Visual Questions: Learning to Assist Blind People and Detect When/Why a Crowd Will Disagree on the Answer

Danna Gurari
University of Texas at Austin
School of Information
September 8, 2018
Task: Answer Visual Questions (VQs)

What food is on the plate?

Which side of the room is the toilet on?

Which currency is this?

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

Asked by sighted and blind people
Human-Based Applications for Answering VQs

- VizWiz
- aira
- Be My Eyes
- BeSpecular
- TapTapSee
Goal: Automatically Answer VQs
Goal: Automatically Answer VQs

Foster community of researchers in academia to advance assistive technology.
Key Challenge #1: Assist All Audiences

What food is on the plate?

Which side of the room is the toilet on?

Which currency is this?

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

Status Quo: VQs from Sighted People

Need: VQs from Blind People
Key Challenge #2: Return Desired Answer

What food is on the plate?
A: pizza

Which side of the room is the toilet on?
A: right

Which currency is this?
A: arabic

Hi there can you please tell me what flavor this is?
A: sweet pepper

Status Quo: 1 answer/VQ
Key Challenge #2: Return Desired Answer

What food is on the plate?
(1) pizza
(2) pizza
(3) pizza
(4) pizza
(5) pizza
(6) pizza
(7) pizza
(8) pizza
(9) pizza
(10) pizza

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

Which currency is this?
(1) arabic
(2) unanswerable
(3) unanswerable
(4) unanswerable
(5) unanswerable
(6) german
(7) unanswerable
(8) unanswerable
(9) unanswerable
(10) euro

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

Status Quo: 1 answer/VQ
Need: All Plausible Answers
This Talk

1: Designing VQA algorithms to meet the needs of blind people.

2: Learning to what extent multiple people will answer a VQ differently.

3: Learning why multiple people will answer a VQ differently.
This Talk

1: Designing VQA algorithms to meet the needs of blind people.

2: Learning to what extent multiple people will answer a VQ differently.

3: Learning why multiple people will answer a VQ differently.
Motivation: Assistive Mobile Application Trend

- VizWiz
- TapTapSee
- Scan Search
- BeMyEyes
- RegionSpeak
- BeSpecular
- LookTel
- Chorus:View
- Aipoly
- Recognizer
- ToolWiz Eyes
- Seeing AI

Timeline: 2010 to 2018
Motivation: VizWiz Mobile Phone Application

Motivation: VizWiz Mobile Phone Application

Motivation: VizWiz Mobile Phone Application

Motivation: VizWiz Mobile Phone Application

Motivation: VizWiz Mobile Phone Application

Motivation: VizWiz Mobile Phone Application

11,045 people asked 72,205 visual questions between 2011 and 2015

Idea: Dataset Creation, Analysis, & Algorithms

11,045 people asked 72,205 visual questions between 2011 and 2015

[Gurari et al; CVPR 2018]
## Dataset Creation: First Dataset Originating from Blind People

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Which Images?</th>
<th>Who Asked?</th>
<th>How Asked?</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAQUAR [28]</td>
<td>NYU Depth V2 [34]</td>
<td>Automatically generated (templates)</td>
<td>—</td>
</tr>
<tr>
<td>KB-VQA [40]</td>
<td>MSCOCO [25]</td>
<td>In-house participants</td>
<td>Typed</td>
</tr>
<tr>
<td>CLEVR [19]</td>
<td>Synthetic Shapes</td>
<td>Automatically generated (templates)</td>
<td>—</td>
</tr>
<tr>
<td>SHAPES [6]</td>
<td>Synthetic Shapes</td>
<td>Automatically generated (templates)</td>
<td>—</td>
</tr>
<tr>
<td><strong>Ours - VizWiz</strong></td>
<td>Blind people use mobile phones to take a picture and ask question</td>
<td>Spoken</td>
<td>—</td>
</tr>
</tbody>
</table>
Dataset Creation

**Goal:** Protect privacy and safety for all individuals involved with the dataset

*(First dataset where real users asked visual questions to support their real daily needs)*

[Gurari et al; CVPR 2018]
Dataset Creation

Anonymization

1. Transcription (removes voice)

2. Re-save (removes metadata)

[Gurari et al; CVPR 2018]
Dataset Creation

Anonymization

1. Transcription (removes voice)

2. Re-save (removes metadata)

Filtering
(Personally Identifying Information, Nudity)
Dataset Creation

Anonymization

1. Transcription (removes voice)

2. Re-save (removes metadata)

Filtering

Answer Collection

[Gurari et al; CVPR 2018]
Dataset Creation

Anonymization

1. Transcription (removes voice)
2. Re-save (removes metadata)

Filtering

Answer Collection

31,173 image/question pairs and 311,730 answers
(Available at http://vizwiz.org/data)

[Gurari et al; CVPR 2018]
Dataset Analysis: What Blind People Ask About

1. Q: Does this foundation have any sunscreen?  
   A: yes

2. Q: What is this?  
   A: 10 euros

3. Q: What color is this?  
   A: green

4. Q: Please can you tell me what this item is?  
   A: butternut squash red pepper soup

5. Q: Is it sunny outside?  
   A: yes

6. Q: Is this air conditioner on fan, dehumidifier, or air conditioning?  
   A: air conditioning

7. Q: What type of pills are these?  
   A: unsuitable image

8. Q: What type of soup is this?  
   A: unsuitable image

9. Q: Who is this mail for?  
   A: unanswerable

10. Q: When is the expiration date?  
    A: unanswerable

11. Q: What is this?  
    A: unanswerable

12. Q: Can you please tell me what the oven temperature is set to?  
    A: unanswerable
Dataset Analysis: What Blind People Ask About

Most common question:

“What is this?”

[Gurari et al; CVPR 2018]
Dataset Analysis: What Blind People Ask About

Many questions begin with a rare word:

“Please…”
“Hi…”
“Okay…”

[Gurari et al; CVPR 2018]
Many images suffer from poor lighting, focus, and framing of the content of interest.
Dataset Analysis: What Blind People Ask About

~29% of visual questions are not answerable
Algorithm Benchmarking: Predict Answer

- Is this shirt clean or dirty?
  - Clean

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

- What type of pills are these?
  - Unanswerable

- What is this?
  - Unanswerable
Algorithm Benchmarking: Predict Answer


<table>
<thead>
<tr>
<th>Method</th>
<th>Acc</th>
<th>CIDEr</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q+I [18]</td>
<td>0.137</td>
<td>0.224</td>
<td>0.000</td>
<td>0.078</td>
</tr>
<tr>
<td>Q+I+A [23]</td>
<td>0.145</td>
<td>0.237</td>
<td>0.000</td>
<td>0.082</td>
</tr>
<tr>
<td>Q+I+BUA [6]</td>
<td>0.134</td>
<td>0.226</td>
<td>0.000</td>
<td>0.077</td>
</tr>
</tbody>
</table>

Existing algorithms generalize poorly as is.

[ Gurari et al; CVPR 2018 ]
Algorithm Benchmarking: Predict Answer


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

Fine-tuning (FT) & training from scratch yield significant improvements!
Algorithm Benchmarking: Predict Answerability

Is this shirt clean or dirty?  
answerable

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

What type of pills are these?  
unanswerable

What is this?  
unanswerable
Algorithm Benchmarking: Predict Answerability

Recall

Precision

Q+C (Caption) [2] (30.6)
Fine-Tuning [2] (56.1)
Train on VizWiz [2] (60.5)
Train on VizWiz [1] (56.0)
Q: 1-layer LSTM (48.9)
C: 1-layer LSTM (46.4)
I: pool5 of ResNet (64.0)
Q+I (71.7)

[1] Anderson et al. CVPR ’18

All methods outperform status quo [2] by 25-41 percentage points!
Algorithm Benchmarking: Predict Answerability

![Graph showing precision vs recall for different models.](image)

- Q+C (Caption) [2] (30.6)
- Fine-Tuning [2] (56.1)
- Train on VizWiz [2] (60.5)
- Train on VizWiz [1] (56.0)
- Q: 1-layer LSTM (48.9)
- C: 1-layer LSTM (46.4)
- I: pool5 of ResNet (64.0)
- Q+I (71.7)

[1] Anderson et al. CVPR ’18

Image information is most predictive and is improved by the question.
ECCV 2018 VizWiz Challenge:
A Difficult Dataset for Modern Vision Algorithms

This Friday September 14 at Theresianum 601 in TU München
(http://www.vizwiz.org/workshop)
This Talk

1: Designing VQA algorithms to meet the needs of blind people.

2: Learning to what extent multiple people will answer a VQ differently.

3: Learning why multiple people will answer a VQ differently.
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

- **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
Key Observation #1: All Answers Match

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
Key Observation #2: All Answers Differ

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
Key Observation #3: Several Shared 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
Idea: Predict Answer Diversity

1 Answer

(Least Diversity)

Does this picture look scary?

(1) yes
(2) yes
(3) yes
(4) yes
(5) yes
(6) yes
(7) yes
(8) yes
(9) yes
(10) yes

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

10 Answers

(Most Diversity)

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
Analyzing If A Crowd’s Answers Differ: Method

10 answers (using AMT) per visual question (VQ):

VizWiz: 1,499 VQs (asked by blind people)  
[Bigham et al; UIST ‘10]

VQA Real: 369,861 VQs (asked by sighted people)  
[Antol et al; ICCV ‘15]

VQA Abstract Scenes: 90,000 VQs (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?
Analyzing If A Crowd’s Answers Differ: Method

Q1: What is unusual about this mustache?
A1: Write your answer here.

Please answer the question using as few words as possible.

Courtesy of Agrawal et al; CVPR 2015
Analyzing If A Crowd’s Answers Differ: Method

Answer post-processing:
- Convert letters to lower case
- Convert numbers to digits
- Remove punctuation/articles
- Spell-check and correct
- ...

Measuring crowd agreement:
- Exact string match (9 from 10 answers)
  - VQ 1: the crowd disagrees
  - VQ 2: the crowd agrees

VQ 1

How many birds can be seen?

Answers
(1) 2  
(2) 1  
(3) 2  
(4) 2  
(5) 1  
(6) 2  
(7) 1  
(8) 1  
(9) 2  
(10) 2

VQ 2

What food is on the plate?

Answers
(1) pizza  
(2) pizza  
(3) pizza  
(4) pizza  
(5) pizza  
(6) pizza  
(7) pizza  
(8) pizza  
(9) pizza  
(10) pizza
Analyzing If A Crowd’s Answers Differ: Results

Answers differ for 47% of the nearly 500,000 visual questions!

3 datasets:

<table>
<thead>
<tr>
<th># VQAs (%)</th>
<th>VizWiz</th>
<th>VQA - Real Images</th>
<th>VQA - Abstract Scenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>At Most One Disagreement</td>
<td>20%</td>
<td>52%</td>
<td>58%</td>
</tr>
</tbody>
</table>

[VizWiz](#) [VQA - Real Images](#) [VQA - Abstract Scenes](#)
Predicting Answer Diversity: Ground Truth

**Answer post-processing:**
- Convert letters to lower case
- Convert numbers to digits
- Remove punctuation/articles
- Spell-check and correct
- ...

**Measuring answer diversity:**
- Entropy measure (exact string match)

\[ E = \sum_{i=1}^{N} -p_i \log p_i \]
Predicting Answer Diversity: Ground Truth

Low Entropy

(1) yes
(2) yes
(3) yes
(4) yes
(5) yes
(6) yes
(7) yes
(8) yes
(9) yes
(10) yes

High Entropy

(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
Predicting Answer Diversity: Approaches

1. Linear Regressor:
   - # salient objects automatically detected
   - Question length
   - First two words of question
   - # Instances of 15 lexical categories
     - e.g., noun, adjectives, verb tense
   - GoogleNet features

2. Deep Learning (DL) classifier:
   - LSTM + CNN (architecture employed by Antol et al. [ICCV 2015])
## Predicting Answer Diversity: Experiments

<table>
<thead>
<tr>
<th></th>
<th>VQA - Real Images</th>
<th>VizWiz</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RC</td>
<td>CC</td>
</tr>
<tr>
<td><strong>Status Quo</strong></td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

**Notes:**

- **RC** = rank coefficient; **CC** = correlation coefficient; **MAE** = mean absolute error

---

Status quo: randomly predict answer entropy
Predicting Answer Diversity: Experiments

<table>
<thead>
<tr>
<th></th>
<th>VQA - Real Images</th>
<th></th>
<th>VizWiz</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RC</td>
<td>CC</td>
<td>MAE</td>
</tr>
<tr>
<td>Status Quo</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.3</td>
</tr>
<tr>
<td>CrowdVerge (Gurari and Grauman 2017)</td>
<td>0.46</td>
<td>0.46</td>
<td>0.38</td>
</tr>
</tbody>
</table>

RC = rank coefficient; CC = correlation coefficient; MAE = mean absolute error

Confidence in prediction of whether a VQ will lead to (dis)agreement
Predicting Answer Diversity: Experiments

<table>
<thead>
<tr>
<th></th>
<th>VQA - Real Images</th>
<th>VizWiz</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RC</td>
<td>CC</td>
</tr>
<tr>
<td>Status Quo</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>CrowdVerge (Gurari and Grauman 2017)</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>Ours: Deep Learning</td>
<td>0.53</td>
<td>0.59</td>
</tr>
<tr>
<td>Ours: Linear Regression</td>
<td>0.63</td>
<td>0.63</td>
</tr>
</tbody>
</table>

RC = rank coefficient; CC = correlation coefficient; MAE = mean absolute error

Our methods improve upon the status quo CC by at least 0.33-0.52!
## Predicting Answer Diversity: Experiments

Most of the predictive power comes from language-based features. (See paper for quantitative analysis)

<table>
<thead>
<tr>
<th></th>
<th>VQA - Real Images</th>
<th>VizWiz</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RC</td>
<td>CC</td>
</tr>
<tr>
<td>Status Quo</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>CrowdVerge (Gurari and Grauman 2017)</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>Ours: Deep Learning</td>
<td>0.53</td>
<td>0.59</td>
</tr>
<tr>
<td>Ours: Linear Regression</td>
<td><strong>0.63</strong></td>
<td><strong>0.63</strong></td>
</tr>
</tbody>
</table>

RC = rank coefficient; CC = correlation coefficient; MAE = mean absolute error

250,000 training VQs, 214,354 test VQs
6,408 training VQs, 1,601 test VQs
### Application: Efficiently Collect Answer Diversity

<table>
<thead>
<tr>
<th>Is my monitor on?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) yes</td>
</tr>
<tr>
<td>(2) yes</td>
</tr>
<tr>
<td>(3) yes</td>
</tr>
<tr>
<td>(4) yes</td>
</tr>
<tr>
<td>(5) yes</td>
</tr>
<tr>
<td>(6) yes</td>
</tr>
<tr>
<td>(7) yes</td>
</tr>
<tr>
<td>(8) yes</td>
</tr>
<tr>
<td>(9) yes</td>
</tr>
<tr>
<td>(10) yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Does this picture look scary?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) yes</td>
</tr>
<tr>
<td>(2) no</td>
</tr>
<tr>
<td>(3) no</td>
</tr>
<tr>
<td>(4) yes</td>
</tr>
<tr>
<td>(5) no</td>
</tr>
<tr>
<td>(6) yes</td>
</tr>
<tr>
<td>(7) yes</td>
</tr>
<tr>
<td>(8) no</td>
</tr>
<tr>
<td>(9) no</td>
</tr>
<tr>
<td>(10) no</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What is in this can?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) unanswerable</td>
</tr>
<tr>
<td>(2) unanswerable</td>
</tr>
<tr>
<td>(3) unanswerable</td>
</tr>
<tr>
<td>(4) tuna</td>
</tr>
<tr>
<td>(5) cat food</td>
</tr>
<tr>
<td>(6) cat food</td>
</tr>
<tr>
<td>(7) unanswerable</td>
</tr>
<tr>
<td>(8) unanswerable</td>
</tr>
<tr>
<td>(9) unanswerable</td>
</tr>
<tr>
<td>(10) pet food</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What is this?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) clothesline</td>
</tr>
<tr>
<td>(2) game net</td>
</tr>
<tr>
<td>(3) backyard</td>
</tr>
<tr>
<td>(4) clothes line pole</td>
</tr>
<tr>
<td>(5) unanswerable</td>
</tr>
<tr>
<td>(6) tiki torch</td>
</tr>
<tr>
<td>(7) garden</td>
</tr>
<tr>
<td>(8) backyard</td>
</tr>
<tr>
<td>(9) compound</td>
</tr>
<tr>
<td>(10) backyard barbecue area</td>
</tr>
</tbody>
</table>
Application: Efficiently Collect Answer Diversity

Redundancy
Wasteful

Redundancy Can Be Valuable

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) yes
(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

Redundancy Can Be Valuable
Application: Efficiently Collect Answer Diversity

Given a batch of VQs, collect answer diversity efficiently:

1. Predict answer entropy for every VQ

2. Solve optimization problem to maximize number of unique answers collected for all VQs given a human budget:

\[
x = \arg\max_x \sum_{k=1}^{n} E_k x_k^1 + E_k x_k^2 + E_k x_k^3 + \ldots + E_k x_k^q \]

Subject to:

\[
c^T x \leq B, \quad x_k^1 + x_k^2 + x_k^3 + \ldots + x_k^q = 1, \forall k = 1, 2, \ldots, n,
\]

\[
x_k^1, x_k^2, x_k^3, \ldots, x_k^q \in \{0, 1\}, \forall k = 1, 2, \ldots, n.
\]
Application: Efficiently Collect Answer Diversity
Application: Efficiently Collect Answer Diversity

CrowdVerge accelerates capturing diversity by 23% compared to VizWiz system.
Application: Efficiently Collect Answer Diversity

Which translates to eliminating over 11 40-hour work weeks and saving $1800!
This Talk

1: Designing VQA algorithms to meet the needs of blind people.

2: Learning to what extent multiple people will answer a VQ differently.

3: Learning why multiple people will answer a VQ differently.
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
Analyzing Why Answers Differ: Taxonomy

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?

Answer Missing

What is in this can?

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?
Analyzing Why Answers Differ: Datasets

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?
Analyzing Why Answers Differ: User Interface

**Question (by blind or sighted person):** Could you please tell me if this is regular coffee or decaffeinated?

**Answers (by 10 different people):**
- decaffeinated
- unsuitable
- unanswerable
- regular
- unanswerable
- unsuitable
- unanswerable
- unsuitable
- unanswerable
- decaffeinated

The answers differ because there are problem(s) with the:

<table>
<thead>
<tr>
<th>Visual Question</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOW QUALITY: image is too small;</td>
<td>SYNONYMS: answers present the same idea, but using different words</td>
</tr>
<tr>
<td>out of focus, of poor quality,</td>
<td>having similar meaning (e.g. “Round” vs. “Circular”)</td>
</tr>
<tr>
<td>or nothing is visible</td>
<td></td>
</tr>
<tr>
<td>ANSWER NOT PRESENT / GUESSWORK:</td>
<td>AMBIGUOUS: good image, and valid question, but taken together they have</td>
</tr>
<tr>
<td>good image, but answer of the</td>
<td>more than one valid interpretation (leading to multiple answers)</td>
</tr>
<tr>
<td>question is not evident from the</td>
<td></td>
</tr>
<tr>
<td>image; so people made guesses</td>
<td></td>
</tr>
<tr>
<td>INVALID: there is no proper</td>
<td>GRANULAR: answers present the same idea, but in different levels of</td>
</tr>
<tr>
<td>question (asks for individual</td>
<td>detail / specialization (e.g. “Plane” vs. “Boeing”)</td>
</tr>
<tr>
<td>opinion (e.g. assessing beauty)</td>
<td></td>
</tr>
<tr>
<td>DIFFICULT: requires domain</td>
<td>SPAM: answers are irrelevant, meaningless, or invalid</td>
</tr>
<tr>
<td>expertise, special skills, or</td>
<td></td>
</tr>
<tr>
<td>immense effort</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analyzing Why Answers Differ: VQA Results

Most common reasons: Ambiguity, synonyms, and granularity

Least common reasons: Other, Spam, & Low Quality Image
Analyzing Why Answers Differ: VizWiz Results

Most common reason: 
*Ambiguity, synonyms, and granularity*

Least common reasons: 
*Spam, Difficulty, & Other*
Predicting Why Answers Differ: Approach

Question
(+ optionally 10 answers)

Image
(variable size)

VGG16
(pretrained on ImageNet)
O/p Dim = 4096

GloVe Word Embedding
Dim=100

LSTM
Dim = 256

Fully Connected Layers
[256 → 128]

Output
(10 multi-labels)
Predicting Why Answers Differ: Experiment

Greatest predictive power: Ambiguity, Granularity, & Synonyms

Lowest predictive power: Other, Low Quality Image, and Spam

(aligns with frequency of these types in the data)
This Talk

1: Designing VQA algorithms to meet the needs of blind people.

2: Learning to what extent multiple people will answer a VQ differently.

3: Learning why multiple people will answer a VQ differently.
Faculty Job Opportunities

https://www.ischool.utexas.edu/facultysearch

• Full Professor
• Tenured Faculty Members
• Assistant Professor
Summary

New dataset and challenge: WORKSHOP ON FRIDAY SEP 14!
- Over 31,000 VQs from blind people.

Novel problems and algorithms:
- Algorithms predict if a crowd will agree on answer and answer diversity.
- Algorithms predict why a crowd will disagree.

Efficient answer collection system:
- 23% acceleration to collect all answers from a crowd.

Answer Diversity Prediction and Collection Team:
Gurari & Grauman, CHI 2017; Yang et. al, HCOMP 2018