AdVQA: Human-Adversarial Visual Question Answering

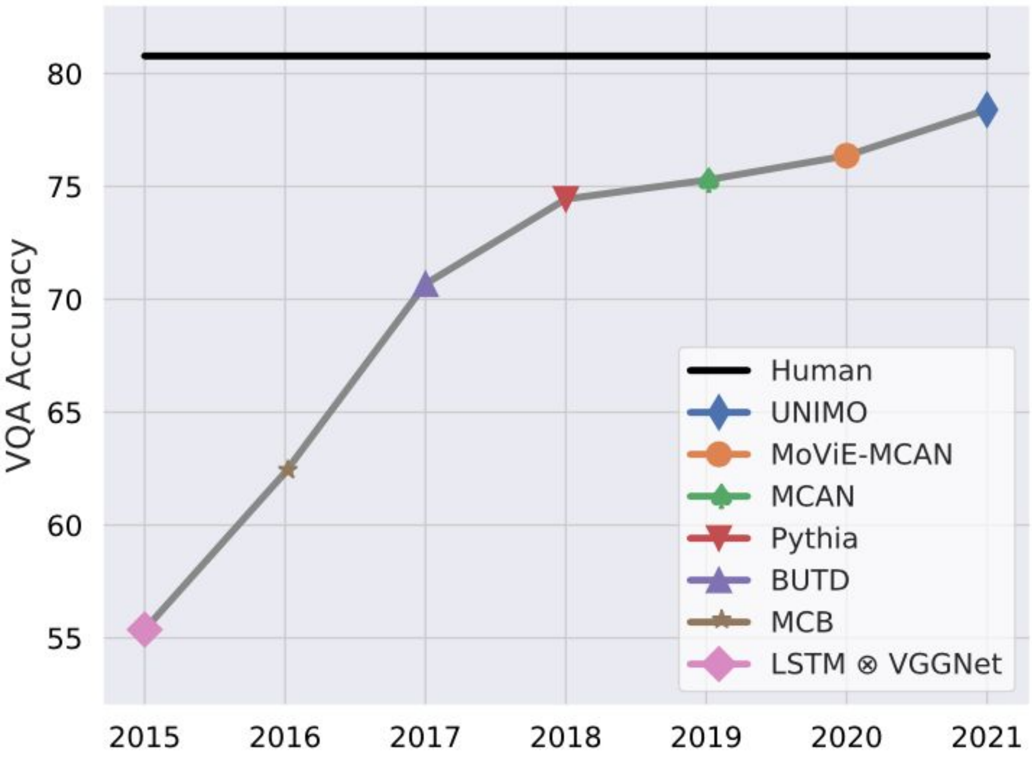
Sasha Sheng*, Amanpreet Singh*, Vedanuj Goswami, Jose Alberto Lopez Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, Douwe Kiela

Facebook AI Research (FAIR)

FACEBOOK AI

* Equal contribution.

Motivation



VQA Accuracy
80
75
70
65
60
55
2015
2016
2017
2018
2019
2020
2021
Human
UNIMO
MoViE-MCAN
MCAN
Pythia
BUTD
MCB
LSTM ⊗ VGGNet

AdVQA Contributions

- AdVQA dataset is considerably **harder** and **trickier** due to its adversarial data collection design. We dynamically collect (SOTA) model fooling questions given the images.
- We evaluate a wide range of existing VQA models on AdVQA and find their performance is significantly lower than on the commonly used VQA2.0 dataset.
- AdVQA dataset can be used not only to shed light on current model shortcomings, but also as an evaluative benchmark to help advance the robustness of the models in the field of Visual Question Answering.
- Prediction file evaluation and model evaluation server available on dynabench, including a public leaderboard.

Check out www.adversarialvqa.org to download our dataset

Examples

Example 1. contrastive examples from VQA and AdVQA

VQA question:
How many cats are in the image?
- **Correct Answer: 2**
- Answer (VisualBERT): 2
- Answer (ViLBERT): 2
- Answer (UniT): 2

AdVQA question:
What brand is the tv?
- **Correct Answer: LG**
- Answer (VisualBERT): sony
- Answer (ViLBERT): samsung
- Answer (UniT): samsung

Images are from COCO2017 - val

Examples

VQA question:
Does the cat look happy?
- **Correct Answer: no**
- Answer (VisualBERT): no
- Answer (ViLBERT): no
- Answer (UniT): no

AdVQA question:
How many cartoon drawings are present on the cat's tie?
- **Correct Answer: 4**
- Answer (VisualBERT): 1
- Answer (ViLBERT): 1
- Answer (UniT): 2

Example 2. contrastive examples from VQA and AdVQA

Examples

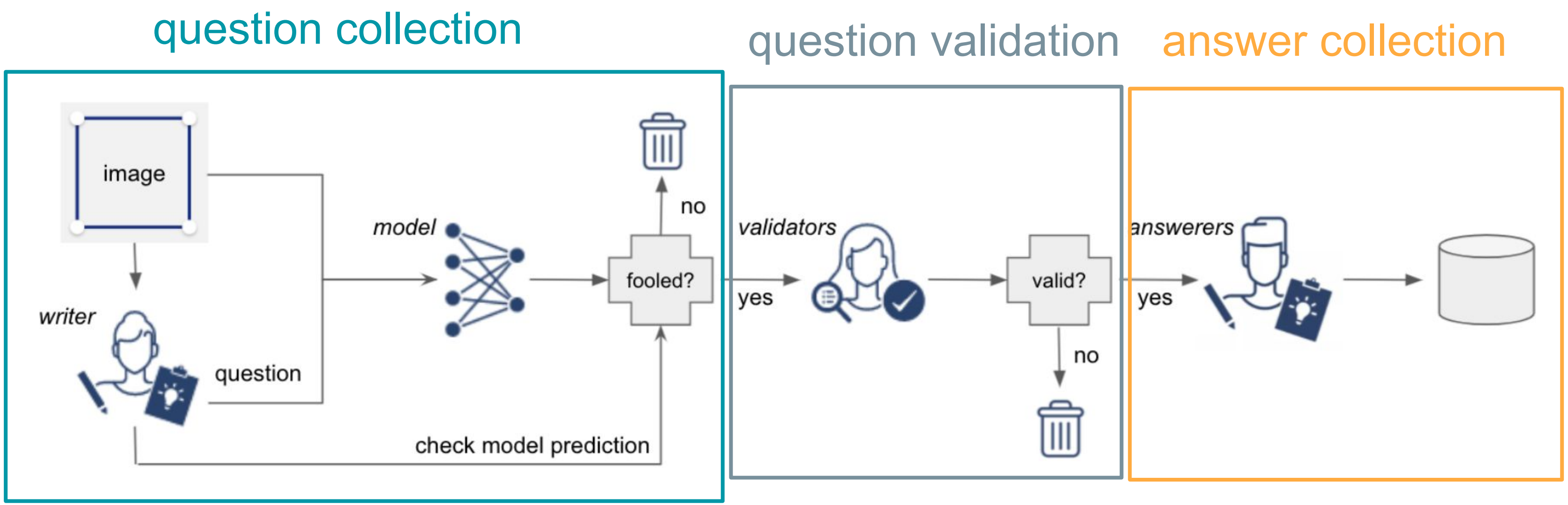
VQA question:
What kind of floor is the man sitting on?
- **Correct Answer: wood**
- Answer (VisualBERT): wood
- Answer (ViLBERT): wood
- Answer (UniT): wood

AdVQA question:
Did someone else take this picture?
- **Correct Answer: no**
- Answer (VisualBERT): yes
- Answer (ViLBERT): yes
- Answer (UniT): yes

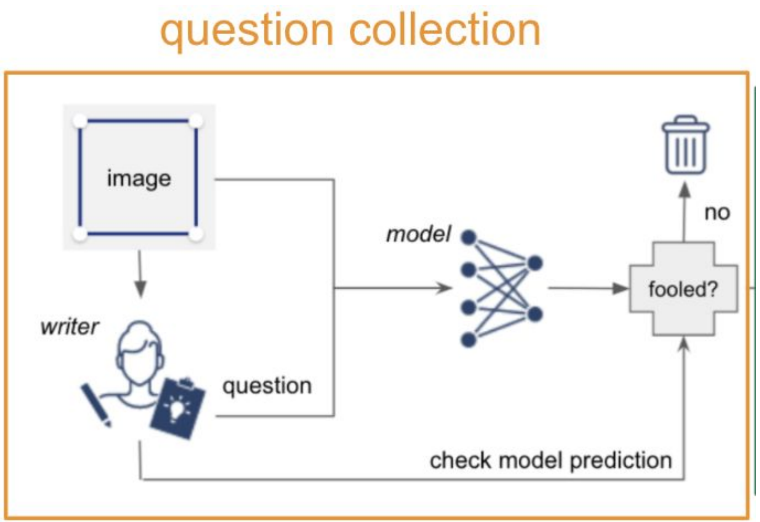
Example 3. contrastive examples from VQA and AdVQA

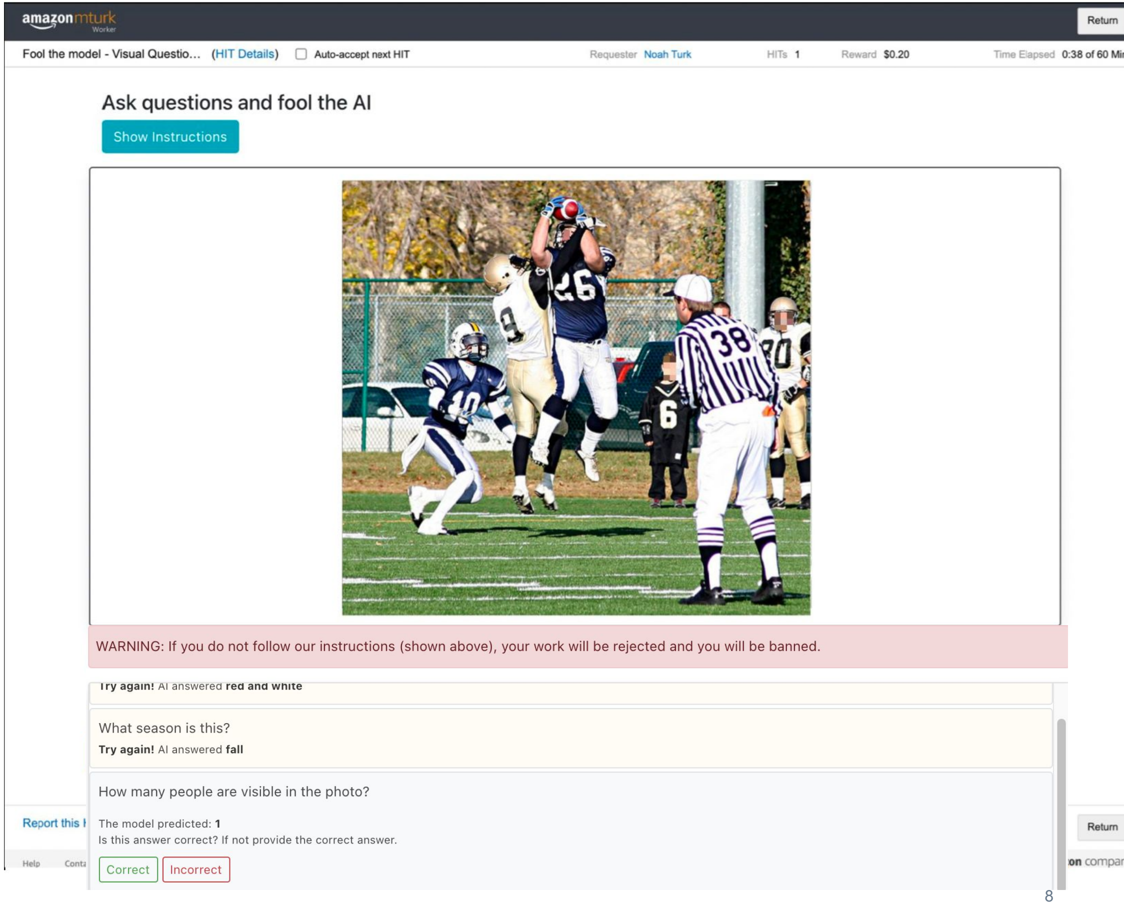
Our Approach

- Dynamic Human in-the-loop adversarial data collection against SOTA model → VQA2.0 winner in 2020: MoViE+MCAN
- 3 stage process:

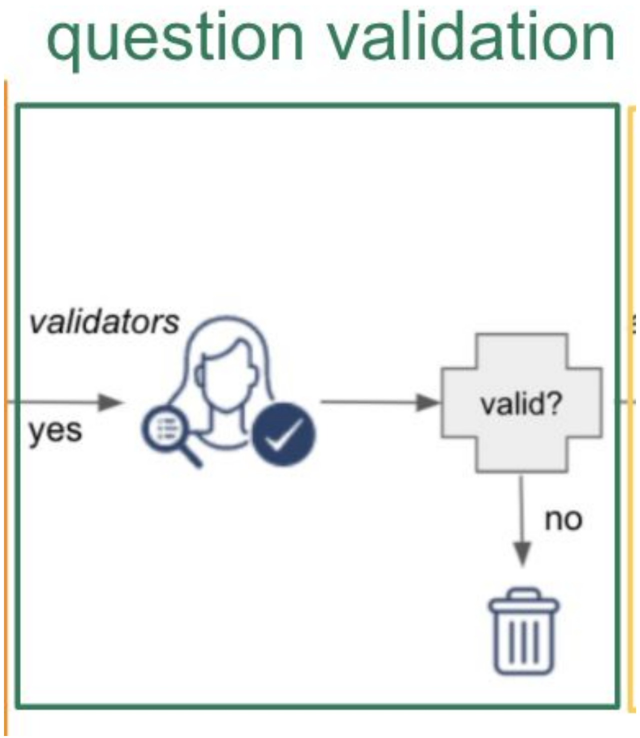


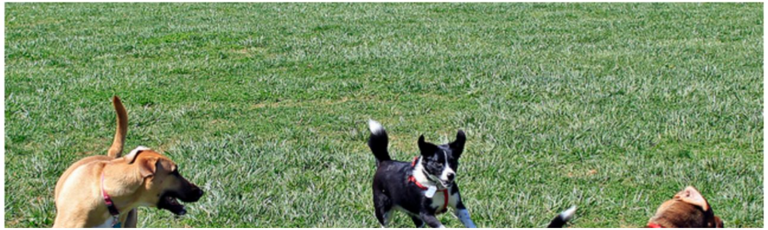
Question Collection





Question Validation





A question is considered **invalid** if any of the following conditions is met:

- **The question does not require the image to answer.**
  ie., What is the capital of the USA?
- **The answer is not commonly known to other people.**
  ie., "What is the name of this plant?" when very few people would be able to recognize the plant and know the name.
- **The question is not based on the scene depicted in the image or the answer could not be provided correctly based on the image.**
  ie., "What is the brand of the soap?" when the brand name of the soap is only partially visible from the image.
  ie., "What is the woman doing?" when there is no woman in the image.

WARNING: If you do not follow our instructions (shown above), your work will be rejected and you will be banned.

IS THE QUESTION BELOW VALID? (SEE INSTRUCTIONS ABOVE TO SEE WHAT WE MEAN BY "VALID")

How many dogs are there?

ACTIONS

- (•) Valid
- ( ) Invalid
- ( ) Flag

DETERMINE IF THE ANSWER IS CORRECT:

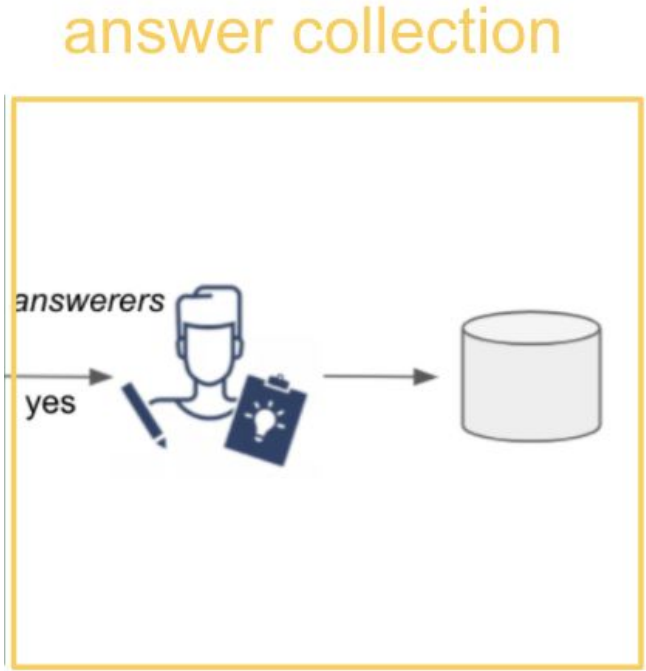
3

ACTIONS

- ( ) Correct
- (•) Incorrect

Submit Validations: 0 / 10.

Answer Collection



- Same interface as the VQA v2.0 answer collection interface
- Collect 10 answers per question
- Added “unanswerable” to:
  - filter bad questions that might have passed through
  - account for ambiguity that can be present in questions
- Filter out bad annotators with gold labels

AdVQA Dataset

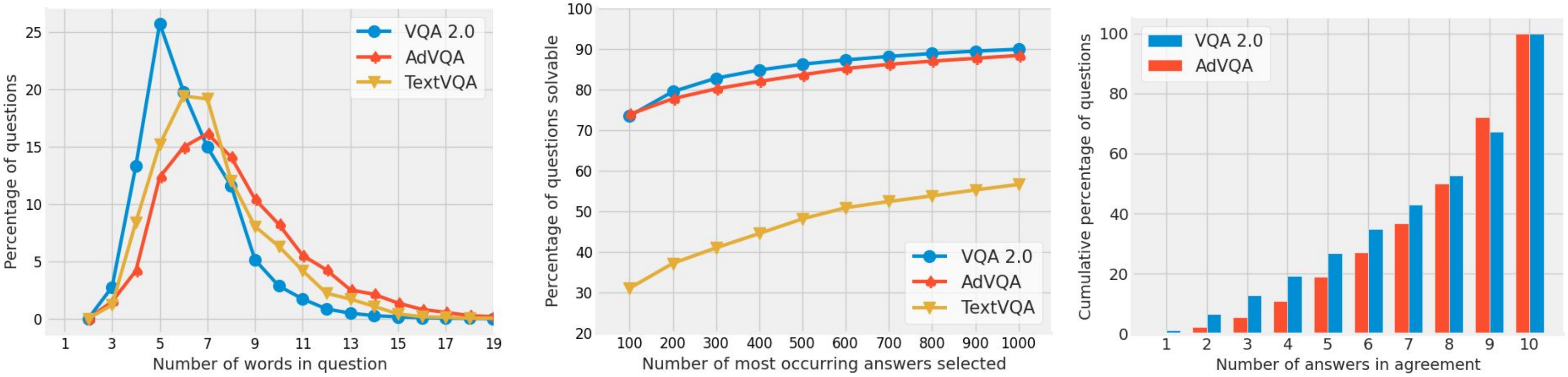
- 46,807 <*question, answers*> pairs (36,087 test, 10k val)
- The data split matches VQA2.0; The 10k val set is released publicly.

Table 2: **AdVQA human-adversarial question collection and dataset statistics.** The model error rate is the percentage of examples where the submitted questions fooled the model (either as claimed during question collection, or after validation). We also report the number of attempts (tries) needed before a validated model-fooling example was found, and how long this took, in seconds.

Total	**Model error rate** claimed	validated	**Tries**	**Time in sec**
			mean/median per ex.	
208,932	40.94% (85,537)	36.17% (75,571)	5.33/4.0	203.22/107.26

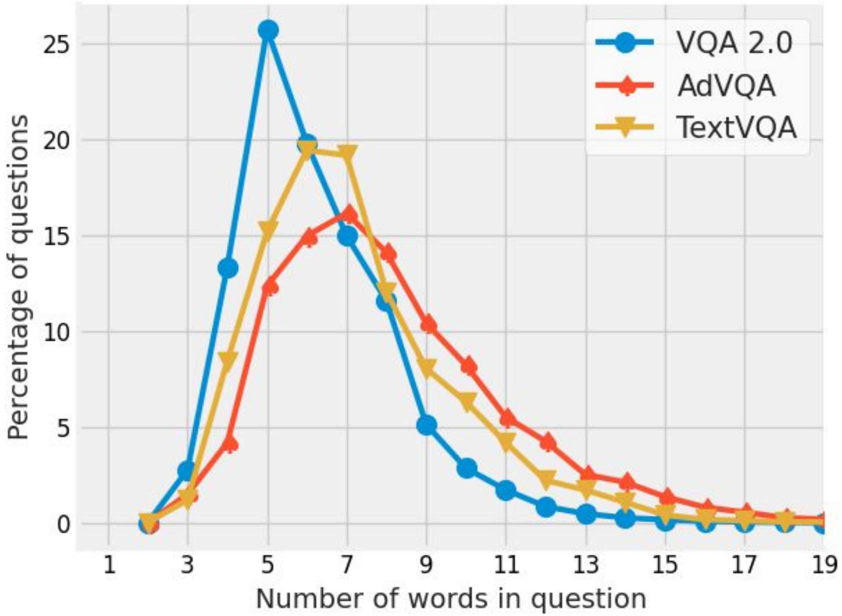
AdVQA Dataset - cont’

- The models will need to understand and reason with rare concepts to do well on AdVQA since 50.9% of the answers in AdVQA val and test sets do not occur in VQA v2 train set.
- 77.2% of the AdVQA val set’s questions are answerable using the original VQA vocab.



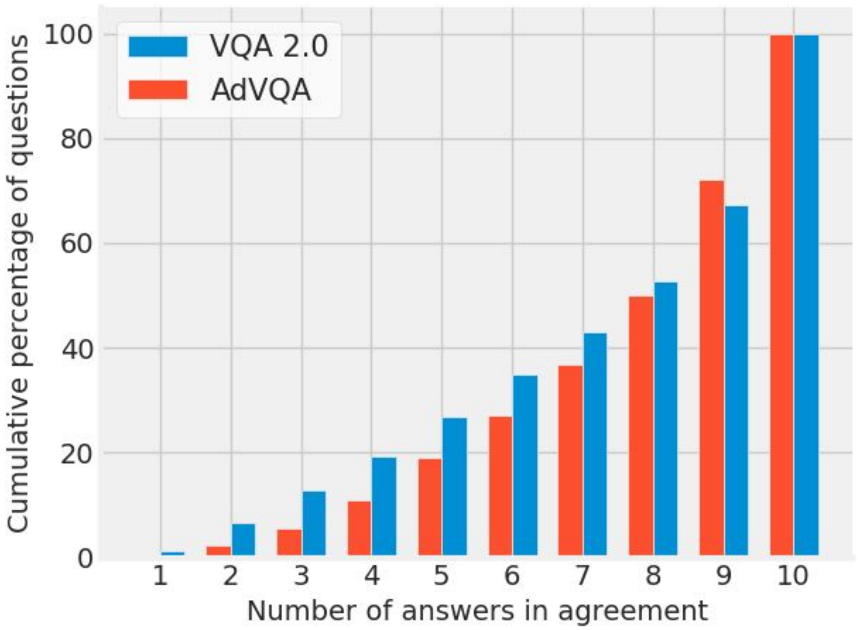
Quantitative statistics for AdVQA val set questions and answers

AdVQA Dataset - cont’



Number of word distribution in questions in AdVQA compared to prior work

AdVQA Dataset - cont’



cumulative percentage of questions w.r.t number of answers in agreement

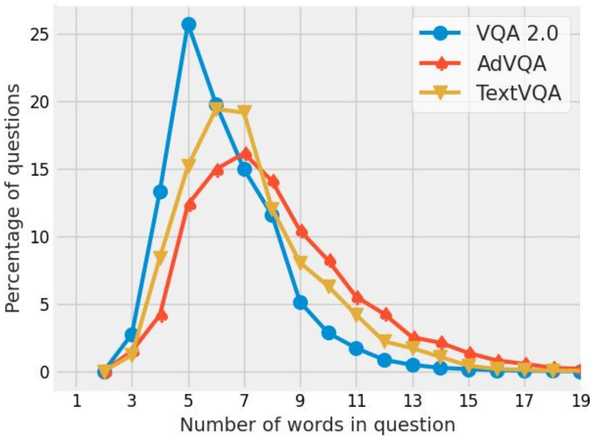
AdVQA Dataset - cont'

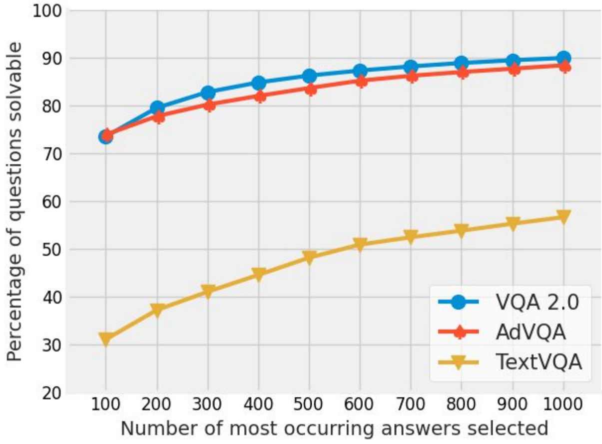
Question Type	VQA test-dev	AdVQA test	VQA val	AdVQA val
yes/no	38.36	23.22	37.70	24.58
number	12.31	35.73	11.48	32.44
others	49.33	41.05	50.82	42.98

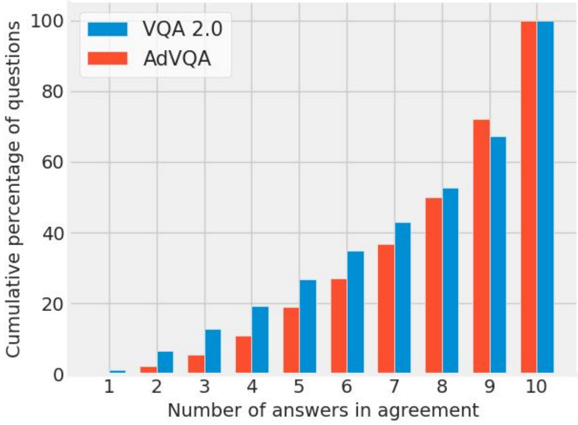
Answer Type Category-wise Distribution

AdVQA Dataset - cont'

- The models will need to understand and reason with rare concepts to do well on AdVQA since 50.9% of the answers in AdVQA val and test sets do not occur in VQA v2 train set.
- 77.2% of the AdVQA val set's questions are answerable using the original VQA vocab.







Quantitative statistics for AdVQA val set questions and answers

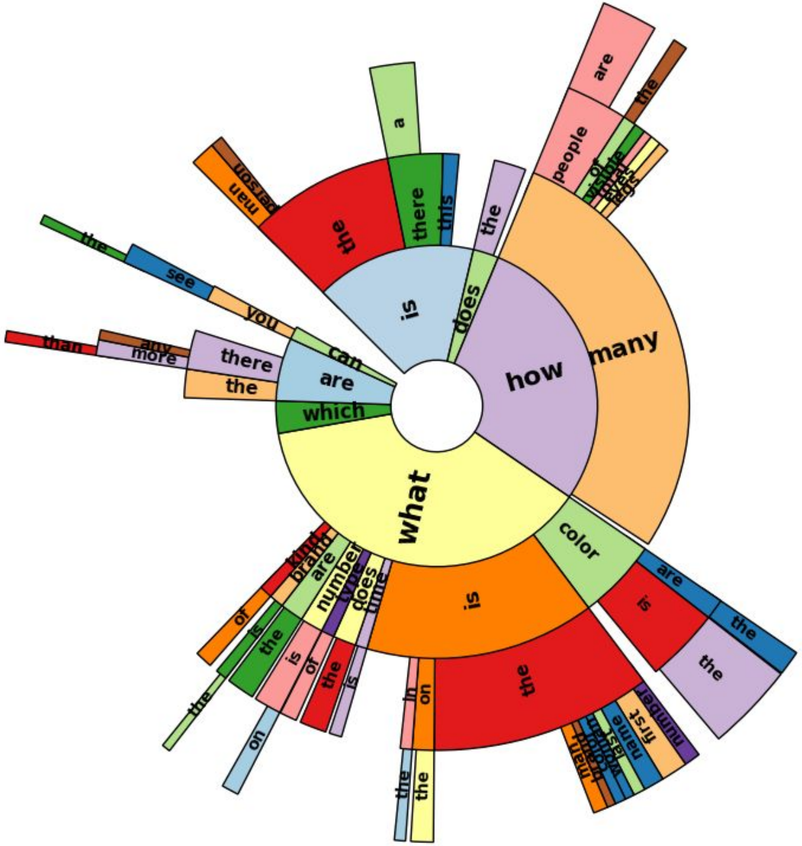


Figure 5: **Sunburst distribution for the first four words in the AdVQA val set questions.** Most questions start with “what” or “how”.

“What” and “how” is the most common
A significant amount of questions can be answered using scene text and counting

Mo

Table 3: **Model performance on VQA v2 and AdVQA** val and test sets. * indicates that this model architecture (but not this model instance) was used in the data collection loop.

Model		**VQA** test-dev	**AdVQA** test	**VQA** val	**AdVQA** val
Human performance		80.78	89.01	84.73	88.46
Majority answer (overall)		-	16.79	24.67	16.98
Majority answer (per answer type)		-	31.86	31.01	33.38
Model in loop	MoViE+MCAN [42]	73.58	13.89	73.51	14.08
Unimodal	ResNet-152 [20]	26.66	20.59	24.85	19.02
	BERT [13]	43.59	30.24	43.71	31.89
Multimodal (unimodal pretrain)	MoViE+MCAN* [42]	69.81	30.02	69.77	31.31
	MMBT [28]	49.27	30.80	49.36	32.57
Multimodal (multimodal pretrain)	VisualBERT [33]	70.40	31.96	69.98	28.09
	ViLBERT [39]	59.45	32.01	59.78	33.67
	ViLT [30]	62.30	31.00	62.33	32.48
	UNITER_{Base} [10]	70.67	27.56	69.30	29.44
	UNITER_{Large} [10]	73.58	29.66	72.82	32.08
	VILLA_{Base} [16]	71.17	27.55	69.87	29.36
	VILLA_{Large} [16]	72.02	28.59	71.1	30.58
Multimodal (unimodal pretrain + OCR)	M4C (TextVQA+STVQA) [23]	32.89	28.86	31.44	29.08
	M4C (VQA v2 train set) [23]	67.66	33.52	66.21	33.33

AdVQA Evaluation Discussions #1

Most multimodal models are unable to beat simple baselines.

AdVQA is difficult. VQA models still have a long way to go to beat those simple baselines.

Table 3: **Model performance on VQA v2 and AdVQA** val and test sets. * indicates that this model architecture (but not this model instance) was used in the data collection loop.

Model		VQA test-dev	AdVQA test	VQA val	AdVQA val
Human performance		80.78	89.01	84.73	88.46
Majority answer (overall)		-	16.79	24.67	16.98
Majority answer (per answer type)		-	31.86	31.01	33.38
Model in loop	MoViE+MCAN [42]	73.58	13.89	73.51	14.08
Unimodal	ResNet-152 [20]	26.66	20.59	24.85	19.02
	BERT [13]	43.59	30.24	43.71	31.89
Multimodal (unimodal pretrain)	MoViE+MCAN* [42]	69.81	30.02	69.77	31.31
	MMBT [28]	49.27	30.80	49.36	32.57
Multimodal (multimodal pretrain)	VisualBERT [33]	70.40	31.96	69.98	28.09
	ViLBERT [39]	59.45	32.01	59.78	33.67
	ViLT [30]	62.30	31.00	62.33	32.48
	UNITER$_{Base}$ [10]	70.67	27.56	69.30	29.44
	UNITER$_{Large}$ [10]	73.58	29.66	72.82	32.08
	VILLA$_{Base}$ [16]	71.17	27.55	69.87	29.36
	VILLA$_{Large}$ [16]	72.02	28.59	71.1	30.58
Multimodal (unimodal pretrain + OCR)	M4C (TextVQA+STVQA) [23]	32.89	28.86	31.44	29.08
	M4C (VQA v2 train set) [23]	67.66	33.52	66.21	33.33

AdVQA Evaluation Discussions #2

MovieMCAN which was not in the loop but trained with a different seed perform similarly to other models.

All VQA models perform poorly on AdVQA, suggesting the examples are by and large representative of shortcomings of VQA techniques overall.

Table 3: **Model performance on VQA v2 and AdVQA** val and test sets. * indicates that this model architecture (but not this model instance) was used in the data collection loop.

Model		VQA test-dev	AdVQA test	VQA val	AdVQA val
Human performance		80.78	89.01	84.73	88.46
Majority answer (overall)		-	16.79	24.67	16.98
Majority answer (per answer type)		-	31.86	31.01	33.38
Model in loop	MoViE+MCAN [42]	73.58	13.89	73.51	14.08
Unimodal	ResNet-152 [20]	26.66	20.59	24.85	19.02
	BERT [13]	43.59	30.24	43.71	31.89
Multimodal (unimodal pretrain)	MoViE+MCAN* [42]	69.81	30.02	69.77	31.31
	MMBT [28]	49.27	30.80	49.36	32.57
Multimodal (multimodal pretrain)	VisualBERT [33]	70.40	31.96	69.98	28.09
	ViLBERT [39]	59.45	32.01	59.78	33.67
	ViLT [30]	62.30	31.00	62.33	32.48
	UNITER_{Base} [10]	70.67	27.56	69.30	29.44
	UNITER_{Large} [10]	73.58	29.66	72.82	32.08
	VILLA_{Base} [16]	71.17	27.55	69.87	29.36
	VILLA_{Large} [16]	72.02	28.59	71.1	30.58
Multimodal (unimodal pretrain + OCR)	M4C (TextVQA+STVQA) [23]	32.89	28.86	31.44	29.08
	M4C (VQA v2 train set) [23]	67.66	33.52	66.21	33.33

AdVQA Evaluation Discussions #3

M4C performs best, whereas VILLA performs worst among the evaluated models.

The ability to read and reason about text is important for AdVQA. Human adversarial examples don't do well on models that are trained on statistically adversarial examples.

Table 3: **Model performance on VQA v2 and AdVQA** val and test sets. * indicates that this model architecture (but not this model instance) was used in the data collection loop.

Model		VQA test-dev	AdVQA test	VQA val	AdVQA val
Human performance		80.78	89.01	84.73	88.46
Majority answer (overall)		-	16.79	24.67	16.98
Majority answer (per answer type)		-	31.86	31.01	33.38
Model in loop	MoViE+MCAN [42]	73.58	13.89	73.51	14.08
Unimodal	ResNet-152 [20]	26.66	20.59	24.85	19.02
	BERT [13]	43.59	30.24	43.71	31.89
Multimodal (unimodal pretrain)	MoViE+MCAN* [42]	69.81	30.02	69.77	31.31
	MMBT [28]	49.27	30.80	49.36	32.57
Multimodal (multimodal pretrain)	VisualBERT [33]	70.40	31.96	69.98	28.09
	ViLBERT [39]	59.45	32.01	59.78	33.67
	ViLT [30]	62.30	31.00	62.33	32.48
	UNITER$_{Base}$ [10]	70.67	27.56	69.30	29.44
	UNITER$_{Large}$ [10]	73.58	29.66	72.82	32.08
	VILLA$_{Base}$ [16]	71.17	27.55	69.87	29.36
	VILLA$_{Large}$ [16]	72.02	28.59	71.1	30.58
Multimodal (unimodal pretrain + OCR)	M4C (TextVQA+STVQA) [23]	32.89	28.86	31.44	29.08
	M4C (VQA v2 train set) [23]	67.66	33.52	66.21	33.33

Summary - AdVQA

- AdVQA dataset is considerably **harder** and **trickier** due to its adversarial data collection design.
- We evaluate a wide range of existing VQA models on AdVQA and report their significantly lower performance than on the commonly used VQA2.0 dataset.
- AdVQA dataset demonstrates the shortcomings of popular VQA models, and it will be used as evaluative benchmark to help advance the state of the art.
- Prediction file evaluation and model evaluation server available on dynabench, including a public leaderboard.

Check out www.adversarialvqa.org to download our dataset

Thank You

Powered by the Dynabench framework