DATASET: https://vizwiz.org/

Why Does a Visual Question Have Different Answers?

Nilavra Bhattacharya [1], Qing Li [2], Danna Gurari [1]

[1] The University of Texas at Austin, [2] University of California, Los Angeles

DATASET: https://vizwiz.org/

Motivation: Different People Can Offer Different Answers to a Visual Question

Is my monitor on?

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

Does this picture look scary?

Answers	
(1) yes	(6) yes
(2) no	(7) yes
(3) no	(8) no
(4) yes	(9) no
(5) no	(10) no

- Approximately a 50/50 split for when a single versus multiple answers are observed for ~500,000 visual questions in three datasets [1]

Goal: Understand Why Answers Differ

- Benefits include empowering system designers and users with knowledge in how to **interpret**, **prevent**, and **resolve** answer differences

- More generally, support developing algorithmic frameworks that can account for the diversity of perspectives in a crowd

References

[1] D. Gurari and K. Grauman. CrowdVerge: Predicting if People Will Agree on the Answer to a Visual Question. CHI 2017.
[2] D. Gurari et al. VizWiz Grand Challenge: Answer Visual Questions from Blind People. CVPR 2018.
[3] Y. Goyal et al. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017
[4] A. Mahendru et al. The Promise of Premise: Harnessing the Question Premises in Visual Question Answering. EMNLP 2017.

Answer-Differences Dataset

Taxonomy of nine reasons for why answers can differ:

What's this?
LOW QUALITY IMAGE

Answers
1. unsuitable
2. unsuitable
3. unanswerable
4. unsuitable
5. unsuitable
6. unsuitable
7. unsuitable
8. nothing
9. jar
10. black

What type of plant is this that is growing in my garden?
DIFFICULT

Answers
1. unanswerable
2. unanswerable
3. green
4. unanswerable
5. milk weed
6. unanswerable
7. princess tree
8. possibly strawberries
9. bell vine
10. tomato

When does this expire?
SYNONYMS

Answers
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

How many people are eating this hotdog?
ANSWER NOT PRESENT / GUESSWORK

Answers
1. 2
2. 0
3. 2
4. 0
5. 2
6. 3
7. 0
8. 2
9. 2
10. 2

What is on the bed?
AMBIGUOUS

Answers
1. pillows
2. pillows
3. pillows
4. pillows
5. blanket
6. pillows
7. pillows
8. blanket
9. sheets
10. pillows

What book is this?
GRANULAR

Answers
1. maxiaids
2. maxiaids
3. maxi aids
4. maxi aids catalog
5. macias catalog
6. maxiaids products for independent living
7. maxiaids
8. maxiaids
9. maxi aids catalog
10. maxi aids 2012 catalog

I just wanted to say thank you for your assistance.
INVALID

Answers
1. unanswerable
2. unanswerable
3. your welcome
4. unanswerable
5. thank you
6. unanswerable
7. you welcome
8. unanswerable
9. no problem
10. thanks

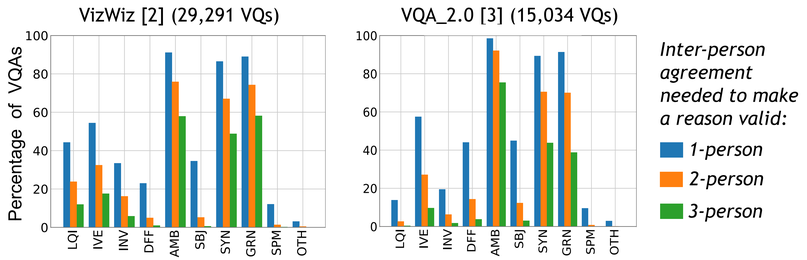
Could the floor use a mopping?
SUBJECTIVE

Answers
1. yes
2. yes
3. yes
4. yes
5. no
6. no
7. no
8. yes
9. no
10. yes

What kind of spice is this?
SPAM

Answers
1. chili powder
2. ewfwfe
3. chili powder
4. chili powder
5. chili powder
6. chili powder
7. chili powder
8. chili powder
9. chili powder
10. chili powder

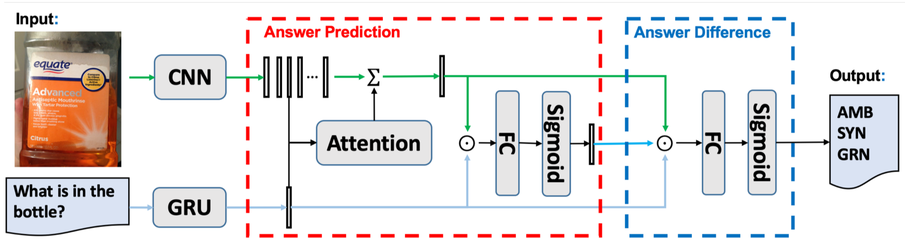
Frequency of each reason leading to answer differences:



Most common reasons: ambiguity, synonymous answers, & varying levels of answer granularity
Least common reason: spam (~1% of VQs)

Novel Task and Algorithm: Given a Visual Question, Predict the Reason(s) Why Answers from a Crowd Will Differ

Algorithm architecture: predicts with image, question, & predicted answers



Evaluation: mean average precision for all reasons

	Model	Overall
VizWiz	Random	30.15
	QI-Relevance [4]	31.71
	VQ answerability [2]	35.31
	I	40.54
	Q	40.5
	Q+I	45.73
	Q+I+A	**50.02**
	Q+I+A_FT	50.01
	Q+I+A_GT	50.68

	Model	Overall
VQA 2.0	Random	30.24
	QI-Relevance [4]	32.23
	I	31.88
	Q	43.47
	Q+I	43.16
	Q+I+A	**44.55**
	Q+I+A_FT	44.46
	Q+I+A_GT	44.09

Random: the best a user can achieve today

QI-Relevance: when VQ is irrelevant, then LQI, IVE, & AMB are reasons

VQ answerability: when VQ is unanswerable, then LQI, IVE, & AMB are reasons

FT: fine-tuned network

GT: uses ground truth instead of predicted answers

Algorithms can anticipate why a crowd will offer different answers, outperforming related baselines by >12 percentage points! The models **perform best** for ambiguity, synonymous answers, & varying answer granularity and **perform worst** for subjective & difficult VQs