

Automatic Generation of Contrast Sets from Scene Graphs: Probing the Compositional Consistency of GQA

Yonatan Bitton[◇] Gabriel Stanovsky[◇] Roy Schwartz[◇] Michael Elhadad[♣]

[◇]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel

[♣] Department of Computer Science, Ben Gurion University, Beer Sheva, Israel

{yonatanbitton,gabis,roys}@cs.huji.ac.il elhadad@cs.bgu.ac.il

Abstract

Recent works have shown that supervised models often exploit data artifacts to achieve good test scores while their performance severely degrades on samples outside their training distribution. *Contrast sets* (Gardner et al., 2020) quantify this phenomenon by perturbing test samples in a minimal way such that the output label is modified. While most contrast sets were created manually, requiring intensive annotation effort, we present a novel method which leverages rich semantic input representation to *automatically* generate contrast sets for the visual question answering task. Our method computes the answer of perturbed questions, thus vastly reducing annotation cost and enabling thorough evaluation of models’ performance on various semantic aspects (e.g., spatial or relational reasoning). We demonstrate the effectiveness of our approach on the popular GQA dataset (Hudson and Manning, 2019) and its semantic scene graph image representation. We find that, despite GQA’s compositionality and carefully balanced label distribution, two strong models drop 13–17% in accuracy on our automatically-constructed contrast set compared to the original validation set. Finally, we show that our method can be applied to the *training* set to mitigate the degradation in performance, opening the door to more robust models.¹

1 Introduction

NLP benchmarks typically evaluate in-distribution generalization, where test sets are drawn *i.i.d* from a distribution similar to the training set. Recent works showed that high performance on test sets sampled in this manner is often achieved by exploiting systematic gaps, annotation artifacts, lexical cues and other heuristics, rather than learning meaningful task-related signal. As a result,

¹Our contrast sets and code are available at <https://github.com/yonatanbitton/AutoGenOfContrastSetsFromSceneGraphs>.

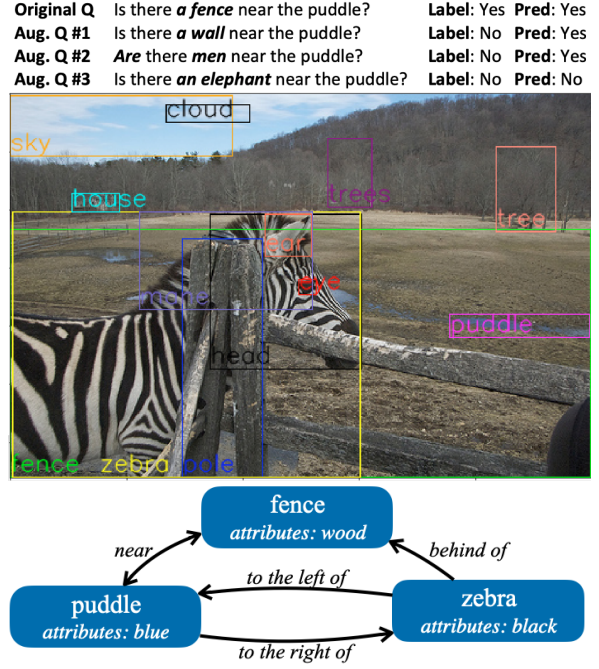


Figure 1: Illustration of our approach based on an example from the GQA dataset. Top: QA pairs and an image annotated with bounding boxes from the scene graph. Bottom: relations among the objects in the scene graph. First line at the top is the original QA pair, while the following 3 lines show our *perturbed* questions: replacing a single element in the question (*a fence*) with other options (*a wall*, *men*, *an elephant*), leading to a change in the output label. For each QA pair, the LXMERT predicted output is shown.

the out-of-domain performance of these models is often severely deteriorated (Jia and Liang, 2017; Ribeiro et al., 2018; Gururangan et al., 2018; Geva et al., 2019; McCoy et al., 2019; Feng et al., 2019; Stanovsky et al., 2019). Recently, Kaushik et al. (2019) and Gardner et al. (2020) introduced the *contrast sets* approach to probe out-of-domain generalization. Contrast sets are constructed via minimal modifications to test inputs, such that their label is modified. For example, in Fig. 1, replacing “a fence” with “a wall”, changes the answer

from “Yes” to “No”. Since such perturbations introduce minimal additional semantic complexity, robust models are expected to perform similarly on the test and contrast sets. However, a range of NLP models severely degrade in performance on contrast sets, hinting that they do not generalize well (Gardner et al., 2020). Except two recent exceptions for textual datasets (Li et al., 2020; Rosenman et al., 2020), contrast sets have so far been built manually, requiring extensive human effort and expertise.

In this work, we propose a method for *automatic* generation of large contrast sets for visual question answering (VQA). We experiment with the GQA dataset (Hudson and Manning, 2019). GQA includes semantic scene graphs (Krishna et al., 2017) representing the spatial relations between objects in the image, as exemplified in Fig. 1. The scene graphs, along with functional programs that represent the questions, are used to *balance* the dataset, thus aiming to mitigate spurious dataset correlations. We leverage the GQA scene graphs to create contrast sets, by automatically computing the answers to question perturbations, e.g., verifying that there is no wall near the puddle in Fig. 1.

We create automatic contrast sets for 29K samples or $\approx 22\%$ of the validation set. We manually verify the correctness of 1,106 of these samples on Mechanical Turk. Following, we evaluate two leading models, LXMERT (Tan and Bansal, 2019) and MAC (Hudson and Manning, 2019) on our contrast sets, and find a 13–17% reduction in performance compared to the original validation set. Finally, we show that our automatic method for contrast set construction can be used to improve performance by employing it during *training*. We augment the GQA training set with automatically constructed *training contrast sets* (adding 80K samples to the existing 943K in GQA), and observe that when trained with it, both LXMERT and MAC improve by about 14% on the contrast sets, while maintaining their original validation performance.

Our key contributions are: (1) We present an automatic method for creating contrast sets for VQA datasets with structured input representations; (2) We automatically create contrast sets for GQA, and find that for two strong models, performance on the contrast sets is lower than on the original validation set; and (3) We apply our method to augment the training data, improving both models’ performance on the contrast sets.

2 Automatic Contrast Set Construction

To construct automatic contrast sets for GQA we first identify a large subset of questions requiring specific reasoning skills (§2.1). Using the scene graph representation, we perturb each question in a manner which changes its gold answer (§2.2). Finally, we validate the automatic process via crowdsourcing (§2.3).

2.1 Identifying Recurring Patterns in GQA

The questions in the GQA dataset present a diverse set of modelling challenges, as exemplified in Table 1, including object identification and grounding, spatial reasoning and color identification. Following the contrast set approach, we create perturbations testing whether models are capable of solving questions which require this skill set, but that diverge from their training distribution.

To achieve this, we identify commonly recurring question templates which specifically require such skills. For example, to answer the question “*Are there any cats near the boat?*” a model needs to identify objects in the image (*cats, boat*), link them to the question, and identify their relative position.

We identify six question templates, testing various skills (Table 1). We abstract each question template with a regular expression which identifies the question types as well as the physical objects, their attributes (e.g., colors), and spatial relations. Overall, these regular expressions match 29K questions in the validation set ($\approx 22\%$), and 80K questions in the training set ($\approx 8\%$).

2.2 Perturbing Questions with Scene Graphs

We design a perturbation method which guarantees a change in the gold answer for each question template. For example, looking at Fig. 2, for the question template *are there X near the Y?* (e.g., “Is there any fence near the players?”), we replace either X or Y with a probable distractor (e.g., replace “fence” with “trees”).

We use the scene graph to ensure that the answer to the question is indeed changed. In our example, this would entail grounding “players” in the question to the scene graph (either via exact match or several other heuristics such as hard-coded lists of synonyms or co-hyponyms), locating its neighbors, and verifying that none of them are “trees.” We then apply heuristics to fix syntax (e.g., changing from singular to plural determiner, see Appendix A.3), and verify that the perturbed sample

Question template	Tested attributes	Example
On which side is the <i>X</i> ?	Relational (left vs. right)	On which side is the <i>dishwasher</i> ? → On which side are the <i>dishes</i> ?
What color is the <i>X</i> ?	Color identification	What color is the <i>cat</i> ? → What color is the <i>jacket</i> ?
Do you see <i>X</i> or <i>Y</i> ?	Compositions	Do you see <i>laptops</i> or cameras? → Do you see <i>headphones</i> or cameras?
Are there <i>X</i> near the <i>Y</i> ?	Spatial, relational	Are there any <i>cats</i> near the boat? → Is there any <i>bush</i> near the boat?
Is the <i>X Rel</i> the <i>Y</i> ?		Is the boy to the <i>right</i> of the man? → Is the boy to the <i>left</i> of the man?
Is the <i>X Rel</i> the <i>Y</i> ?		Is the <i>boy</i> to the right of the man? → Is the <i>zebra</i> to the right of the man?

Table 1: Question templates with original question examples, and generated perturbations modifying the answer. Italic text indicates variables, bold text indicates the perturbed atoms.

does not already exist in GQA. The specific perturbation is performed per question template. In question templates with two objects (*X* and *Y*), we replace *X* with *X'*, such that *X'* is correlated with *Y* in other GQA scene graphs. In question templates with a single object *X*, we replace *X* with a textually-similar *X'*. For example in the first row in Table 1 we replace *dishwasher* with *dishes*. Our perturbation code is publicly available.

This process may yield an arbitrarily large number of contrasting samples per question, as there are many candidates for replacing objects participating in questions. We report experiments with up to 1, 3 and 5 contrasting samples per question.

Illustrating the perturbation process. Looking at Fig. 1, we see the scene-graph information: *objects* have bounding-boxes around them in the image (e.g., *zebra*); *Objects* have *attributes* (*wood* is an attribute of the *fence* object); and there are *relationships* between the objects (the puddle is to the *right* of the zebra, and it is *near* the fence). The original (question, answer) pair is (“is there a fence near the puddle?”, “Yes”). We first identify the question template by regular expressions: “Is there *X* near the *Y*”, and isolate *X=fence*, *Y=puddle*. The answer is “Yes”, so we know that *X* is indeed near *Y*. We then use the existing information given in the scene-graph. We search for *X'* that is not near *Y*. To achieve this, we sample a random object (*wall*), and verify that it doesn’t exist in the set of scene-graph objects. This results in a perturbed example “Is there a *wall* near the puddle?”, and now the ground truth is computed to be “No”. Consider a different example: (“Is the puddle to the left of the zebra?”, “Yes”). We identify the question template “Is the *X Rel* the *Y*”, where *X=puddle*, *Rel=to the left*, *Y=zebra*. The answer is “Yes”. Now we can easily change *Rel'=to the right*, resulting in the (question, answer) pair (“Is the puddle to the right

of the zebra?”, “No”).

We highlight the following: (1) This process is done entirely automatically (we validate it in Section 2.3); (2) The answer is deterministic given the information in the scene-graph; (3) We do not produce unanswerable questions. If we couldn’t find an alternative atom for which the presuppositions hold, we do not create the perturbed (question, answer) pair; (4) Grounding objects from the question to the scene-graph can be tricky. It can involve exact match, number match (*dogs* in the question, and *dog* in the scene-graph), hyponyms (*animal* in the question, and *dog* in the scene-graph), and synonyms (*motorbike* in the question, and *motorcycle* in the scene-graph). The details are in the published code; (5) The only difference between the original and the perturbed instance is a single atom: an object, relationship, or attribute.

2.3 Validating Perturbed Instances

To verify the correctness of our automatic process, we sampled 553 images, each one with an original and perturbed QA pair for a total of 1,106 instances ($\approx 4\%$ of the validation contrast pairs). The (image, question) pairs were answered independently by human annotators on Amazon Mechanical Turk (see Fig. 3 in Appendix A.4), oblivious to whether the question originated from GQA or from our automatic contrast set. We found that the workers were able to correctly answer 72.3% of the perturbed questions, slightly lower than their performance on the original questions (76.6%).² We observed high agreement between annotators ($\kappa = 0.679$).

Our analysis shows that the human performance difference between the perturbed questions and the original questions can be attributed to the scene

²The GQA paper reports higher human accuracy (around 90%) on their original questions. We attribute this difference to the selection of a subset of questions that match our templates, which are potentially more ambiguous than average GQA questions (see Section 3).



The *bat* the *batter* is *holding* has what color? Brown →
 The *helmet* has what color? Blue
 Is there any *fence* near the *players*? Yes →
 Are there any *trees* near the *players*? No
 Do you see either *bakers* or *photographers*? No →
 Do you see either *spectators* or *photographers*? Yes
 Is the *catcher* to the *right* of an *umpire*? No →
 Is the *catcher* to the *right* of a *batter*? Yes
 Is the *catcher* to the *right* of an *umpire*? No →
 Is the *catcher* to the *left* of an *umpire*? Yes

Figure 2: GQA image (left) with example perturbations for different question templates (right). Each perturbation aims to change the label in a predetermined manner, e.g., from “yes” to “no”.

Model	Training set	Original	Augmented
MAC	Baseline	64.9%	51.5%
	Augmented	64.4%	68.4%
LXMERT	Baseline	83.9%	67.2%
	Augmented	82.6%	77.2%

Table 2: Model accuracy on the original validation set and on our generated contrast sets with maximum of 5 augmentations. *Baseline* refers to the original models, *augmented* refers to the models trained with our augmented training contrast sets.

	MAC		LXMERT	
	Original	Aug.	Original	Aug.
On which side is the <i>X</i> ?	68%	57%	94%	81%
What color is the <i>X</i> ?	49%	49%	69%	62%
Are there <i>X</i> near the <i>Y</i> ?	85%	66%	98%	79%
Do you see <i>X</i> or <i>Y</i> ?	88%	53%	95%	65%
Is the <i>X</i> <i>Rel</i> the <i>Y</i> ?	85%	44%	96%	69%
Is the <i>X</i> <i>Rel</i> the <i>Y</i> ?	71%	38%	93%	55%
Overall	65%	52%	84%	67%

Table 3: Model accuracy on the original and augmented validation set by question template for a maximum 5 augmentations per instance.

graph annotation errors in the GQA dataset: 3.5% of the 4% difference is caused by a discrepancy between image and scene graph (objects appearing in the image and not in the graph, and vice versa). Examples are available in Fig. 5 in Appendix A.5.

3 Experiments

We experiment with two top-performing GQA models, MAC (Hudson and Manning, 2018) and LXMERT (Tan and Bansal, 2019),³ to test their generalization on our automatic contrast sets, leading to various key observations.

Models struggle with our contrast set. Table 2 shows that despite GQA’s emphasis on dataset balance and compositionality, both MAC and LXMERT degraded on the contrast set: MAC 64.9% → 51.5% and LXMERT 83.9% → 67.2%, compared to only 4% degradation in human performance. Full breakdown of the results by template is shown in Table 3. As expected, question templates that reference two objects (*X* and *Y*) result in larger performance drop compared to those containing a single object (*X*). Questions about colors

had the *smallest* performance drop, potentially because the models performance on such multi-class, subjective questions is relatively low to begin with.

Training on perturbed set leads to more robust models. Previous works tried to mitigate spurious datasets biases by explicitly balancing labels during dataset construction (Goyal et al., 2017; Zhu et al., 2016; Zhang et al., 2016) or using adversarial filtering (Zellers et al., 2018, 2019). In this work we take an inoculation approach (Liu et al., 2019) and augment the original GQA training set with contrast training data, resulting in a total of 1,023,607 training samples. We retrain both models on the augmented training data, and observe in Table 2 that their performance on the *contrast set* almost matches that of the original validation set, with no loss (MAC) or only minor loss (LXMERT) to original validation accuracy.⁴ These results indicate that the perturbed *training* set is a valuable signal, which helps models recognize more patterns.

Contrast Consistency. Our method can be used to generate many augmented questions by simply sampling more items for replacement (Section 2).

³MAC and LXMERT are the top two models in the GQA leaderboard with a public implementation as of the time of submission: <https://github.com/airsplay/lxmert> and <https://github.com/stanfordnlp/mac-network/>.

⁴To verify that this is not the result of training on more data, we repeated this experiment, removing the same amount of original training instances (so the final dataset size is the same as the original one), and observed very similar results.

Augmentations per instance	Contrast sets	Acc.	Consistency
1	11,263	66%	63.4%
3	23,236	67%	51.1%
5	28,968	67%	46.1%

Table 4: Accuracy and consistency results for the LXMERT model on different contrast set sizes.

This allows us to measure the *contrast consistency* (Gardner et al., 2020) of our contrast set, defined as the percentage of the contrast sets for which a model’s predictions are correct for all examples in the set (including the original example). For example, in Fig. 1 the set size is 4, and only 2/4 predictions are correct. We experiment with 1, 3, and 5 augmentations per question with the LXMERT model trained on the original GQA training set. Our results (Table 4) show that sampling more objects leads to similar accuracy levels for the LXMERT model, indicating that quality of our contrast sets does not depend on the specific selection of replacements. However, we observe that consistency drops fast as the size of the contrast sets per QA instance grows, indicating that model success on a specific instance does not mean it can generalize robustly to perturbations.

4 Discussion and Conclusion

Our results suggest that both MAC and LXMERT under-perform when tested out of distribution. A remaining question is whether this is due to model architecture or dataset design. Bogin et al. (2020) claim that both of these models are prone to fail on compositional generalization because they do not decompose the problem into smaller sub-tasks. Our results support this claim. On the other hand, it is possible that a different dataset could prevent these models from finding shortcuts. Is there a dataset that can prevent *all* shortcuts? Our automatic method for creating contrast sets allows us to ask those questions, while we believe that future work in better training mechanisms, as suggested in Bogin et al. (2020) and Jin et al. (2020), could help in making more robust models.

We proposed an automatic method for creating contrast sets for VQA datasets that use annotated scene graphs. We created contrast sets for the GQA dataset, which is designed to be compositional, balanced, and robust against statistical biases. We observed a large performance drop between the original and augmented sets. As our contrast sets

can be generated cheaply, we further augmented the GQA training data with additional perturbed questions, and showed that this improves models’ performance on the contrast set. Our proposed method can be extended to other VQA datasets.

Acknowledgements

We thank the reviewers for the helpful comments and feedback. We thank the authors of GQA for building the dataset, and the authors of LXMERT and MAC for sharing their code and making it usable. This work was supported in part by the Center for Interdisciplinary Data Science Research at the Hebrew University of Jerusalem, and research gifts from the Allen Institute for AI.

References

- Ben Bogin, Sanjay Subramanian, Matt Gardner, and Jonathan Berant. 2020. Latent compositional representations improve systematic generalization in grounded question answering. *arXiv preprint arXiv:2007.00266*.
- Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. [Misleading failures of partial-input baselines](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5533–5538, Florence, Italy. Association for Computational Linguistics.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. [Evaluating models’ local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.
- Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. [Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.
- 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.

- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Drew Arad Hudson and Christopher D. Manning. 2018. [Compositional attention networks for machine reasoning](#). In *International Conference on Learning Representations*.
- Robin Jia and Percy Liang. 2017. [Adversarial examples for evaluating reading comprehension systems](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.
- Xisen Jin, Junyi Du, Arka Sadhu, Ram Nevatia, and Xiang Ren. 2020. [Visually grounded continual learning of compositional phrases](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2018–2029, Online. Association for Computational Linguistics.
- Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.
- Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. 2020. [Linguistically-informed transformations \(LIT\): A method for automatically generating contrast sets](#). In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135, Online. Association for Computational Linguistics.
- Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. [Inoculation by fine-tuning: A method for analyzing challenge datasets](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. [Semantically equivalent adversarial rules for debugging NLP models](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.
- Shachar Rosenman, Alon Jacovi, and Yoav Goldberg. 2020. [Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3702–3710, Online. Association for Computational Linguistics.
- Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. [Evaluating gender bias in machine translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.
- Hao Tan and Mohit Bansal. 2019. [LXMERT: Learning cross-modality encoder representations from transformers](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. [SWAG: A large-scale adversarial dataset for grounded commonsense inference](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022.
- Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages
4995–5004.

A Appendix

Ethical Considerations

We created contrast sets automatically, and verified their correctness via the crowdsourcing annotation of a sample of roughly 1K instances. Section 2.3 describes the annotation process on Amazon Mechanical Turk. The images and original questions were sampled from the public GQA dataset (Hudson and Manning, 2019), in the English language. Fig. 3 in Appendix A.4 provides example of the annotation task. Overall, the crowdsourcing task resulted in ≈ 6 hours of work, which paid an average of 11USD per hour per annotator.

Reproducibility The augmentations were performed with a MacBook Pro laptop. Augmentations for the validation data takes < 1 hour per question template, and for the training data < 3 hours per question template. Overall process, < 24 hours.

The experiments have been performed with the public implementations of MAC (Hudson and Manning, 2018) and LXMERT (Tan and Bansal, 2019), models: <https://github.com/airsplay/lxmert>, <https://github.com/stanfordnlp/mac-network/>. The configurations were modified to not include the validation set in the training process. The experiments were performed with a Linux virtual machine with a NVIDIA’s Tesla V100 GPU. The training took ~ 1 -2 days in each model. Validation took ~ 30 minutes.

A.1 Generated Contrast Sets Statistics

Table 5 reports the basic statistics of automatic contrast sets generation method when applied on the GQA validation dataset. It shows the overall number of images and QA pairs that matched the 6 question types we identified. Table 6 shows the statistics per question type, indicating how productive each augmentation method is. Tables 7 and 8 shows the same statistics for the GQA Training dataset.

# Aug. QA pairs			
	Max 1	Max 3	Max 5
# Images	10,696	10,696	10,696
# QA pairs	132,062	132,062	132,062
# Aug. QA pairs	12,962	26,189	32,802
# Aug. images	6,166	6,166	6,166
% Aug. images	57.6%	57.6%	57.6%
% Aug. QA pairs	9.8%	19.8%	24.8%

Table 5: Validation data augmentation statistics

Question template	# Aug. QA pairs		
	Max 1	Max 3	Max 5
On which side is the <i>X</i> ?	2,516	4,889	5,617
What color is the <i>X</i> ?	4,608	10,424	12,414
Are there <i>X</i> near the <i>Y</i> ?	382	867	1,320
Do you see <i>X</i> or <i>Y</i> ?	1,506	4,514	7,516
Is the <i>X Rel</i> the <i>Y</i> ?	766	1,314	1,392
Is the <i>X Rel</i> the <i>Y</i> ?	1,417	1,416	1,416

Table 6: Augmentation statistics per question template for the validation data

A.2 Models Performance Breakdown by Question Type and Number of Augmentations

Table 3 shows the breakdown of the performance of the MAC and LXMERT models per question type, on both the original GQA validation set and on the augmented contrast sets on validation.

The LXMERT model has two stages of training: pre-training on several datasets (which includes GQA training and validation data) and fine-tuning. To avoid inflating results on the validation data, we re-trained the pre-training stage without the GQA data, and fine-tuned on the training sets. Table 2. We discovered lower performance on the original set ($\sim 5\%$) with both models, but the same improvement on the augmented set ($\sim 10\%$).

# Images	72,140
# QA pairs	943,000
# Aug. QA pairs	89,936
# Aug. images	43,463
% Aug. images	60.2%
% Aug. QA pairs	9.5%

Table 7: Training data augmentation statistics

A.3 Linguistic Heuristics for Questions Generation

For each question type, we select an object in the image scene graph, and update the question by substituting the reference to this object by another object. When substituting one object by another, we need to adjust the question to keep it fluent. Table 10 shows the specific linguistic rules we verify when performing this substitution.

A.4 Annotation Task for Verifying Generated Contrast Sets

Fig. 3 shows the annotation task that is shown to Turkers to validate the QA pairs generated by our method.

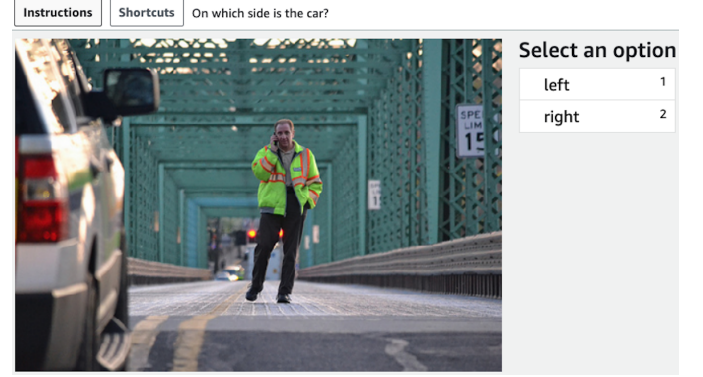


Figure 3: Example of the annotation task at the Amazon Mechanical Turk website

Question template	# Aug. QA pairs	# Aug. images	% Aug. questions
On which side is the <i>X</i> ?	17,935	16,224	2.2%
What color is the <i>X</i> ?	32,744	27,704	4.1%
Are there <i>X</i> near the <i>Y</i> ?	2,682	2,323	0.3%
Do you see <i>X</i> or <i>Y</i> ?	10,666	9,704	1.1%
Is the <i>X Rel</i> the <i>Y</i> ?	6,302	5,479	0.6%
Is the <i>X Rel</i> the <i>Y</i> ?	9,938	8,007	1.1%

Table 8: Augmentation statistics per question template for the training data

	Original Dataset			Aug. dataset								
				Max 1		Max 3			Max 5			
	Size	MAC	LXMERT	MAC	LXMERT	Size	MAC	LXMERT	Size	MAC	LXMERT	
On which side is the X ?	2,538	68%	94%	56%	79%	4,927	57%	80%	5,662	57%	81%	
What color is the X ?	4,654	49%	69%	48%	62%	10,506	49%	62%	12,498	49%	62%	
Are there X near the Y ?	382	85%	98%	72%	84%	867	69%	80%	1,320	66%	79%	
Do you see X or Y ?	1,506	88%	95%	53%	63%	4,205	53%	64%	6,679	53%	65%	
Is the X <i>Rel</i> the Y ?	766	85%	96%	42%	67%	1,314	44%	69%	1,392	44%	69%	
Is the X <i>Rel</i> the Y ?	1,417	71%	93%	38%	55%	1,417	38%	55%	1,417	38%	55%	
Overall	11,263	65%	84%	50%	66%	23,236	51%	67%	28,968	52%	67%	

Table 9: Model accuracy by question template and maximum number of augmentations.

Italic text indicates variables, bold text indicates the perturbed atoms.

A.5 Examples

Linguistic rule	Explanation	Examples
Singular vs. plural	If the noun is singular and countable: add “a” or “an” If needed, replace “Are” and “Is”	“a fence”, “men” “a boy”, “an elephant”
Definite vs. indefinite	Do not change definite articles to indefinite articles, and vice versa	“is there any fence near the boy” suggests that there is a boy in the scene graph, which is not always correct
General vs. specific	Meaning can be changed When replacing to general or specific terms	“Cats in the image” => “Animals in the image”, “Animals not in the image” => “cats not in the image”, The opposite directions not necessarily holds
Countable vs. uncountable	If the noun is uncountable, do not add “a” or “an”	“A cat”, “water”

Table 10: Partial linguistic rules to notice using our method.



	Original QA	Augmented QA
	On which side is the <i>blanket</i> ? Right	On which side is the <i>ornament</i> ? Left
	What color is the <i>teddy bear to the right of the pillow</i> ? Brown	What color is the <i>christmas lights</i> ? Yellow
Figure 4:	Is there a <i>couch</i> near the <i>blanket</i> ? Yes	Is there a <i>cat</i> near the <i>blanket</i> ? No
	Do you see a <i>pillow</i> or <i>couch</i> there? Yes	Do you see a <i>dress</i> or a <i>carpet</i> there? No
	If the <i>pillow</i> to the <i>left</i> of a <i>cat</i> ? No	Is the <i>pillow</i> to the <i>left</i> of a <i>teddy bear</i> ? Yes
	Is the <i>pillow</i> to the <i>left</i> of a <i>cat</i> ? No	No aug. - No relation between (pillow, cat)

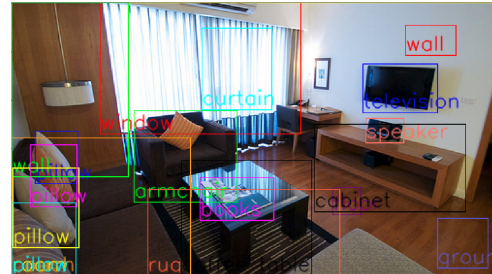


(a) First case example - multiple objects

Augmented question: On which side of the photo are the bananas?

Expected answer: right

“bananas” are annotated in green text color in the right side of the image, but it also appears in additional locations

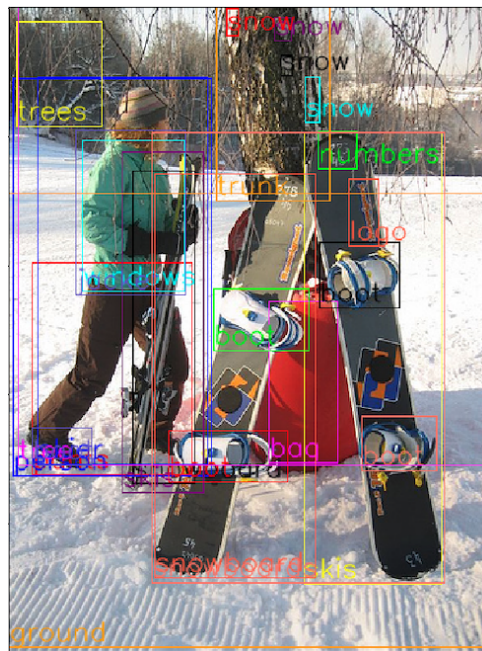


(b) Second case example - missing annotation

Augmented question: Do you see either a brown chair or couch in this picture?

Expected answer: No

We can see a couch in the left side of the image which is not annotated in the scene graph



(c) Third case example - incorrect annotation

Augmented question: Do you see either any windows or fences?

Expected answer: Yes

We can see an incorrect annotation of “windows” on the person shirt in azure text color.

Figure 5: Scene graph annotation mistakes