VizWiz Grand Challenge: Answering Visual Questions from Blind People

Danna Gurari[1], Qing Li[2], Abigale J. Stangl[3], Anhong Guo[4], Chi Lin[1], Jiebo Luo[5], Kristen Grauman[1], Jeffrey P. Bigham[4]

[1] University of Texas at Austin, [2] University of Science and Technology of China, [3] University of Colorado Boulder, [4] Carnegie Mellon University, [5] University of Rochester

Goal: Answer Visual Questions from Blind People

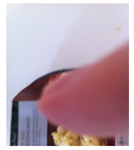
Q: Does this foundation have any sunscreen?
A: Yes

Q: What is this?
A: 10 euros

Q: What color is this?
A: green

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

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



Q: What is this?
A: unanswerable

Q: Who is this mail for?
A: unanswerable

Q: When is the expiration date?
A: unanswerable

Prior work: 11,045 people used the VizWiz mobile phone application to ask crowd workers 72,205 visual questions (Bigham et al. UIST 2010)

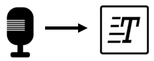
Idea: teach machine to automatically answer the visual questions

VizWiz vs existing VQA: first dataset to originate from blind people and to capture interests of real users of a VQA system in natural settings

Key Contribution #1: VizWiz Dataset Creation

1. Anonymization

Transcription (remove voice)



Re-save image (remove metadata)

2. Filtering (14,796 VQs removed)

Filter	# of VQs
Question Missing	7,477
Crowd Workers	4,626
In-House Experts	2,693
- Personally-Identifying Information	895
- Location	377
- Indecent Content	55
- Suspicious Complex Scene	725
- Suspicious Low Quality Image	578
- Other	63

3. Answer Collection

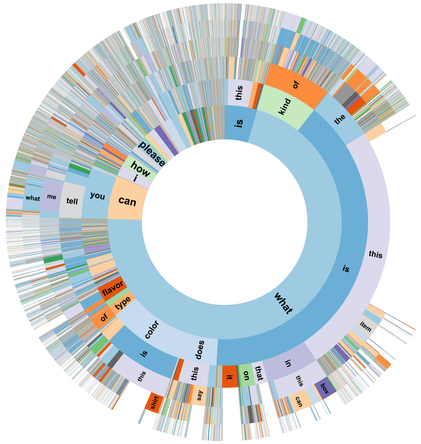
10 answers per visual question

Dataset has 31,173 image/question pairs and 311,730 answers.

Key Contribution #2: VizWiz Dataset Analysis

VizWiz is unique because (1) questions are spoken and so can be more conversational or have audio recording errors, (2) images are captured by blind people and so often are poor quality, and (3) many visual questions are not answerable.

Question Diversity

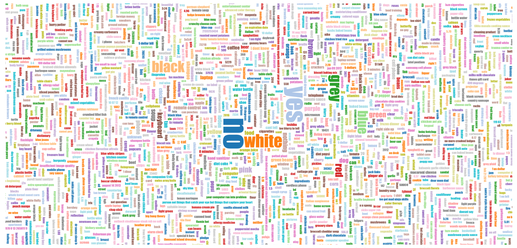


Most common question: "What is this?"

Image Diversity

(average image excluding "unanswerable" visual questions)

Answer Diversity



(excludes "unanswerable" and "unsuitable image")

~29% of visual questions are unanswerable

Dataset	Which Images?	Who Asked?	How Asked?
DAQUAR	NYU Depth V2	In-house participants, Automatically generated (templates)	———
VQA v1.0: Abstract	Abstract Scenes	Crowd workers (AMT)	Typed
VQA v1.0: Real	MSCOCO	Crowd workers (AMT)	Typed
Visual Madlibs	MSCOCO	Automatically generated (templates)	———
FM-IQA	MSCOCO	Crowd workers (Baidu)	Typed
KB-VQA	MSCOCO	In-house participants	Typed
COCO-QA	MSCOCO	Automatically generated (captions)	———
VQA v2.0: Real	MSCOCO	Crowd workers (AMT)	Typed
Visual7W	MSCOCO	Crowd workers (AMT)	Typed
CLEVR	Synthetic Shapes	Automatically generated (templates)	———
SHAPES	Synthetic Shapes	Automatically generated (templates)	———
Visual Genome	MSCOCO & YFCC100M	Crowd workers (AMT)	Typed
FVQA	MSCOCO & ImageNet	In-house participants	Typed
TDIUC	MSCOCO & YFCC100M	Crowd workers (AMT), In-house participants, Automatically generated	Typed
Ours - VizWiz	Blind people use mobile phones to take a picture and ask question		Spoken

Key Contribution #3: Algorithm Benchmarking

Task 1: Predict Answer to a Visual Question

Evaluation Metric: $\text{accuracy} = \min(\frac{\text{number of humans that provided that answer}}{3}, 1)$

[1] Goyal et al. CVPR '17, [2] Kazemi et al. arXiv '17, [3] Anderson et al. CVPR '18

	All	Yes/No	Number	Unanswerable	Other
Q+I [1]	0.137	0.598	0.045	0.070	0.142
Q+I+A [2]	0.145	0.605	0.068	0.071	0.155
Q+I+BUA [3]	0.134	0.582	0.071	0.060	0.143
Train on VizWiz [1]	0.465	0.597	**0.262**	**0.805**	0.264
Train on VizWiz [2]	0.469	0.608	0.218	0.802	0.274
Train on VizWiz [3]	0.469	0.596	0.210	**0.805**	0.273
Fine-Tuning [1]	0.466	0.675	0.220	0.781	0.275
Fine-Tuning [2]	0.469	**0.681**	0.213	0.770	0.287
Fine-Tuning [3]	**0.475**	0.669	0.220	0.776	**0.294**

Training from scratch (rows 4-6) and fine-tuning (rows 7-9) yield significant improvements over relying on existing algorithms as is.

Task 2: Predict if a Visual Question Can Be Answered

[4] Mahendru et al. EMNLP '17

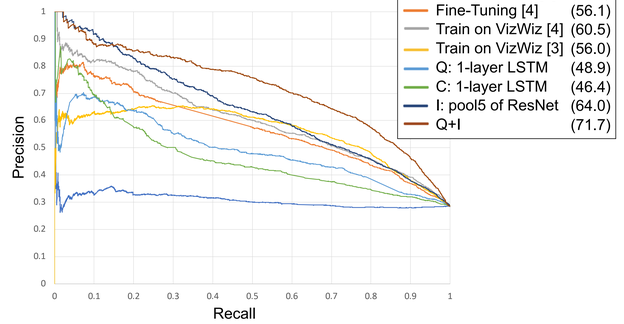


Image information provides the greatest predictive power (i.e., AP = 64) and is solidly improved by adding the question information (i.e., AP = 71.7).

VizWiz is a difficult dataset for modern vision algorithms.

Dataset

http://vizwiz.org/data

ECCV 2018 Challenge

http://vizwiz.org/workshop