Human-Adversarial Visual Question Answering

Anonymous Author(s)
Affiliation
Address
email

Abstract

Performance on the most commonly used Visual Question Answering dataset (VQA v2) is starting to approach human accuracy. However, in interacting with state-of-the-art VQA models, it is clear that the problem is far from being solved. In order to stress test VQA models, we benchmark them against human-adversarial examples. Human subjects interact with a state-of-the-art VQA model, and for each image in the dataset, attempt to find a question where the model's predicted answer is incorrect. We find that a wide range of state-of-the-art models perform poorly when evaluated on these examples. We conduct an extensive analysis of the collected adversarial examples and provide guidance on future research directions. We hope that this Adversarial VQA (AdVQA) benchmark can help drive progress in the field and advance the state of the art.[1]

1 Introduction

Visual question answering (VQA) is widely recognized as an important evaluation task for vision and language research. Besides direct applications such as helping the visually impaired or multimodal content understanding on the web, it offers a mechanism for probing machine understanding of images via natural language queries. Making progress on VQA requires bringing together different subfields in AI – combining advances from natural language processing (NLP) and computer vision together with those in multimodal fusion – making it an exciting task in AI research.

Over the years, the performance of VQA models has started to plateau on the popular VQA v2 dataset [18] – approaching inter-human agreement – as evidenced by Fig. 1. This raises important questions for the field: To what extent have we solved the problem? If we haven't, what are we still missing? How good are we really?

An intriguing method for investigating these questions is dynamic data collection, where human annotators and state-of-the-art models are put "in the loop" together to collect data adversarially. Annotators are tasked with and rewarded for finding model-fooling examples, which are then verified by other humans. The easier it is to find such examples, the worse the model's performance can be said to be. The collected data can be used to "stress test" current VQA models and can serve as the next iteration of the VQA benchmark that can help drive further progress.

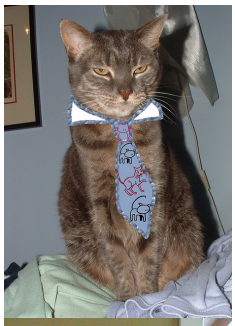
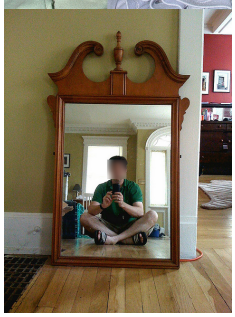
The commonly used VQA dataset [18] dataset was collected by instructing annotators to "ask a question about this scene that [a] smart robot probably can not answer" [3]. One way of thinking about our proposed human-adversarial data collection is that it explicitly ensures that the questions can not be answered by today's "smartest" models.

This work is, to the best of our knowledge, the first to apply this human-adversarial approach to any multimodal problem. We introduce Adversarial VQA (AdVQA), a large evaluation dataset of 28,522 examples in total, all of which fooled the VQA 2020 challenge winner, MoViE+MCAN [40] model.

[1] We plan to make an evaluation server available for community to evaluate on AdVQA

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Table 1: **Contrastive examples from VQA and AdVQA**. Predictions are given for the VisualBERT, ViLBERT and UniT models, respectively. Models can answer VQA questions accurately, but consistently fail on AdVQA questions.

Image	VQA	AdVQA
	Q: How many cats are in the image? **A**: 2 **Model**: 2, 2, 2	**Q**: What brand is the tv? **A**: lg **Model**: sony, samsung, samsung
	**Q**: Does the cat look happy? **A**: no **Model**: no, no, no	**Q**: How many cartoon drawings are present on the cat's tie? **A**: 4 **Model**: 1, 1, 2
	**Q**: What kind of floor is the man sitting on? **A**: wood **Model**: wood, wood, wood	**Q**: Did someone else take this picture? **A**: no **Model**: yes, yes, yes

We evaluate a wide range of existing VQA models on AdVQA and find that their performance is significantly lower than on the commonly used VQA v2 dataset [18] (see Table 1). Furthermore, we conduct an extensive analysis of AdVQA characteristics, and contrast with the VQA v2 dataset.

We hope that this new benchmark can help advance the state of the art by shedding important light on the current model shortcomings. Our findings suggest that there is still considerable room for continued improvement, with much more work remaining to be done.

2 Related Work

Stress testing VQA. Several attempts exist for stress testing VQA models. Some examine to what extent VQA models' predictions are grounded in the image content, as opposed to them relying primarily on language biases learned from the training dataset. The widely used VQA v2 dataset [18] was one attempt at this. The dataset contains pairs of similar images that have different answers to the same question, rendering a language-based prior inadequate. VQA under changing priors (VQA-CP) [1] is a more stringent test where the linguistic prior not only is weak in the test set, but is adversarial relative to the training dataset (*i.e.*, answers that are popular for a question type during training are rarer in the test set and vice versa). Datasets also exist that benchmark specific capabilities in VQA models, such as reading and reasoning about text in images [49], leveraging external knowledge [39], and spatial [24] or compositional reasoning [27]. Other vision and language datasets, such as Hateful Memes [30], have also tried to make sure the task involves true multimodal reasoning, as opposed to any of the individual modalities sufficing for arriving at the correct label.

Saturating prior work. The VQA v2 [18] challenge has been running yearly since 2016 and has seen tremendous progress, as can be seen in Figure 1. From simple LSTM [21] and VGGNet [46] fused models to more advanced fusion techniques (MCB [15], Pythia [25]) to better object detectors (MCAN [53], MoViE+MCAN [40], BUTD [2]) to transformers [51, 48] and self-supervised pretraining (MCAN [53], UNIMO [34]), the community has brought the performance of models on the VQA v2 dataset close to human accuracy, thus starting to saturate progress on dataset. As we show through AdVQA, however, saturation is far from achieved on the overall task, and hence there is a need for new datasets to continue benchmarking progress in vision and language reasoning.

Adversarial datasets. As AI models are starting to work well enough in some narrow contexts for real world deployment, there are increasing concerns about the robustness of these models. This has led to fertile research both in designing adversarial examples to "attack" models (e.g., minor imperceptible noise added to image pixels that significantly changes the model's prediction [17]), as well as in approaches to make models more robust to "defend" against these attacks [8, 11]. The human-in-the-loop adversarial setting that we consider in this paper is qualitatively different from the statistical perturbations typically explored when creating adversarial examples. This type of human-in-the-loop adversarial data collection have been explored in NLP e.g., natural language inference [41], question answering [4, 44], sentiment analysis [43], hate speech detection [52] and dialogue safety [14]. To the best of our knowledge, ours is the first work to explore a multimodal human-adversarial benchmark.

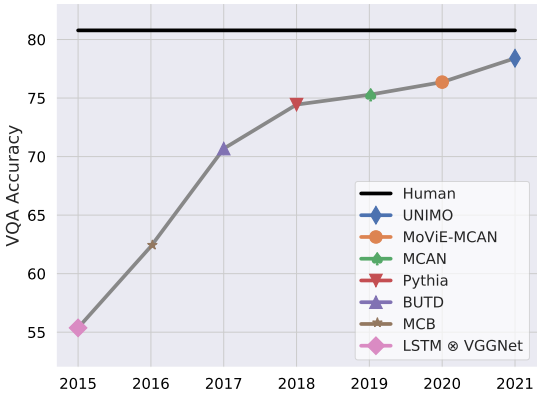


Figure 1: Progress on the VQA v2 dataset over time.

3 AdVQA

The aim of this work is to investigate state of the art VQA model performance via human-adversarial data collection. In this human-and-model-in-the-loop paradigm, human annotators are tasked with finding examples that fool the model. In this case, annotators are shown an image and are tasked with asking difficult but valid questions that the model answers incorrectly. We collected our dataset using Dynabench [28], a platform for dynamic adversarial data collection and benchmarking. For details on our labeling user interface, please refer to the supplementary material.

The VQA v2 dataset is based on COCO [35] images. We collected adversarial questions for subsets of the val2017 COCO images (4,095) and test COCO images (22,259), respectively. The data collection involves three phrases (Figure 2): First, in the **question collection** phase, Amazon Mechanical Turk workers interact with a state-of-the-art VQA model and self-report image-question pairs that the model failed to answer correctly. Second, in the **question validation** phase, a new set of workers validates whether the answers provided by the VQA model for the image-question pairs collected in the first phrase are indeed incorrect. Finally, in the **answer collection** phase, we collect 10 ground truth answers for the image-question pairs validated in the second phase. In what follows, we provide further details for each of these steps.

3.1 Question Collection

In this phase, annotators on Amazon MTurk are shown an image and are tasked with asking questions about this image. The interface has a state-of-the-art VQA model in the loop. For each question the annotator asks, the model produces an answer. Annotators are asked to come up with questions that fool the model. Annotators are done (*i.e.*, can submit the HIT) once they successfully fooled the model or after they tried a minimum of 10 questions (whichever occurs first). To account for the fact that it may be hard to think of many questions for some images, annotators are given the option to skip an image after providing non-fooling three questions. We use the VQA Challenge 2020 winner –

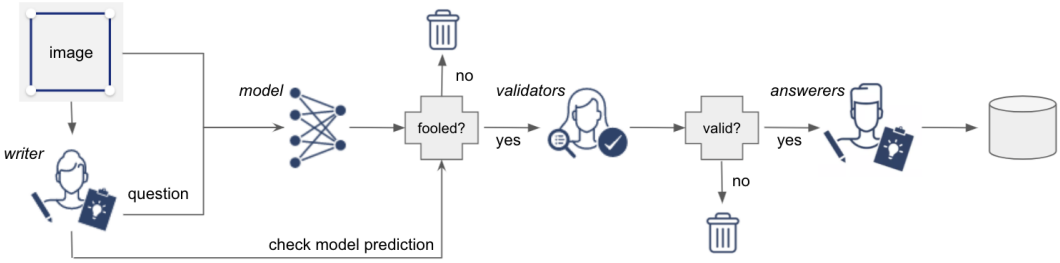
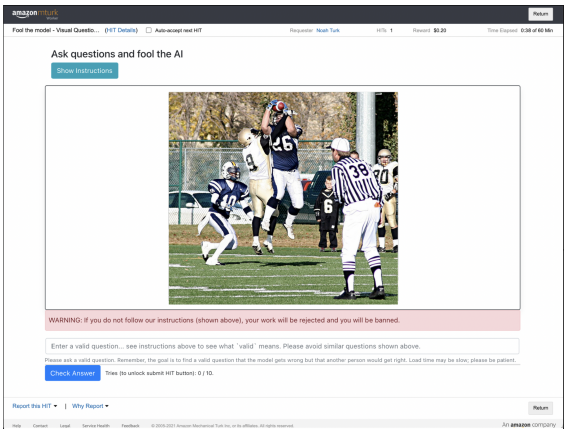


Figure 2: Diagram of the AdVQA human-adversarial data collection flow.



(a) **Main interface**. The annotators follow the instructions provided to write the question in the text field provided which is then answered by the model-in-the-loop.

(b) **Magnifier,** available both in question collection, validation and answer collection helps annotators ask and answer more specific questions.

Figure 3: **Question collection interface** showing the first stage of AdVQA collection setup.

MoViE+MCAN [40] – trained on the COCO 2017 train set as the model in the loop. This is to ensure that the adversarial benchmark we collect is challenging for the current state-of-the-art in VQA.

The interface provides a digital magnifier to zoom into an area of the image when the mouse hovers over it, allowing workers to examine the image closely if necessary. See Figure 3b. Keyboard shortcuts were provided for convenience.

3.2 Question Validation

Note that in the question collection phase, annotators self-report when they have identified a question that is incorrectly answered by the model. Whether or not this was actually the case is verified in the question validation phase. Two different annotators are shown the image, question, and answer predicted by the model, and asked whether the model's answer is correct. As an additional quality control, the annotator is also asked whether the question is "valid". A valid question is one where the image is necessary and sufficient to answer the question. Examples of invalid questions are: "What is the capital of USA?" (does not need an image) or "What is the person doing?" when the image does not contain any people or where there are multiple people doing multiple things. If the two validators disagree, a third annotator is used to break the tie. Examples are added to the dataset only if at least two annotators agree that the question is valid.

3.3 Answer Collection

In the final stage of data collection, following [18, 49], we collect 10 answers per question providing instructions similar to those used in [18] making sure that no annotator sees the same question twice. In addition, as an extra step of caution, to filter bad questions that might have passed through last two stages and to account for ambiguity that can be present in questions, we allow annotators to

select *"unanswerable"* as an answer. Further, to ensure superior data quality, we occasionally provide annotators with hand-crafted questions for which we know the non-ambiguous single true answer, as a means to identify and filter out annotators providing poor quality responses.

3.4 Human-Adversarial Annotation Statistics

The statistics for the first two stages (question collection and validation) are shown in Table 2. We find that annotators took about 5 tries and on average around 4 minutes to find a model-fooling example. For computing the model error rate, we can look at the instances where annotators claimed they had fooled the model, and where annotators were verified by other annotators to have fooled the model. The latter has been argued to be a particularly good metric for measuring model performance [28]. We also further confirm that the model was indeed fooled by running the model-in-the-loop on a subset of examples in which human agreement was 100% and found the accuracy to be 0.08% on val, and 0.1% on test splits of AdVQA respectively.

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**		**Tries**	**Time in sec**
	claimed	validated	mean/median per ex.	
75,669	43.71% (33,074)	37.69% (28,522)	5.25/4.0	226.17/166.66

Interestingly, the "claimed" model error rate, based on self-reported model-fooling questions, is similar to that of text-based adversarial question answering, which was at 44.0% for RoBERTa and 47.1% for BERT [4]. The validated error rate for our task is much higher than e.g. for ANLI [41], which was 9.52% overall, suggesting that fooling models is a lot easier for VQA than for NLI. It appears to be not too difficult for annotators to find examples that the model fails to predict correctly.

4 Model Evaluation

4.1 Baselines and Methods

We analyze the performance of several baselines and a wide variety of state-of-the-art VQA models on the AdVQA dataset. We evaluate the same set of models on VQA v2 dataset as a direct comparison.

Prior baselines. We start by evaluating two prior baselines: answering based on the overall majority answer, or the per answer type majority answer in the validation dataset. We use the same answer types as in [3]. The overall majority answer in AdVQA is *no*. The majority answer is *no* for "yes/no", *2* for "numbers" and *unanswerable* for "others". See Section 5 for more details on answer types.

Unimodal baselines. Next, we evaluate two unimodal pretrained models: i) ResNet-152 [20] pretrained on Imagenet [12]; and ii) BERT [13], both finetuned on the task using the visual (image) or textual (question) modality respectively, while ignoring the other modality. We observe that the unimodal text model performs better than unimodal image for both VQA and AdVQA.

Multimodal methods. We evaluate two varieties of multimodal models: i) unimodal pretrained and ii) multimodal pretrained. In the unimodal pretrained category we explore MMBT [29], MoViE+MCAN [40] and UniT [22]. These models are initialized from unimodal pretrained weights: BERT pretraining for MMBT; Imagenet + Visual Genome [32] detection pretraining for MoViE+MCAN; and Imagenet + COCO [36] detection pretraining for the image encoder part and BERT pretraining for the text encoder part in UniT. In the multimodal pretrained category, we explore VisualBERT [33], VilBERT [37, 38], VilT [31], UNITER[10] and VILLA [16]. These models are first initialized from pretrained unimodal models and then pretrained on different multimodal datasets on proxy self-supervised/semi-supervised tasks before finetuning on VQA. VisualBERT is pretrained on COCO Captions [9]; VilBERT is pretrained on Conceptual Captions [45]; ViLT, UNITER and VILLA models are pretrained on COCO Captions [9] + Visual Genome [32] + Conceptual Captions [45] +

SBU Captions [42] datasets. We find that multimodal models in general perform better than unimodal models, as we would expect given that both modalities are important for the VQA task.

Multimodal OCR methods. As we will see in Section 5, a significant amount of questions in AdVQA can be answered using scene text. We test a state-of-the-art TextVQA [47] model, M4C [23] on AdVQA. We evaluate two versions: (i) trained on VQA 2.0 dataset, and (ii) trained on TextVQA [47] and STVQA [5]. In both cases, we use OCR tokens extracted using the Rosetta OCR system [6]. We also use the same answer vocabulary used by other models for fair comparison.

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	91.18	84.73	87.53
Majority answer (overall)		-	13.38	24.67	11.65
Majority answer (per answer type)		-	27.39	31.01	29.24
Model in loop	MoViE+MCAN [40]	73.56	10.33	73.51	10.24
Unimodal	ResNet-152 [20]	26.37	10.85	24.82	11.22
	BERT [13]	39.47	26.9	39.40	23.81
Multimodal (unimodal pretrain)	MoViE+MCAN* [40]	71.36	26.64	71.31	26.37
	MMBT [29]	58.00	26.70	57.32	25.78
	UniT [22]	64.36	28.15	64.32	27.55
Multimodal (multimodal pretrain)	VisualBERT [33]	70.37	28.70	70.05	28.03
	ViLBERT [37]	69.42	27.36	69.27	27.36
	ViLT [31]	64.52	27.11	65.43	27.19
	UNITER$_{Base}$ [10]	71.87	25.16	70.50	25.20
	UNITER$_{Large}$ [10]	73.57	26.94	72.71	28.03
	VILLA$_{Base}$ [16]	70.94	25.14	69.50	25.17
	VILLA$_{Large}$ [16]	72.29	25.79	71.40	26.18
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

4.2 Discussion

Our model evaluation results show some surprising findings. We discuss our observations and hypotheses around these findings in this section.

Baseline comparison. Surprisingly, most multimodal models are unable to outperform a unimodal baseline (BERT [13]) and simple majority answer (per answer type) prior baseline which only predicts the most-frequently occurring word for the question's category (*no* for "yes/no", *2* for "numbers" and *unanswerable* for "others"). A detailed breakdown of category-wise performance provided in Table 4 suggests that even though "yes/no" questions are often considered to be easy, the baseline multimodal models we evaluate are unable to beat the majority answer prediction. We also observe varied but close to majority answer performance in the "numbers" category, which is a question type known to be difficult for VQA models [40]. In contrast, the models outperform the majority answer in the "others" category even though "unanswerable" (the majority answer) is not in the answer vocabulary used. Interestingly, M4C outperforms all on "numbers" and "others" categories possibly thanks to its text reading capabilities. These trends showcase the difficulty of AdVQA and suggest that we have a long way to go still, given that we are as yet apparently unable to beat such simple baselines.

Model rankings. First, M4C [23] performs the best among the evaluated models. Interestingly, it is smaller than many of the more sophisticated model architectures that score higher on VQA. This is probably due to the importance of the ability to read and reason about text in the image for answering some AdVQA questions. Second, among models that can't read text, VisualBERT [33] is the best model, despite (or perhaps because of?) it being simple and having fewer parameters compared to other models. Third, the adversarially-trained VILLA [16] model performs surprisingly poorly. While it may be more robust to statistically-generated adversarial examples, it appears to be less

so against human-adversarial examples. Fourth, we find that all of these models perform poorly compared to humans, while model performance on VQA is much closer to that of humans.

Human performance. Another surprising finding is that inter-human agreement is higher on the AdVQA dataset than on VQA. This could be due to different data annotation procedures, requirements on annotators or just statistical noise. Human-adversarial questions may also be more specific due to annotators having to make crisp decisions about the model failing or not.

Table 4: **The category-wise performance of VQA models.** The state-of-the-art VQA models perform very close to the majority class prior, illustrating the challenge and difficulty of AdVQA.

Model	Question Type		
	yes/no	numbers	others
Majority Class	62.28	31.11	9.29
ResNet-152	62.81	0.18	0.51
BERT	67.58	26.87	9.25
VisualBERT	55.51	32.29	17.66
ViLBERT	55.58	29.49	16.67
MoViE+MCAN*	52.74	33.62	14.56
M4C (VQA2)	56.67	38.04	22.73

Model-in-the-loop's performance. Interestingly, the MoViE+MCAN model that was *not* used in the loop and trained with a different seed, performs very similarly to other models. This suggests that to some extent, annotators overfit to the model instance. An alternative explanation is that model selection for all evaluated models was done on the AdVQA validation set, which was (obviously) not possible for the model in the loop used to construct the dataset. In Adversarial NLI [41], the entire model class of the in-the-loop model was affected. Note however that all VQA models perform poorly on AdVQA, suggesting that the examples are by and large representative of shortcomings of VQA techniques overall, and not of an individual model instance or class.

Train vs Test Distribution. We experiment with finetuning VisualBERT and VilBERT further on the AdVQA *val* set, finding that this improved test accuracy from 28.70 to 33.10 and 27.36 to 32.85, respectively, suggesting a difference in the VQA and AdVQA distributions, as we would expect.

4.3 Training Details

We train all the models in Table 3 on the VQA *train* + *val* split excluding the COCO 2017 validation images. We also add questions from VG [32] corresponding to the images that overlap with the VQA training set to our training split. The VQA *val* set in our results contains all questions associated with the images in the COCO 2017 validation split. This is also consistent with the training split used for the model in the loop. We collect our AdVQA validation set on the images from the COCO 2017 *val* split, ensuring there is no overlap with the training set images. We choose the best checkpoint for each model by validating on the AdVQA validation set.

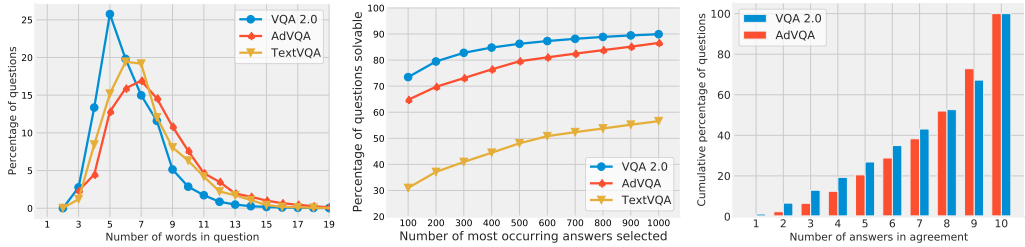
For all our experiments, we use the standard architecture for the models as provided by their authors. For the models that are initialized from pretrained models (whether unimodal or multimodal pretrained), we use off-the-shelf pretrained model weights and then finetune on our training set. We do not do any hyperparamter search for these models and use the best hyperparams as provided by respective authors. We finetune each model with three different seeds and report average accuracy.

We run most of our experiments on NVIDIA V100 GPUs. The maximum number of GPUs used for training is 8 for larger models. Maximum training time is 2 days. More details about hyperparameters, number of devices and time for training each model are provided in the Appendix.

5 Dataset Analysis

Questions. We first analyze the question diversity in AdVQA and compare them with popular VQA datasets. AdVQA contains 28,522 questions (5,123 in val and 23,399 in test) each with 10 answers.

Fig. 4a shows the distribution of question length compared with [18, 49]. The average question length in AdVQA is 7.8, which is higher than the VQA v2 (6.3), and TextVQA (7.2). The workers often need to get creative and more specific to fool the model-in-the-loop, leading to somewhat longer questions (*e.g.* specifying a particular person to ask about). Fig. 6a shows the top 15 most occurring questions from the val set, showcasing that questions involving text (*e.g.* time, sign) and counting (*e.g.*



(a) (b) (c)

Figure 4: **Quantitative statistics for AdVQA val set questions and answers** showing longer question length, more diversity and better human-agreement. (a) Percentage of questions solvable with a particular question length in AdVQA val set. We see that the average question length (7.82) in AdVQA is higher than the prior work. (b) Percentage of questions solvable when a particular number of top k most occurring answers are selected from each dataset. The plot suggests that AdVQA has a much more diverse answer vocabulary compared to [18] but also not quite as challenging as [49]. (c) Cumulative human agreement scores. We see that human agreement is better and higher on AdVQA compared to [18].

how many) are major failures for current state-of-the-art VQA models, corroborating the findings of prior work in [49, 19, 50, 26]. Fig. 5 shows a sunburst plot for the first 4 words in the AdVQA val set questions. We can observe that questions in AdVQA often start with "what" or "how" frequently inquiring about aspects like "many" (count) and "brand" (text).

Answers. In AdVQA val set, only 66.6% (3,856) answers occur more than twice compared to 94.8% in VQA v2 [18], suggesting that the diversity of possible answers is much larger in AdVQA compared to VQA v2. Fig. 4b shows the percentage of questions that are solvable with a vocabulary of top k most occurring answers. We observe that more questions in VQA v2 while fewer question in TextVQA are solvable with smaller vocabulary compare to AdVQA. This suggests that AdVQA is more diverse and difficult than VQA v2 but not as narrowly focused as TextVQA, making it a great testbed for future VQA models. We also showcase more qualitative examples in our supplementary to demonstrate that AdVQA's diversity doesn't lead to unnatural questions. Fig. 4c shows the cumulative percentage of questions where more than a particular number of annotators agree with each other. Fig. 6b shows the top 500 most occurring answers in the AdVQA val set; starting from very common answers such as "no" and counts ("1", "9"), to gradually more specific and targeted answers like "21", "teddy bear" and "center".

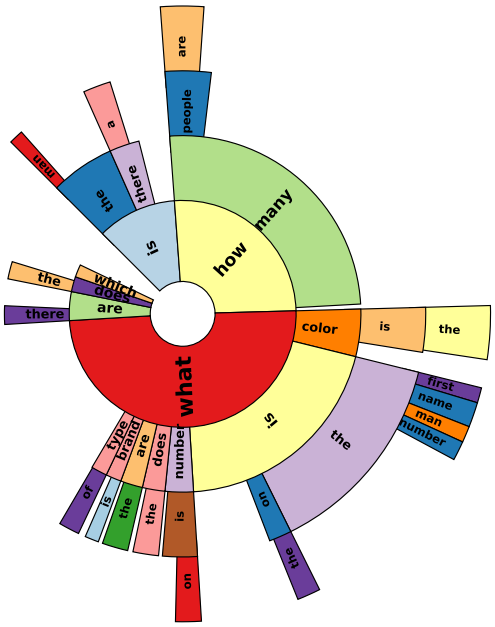
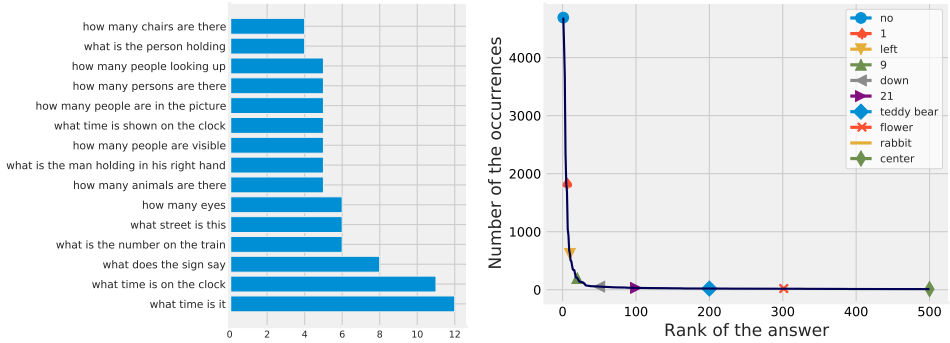


Figure 5: **Sunburst distribution for the first four words in the AdVQA val set questions.** Most questions start with "what" or "how".

Answer Vocabulary. To showcase the challenge of AdVQA, we take the original VQA v2 vocabulary used in the Pythia v0.3 model [47]. We find that 77.2% of the AdVQA val set's questions are answerable using this vocabulary, suggesting that a model with powerful reasoning capability won't be heavily limited by vocabulary on AdVQA. But, we also note that for high performance on AdVQA, a model will need to understand and reason about rare concepts, as 50.9% of the answers in AdVQA val and test sets don't occur in VQA v2 train set.



(a) (b)

Figure 6: **Qualitative statistics for AdVQA val set questions and answers** showing top questions, answers and word distribution. (a) **Top 15 most occurring questions in AdVQA val.** Most of the top questions start with "what". (b) **Total occurrences for 500 most common answers** with markers for particular ranks.

Question Types. Table 5 shows the category-wise distribution for AdVQA questions compared with VQA v2 [18]. We can observe a shift from more easy questions of the "yes/no" category in VQA v2 dataset to more difficult questions in "numbers" category (as suggested in prior work [40]) in AdVQA.

Human Agreement. In the AdVQA val set, 27.2% human annotators agree on all 10 answers while 3 or more annotators agree an answer for 97.7%, which is higher compared to 93.4% on VQA v2 [18], even though as discussed AdVQA contains a large number of rare concepts.

Table 5: **The category-wise distribution of answers.** Compared to VQA, AdVQA contains more "number"based and lesser "yes/no" questions supporting the prior work's observations around failure of VQA models to count and read text.

Question Type	VQA test-dev	AdVQA test	VQA val	AdVQA val
yes/no	38.36	17.89	37.70	17.90
number	12.31	41.91	11.48	31.80
others	49.33	40.20	50.82	50.30

Relationship to TextVQA [47]. To understand if the ability to read text is crucial for AdVQA, we extract Rosetta [6] tokens on the AdVQA val set and determine how many questions can be answered using OCR tokens at an edit distance of 1 to account for OCR errors. We find that 15.4% of the AdVQA val questions are solvable using OCR tokens suggesting that ability to read scene text would play a crucial role in solving AdVQA.

6 Conclusion, Limitations & Outlook

We introduced the AdVQA dataset, a novel human-adversarial multimodal dataset designed for accelerating progress on Visual Question Answering (VQA). Current VQA datasets have started plateauing and are approaching saturation with respect to human performance. In this work, we demonstrate that the problem is far from solved.

In particular, our analysis and model evaluation results suggest that current state-of-the-art models underperform on AdVQA due to a reasoning gap incurred from a combination of (i) inability to read text; (ii) inability to count; (iii) heavy bias towards the VQA v2 question and answer distribution; (iv) external knowledge; (v) rare unseen concepts; and (vi) weak multimodal understanding. We've shown the gap is unlikely to be due to (i) limited answer vocabulary; (ii) language representation (BERT performance compared to other); (iii) no pretraining (UniT); or (iv) lack of adversarial training (VILLA performance). We will provide an evaluation benchmark for community to evaluate on AdVQA and hope that AdVQA will help bridge the gap by serving as a dynamic new benchmark for visual reasoning with a large amount of headroom for further progress in the field.

In future work, it would be interesting to continue AdVQA as a dynamic benchmark. If a new state of the art emerges, those models can be put in the loop to examine how we can improve even further.

Broader Impact

This work analyzed state-of-the-art Visual Question Answering (VQA) models via a dataset constructed using a dynamic human-adversarial approach. We hope that this work can help make VQA models more robust.

VQA datasets contain biases, both in the distribution of images represented in the datasets, as well as the corresponding questions and answers. These biases are likely amplified by the VQA models trained on these datasets. Biases studied in the context of image captioning [7, 54] are also relevant for VQA. English is the only language represented in this work.

VQA models can be useful for aiding visually impaired users. The commonly used VQA datasets are not representative of the needs of visually impaired users – both in terms of the distribution of images (typically consumer photographs from the web), and in terms of the questions contained in the datasets (typically asked by sighted individuals while looking at the image). In contrast, the VizWiz dataset [19] contains questions asked by visually impaired users on images taken by them to accomplish day-to-day tasks. It would be beneficial for the community to collect larger datasets in this context to enable progress towards relevant technology. It would be important to do this with input from all relevant stakeholders, and in a responsible and privacy-preserving manner. VQA models are currently far from being accurate enough to be uesful or safe in these contexts.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering, 2018.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[4] Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020.

[5] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4291–4301, 2019.

[6] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 71–79, 2018.

[7] Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[11] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863. PMLR, 2017.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. *arXiv preprint arXiv:1908.06083*, 2019.

[15] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[16] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020.

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[22] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021.

[23] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002, 2020.

[24] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019.

[25] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[26] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[27] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

[28] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in nlp, 2021.

[29] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

[30] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes, 2021.

[31] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021.

[32] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[33] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[34] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[37] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[38] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[39] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019.

[40] Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. Revisiting modulated convolutions for visual counting and beyond. *arXiv preprint arXiv:2004.11883*, 2020.

[41] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding, 2020.

[42] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011.

[43] Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. Dynasent: A dynamic benchmark for sentiment analysis, 2020.

[44] Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv:1904.04792*, 2019.

[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[47] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`, 2020.

[48] Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. Are we pretraining it right? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*, 2020.

[49] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019.

[50] Alexander Trott, Caiming Xiong, and Richard Socher. Interpretable counting for visual question answering. *arXiv preprint arXiv:1712.08697*, 2017.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[52] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*, 2020.

[53] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.

[54] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the relevant section.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the Broader Impact statement.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 4.3.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.3.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes] For the license of the images derived from the COCO dataset, see https://cocodataset.org/#termsofuse.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include the validation set as supplementary material. The test set will remain hidden.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We used crowdworkers who were remunerated for their services.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] To the best of our knowledge, the dataset does not contain PII or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix A.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We do not foresee any participant risk in this study.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix A.