example Questions</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Synonyms</td>
<td>How are the water conditions?</td>
</tr>
<tr>
<td>Subjective</td>
<td>Could the floor use a mopping?</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>Which side of the room is the toilet on?</td>
</tr>
<tr>
<td>Difficult</td>
<td>How many sheep are there?</td>
</tr>
<tr>
<td>Low Quality Image</td>
<td>Can you see the label and tell me what it is please?</td>
</tr>
<tr>
<td>Invalid VQ</td>
<td>I just wanted to say thank you for your assistance.</td>
</tr>
<tr>
<td>Granularity</td>
<td>What book is this?</td>
</tr>
</tbody>
</table>

Identify what crowd aggregation scheme to use to choose a best answer
<table>
<thead>
<tr>
<th>Why Useful to Predict Reasons?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Synonyms</strong></td>
</tr>
<tr>
<td>How are the water conditions?</td>
</tr>
<tr>
<td><strong>Subjective</strong></td>
</tr>
<tr>
<td>Could the floor use a mopping?</td>
</tr>
<tr>
<td><strong>Ambiguous</strong></td>
</tr>
<tr>
<td>Which side of the room is the toilet on?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many sheep are there?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Low Quality Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can you see the label and tell me what it is please?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Invalid VQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>I just wanted to say thank you for your assistance.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>What book is this?</td>
</tr>
</tbody>
</table>
This Talk

1: Learning to anticipate *when* a crowd will answer a VQ differently towards *efficiently capturing all plausible answers* for a batch of VQs

2: Learning to anticipate *why* a crowd will answer a VQ differently towards *supporting users to resolve answer differences*
Acknowledgments

Chun-Ju Yang  Dr. Kristen Grauman  Nilavra Bhattacharya  Qing Li  Crowd Workers

Funding
Conclusions

A crowd’s answers often differ:
- Answer differences arise for ~50% of nearly 500,000 VQs

Novel problem and algorithms:
- Algorithms can predict when, to what extent, and why answers will differ

Answers can be collected more efficiently:
- Over 23% acceleration to collect all plausible answers from a crowd

Gurari & Grauman, CHI 2017; Yang, Grauman, & Gurari, HCOMP 2018; Under Review