FVQA 2.0: Introducing Adversarial Samples into Fact-based Visual Question Answering

Weizhe Lin
Department of Engineering
University of Cambridge
United Kingdom
wl356@cam.ac.uk

Zhilin Wang
Department of Linguistics
University of Washington
United States
zhilinw@uw.edu

Bill Byrne
Department of Engineering
University of Cambridge
United Kingdom
bill.byrne@eng.cam.ac.uk

Abstract

The widely used Fact-based Visual Question Answering (FVQA) dataset contains visually-grounded questions that require information retrieval using common sense knowledge graphs to answer. It has been observed that the original dataset is highly imbalanced and concentrated on a small portion of its associated knowledge graph. We introduce FVQA 2.0 which contains adversarial variants of test questions to address this imbalance. We show that systems trained with the original FVQA train sets can be vulnerable to adversarial samples and we demonstrate an augmentation scheme to reduce this vulnerability without human annotations.

1 Introduction

Knowledge-based Visual Question Answering (KB-VQA) lies at the intersection of Computer Vision, Natural Language Processing, and Information Retrieval. A KB-VQA system must access external knowledge sources to find a correct and complete answer, a task that is sometimes hard for humans.

Fact-based Visual Question Answering (FVQA) (Wang et al., 2017) is a VQA task in which visually-grounded questions and answers about images are grounded by knowledge-graph (KG) triplets taken from several 'common sense' knowledge bases, such as ConceptNet (Speer et al., 2017), Webchild (Tandon et al., 2017), and DBpedia (Auer et al., 2007). For instance, "Question: Which thing in the image can be used for scooping food? Answer: spoon" is associated with the KG triplet "spoon - UsedFor - scooping food". These questions are challenging in that retrieving information from external KGs is necessary.

The original FVQA dataset (Wang et al., 2017) has several readily observed limitations. First, the dataset is small (5486 samples) and the annotations are limited to a single answer per question, ignoring other correct answers. This limitation arises from the FVQA creation process in which annotators were first asked to select a KG triplet on which they would ask a question about an image. This approach prevented the annotators from labeling other valid KG triplets. Secondly, the dataset is highly imbalanced. Some triplets and answers are frequently used, but other KG triplets and answers are severely underrepresented in training. For example, there are 1,129 possible answers in total, but over 90% of questions focus on only a half of them; 792 (70%) answers appear less than 3 times; only 4,216 out of ∼220k triplets are used.

These limitations lead to a potential problem: KB-VQA systems trained on this dataset overfit on these frequently used triplets and perform poorly on variants that contain other valid triplets or other images. Also, extensive overlap between training and test can lead to an unrealistically high question answering baseline performance. We noted that a question with a triplet unseen in training is often answered with 'person', since it is the most frequent answer in the original data distribution.

To overcome these limitations, we introduce an enlarged test set that contains two types of adversarial samples (as shown in Fig. 1): (1) *FixQ*: the question remains the same, but is associated with a different image and a different correct answer. This ensures that a system is less able to achieve high performance if it is biased by language patterns in questions; (2) *FixA*: the answer remains the same, but the question is asked in a different way. This favours systems that do more than make straightforward associations between questions and answers based on the training data. In contrast to the original test set, this new set further challenges KB-VQA system to retrieve knowledge from KBs and answer questions without being biased towards frequent answers in the original dataset. We show that models trained on the original FVQA training sets are significantly less robust on these adversarial test samples.

Findings of the Association for Computational Linguistics: EACL 2023, pages 149–157
May 2-6, 2023 ©2023 Association for Computational Linguistics

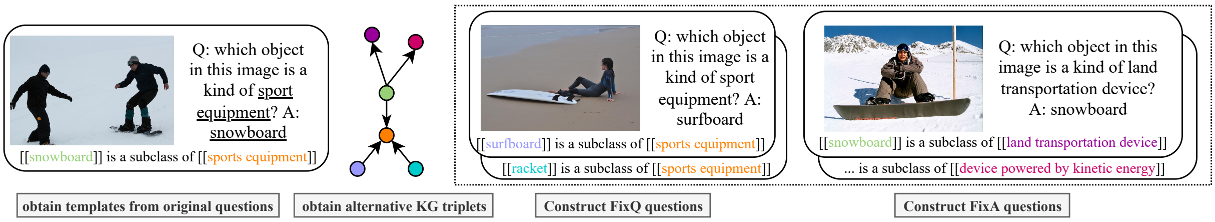

Q: which object in this image is a kind of sport equipment? A: snowboard
[[snowboard]] is a subclass of [[sports equipment]]
Q: which object in this image is a kind of sport equipment? A: surfboard
[[surfboard]] is a subclass of [[sports equipment]]
[[racket]] is a subclass of [[sports equipment]]
Q: which object in this image is a kind of land transportation device? A: snowboard
[[snowboard]] is a subclass of [[land transportation device]]
... is a subclass of [[device powered by kinetic energy]]
obtain templates from original questions
obtain alternative KG triplets
Construct FixQ questions
Construct FixA questions

Figure 1: The workflow of constructing adversarial samples (FixQ and FixA questions) from the original test set questions.

Given that it is hard to guarantee a good triplet coverage during annotation, we explore an augmentation scheme to address this problem without costly human annotation of large-scale adversarial training samples. Our scheme generates slightly noisy adversarial samples that improve the coverage of valid KG triplets to enhance model training.

Our contributions are:

(1) We introduce FVQA 2.0, which adds an adversarial test set that challenges KB-VQA system robustness to adversarial variants of questions.

(2) We demonstrate the performance gap between the original test set and the adversarial test set, showing that considering adversarial samples is important for better realistic KB-VQA performance.

(3) To further demonstrate the importance of adversarial samples, we leverage a semi-automated augmentation scheme to improve system robustness on the adversarial test through the creation of large-scale noisy adversarial examples.

2 Related Work

KB-VQA questions can focus on facts and concepts, as in FVQA (Wang et al., 2017) and OK-VQA (Marino et al., 2019). Such questions challenge the information retrieval ability of systems. KB-VQA questions can also require commonsense reasoning, as in parts of OK-VQA and A-OKVQA (Schwenk et al., 2022). In particular, S3VQA (Jain et al., 2021) is an augmented version of OKVQA, improving both the quantity and quality of some question types. A-OKVQA has shifted its core task to "reasoning questions". Only 18% of questions in A-OKVQA require answers from an external knowledge base.

VQA 2.0 (Goyal et al., 2017) collects 'complementary images' such that each question is associated with a pair of images that result in different answers. Jain et al. (2021) derive new S3VQA questions from manually defined question templates. They annotated spans of objects that could be replaced, and then substituted them with a complicated substitute-and-search system. In contrast to their labour-intensive annotation work, our adversarial samples are collected through a semi-automatic approach that fully leverages the structural information in KGs to significantly reduce the human work required.

More broadly, in Knowledge-Graph Question Answering (KG-QA), work has exploited KG to generate synthetic data in unseen domains (Linjordet, 2020; Trivedi et al., 2017; Linjordet and Balog, 2020). Our work extends visually-grounded questions with valid common sense KG triplets.

3 Method

Extracting Question Templates. We extract question templates that can be used to reconstruct new questions using other valid KG triplets. We apply a rule-based system to replace KG entities that appear in the questions. For example, 'what is used for storing liquid in this image?' is transformed to 'what is used for <t> in this image?' given that the associated KG triplet is "bottle (<h>) - /r/UsedFor (<r>) - store liquid (<t>)".

For each template, we construct new question-answer pairs by exploring the node structure of the KG. For example, "bottle - /r/UsedFor - hold water" is also a valid triplet from ConceptNet, whose head and relation are the same as the original triplet. A new question "Q: what is used for holding water in this image? A: bottle" can now be constructed.

Template Filtering. We focus on questions about object concepts that are transferable to other images, ignoring a small portion (<10%) of FVQA questions to which the answers are based on particular scenes (e.g. 'what can you often find in the place shown in this picture?').

Human annotators are employed to filter out non-transferable templates, such as questions that contain specific object positioning ("what is the object in the lower right of this image used for?"). This process takes around 1 hour with two annotators to obtain 440 valid templates after removing highly similar templates.

Matching Suitable Images. We use 619 of FVQA images[1] that are also present in the Visual Genome dataset (Krishna et al., 2017). Using the object annotations of the VG dataset to determine if an image contains the object being asked, we employ a rule-based system to assign a suitable image to each generated adversarial sample, within which process all images are assigned to approximately the same number adversarial samples by a simple approach described in Appendix B. We limit the number of FixQ and FixA questions generated by each template to 5, which guarantees a reasonable dataset size. 3,805 questions are generated.

Manual Verification. We conduct manual verification to rule out samples that are incorrectly generated. 432 counter-intuitive KG triplets are removed in this step. Finally, we obtain 2,820 adversarial samples, offering 1,671 new valid triplets from the KG. Around 75% samples are verified as correct, showing that the rule-based generation works well. The remainder are discarded.

The official FVQA evaluation performs 5-fold validation: each split preserves around half its samples for testing. As a naming convention, under each split, the templates extracted from the original training samples are called 'train templates' while the rest are 'test templates'. Since the train templates may contain language patterns that have been learned in training, we ensure that only questions derived from test templates are used in the adversarial testing. As a result, we have 1,376 adversarial test samples per split on average, with 1,129 FixQ and 246 FixA questions.

Augmentation with Adversarial Data. We explore an augmentation scheme to augment the training data with slightly noisy but auto-generated adversarial samples, which avoids heavy annotation work. In each split, **only the train templates** (defined in the above paragraph) are used to generate adversarial samples for training such that no information of test samples is leaked to training. This avoids biasing the training to the test sets, which would make the test sets less indicative of true system performance. We obtain an augmentation set with 2,262 questions per split on average semi-automatically, which would otherwise cost hundreds of hours to build from scratch. The origins of these adversarial samples are referred to as '*Originating Questions*'. There are 435 such questions per split. In training, these questions are randomly replaced by their adversarial variants.

4 FVQA 2.0 Statistics

Set Name	#Samples	std
Standard Train Set	2,927	69
Standard Test Set	2,899	69
Originating Questions Set	435	52
Adversarial Test Set	1,376	193
- FixA Questions	1,129	157
- FixQ Questions	246	38
Augmentation data	2,262	267

Table 1: Dataset Statistics. #Samples: average number of samples across 5 folds; std: the standard deviation over 5 folds.

The numbers of samples in each set are provided in Table 1. The official FVQA dataset creates 5 folds by splitting the images being used. Half of these images are used in training while the other half are reserved for testing. In all our new sets, under each split, questions for training are not leaked to testing. The 'Originating Question Set' is a subset of Standard Test Set by its definition (Sec. 3). The Adversarial Test Set is formed by FixA questions and FixQ questions; it is created by automatically generating adversarial question variants from the questions in the 'Originating Question Set'. It covers relationships such as /r/RelatedTo, /r/IsA, /r/PartOf, /r/HasA, /r/UsedFor, /r/CapableOf, /r/AtLocation, /r/Desires, /r/MadeOf. The augmentation data consists of adversarial variants that are derived from the questions in the Standard Train Set.

5 Experiments

Baseline Systems We use several FVQA systems for comparison[2]: FVQA (Wang et al., 2017), the baseline system provided in the official FVQA dataset paper; GCN (Narasimhan et al., 2018), a model that leverages graph convolutional networks (GCNs) to aggregate features from visual/language/fact modalities; Mucko (Zhu et al., 2020), the current state-of-the-art system that uses GCNs to combine visual, fact, and semantic graphs.

We test our augmentation scheme on several systems that have code available: **RAVQA-NoDPR**

[1] FVQA images are from Microsoft COCO (Lin et al., 2014) and ILSVRC (Russakovsky et al., 2015).

[2] Since many recent FVQA systems are not open-sourced, we additionally include KB-VQA systems from OKVQA.

and **RAVQA-DPR** (Lin and Byrne, 2022), T5 (Raffel et al., 2020)-based models that transform images into texts (e.g. objects, attributes, and image captions) and the DPR version additionally uses Dense Passage Retrieval (Karpukhin et al., 2020) to retrieve documents from knowledge bases[3]; **TRiG** (Gao et al., 2022), a model that is similar to RAVQA-DPR but different in embedding fusion; **ZS-F-VQA** (Chen et al., 2021), an FVQA system that obtains the final prediction by fusing the individual predictions in answer/fact/relation graphs.

Metrics. We report accuracy and standard deviation over 5 splits (Sec. 4). In calculating accuracy for open-ended generation systems (RAVQA/TRiG), a question is considered successfully answered if the generated answer string is an exact match to the ground-truth answer node, which is the closest KG node to the ground-truth answer string (shortest in Levenshtein distance computed from node names).

Performance and Discussion. Table 2 shows that the systems used for evaluating the new adversarial set are sufficiently strong (e.g. 69.56% accuracy by RAVQA-DPR) in comparison with the three models that do not have code available, which achieve 58.76% (FVQA), 69.35% (GCN), and 73.06% (Mucko, current state-of-the-art) respectively. RAVQA-NoDPR achieves 84.59% accuracy on the originating questions but obtains only 71.48% accuracy on the adversarial samples derived from them. Such performance gaps are readily observed on all systems. Systems trained on the original training sets fail to perform equally well on the two sets, showing that the original FVQA training data does not contain adversarial variants and the resulting systems are vulnerable to them.

By incorporating adversarial variants in training, all systems achieve much better performance on the challenging adversarial set, e.g. RAVQA-NoDPR is improved from 71.48% to 82.38% (+10.9%). The performance on the standard and adversarial test sets now match well, with the gap reduced from more than 10% to ∼3%, showing that the augmentation scheme significantly improves systems' reliability and robustness. The relative improvement is slightly less (+8.1%) for RAVQA-DPR, which is expected given that it is a retrieval-based system designed to answer both seen and unseen questions with its strong retrieval ability. ZS-F-VQA benefits greatly from augmentation: its adversarial performance is improved by 24.09%. This is because its model size is much smaller and it can easily be biased by language patterns, images, and frequent answers seen in training.

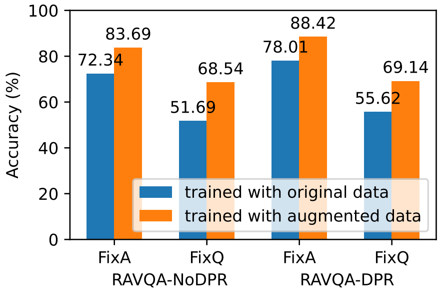


Figure 2: Performance on FixQ and FixA questions.

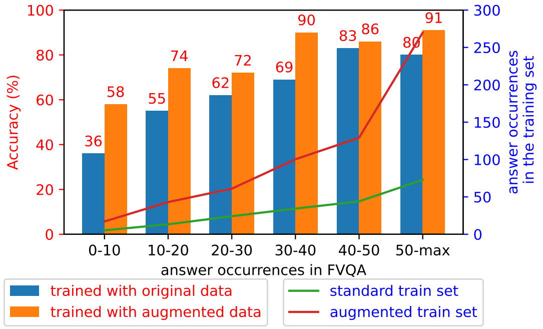


Figure 3: RAVQA-DPR accuracy on adversarial questions and answer occurrences in the standard/augmented training sets. They are grouped by the number of answer occurrences in the original FVQA dataset. For example, a question is counted towards the '0-10' group if its answer appears less than 10 times in the original dataset.

In summary, systems trained on the original training sets are vulnerable to adversarial variants of the test questions. We show that through generating adversarial samples for data augmentation, systems become much more robust to these variants.

Analysis of Model Vulnerability. As shown in Fig. 2, RAVQA systems trained with original training sets perform better on FixA questions (∼88%) than on FixQ questions (∼69%). This suggests that systems perform worse when asked the same questions on different images. This is potentially because the language patterns seen in training bias the models to frequent choices, lowering the FixQ generalizability. In contrast, systems are less distracted by different ways of asking for the same answer, potentially due to the strong language modelling capability of T5 used by them. The augmentation scheme improves systems on both types of questions significantly (by ∼10% on each), showing the value of adversarial samples in training.

[3] In our experiments, the knowledge base consists of surface texts of triplets (e.g. "[car] has [4 wheels]").

Test on:	Standard Test Set		Originating Question Set		Adversarial Test Set	
Trained on:	Original	Augmented	Original	Augmented	Original*	Augmented (improv. over *)
ZS-F-VQA	48.16 ±1.03	48.57 ±1.00	63.67 ±0.88	64.63 ±0.81	49.97 ±2.37	74.06 ±1.92 +24.09
TRiG	64.94 ±0.93	65.73 ±0.33	81.67 ±1.12	83.48 ±1.89	68.86 ±3.26	79.79 ±1.34 +10.93
RAVQA-NoDPR	66.19 ±1.15	66.70 ±1.00	84.59 ±1.24	85.75 ±0.90	71.48 ±2.08	82.38 ±1.65 +10.90
RAVQA-DPR	69.56 ±0.78	69.90 ±0.56	87.52 ±1.68	88.33 ±1.40	76.91 ±1.93	85.05 ±1.15 +8.14

Table 2: Model performance on the standard test set, originating questions (from which the adversarial questions are derived), and adversarial test set. Results are reported as the average of 5 folds with standard deviations.

Fig. 3 plots the RAVQA-DPR performance on the adversarial test set questions that are grouped by their answer occurrences in the original FVQA dataset. The answer distribution of the original dataset affects adversarial performance greatly: systems perform much worse on questions whose answers appear less frequently in FVQA. In contrast, the performance deterioration that arises from answer rarity is mitigated significantly after augmentation. The augmentation scheme (red v.s. green curve in Fig. 3) compensates for the imbalanced answer distribution by providing more question variants so that systems are trained on both popular and rare answers.

6 Conclusion

We show that the FVQA test sets are not sufficiently indicative of true system performance through providing a new human-verified adversarial test set that contains adversarial variants of the original test set questions. We show the value of adversarial samples in KB-VQA datasets by showing an augmentation scheme that leverages structural information in KGs to create augmentation questions for training, which improves models' robustness to adversarial variants.

We release the dataset and the codes in Github (https://github.com/LinWeizheDragon/Retrieval-Augmented-Visual-Question-Answering).

7 Limitations

The adversarial test set was firstly generated from the original FVQA dataset by a rule-based system and then filtered by human annotators. As a result, the new set is limited with respect to the question types, language patterns, and knowledge triplets used in FVQA. One potential solution to overcome this limitation is to invest more human effort to generate adversarial questions from scratch, which is, however, much more expensive than the semi-automatic approach presented here.

The proposed augmentation approach also relies on the relationships encoded in the knowledge base (e.g. ConceptNet (Speer et al., 2017)). These will influence the quality and diversity of the augmented data, with the expectation that improvements in KG scope and quality will improve data augmentation.

The number of adversarial examples introduced in this work is sufficiently large for investigating the performance discrepancies (on the original and adversarial test sets) and demonstrating the necessity of KB-VQA adversarial samples. However, it is considered beneficial to introduce adversarial samples on a larger scale by considering them in the design of future KB-VQA datasets.

8 Ethics Statement

Our dataset was created semi-automatically from the FVQA dataset and ConceptNet, a crowd sourced common sense knowledge graph. Though we have included human annotators in the loop to remove sexual, offensive, and other inappropriate data samples that were automatically generated (we removed ∼200 inappropriate knowledge graph triplets during annotation), we recognize that the dataset may still contain a small number of inappropriate samples. Any developers who replicate the semi-automatic methodology described in the paper to extend the datasets should include a similar review step in the manual work flow. We also recognize that the systems trained on this dataset may convey such inappropriate information to users in real-life applications. Therefore, extra care must be taken when using this dataset in applications that interact directly with real users.

9 Acknowledgement

W. Lin was supported by a Research Studentship funded by Toyota Motor Europe (RG92562(24020)). We thank our colleagues, Daniel Olmeda Reino (Toyota Motor Europe) and Jonas Ambeck (Toyota Motor Europe), who provided insight and expertise in this project.

We thank Alexandru Coca (University of Cambridge) for comments that greatly improved the manuscript. We would also like to thank all the reviewers for their knowledgeable reviews.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z Pan, Zonggang Yuan, and Huajun Chen. 2021. Zero-shot visual question answering using knowledge graph. In *International Semantic Web Conference*, pages 146–162. Springer.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2491–2498.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Trond Linjordet. 2020. Neural (knowledge graph) question answering using synthetic training data. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, page 3245–3248, New York, NY, USA. Association for Computing Machinery.

Trond Linjordet and Krisztian Balog. 2020. Sanitizing synthetic training data generation for question answering over knowledge graphs. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 121–128.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv preprint arXiv:2206.01718*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Niket Tandon, Gerard De Melo, and Gerhard Weikum. 2017. Webchild 2.0: Fine-grained commonsense knowledge distillation. In *Proceedings of ACL 2017, System Demonstrations*, pages 115–120.

Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. Lc-quad: A corpus for complex question answering over knowledge graphs. In *International Semantic Web Conference*, pages 210–218. Springer.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1097–1103. International Joint Conferences on Artificial Intelligence Organization. Main track.

A Training Details

ZS-F-VQA: The experiments were performed on 1 × Nvidia RTX 3090. We used the code from the official repository[4]. The original paper dropped questions that have rare answers. For fair comparison with other models, we added these rare answers back and performed training and testing. We chose to report the performance of the system which uses 'SAN' as the base model (details are in the paper and the repository), since this setting has achieved the best performance. The hyperparameters for training are kept the same as the original paper. In testing, we selected $k_e = 10; k_r = 1; score = 10$ by grid search (search range: $0 \leq k_e \leq 20; 0 \leq k_r \leq 20; 0 \leq score \leq 20$).

RA-VQA-NoDPR/RAVQA-DPR/TRiG: All experiments were performed on 1 × Nvidia A-100 GPU. We chose Adam (Kingma and Ba, 2015) as the optimizer. When the model has a DPR component, we trained the DPR component for 4 epochs with a constant learning rate 10^{-5}. In training the answer generator, the learning rate linearly decays from 6×10^{-5} to 0 after 10 epochs, as suggested in the original paper. For each split, the checkpoints at global step 2k (around 3.5 epochs) were used in testing. We retrieve 5 best documents when predicting the answer ($K_{\text{train}} = 5$), since this number was reported to best balance the computation and performance (Lin and Byrne, 2022).

We obtained the pre-trained model parameters (T5-large and BERT-base) from Huggingface (Wolf et al., 2020). These systems are implemented with Huggingface Python libraries (under Apache License 2.0). The FAISS (Johnson et al., 2019) system is under MIT License.

B Balancing Images in Adversarial Variants

In assigning suitable images to question templates, it is necessary to ensure the diversity of images being used. We achieve this by controlling the number of assignments per image with a simple approach so that the numbers are approximately the same for all images.

In the process, for each new question-answer pair that needs an image, we rank all the images that contain the object being asked in the the question by their current total number of assignment. We select the image that satisfies the conditions as well as having the fewest number of assignment as the associated image of the new sample. We found that by applying this simple yet effective strategy, the assigned images present a good diversity.

C Annotation Details

Two annotators (volunteers in the research group) worked independently to rule out incorrectly generated examples. An example was accepted only if the two annotators achieved consensus. The annotators attempted to fix grammar errors that caused severe misunderstanding, while mild errors were kept (for example, 'is used for carry people' does not prevent models/people from understanding the question, and thus the annotators are not required to fix them).

[4] https://github.com/China-UK-ZSL/ZS-F-VQA

In particular, questions that might contain information of individuals / private information were dropped, though it is a very rare case.

Questions with multiple answers: when a question can be answered with multiple instances in an image, all possible answers are included. During annotation, incorrect answers were dropped from the list. In evaluation, answering any correct answer is considered successful. There are around 11% multiple-answer questions at the end.

D Additional Results

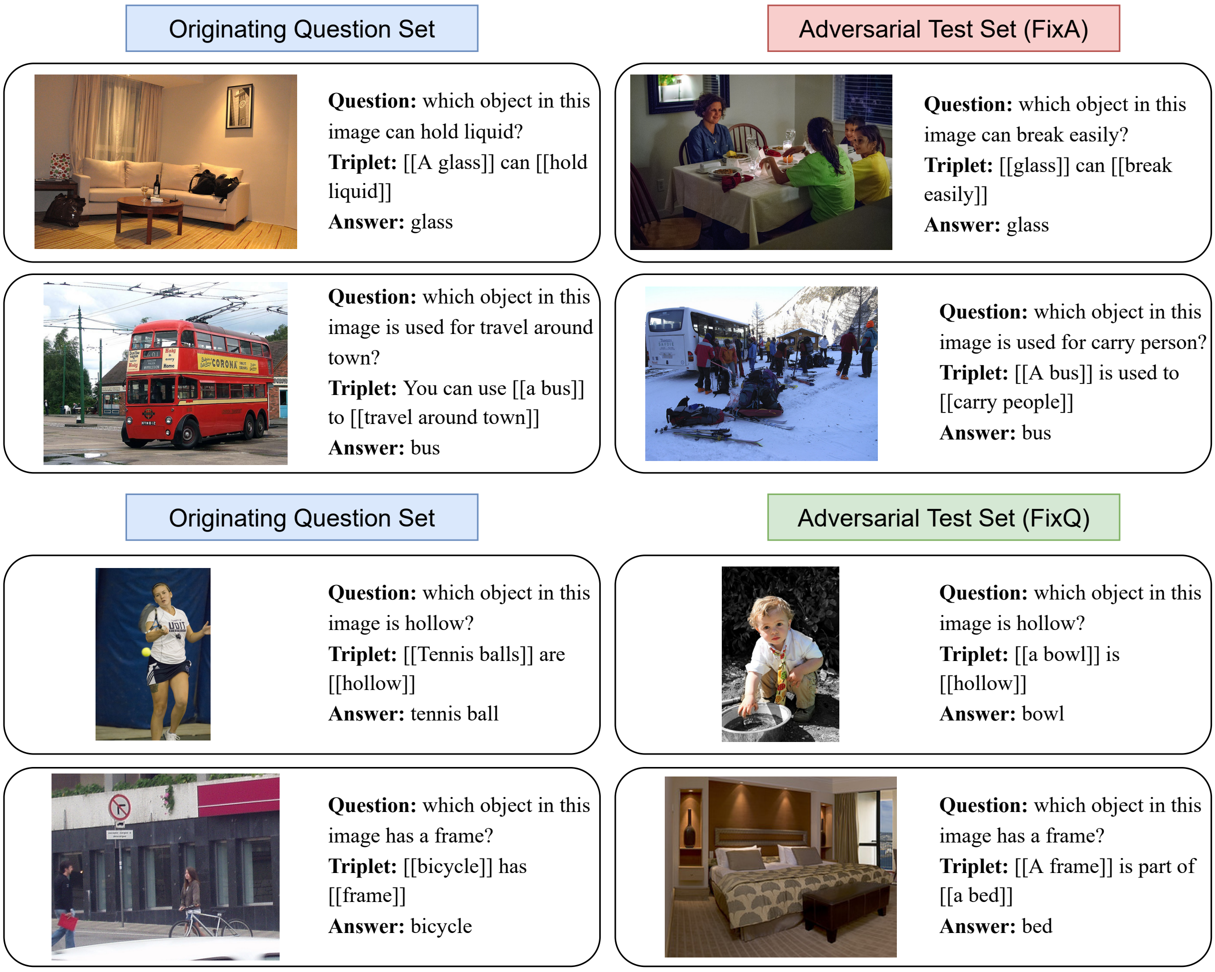
We include some additional baseline performance in Table 3. It can be easily seen that the performance on originating questions (the original FVQA questions that are used to derive the adversarial samples) is very high even when images are excluded. This further supports our argument that the original dataset is heavily biased to frequent answers. The performance on the adversarial set is lower, showing that this new test set is more challenging and less biased toward language patterns.

E More Examples of FVQA 2.0

We demonstrate some more examples from the new Adversarial Test Set in Fig. 4.

Models	Standard Test Set	Originating Question Set	Adversarial Test Set
RAVQA-DPR	69.56 ±0.78	87.52 ±1.68	76.91 ±1.93
(without triplets)	66.19 ±1.15	84.59 ±1.24	71.48 ±2.08
(without images)	43.83 ±0.68	57.53 ±2.93	50.02 ±1.00
(without triplets and images)	40.29 ±1.60	51.41 ±3.25	42.55 ±0.90

Table 3: The performance of some additional baseline systems on the standard test set, originating questions (from which the adversarial questions are derived), and adversarial test set. Results are reported as the average of 5 folds with standard deviations.



Question: which object in this image can hold liquid?
Triplet: [[A glass]] can [[hold liquid]]
Answer: glass

Question: which object in this image can break easily?
Triplet: [[glass]] can [[break easily]]
Answer: glass

Question: which object in this image is used for travel around town?
Triplet: You can use [[a bus]] to [[travel around town]]
Answer: bus

Question: which object in this image is used for carry person?
Triplet: [[A bus]] is used to [[carry people]]
Answer: bus

Question: which object in this image is hollow?
Triplet: [[Tennis balls]] are [[hollow]]
Answer: tennis ball

Question: which object in this image is hollow?
Triplet: [[a bowl]] is [[hollow]]
Answer: bowl

Question: which object in this image has a frame?
Triplet: [[bicycle]] has [[frame]]
Answer: bicycle

Question: which object in this image has a frame?
Triplet: [[A frame]] is part of [[a bed]]
Answer: bed

Figure 4: More examples taken from the FVQA 2.0 adversarial test set. The questions in the left column are from the official FVQA test set. They are used to derive the adversarial questions in the right column. FixA: the answer remains the same while the way of asking for the answer is different; FixQ: the question remains the same, but the answer changes in a different image. More details are presented in Sec. 1.