Adversarial VQA: A New Benchmark for Evaluating the Robustness of VQA Models

adversarialvqa.github.io

Linjie Li[1], Jie Lei[2], Zhe Gan[1], Jingjing Liu[3]
[1]Microsoft [2]UNC Chapel Hill [3]Tsinghua University
{lindsey.li, zhe.gan}@microsoft.com
jielei@cs.unc.edu, JJLiu@air.tsinghua.edu.cn

Abstract

Benefiting from large-scale pre-training, we have witnessed significant performance boost on the popular Visual Question Answering (VQA) task. Despite rapid progress, it remains unclear whether these state-of-the-art (SOTA) models are robust when encountering examples in the wild. To study this, we introduce Adversarial VQA, a new large-scale VQA benchmark, collected iteratively via an adversarial human-and-model-in-the-loop procedure. Through this new benchmark, we discover several interesting findings. (i) Surprisingly, we find that during dataset collection, non-expert annotators can easily attack SOTA VQA models successfully. (ii) Both large-scale pre-trained models and adversarial training methods achieve far worse performance on the new benchmark than over standard VQA v2 dataset, revealing the fragility of these models while demonstrating the effectiveness of our adversarial dataset. (iii) When used for data augmentation, our dataset can effectively boost model performance on other robust VQA benchmarks. We hope our Adversarial VQA dataset can shed new light on robustness study in the community and serve as a valuable benchmark for future work.

1. Introduction

Visual Question Answering (VQA) [4] is a task where given an image and a question about it, the model provides an open-ended answer. A successful VQA system can be applied to real-life scenarios such as a chatbot that assists visually impaired people. In these applications, the VQA models are expected to handle diverse question types from recognition to reasoning, and answer questions faithfully based on the evidence in the image.

While model performance on the popular VQA dataset [14] has been advanced in recent years [4, 19, 3, 50, 9, 43, 54], with better visual representations [18, 54], more sophisticated model designs [12, 27], large-scale pre-training [30, 41, 7, 42, 55] and adversarial training [11], today's VQA models are still far from being robust enough for practical use. There are some works studying the robustness of VQA models, such as their sensitivity to visual content manipulation [1], answer distribution shift [2], linguistic variations in input questions [39], and reasoning capabilities [13, 38]. However, current robust VQA benchmarks mostly suffer from three main limitations: (*i*) designed with heuristic rules [13, 2, 1]; (*ii*) focused on a single type of robustness [38, 39, 13]; (*iii*) based on VQA v2 [14] images (or questions), which state-of-the-art (SOTA) VQA models are trained on [13, 2, 1, 38, 39]. The images [1] or questions [13, 17] are often synthesized, not provided by human.

In addition, previous data collection procedures on VQA benchmarks are often *static*, meaning that the data samples in these datasets do not evolve, and model performance can saturate on the fixed dataset without good generalization. For example, model accuracy on VQA v2 has been improved from 50% [4] to 76% [54] since inception. Similarly, on robust VQA benchmarks, a recent study [28] has

Figure 1: Illustration of data collection examples. The workers try to attack the VQA model for at most 5 times by asking *hard* questions about the image, and succeeds at the last attempt. Green (red) indicates a correct (wrong) answer.

found that pre-trained models can greatly lift state of the art. Yet it remains unclear whether such high performance can be maintained when encountering examples in the wild.

To build an organically evolving benchmark, we introduce Adversarial VQA (AVQA), a new large-scale VQA dataset *dynamically* collected with Human-And-Model-in-the-Loop Enabled Training (HAMLET) [47]. AVQA is built on images from different domains, including web images from Conceptual Captions [40], user-generated images from Fakeddit [32], and movie images from VCR [52]. Our data collection is iterative and can be perpetually going. We first ask human annotators to create examples that current best models cannot answer correctly (Figure 1). These newly annotated examples expose the model's weaknesses, and are added to the training data for training a stronger model. The re-trained model is subjected to the same process, and the collection can iterate for several rounds. After each round, we train a new model and set aside a new test set. In this way, not only is the resultant dataset more challenging than existing benchmarks, but this process also yields a "moving post" target for VQA systems, rather than a static benchmark that will eventually saturate.

With this new benchmark, we present a thorough quantitative evaluation on the robustness of VQA models along multiple dimensions. First, we provide the first study on the vulnerability of VQA models when under adversarial attacks by human. Second, we benchmark several SOTA VQA models on the proposed dataset to reveal the fragility of VQA models. We observe a significant and universal performance drop when compared to VQA v2 and other robust VQA benchmarks, which corroborates our belief that existing VQA models are not robust enough. Meanwhile, this also demonstrates the transferability of these adversarial examples – data samples collected using one set of models are also challenging for other models. Third, as our annotators can ask different types of questions for different types of robustness, our analyses show that SOTA models suffer across various questions types, especially counting and reasoning.

Our main contributions are summarized as follows. (i) For better evaluation of VQA model robustness, we introduce a new VQA benchmark dynamically collected with a Human-and-Model-in-the-Loop procedure. (ii) Despite rapid advances on VQA v2 and robust VQA benchmarks, the evaluation on our new dataset shows that SOTA models are far from being robust. In fact, they are extremely vulnerable when attacked by human annotators, who can succeed within 2 trials on average. (iii) We provide a thorough analysis to share insights on the shortcomings of current models as well as comparison with other robust VQA benchmarks.

2. Related Work

Robust VQA Benchmarks There has been a growing interest in building new benchmarks to study the robustness of VQA models. VQA-CP [2], the first robust VQA benchmark constructed via reshuffling examples in VQA v2 [14], is proposed to evaluate question-oriented language bias in VQA models. GQA-OOD [22] improves from VQA-CP, and proposes to evaluate the performance differences between in-distribution and out-of-distribution split. Besides language bias, VQA-Rephrasings [39] exposes the brittleness of VQA models to linguistic variations in questions by collecting human-written rephrasings of VQA v2 questions. Causal VQA [1] studies robustness against semantic image manipulations, and tests for prediction consistency to questions on clean images and corresponding edited images. Further studies investigate robustness against reasoning. For instance, [38] collects perception-related subquestions per question for a new reasoning split of VQA dataset. [13] tests model's ability to logical reasoning through logical compositions of yes/no questions in VQA v2. GQA [17] provides large-scale rule-based questions from ground-truth scene graphs, that can test VQA model's ability on positional reasoning and relational reasoning.

Despite the continuous efforts in evaluating robustness of VQA models, these works mostly focus on a single type of robustness, and are based on the original VQA v2 dataset via either another round of question collection given the existing VQA examples, or automatic transformation or manipulation of current examples. In comparison, we use different image sources, and collect a new challenging VQA benchmark by allowing human annotators to directly attack current state-of-the-art VQA models.

Model-in-the-Loop Data Collection Dataset collection with a *model-in-the-loop* setting has received increasing attention in recent years in the NLP community. In this setting, models are used in the collection process to identify wrongly predicted, thus more challenging examples. These models are used either as a post-processing filter [53, 5] or directly during annotation [49, 34, 5]. In ANLI [34], the *model-in-the-loop* strategy is extended to a *Human-And-Model-in-the-Loop Enabled Training* (HAMLET) setting, where the data collection happens in multiple rounds, and in each round, the models are updated to stronger versions by training with examples collected from previous rounds. The goal of ANLI is to create a natural language inference (NLI) dataset that can grow along with the rapid advance of model capabilities [10, 29, 48, 24]. In contrast to static datasets that will eventually saturate as models become stronger, datasets created with the HAMLET procedure are dynamic – if the test set saturates with a more powerful model, one can use this more powerful model to assist the collection of a new set of difficult examples, leading to a never-ending challenge for the community. Meanwhile, the adversarial nature of the HAMLET procedure also helps to identify the weaknesses and vulnerabilities of existing models, and the biases or annotation artifacts [15, 35, 26] in ex-

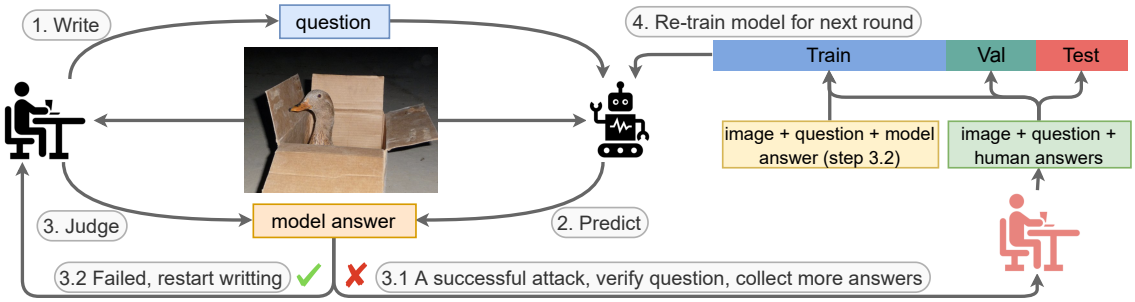


Figure 2: Overview of our adversarial data collection process, for a single round. The process can be considered as a game played by two parties, a human annotator and a well-trained model. Given an image, the annotator tries to attack the model by writing a tricky question (*step 1*), the model then predicts an answer to the question (*step 2*). Next, the human annotator judges the correctness of the model answer (*step 3*). If the model answer is judged as "definitely wrong" ✗, meaning the attack is successful, then we verify the question and collect more answers for it (*step 3.1*). Otherwise, the attack is failed, the annotator needs to write another question to attack the model (*step 3.2*). The *val* and *test* splits contain only successfully attacked questions, while *train* split contains also the failed questions.

isting datasets [6, 46, 26]. Beyond its application to the NLI task, the HAMLET procedure is also proved to be useful in collecting more challenging examples for the video-and-language future prediction task [26].

3. Adversarial VQA Dataset

In this section, we introduce the AVQA dataset in detail. Sec. 3.1 explains the data collection pipeline. Sec. 3.2 and Sec. 3.3 present data statistics and the comparison with other datasets.

3.1. Data Collection Pipeline

The HAMLET data collection procedure can be considered as a game played by two parties: a human annotator and a well-trained model. The human annotator competes against the model as an adversary and tries to design adversarial examples to identify its vulnerabilities. After collecting enough examples, the model augments its training with the collected data to defend similar attacks. For VQA, we define the adversarial example as an adversarial question on a natural image that the model answers incorrectly.

As shown in Figure 2, given an image, the human annotator tries to write a *tricky* question that the VQA model may fail. Once the question is submitted, an online model prediction will be displayed immediately to the workers. The model answer is then judged by the same annotator as either "definitely correct", "definitely wrong", or "not sure". If the model prediction is "definitely wrong", then the attack is successful, and we further ask the annotator to provide a correct answer. Otherwise, the annotator needs to write another question until the model predicts a wrong answer, or the number of tries exceeds a threshold (5 tries). To avoid obviously invalid questions caused by the annotator taking shortcuts (*e.g.*, untruthful judgement on model predictions, questions irrelevant to the image content), we also launch an answer annotation task. Successfully attacked questions are provided to 9 other annotators to collect extra answers, as well as their confidence level ("confident", "maybe" and "not confident") of their answer. The questions that receive less than 6 "confident" answers and have no agreement in answers among 10 annotators are removed during post-processing. In the end, each image is presented to 3 workers for question collection, and each image-question pair is shown to 10 annotators for answer collection.

This procedure can be continuously deployed for multiple rounds. At each round, we strengthen the models as we re-train them with extra data collected from previous rounds. This "dynamic" evolution of attacked models allows the collection of "harder" questions in the later rounds. In our setup, we launch the data collection for 3 rounds on Amazon Mechanical Turk. However, this data collection can be a never-ending process, as we can always replace the attacked model with a stronger model trained on newly collected data or better architectures developed in the future.

Round 1 (R1) For the first round, we employ VQA models trained on examples from VQA v2 [14] and VGQA [23] as our starting point. To avoid the collected questions overfitting to the vulnerabilities of a single model or a single architecture, for each user question, we randomly sample one model from LXMERT [43], UNITER-B [9] and UNITER-L [9] as the attacked model to generate the answer. We choose LXMERT and UNITER as representatives of two-stream and single-stream pre-trained V+L models, due to their strong performance on VQA v2. We use images sampled from Conceptual Captions [40] for annotation. In total, we collected 38.7K verified[1] questions and 28.2K unverfied questions over 13.7K images, and split the verified examples into 60%/10%/30% for train/val/test splits. All unverified examples are also added to the training split.

Round 2 (R2) For the second round, we re-train our models with questions from VQA v2, VGQA and R1's train split, and select the best model checkpoints of LXMERT,

[1] Verified questions are all successfully attacked questions.

Dataset	Image Source	#Image	IsCollected	#IQ	Model error rate (%)	#Tries	Time (sec.)	Data Split
				Total/Verified	Total/Verified	Mean/Median	per verified ex.	Train/Val/Test
Previous Robust VQA Datsets								
VQA-Reph.	COCO	-	✓	162K/-	-	-	-	-/162K/-
VQA-Intro.		-	✓	238K/-	-	-	-	222K/-/93K
VQA-LOL Comp.		-	✗	1.25M/-	-	-	-	916M/43K/291K
VQA-LOL Supp.		-	✗	2.55M/-	-	-	-	1.9M/9k/669K
VQA-CP v2		-	✗	-/-	-	-	-	438K/-/220K
IV-VQA	COCO†	357K	✗	376K/-	-	-	-	257K/11.6K/108K
CV-VQA		18.0K	✗	12.7K/-	-	-	-	8.5K/0.4K/3.7K
Ours								
R1	CC	13.7K	✓	93.1K/45.6K	48.9/35.2	1.6/1	71.0	53.6K/3.3K/10.0K
R2	CC	13.1K	✓	70.4K/37.8K	56.1/49.0	1.5/1	54.2	42.8K/2.7K/8.3K
R3	Various	11.1K	✓	79.5K/40.3K	50.7/34.4	1.6/1	57.3	45.9K/2.7K/8.1K
AVQA	Various	37.9K	✓	243.0K/123.7K	50.9/38.1	1.6/1	61.3	142.1K/8.7K/26.4K

Table 1: Data statistics. 'Model error rate' is the percentage of examples that the model gets wrong; 'Verified' is the questions with 10 answer annotations. Images for R3 are from various domains: Conceptual Captions (CC) [40], VCR [52] and Fakeddit [32]. We compare our dataset against previous robust VQA datasets, based on COCO [8] images. For number of image-question pairs (#IQ) and images (#Image), we only report the number of new examples generated/collected in each dataset. † indicates that the images are not natural, but edited. 'IsCollected' indicates whether the data is collected via crowdsourcing.

UNITER-B and UNITER-L based on R1's val set. Similarly, we randomly sample one model at a time for the workers to attack. A new set of non-overlapping Conceptual Captions images are used. In total, we collected 23.5K verified questions and 19.3K unverified question over 13.1K images, and split the data in a similar manner to R1.

Round 3 (R3) For the third round, we include more diverse images from different domains: (i) web images from Conceptual Captions [40]; (ii) user-generated images from Fakeddit [32]; and (iii) movie frame images from VCR [52]. The attacked model is still randomly sampled from LXMERT, UNITER-B and UNITER-L, but we add the training set from R1 and R2 to the training data.

Summary Finally, combining data collected in R1, R2 and R3 produces our proposed AVQA dataset. In the end, we collected 243.0K questions over 37.9K images, with 142.1K/8.7K/26.4K images in the train/val/test split.

3.2. Data Statistics

The data statistics of the new dataset are summarized in Table 1. The number of examples we collected per image varies per round, starting with approximately 6.8 questions/image for R1, to around 5.4 for R2 and 7.2 for R3. Under the same image domain for R1 and R2, we suspect that the annotators learn to identify model vulnerabilities more rapidly than the models learn to defend itself from the adversarial examples. We provide analyses in Sec. 4.1 and 4.4 for further investigation. On the one hand, the annotators are getting better at identifying vulnerabilities of these models. Analyses of question types per round in Sec. 4.4 show that the workers tend to ask more questions in certain categories, such as "count", "OCR" and "commonsense reasoning", that the model is more likely to fail. On the other hand, although the attacked model is strengthened through data augmentation, the model does not seem to learn from the adversarial examples effectively.

For each round, we report the model error rate, both on verified and all examples. The model error rate reported under "Total" captures the percentage of examples where the writer disagrees with the model's answer during question collection, but where we are not yet sure that the example is correct. The verified model error rate is the percentage of model errors from examples that we further collected 9 additional answers from other workers. We observe an increase in model error rate from R1 to R2. Assuming constant image domain difficulty in R1 and R2, the higher model rate suggests that the models in the later rounds are not significantly stronger, or the annotators are getting better at fooling the state-of-the-art models. In R3, where we included images from more diverse domains, the model error rate decreases from 49.0% to 34.4%. We suspect it is because the movie images from VCR are mostly human-centric, which is commonly observed in COCO.

We also report the average number of attempts ("#Tries" in Table 1) that a worker needed to complete the annotation process for each image, *i.e.*, to successfully attack the model or exceed the limits on number of tries. Surprisingly, although the VQA models used in the later rounds are trained with more data, the number of tries needed to successfully attack them does not increase. On average, it takes less than 2 tries to successfully attack a VQA model. Similarly, the average time needed per successful attack decreases by 15 seconds as data collection progresses.

3.3. Comparison with Other Datasets

Our Adversarial VQA dataset sets a new benchmark for evaluating the robustness of VQA models. It improves upon existing robust VQA benchmarks in several ways. First, the dataset by design is more difficult than previous

Model	Training Data	R1	R2	R3	AVQA	VQA v2	Δ(v2, AVQA)
		val/test	val/test	val/test	val/test	test-dev	test-dev, test
BUTD	VQA v2 +VGQA	20.80/19.28	18.77/18.85	20.63/21.10	20.12/19.71	67.60	47.89
	ALL	24.96/22.11	22.62/22.78	23.92/23.61	23.91/22.78	67.52	**44.74**
UNITER-B	VQA v2 +VGQA	20.60/17.91	17.86/18.55	20.71/20.17	19.79/18.81	72.70	53.89
	+R1	26.03/22.94	17.30/17.36	20.56/20.61	21.62/20.47	72.98	52.51
	+R1+R2	26.60/24.76	23.21/23.86	19.26/18.73	23.26/22.62	72.75	50.13
	ALL	26.85/24.93	23.38/23.92	24.48/23.27	25.04/24.10	72.66	48.56
UNITER-L	VQA v2 +VGQA	25.04/23.72	17.82/17.49	19.63/19.77	21.12/20.55	73.82	53.27
	+R1	29.31/26.63	19.34/18.66	19.78/18.99	23.25/21.78	73.89	52.11
	+R1+R2	30.13/28.15	23.11/23.54	17.35/17.05	23.97/23.29	73.77	50.48
	ALL	**30.80/28.45**	22.95/23.11	24.08/21.97	**26.27/24.78**	**74.15**	49.37
LXMERT	VQA v2 +VGQA	19.76/18.15	18.98/18.79	21.08/21.27	19.93/19.31	72.31	53.00
	+R1	23.89/22.65	19.01/17.91	21.64/21.42	21.68/20.78	72.51	51.73
	+R1+R2	26.76/24.86	23.28/**24.11**	19.39/19.57	23.38/23.00	72.61	49.61
	ALL	26.35/24.55	**23.84**/24.02	**25.27/23.71**	25.24/24.13	72.42	48.29

Table 2: Model performance of various models under different settings. AVQA / ALL refers to R1+R2+R3 / VQA v2+VGQA+AVQA.

datasets. During collection, we do not constrain the worker to ask questions that only fall into a single robustness type (Sec. 4.4). As a result, our dataset is helpful in defending model robustness against several robust VQA benchmarks (Sec. 4.3). Second, most robust VQA datasets are based on VQA v2 validation set, which state-of-the-art models use for training or hyper-parameter tuning. Thus, it is difficult to analyze the robustness of the best-performing models due to this data leakage. Our dataset is built on non-overlapping images from diverse domains, which naturally resolves it. Lastly, our dataset is composed of human-written questions on natural images, rather than rule-based questions in [13, 17] or manipulated images in [1]. A detailed comparison on data statistics is provided in Table 1.

Our work is inspired by ANLI [34]. While ANLI focuses on the pure text task of natural language inference, our work targets at the multi-modal task of visual question answering. However, due to the open-ended nature of VQA problem, the construction of AVQA is more challenging. Instead of giving the worker a target label when collecting adversarial questions, we first ask the worker to judge whether the model prediction is correct, then provide a ground-truth answer. Our verification process is also different from ANLI. In order to evaluate model performance under the same criteria as VQA v2 [14], we collect 10 answers from workers in total. Unlike the observations on ANLI, where the adversarial robustness of NLI models can be improved in a large extent through data augmentation of ANLI, our analysis on AVQA in Sec. 4 will show that it is more difficult to defend against adversarial attacks for VQA models.

4. Experiments and Analysis

In this section, we conduct extensive experiments to study the AVQA dataset. Specifically, Sec. 4.1 and Sec. 4.2 evaluate different model architectures with different modality inputs on AVQA; Sec. 4.3 examines how AVQA can help over other popular robust VQA benchmarks; Sec. 4.4 explores the question types that can fool the models; and Sec. 4.5 compares our data collection with automatic adversarial attack methods both qualitatively and quantitatively.

4.1. Model Evaluation

Table 2 reports the main results. In addition to UNITER-B, UNITER-L [9] and LXMERT [43], we also include BUTD [3] as an example of task-specific model with different model architecture, prior to the large-scale pre-training era. We show performance on the AVQA test sets per round, the total AVQA test set, and VQA v2 test-dev set. Our key observations are summarized as follows.

O1: Adversarial examples are transferrable across models. Both LXMERT and UNITER are variants of Transformer [45] architecture. We use BUTD as an example to investigate whether the adversarial examples are transferrable among the three models. The ∼20 performance of BUTD (trained on VQA v2+VGQA) on test set of each round indicates that workers did not find vulnerabilities specific to a single model architecture, but generally applicable ones across different model architectures.

O2: The difficulty level of rounds does not decrease. Under the same training data, we observe that the model achieves comparable or even lower performance on later rounds. As aforementioned in data statistics, the increased model error rates and the decreased average tries annotators needed suggest that the later rounds contain more difficult examples.

O3: Training with more rounds help defend robustness... Generally, our results indicate that training on more rounds improves model performance.

...but data augmentation alone is not effective. To investigate how much improvements are from adversarial examples, we show comparison of UNITER-B results on verified

Data	R1	R2	R3
Verified	25.63	22.84	23.63
Combined	26.85	22.82	24.38

Table 3: Comparison of verified and combined data. Results are reported on val split from UNITER-B trained on training data of each round, VQA v2 and VGQA.

Training Data	Lang. Only	R1 test	R2 test	R3 test	VQA v2 test-dev
VQA v2+VG	✗	17.91	18.55	20.17	**72.70**
AVQA-only	✗	**25.66**	**24.91**	**24.75**	59.99
ALL	✗	24.93	23.92	23.27	72.66
VQA v2+VG	✓	17.82	17.03	21.32	45.81
AVQA-only	✓	20.37	21.49	22.89	38.21
ALL	✓	19.75	20.75	22.81	46.23

(a) Language-only model performance.

Model	AVQA val	VQA-CP v2 test
BUTD	**23.91**	40.62 (38.82 [44])
+ [44]	23.79	**43.96**
UNITER-B	**25.04**	47.02 (46.93 [28])
+ [44]	24.70	**47.12**

(b) Model performance with a VQA-CP baseline from [44].

Table 4: Analysis on language bias.

and combined data in Table 3. In addition to verified data, the combined data include examples that the worker thinks the model has answered correctly. Even with almost doubled data size, results on combined data are not significantly better. This implies that simply training on more examples that the model correctly answers can hardly help the model be robust to adversarial attacks.

O4: Large model does not possess a clear advantage. Although outperforming UNITER-B and LXMERT on R1, UNITER-L does not show a clear advantage over R2 and R3. Overall, these three models achieve similar performance across rounds and on AVQA. When trained with "ALL" data, the performance gain from UNITER-L over BUTD is only +2.00 on AVQA, even though UNITER-L is pre-trained with extensive amount of image-text pairs.

4.2. Key Factor Analysis

We dive deeper into the key factors behind the low performance of state-of-the-art models on AVQA, and try to answer the following questions.

Q1: Is the language in AVQA biased? Starting from VQA-CP [2], concerns have been raised about the propensity of models to pick up on spurious artifacts that are present just in the co-occurrence of question-answer pairs, without actually paying attention to the image content. We compare full models trained with both images and questions to models trained only on questions by zeroing out image features in Table 4a. The results show that language-only models perform poorly on AVQA, and similarly on VQA v2.

Model	Training Data	AVQA test	VQA v2 test-dev
UNITER-B	VQA v2 +VGQA	18.81	72.70
	ALL	24.10	72.66
ClipBERT	VQA v2 +VGQA	21.16	69.08
	ALL	24.35	69.17
VILLA-B	VQA v2 +VGQA	19.68	73.37
	ALL	**26.08**	**74.28**

Table 5: Evaluation of grid-feature-based method ClipBERT [25], and adversarial-training-based method VILLA [11]. 'ALL' refers to VQA v2+VGQA+AVQA.

Language-only model performance decreases over rounds for AVQA. However, UNITER-B is not much better than language-only on AVQA. Obviously, without manual intervention, some bias remains in how annotators phrase questions. For example, there might be more counting questions with answers other than 2, which is the majority answer in VQA v2. Therefore, models trained on AVQA only performs slightly higher for both UNITER-B and Language-only model. However, we also observe the significant drop in VQA v2 performance is out of proportion to the slight performance improvement on AVQA.

We further investigate if the low performance is due to the difference in answer distribution between training and testing split. Due to the large number of answer candidates (more than 3000 for VQA v2), it is impossible to evenly balance the possibility of each answer. Therefore, we test out this hypothesis by adopting a simple yet effective baseline method on VQA-CP [44]: adding a regularization term by replacing the image with a randomly sampled one. The intuition is that the answer to a question corresponding to a given image is very unlikely to be correct for a randomly sampled image. As reported in Table 4b, although effective on VQA-CP, adding such regularization hurts the performance on AVQA for both BUTD and UNITER-B. In addition, when applied to a stronger model on VQA-CP, *i.e.*, UNITER-B, the regularization term is less effective.

Q2: Is AVQA transferrable to different visual features? The AVQA dataset is collected with the assistance of models trained on Faster R-CNN [36] region features [3]. To investigate whether these collected adversarial examples are transferrable to different image features, we conduct experiments using another type of feature, *i.e.*, grid features [18] from CNNs, which have shown to be effective for VQA tasks [18, 16, 33, 25]. Specifically, we consider ClipBERT [25], an end-to-end pre-trained model that directly takes in raw images and questions, and the images are represented by grid features as in [18]. Meanwhile, ClipBERT's end-to-end training strategy may also help to defend potential attacks to fixed feature representations widely used in previous work [9, 43, 3]. Table 5 compares ClipBERT against UNITER-B. The poor performance of ClipBERT on AVQA suggests that adversarial examples in AVQA

Model	Training Data	VQA-Rep.	VQA-LOL Comp.	VQA-LOL Supp.	VQA-Intro.	CV-VQA	IV-VQA
		Acc. ↑	Acc. ↑	Acc. ↑	M✓ S✓ ↑	#flips ↓	#flips ↓
Previous models	VQA v2 Train	56.59 [39]	49.88 [13]	50.54 [13]	50.05 [38]	7.53 [1]	78.44 [1]
UNITER-B [28]	VQA v2 Train	64.66	54.16	49.89	56.69	8.47	40.67
UNITER-B (ours)	VQA v2 Train	64.56	54.54	50.00	56.80	8.44	39.97
	+AVQA	**65.42**	**55.10**	**51.36**	**57.93**	**8.43**	**38.40**

Table 6: Model performance on recent robust VQA benchmarks.

Round	Count	OCR	Reasoning				Visual Concept Recognition				
			Position	Relation	Common-sense	Other	Low-level	Action	Small Object	Occlusion	Abstract
R1	23.3%	10.7%	14.7%	8.3%	17.3%	0.7%	9.7%	4.3%	13.3%	14.7%	6.3%
R2	30.0%	22.7%	12.0%	27.7%	20.0%	4.3%	12.7%	9.3%	22.7%	10.0%	15.3%
R3	35.3%	13.0%	13.0%	28.3%	25.0%	6.3%	11.7%	4.3%	20.0%	20.0%	6.0%
Ave.	29.6%	15.4%	13.2%	21.4%	20.8%	3.8%	11.3%	6.0%	18.7%	14.9%	9.2%

Table 7: Analysis of 300 randomly sampled AVQA examples per round and on average. Low-level visual concepts include color, shape, and texture. A question may belong to multiple different categories.

are transferrable to different image representations. However, ClipBERT performs comparably to UNITER-B on AVQA, although it significantly under-performs UNITER-B on VQA v2, which suggests that VQA v2 may not be reliable for evaluating model robustness.

Q3: How effective is adversarial training on AVQA? We examine the effectiveness of adversarial training by adopting PGD-based adversarial training method VILLA in [11]. VILLA-B is both adversarially pre-trained on large-scale image-text data and adversarially finetuned on the respective dataset. We compare its performance against UNITER-B on both AVQA and VQA v2 in Table 5. Adversarial training brings slight performance improvement. However, the performance gap between AVQA and VQA v2 is still very significant. Note that VILLA-B crafts adversarial examples during training by adding adversarial perturbations to the embedding space. These adversarial perturbations can hardly change the intrinsic statistics of training data, such as the distribution of question types and relevant objects in the image. Our analysis of question types and visual recognition concepts in Sec. 4.4 will show that AVQA is hard because it requires the model to have the ability to reason, count and recognize different visual concepts.

4.3. Evaluation on Other Datasets

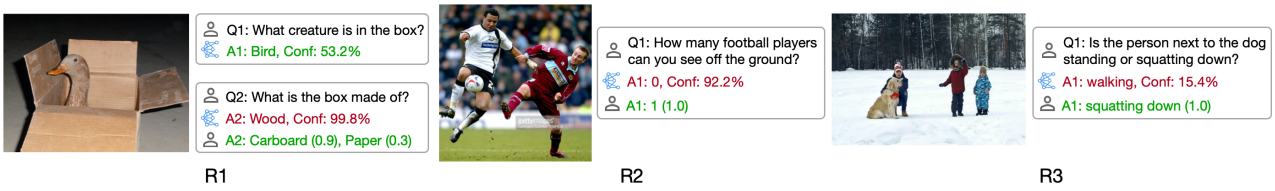
We also test models on recent robust VQA benchmarks including: VQA-Rephrasings [39] for linguistic variations, VQA-LOL [13] Complement/Supplement for logical reasoning, VQA-Introspect [38] for consistency of model predictions in perceptual sub-questions and main reasoning questions, CV-VQA [1] and IV-VQA [1] for model robustness to image manipulations. Results are summarized in Table 6. We observe that UNITER-B can already outperform previous models for most of the benchmarks, which is consistent with observations in [28]. Training on AVQA is helpful in improving model performance on robustness benchmarks. Particularly, AVQA helps to boost model reasoning capability across 3 datasets. It is likely that AVQA exposes the model training to more diverse question templates, hence improves on VQA-Rephrasings. On IV-VQA, which focuses on counting questions, AVQA helps to improve performance despite of the significant performance gain UNITER-B has already achieved.

4.4. Analysis on Question Types

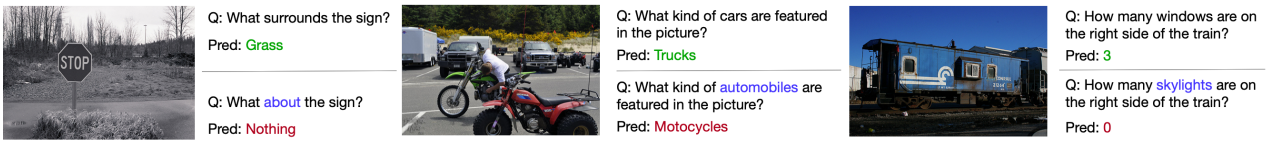
We manually annotate 300 randomly sampled examples from each round to investigate: *which types of questions do workers employ to fool the models*, and *how they evolve as the rounds progress*.

Results are summarized in Table 7. Questions are categorized into 4 meta-categories: counting, OCR, reasoning, and visual concept recognition. Although OCR and counting can be considered as visual concept and quantitative reasoning, we separate them out as they contribute a large portion per round, to almost 50% in the later rounds. There are three main reasoning questions: positional reasoning (*i.e.*, the relative/absolute position of an object), relational reasoning (*i.e.*, semantic relationship between two or more objects), and commonsense reasoning (*i.e.*, visual commonsense reasoning, *e.g.*, "Is the water more likely to be a lake or an ocean", given an image showing a body of water surrounded by mountains.). Other reasoning questions include comparative reasoning (*e.g.*, "which person is taller?") and logical reasoning (*e.g.*, negation). For visual concept recognition, we roughly divide them into low-level visual concepts (*e.g.*, color, shape, texture), action (*e.g.*, "what is the person doing"), small objects, occluded objects, and abstract objects (*e.g.*, objects in painting).

We observe that annotators rely heavily on counting questions to attack the models – nearly 30% of the sampled



(a) Visualization of examples collected per round in AVQA. Each ground-truth answer (VQA score) is collected from 10 workers.



(b) Visualization of examples generated via textual adversarial attack methods. Blue indicates the changes made in adversarial questions.

Figure 3: Illustration of adversarial examples from (a) AVQA and (b) textual adversarial attack methods: Sears [37], Textfooler [20] and Sememe+PSO [51]. Green (red) indicates a correct (wrong) answer.

Method	#Tries	Error Rate	Orig. Acc.	Adv. Acc.
Sears [37]	3.0	11.6%	69.1	63.0
Textfooler [20]	39.5	1.4%	69.1	67.8
Sememe+PSO [51]†	35.9	88.6%	84.9	12.5
AVQA	1.6	38.1%	-	-

Table 8: Comparison to adversarial attack methods. Orig. Acc. (Adv. Acc.) is the accuracy on original (adversarial) examples. (†) Note that Sememe+PSO only attacks questions longer than 10 words, so 94.8% examples are not being attacked.

questions across all rounds fall into this category. While R1 questions are mostly on objects that are of normal sizes and less occluded, we found that the counting questions become harder in R2 and R3 as many of them are about small and occluded objects. There is also a surge in abstract and OCR questions for R2, due to the increase in the number of abstract images and images that contain scene text. The percentage of reasoning questions, especially relational reasoning and commonsense reasoning, increases drastically from R1 to R2 and R3. Visualizations in Figure 3a show that questions in later rounds are indeed more complicated, with more detailed relational and positional descriptions when referring to an object. Overall, these findings are compatible with the idea that VQA models are not robust enough to various types of questions.

4.5. Why Human-in-the-Loop?

Textual adversarial attack methods [31, 20, 51] have been widely explored in NLP. The goal is to alter model predictions with minor changes to the input textual queries, so that adversarial examples can be generated and model vulnerabilities can be identified automatically. We investigate whether we can directly apply these methods to generate adversarial examples in high quality and compare the generated examples to AVQA. In total, we consider 3 different textual adversarial attack methods, including Sears [37] via bask-translation for sentence-level attacks, Textfooler [20] and Sememe+PSO [51] by replacing words with its synonyms or words that share the same sememe annotations for word-level attacks. The adversarial attacks are performed to all questions on 5000 images in the Karpathy split [21]. We visualize examples in Figure 3b. Without human-in-the-loop, the generated adversarial questions share similar problems: (i) the adversarial question does not share the same answer with the original question, therefore additional answer annotations may need to be collected; (ii) model prediction to the adversarial question is not necessarily incorrect when it is different from answers to the original question; (iii) word similarity may not hold when it needs to be grounded to the image (*e.g.*, window vs. skylights). In addition, we compare these methods against the AVQA dataset quantitatively in Table 8. Generally, humans take much fewer tries and have a higher successful rate when attacking VQA models. How to design effective adversarial attack methods to generate high-quality VQA examples can be an interesting future research direction.

5. Conclusion

In this work, we collect a new benchmark Adversarial VQA (AVQA) to evaluate the robustness of VQA models. It is collected iteratively for 3 rounds via a human-and-model-in-the-loop enabled training paradigm, on images from different domains. AVQA questions cover diverse robustness types, enabling a more comprehensive evaluation on model robustness. Our analysis shows that state-of-the-art models cannot maintain decent performance on AVQA, despite of large-scale pre-training, adversarial training, sophisticated model architecture design, and stronger visual features. AVQA brings a new challenge to the community on how to design more robust VQA models that are ready to deploy in real-life applications.

References

[1] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *CVPR*, 2020. 1, 2, 5, 7

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018. 1, 2, 6

[3] 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 *CVPR*, 2018. 1, 5, 6

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 1

[5] Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the ai: Investigating adversarial human annotation for reading comprehension. *TACL*, 2020. 2

[6] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015. 3

[7] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *ECCV*, 2020. 1

[8] 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. 4

[9] 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 *ECCV*, 2020. 1, 3, 5, 6

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[11] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPs*, 2020. 1, 6, 7

[12] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *CVPR*, 2019. 1

[13] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *ECCV*, 2020. 1, 2, 5, 7

[14] 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 *CVPR*, 2017. 1, 2, 3, 5

[15] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *NAACL*, 2018. 2

[16] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 6

[17] Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. In *CVPR*, 2019. 1, 2, 5

[18] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *CVPR*, 2020. 1, 6

[19] 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. 1

[20] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI*, 2020. 8

[21] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 8

[22] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? In *CVPR*, 2021. 2

[23] 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. *IJCV*, 2017. 3

[24] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*, 2020. 2

[25] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021. 6

[26] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *EMNLP*, 2020. 2, 3

[27] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *ICCV*, 2019. 1

[28] Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*, 2020. 1, 6, 7

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1

[31] John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *EMNLP: System Demonstrations*, 2020. 8

[32] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*, 2019. 2, 4

[33] Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. Revisiting modulated convolutions for visual counting and beyond. In *ICLR*, 2021. 6

[34] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *ACL*, 2020. 2, 5

[35] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 2018. 2

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 6

[37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *ACL*, 2018. 8

[38] Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Interrogating vqa models with sub-questions. In *CVPR*, 2020. 1, 2, 7

[39] M Shah, X Chen, M Rohrbach, and D Parikh. Cycle-consistency for robust visual question answering. In *CVPR*, 2019. 1, 2, 7

[40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2, 3, 4

[41] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 1

[42] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *NAACL-HLT*, 2021. 1

[43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 1, 3, 5, 6

[44] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. In *NeurIPS*, 2020. 6

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5

[46] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*, 2018. 3

[47] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *CVPR*, 2020. 2

[48] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2020. 2

[49] Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H Miller, Arthur Szlam, Douwe Kiela, and Jason Weston. Mastering the dungeon: Grounded language learning by mechanical turker descent. In *ICLR*, 2018. 2

[50] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 1

[51] Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. Word-level textual adversarial attacking as combinatorial optimization. In *ACL*, 2020. 8

[52] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 2, 4

[53] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*, 2018. 2

[54] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 1

[55] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *CVPR*, 2021. 1