Sunny and Dark Outside?! Improving Answer Consistency in VQA through Entailed Question Generation

[1]Arijit Ray, [1]Karan Sikka, [1]Ajay Divakaran, [*]Stefan Lee, [1]Giedrius Burachas
[1]first.last@sri.com, [*]steflee@gatech.edu
[1]SRI International, [*]Georgia Institute of Technology

Abstract

While models for Visual Question Answering (VQA) have steadily improved over the years, interacting with one quickly reveals that these models lack consistency. For instance, if a model answers "red" to "What color is the balloon?", it might answer "no" if asked, "Is the balloon red?". These responses violate simple notions of entailment and raise questions about how effectively VQA models ground language. In this work, we introduce a dataset, ConVQA, and metrics that enable quantitative evaluation of consistency in VQA. For a given observable fact in an image (e.g. the balloon's color), we generate a set of logically consistent question-answer (QA) pairs (e.g. Is the balloon red?) and also collect a human-annotated set of common-sense based consistent QA pairs (e.g. Is the balloon the same color as tomato sauce?). Further, we propose a consistency-improving data augmentation module, a Consistency Teacher Module (CTM). CTM automatically generates entailed (or similar-intent) questions for a source QA pair and fine-tunes the VQA model if the VQA's answer to the entailed question is consistent to the source QA pair. We demonstrate that our CTM-based training improves the consistency of VQA models on the ConVQA datasets and is a strong baseline for further research.

1 Introduction

> "A skeptic, I would ask for consistency first of all."
>
> *Sylvia Plath (Plath, 2007)*

Visual Question Answering (VQA) (Antol et al., 2015) involves answering natural language questions about images. Despite the recent progress on VQA, we observe that existing methods are prone to making blatant mistakes while answering questions regarding the same visual fact but from slightly different perspectives (Figure 1). This reveals a critical limitation of the state-of-the-art models in maintaining consistency.

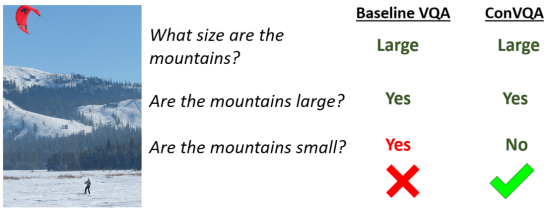
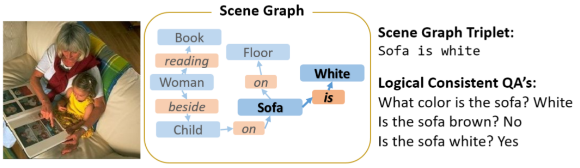


Figure 1: Current VQA models often fail at consistently answering semantically rephrased questions. To address this limitation, we construct a consistent VQA (ConVQA) dataset with diverse QA pairs that query the same visual fact. We also propose a Consistency Teacher Module (CTM) that improves VQA consistency via rewarding consistent behavior.

In particular, we motivate our definition of consistency based on classical deductive logic (Tarski and Tarski, 1994) that defines a consistent theory as one that does not entail a contradiction. Correspondingly, we define consistency, in the context of VQA, as being able to answer questions posed from different semantic perspectives about a certain fact without any contradiction. In addition, consistent Question-Answer (QA) pairs can be derived based on simple notions of logic or by commonsense reasoning. For instance, say an image contains a "large building". Logic-based QA pairs can be "is the building small? no" and "what size is the building? large". On the other hand, if an image contains "vegetarian pizza", commonsense-based QA pairs can be "is it a vegetarian pizza? yes" and "is there pepperoni on the pizza? no", which requires commonsense knowledge that "pepperoni" is not vegetarian.

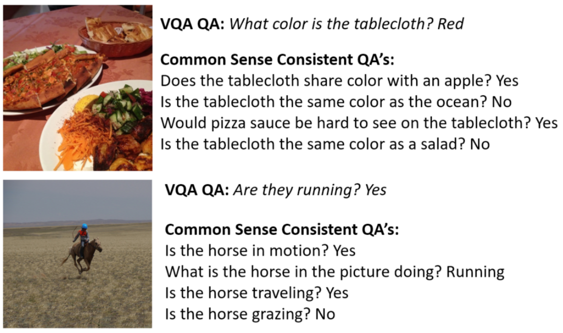
While attempts have been made to construct logic-based consistent VQA datasets (Hudson and Manning, 2019), they still fall short on commonsense-based consistency. To this end, our ConVQA Dataset consists of two subsets: 1) a challenging human-annotated set comprised of commonsense-based consistent QA's (shown in



Scene Graph Triplet:
Sofa is white

Logical Consistent QA's:
What color is the sofa? White
Is the sofa brown? No
Is the sofa white? Yes

Figure 2: Qualitative examples from our ConVQA dataset derived based on logic.



VQA QA: *What color is the tablecloth? Red*

Common Sense Consistent QA's:
Does the tablecloth share color with an apple? Yes
Is the tablecloth the same color as the ocean? No
Would pizza sauce be hard to see on the tablecloth? Yes
Is the tablecloth the same color as a salad? No

VQA QA: *Are they running? Yes*

Common Sense Consistent QA's:
Is the horse in motion? Yes
What is the horse in the picture doing? Running
Is the horse traveling? Yes
Is the horse grazing? No

Figure 3: Qualitative examples from our human-annotated commonsense-based ConVQA Dataset.

Figure 3), and 2) an automatically generated logic-based consistent QA dataset (shown in Figure 2).

To improve the consistency of VQA models, we propose a Consistency Teacher Module (CTM), which consists of a Question Generator that synthesizes entailed (or similar-intent) questions given a seed QA pair and a Consistency Checker that examines whether answers to those similar-intent questions are consistent. For training a consistent VQA model, our CTM acts as a consistency-based data augmentation scheme that trains a VQA model with consistent answers to entailed questions. We demonstrate that our approach improves the performance of a baseline VQA model on our ConVQA testing sets in terms of both accuracy and consistency. Our datasets and models will be available at https://bit.ly/32exlM7.

2 Related Work

Checking for consistency can be considered as an interrogative Turing Test (Radziwill and Benton, 2017) for linguistic robustness (Stede, 1992), (Li et al., 2017). Works such as Xu et al. (2018) explore the robustness of VQA with respect to image variations, whereas works such as Ray et al. (2016) and Mahendru et al. (2017) focus on the understanding of the premise of a question instead of relying on dataset biases (Agrawal et al., 2017) (Goyal et al., 2017) or linguistic biases (Ramakrishnan et al., 2018).

Recently, the research community has shown great interest in evaluating VQA for consistency and plausibility. GQA (Hudson and Manning, 2019) is established as a scene-graph based QA dataset. Their questions, similar to Johnson et al. (2017), require multiple hops of reasoning, and are not validated or annotated by humans. Our ConVQA differs from GQA in the following two aspects. First, we provide a human-validated test set of the automatically generated logic-based consistent QA's for a more accurate performance evaluation. Second, we collect human-annotated QA pairs based on common-sense in addition to the logic-based QA's. The most relevant work to ours is Shah et al. (2019). However, they focus strictly on question paraphrases that maintains the same answers as the source question. We, however, focus on generating questions which can have different answers, but are about the same visual fact, which greatly increases the diversity of the resulting QA pairs. To the best of our knowledge, the proposed ConVQA dataset is the first consistent QA dataset that contains human-annotated consistent QA's based on common-sense.

Other works have also looked into question generation (Zhang et al., 2016), (Mostafazadeh et al., 2016) for training better VQA models. In Misra et al. (2017), QA pairs are obtained from an oracle in a simulated environment. In contrast, our CTM-based training operates on real images and uses a learned consistency measure to train the VQA module with consistent QA's.

3 ConVQA Datasets

The consistent QA pairs in our ConVQA are generated automatically based on simple notions of logical consistency or are human-annotated using commonsense reasoning.

Logic-based Consistent QA. (L-ConVQA) Consider the Visual Genome (Krishna et al., 2017) scene graph in Figure 2 consisting of objects, attributes, and their relationships. We consider each triplet to encode a single 'visual fact', for instance, that the sofa is white. We employ slot-filler NLP techniques to generate a set of QA pairs for each triplet (object-relation-subject) in the scene graph. Currently, we focus on attribute (e.g., color, size), existential (e.g., is there) and relational (e.g., sofa on floor) consistency. We leverage WordNet (Miller, 1995) and a manually generated list of antonyms (e.g., white vs. black) and hypernyms (e.g., white → color) to generate these QA pairs.

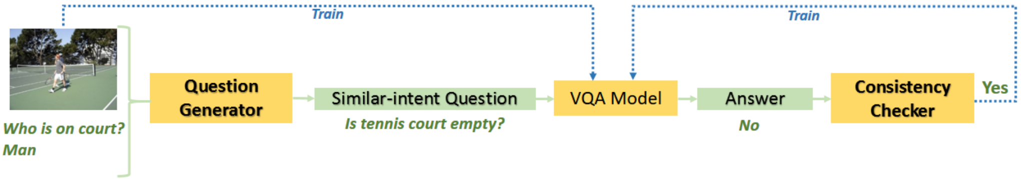


Figure 4: Block diagram of the proposed CTM including a Question Generator that synthesizes questions with similar intent and a Consistency Checker that classifies QA pairs as consistent, unrelated, or contradictory. CTM finetunes VQA models via reinforcement learning with the answer consistency as the reward to encourage VQA models to answer rephrased questions more consistently. The examples shown are from a real run.

For example, for the attribute "white" of an object "cup", we generate QA pairs such as "is the cup white? yes", "is the cup black? no" and "what color is cup? white". We also filter objects and relationships by frequency and saliency (e.g., based on bounding boxes) to avoid non-salient and infrequent objects or noisy relationships. We have a total of 880,141 QA pairs in 255,910 sets on 70,292 images. We split the data into a training set with 47,999 images, a validation set with 9,993 images, and a test set with 12,300 images. Notably, we create a smaller clean test set (12,325 QA pairs on 725 images) using Amazon Mechanical Turk (AMT) where three independent workers were asked to remove incorrect or unnatural QA's.

Commonsense-based Consistent QA. While the logic-based consistent QA set provides a first step into large-scale examination of VQA consistency, the generated questions require limited reasoning and commonsense and are, therefore, frequently simpler than human-annotated ones. Hence, we collect more challenging QA pairs based on commonsense (**CS-ConVQA**) by asking AMT workers to write intelligent rephrases of QA pairs sampled from the VQA2.0 (Goyal et al., 2017) validation dataset. AMT workers were instructed to avoid simple word paraphrases and instead to write rephrases that require commonsense reasoning in order to answer the question consistently. We collect approximately 3.5 consistent QA pairs per image for 6439 images. After filtering images that overlap with the training set of the L-ConVQA subset, we split this data into a training set (1568 images), a validation set (450 images), and a test set (1590 images).

4 Approach

To improve VQA consistency, we propose training a VQA model using a Consistency Teacher Module (CTM) that generates entailed questions and performs a consistency-based data augmentation. More specifically, CTM consists of two trainable components – an entailed question generator and a consistency checker.

Entailed Question Generator. For a given a source question-answer pair, we define entailed questions as those for which the answer should be obvious given the source QA pair. For example, given the source QA pair "Who is on court? Man", an appropriate entailed question might be "Is the tennis court empty?". We train a question generation model that given representations of the image and the source QA pair, generates a new question. Specifically, our question generator concatenates the deep features of an image (extracted using a ResNet152 (He et al., 2016) network) and a QA pair (extracted using a 1-layer LSTM (Hochreiter and Schmidhuber, 1997)) to represent a visual fact. These features are fed into another LSTM model to generate a similar-intent question. We train this module on the automatically generated Logical L-ConVQA train set. We also include some closely related (according to averaged Word2Vec (Mikolov et al., 2013) distance) Visual Genome (Krishna et al., 2017) QA pairs in the training of the question generator to add some diversity to the generated questions.

Consistency Checker. Once the VQA model produces an answer for the generated entailed question, it may or may not be consistent with the source question. To evaluate this and provide feedback to the VQA model, we train a consistency checker that processes the image and both the source and entailed QA pairs. Similar to the question generator architecture, the consistency checker takes in deep features of the image and two QA pairs and classifies the QA pairs as consistent, inconsistent or unrelated. This model is trained using the automatically generated L-ConVQA train set alone. Inconsistent examples in the L-ConVQA set are made using simple techniques such as flipping yes/no answers and replac-

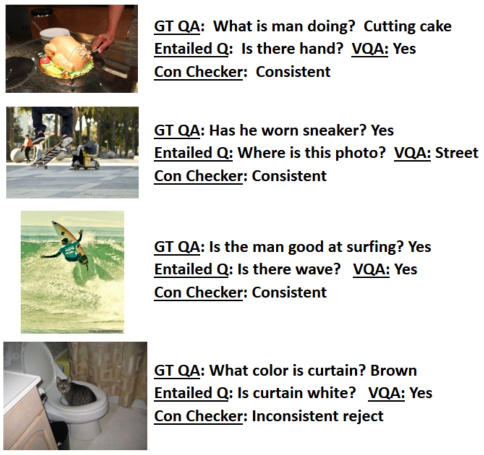

GT QA: What is man doing? Cutting cake
Entailed Q: Is there hand? VQA: Yes
Con Checker: Consistent
GT QA: Has he worn sneaker? Yes
Entailed Q: Where is this photo? VQA: Street
Con Checker: Consistent
GT QA: Is the man good at surfing? Yes
Entailed Q: Is there wave? VQA: Yes
Con Checker: Consistent
GT QA: What color is curtain? Brown
Entailed Q: Is curtain white? VQA: Yes
Con Checker: Inconsistent reject

Figure 5: Qualitative examples of entailed question generation and consistency checking for training the VQA.

ing entities in the scene graph triplets.

Consistency Teacher Module (CTM). Putting these two components together, we can train a VQA model based on its consistency on generated entailed questions. Figure 4 shows our pipeline. During training, for each source VQA QA pair ("Who is on court? Man"), we generate an entailed question ("Is tennis court empty?") and produce the VQA model's answer ("No"). We then run the consistency checker to determine if the generated answer is consistent with the source QA (in this case, "Who is on court? Man"). If the answer is consistent (and VQA confidence > 0.7), we treat it as the ground truth for the entailed question and update the VQA model as if this example were part of the original dataset. Likewise if it is deemed inconsistent, or if the question is deemed unrelated, it is unclear what the correct answer should be, so we do not update the model.

5 Experiments

To evaluate our approach, we apply the Consistency Teacher Module (CTM) module to a state-of-the-art VQA model trained on VQAv2 and evaluate performance on the ConVQA datasets. We describe training procedures, metrics, and baselines in this section.

Consistency Teacher Module Training. We train the components of CTM – Entailed Question Generation and Consistency Checker – using only the synthetic L-ConVQA train set (referred as the standard **CTM**) or a mix of Visual Genome and L-ConVQA train (referred as **CTMvg**) and keep them frozen when fine-tuning the VQA model. When we train the VQA, the sets used to finetune the VQA or seed the CTM come from splits not seen during training of the CTM - val split of L-ConVQA and train split of CS-ConVQA.

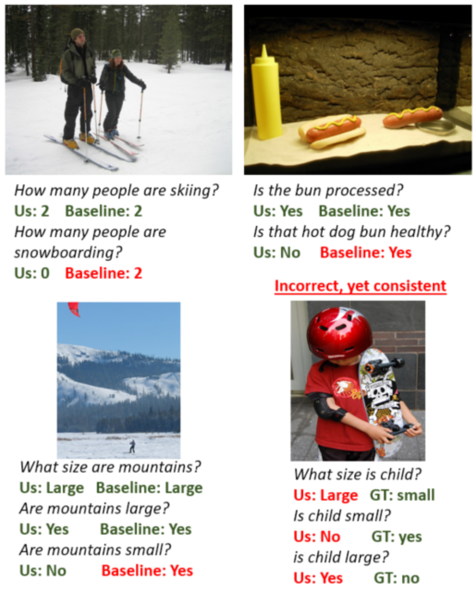


Figure 6: Qualitative examples of our improved VQA consistency by CTM training compared to baseline bottom-up VQA. Green is correct, Red is wrong. GT is Ground Truth.

Qualitative examples of our Entailed Question Generator trained only on L-ConVQA are shown in Figure 5. Despite only being trained on the automatically-generated L-ConVQA data, it generates reasonably well-entailed questions on human-annotated questions.

Our Consistency Checker has a high accuracy of classification on the L-ConVQA test set (90%). However, when tested with a mix of commonsense-based CS-ConVQA, the accuracy drops to 64% (chance is 33% for 3 classes). During training the VQA using the pre-trained Consistency Checker, precision is more important. Hence, we use the classifier at above 90% confidence threshold, where the precision is 70.38%.

Evaluation Metrics for ConVQA. We report three metrics for ConVQA – capturing notions of consistency and performance.

- **Perfect-Consistency (Perf-Con).** A model is perfectly consistent for a question set if it answers all questions in the set correctly. We report the percentage of such sets as Perf-Con.
- **Average Consistency (Avg-Con).** We also report the average accuracy within a consistent question set over the entire dataset as Avg-Con.
- **Accuracy (top-1).** Finally, we report the top-1 accuracy over all questions in the dataset.

Baselines. We compare to a number of baseline models to put our CTM results into context:

Table 1: Performance comparison of baseline VQA trained on VQA2.0, baseline VQA finetuned on ConVQA, and VQA trained using our CTM. **L-ConVQA** is the human-cleaned Logical Consistent QA dataset, **CS-ConVQA** is the human annotated Common-sense Consistency Dataset and **VG** is Visual Genome. CTM-based training produces the best results in terms of overall accuracy and consistency. **DATA** denotes the data used to fine-tune VQA or seed the CTM question generator.

		DATA	**L-ConVQA**			**CS-ConVQA**				
			Perf Con	**Avg Con**	**Top1**	**Perf Con**	**Avg Con**	**Top1**	**Yes/No**	**Num**
a)	VQA	VQA2.0	36.25	71.36	70.34	26.13	59.61	60.03	65.49	31.39
b)	FineTune	CS-ConVQA	34.54	70.39	69.48	26.39	59.65	60.07	65.80	35.92
c)	FineTune	L/CS-ConVQA	**54.68**	**83.42**	**83.16**	24.70	59.30	59.60	65.14	33.33
d)	**+CTM**	L/CS-ConVQA	54.6	83.23	82.79	25.94	**60.39**	**60.78**	**66.63**	**36.89**
e)	FineTune	L/CS-ConVQA,VG	36.40	71.60	70.94	25.22	59.19	59.56	65.30	31.39
f)	**+CTMvg**	L/CS-ConVQA,VG	51.41	81.66	81.37	**27.49**	59.75	60.15	66.41	34.95

– **VQA.** We take the bottom-up top-down VQA model (Anderson et al., 2018) as our base model for these experiments. To evaluate consistency in existing models, we present results on ConVQA of this model pretrained on VQA2.0.

– **Finetuned models:** We present results for models finetuned on ConVQA and Visual Genome – **Finetune CS-ConVQA** finetuned on the commonsense ConVQA dataset, **Finetune L/CS-ConVQA** on both logical and commonsense ConVQA, and **Finetune L/CS-ConVQA,VG** extending to Visual Genome questions.

When we apply our CTM model to the finetuned baselines above, we seed the question generator with the associated dataset.

6 Results and Analysis

Table 1 shows quantitative results on our L-ConVQA and CS-ConVQA datasets. We make a number of observations below.

The state-of-the-art VQA has low consistency. The baseline VQA system **(row a)** retains similarly high top-1 accuracy on the ConVQA splits (63.58% on VQAv2 vs 70.34% / 60.03% on L-ConVQA / CS-ConVQA); however, it achieves only 26.13% perfect consistency on the human generated CS-ConVQA questions.

Finetuning is an effective strategy for the synthetic L-ConVQA split. Finetuning on L-ConVQA train results in 18.43% gains in perfect consistency on L-ConVQA test (**row c vs a**). This is unsurprising given the templated questions and simple concepts in L-ConVQA; however, perfect consistency is low in absolute terms at 54.68%.

Finetuning does not lead to significant gains in consistency for human-generated questions. Finetuning the VQA model on CS-ConVQA **(row b)** leads to an improvement in consistency of only 0.26%. Likewise, adding L-ConVQA **(row c)** and extra Visual Genome questions **(row e)** actually reduces consistency.

CTM-based training preserves or improves consistency when leveraging additional data. When we apply CTM to the Finetuned L/CS-ConVQA model, we improve CS-ConVQA perfect consistency by 1.24% **(row d vs c)** while modestly improving other metrics. Extending to Visual Genome questions, the CTM augmented model improves perfect consistency in CS-ConVQA by 2.27% over the finetuned model **(row f vs e)**. Interestingly, the CTM modules were never trained with the human-annotated CS-ConVQA questions and yet lead to this improvement on CS-ConVQA by acting as an intelligent data augmenter/regularizer.

7 Conclusion and Discussion

In this paper, we introduced a ConVQA dataset consisting of logic-based and commonsense-based consistent QA pairs about visual facts in an image. We also proposed a Consistency Teacher Module that acts as a consistency-based data augmenter to teach VQA models to answer consistently. As future work, we plan to look into improving our automatically generated consistent QA pairs using external knowledge-bases.

Acknowledgements

We thank Yi Yao, Xiao Lin, Anirban Roy, Avi Ziskind and the anonymous reviewers for their helpful comments. This research was developed under DARPA XAI. The views, opinions and/or findings expressed are those of the authors' and should not be interpreted as representing the official views/policies of the DoD / U.S. Govt.

References

Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. 2017. C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. *arXiv preprint arXiv:1704.08243*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

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 *CVPR*, volume 1, page 3.

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

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

Drew A Hudson and Christopher D Manning. 2019. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*.

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

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.

Yitong Li, Trevor Cohn, and Timothy Baldwin. 2017. Robust training under linguistic adversity. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 21–27.

Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. 2017. The promise of premise: Harnessing question premises in visual question answering. *arXiv preprint arXiv:1705.00601*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ishan Misra, Ross Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, and Laurens van der Maaten. 2017. Learning by asking questions. *arXiv preprint arXiv:1712.01238*.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*.

Sylvia Plath. 2007. *The unabridged journals of Sylvia Plath*. Anchor.

Nicole M Radziwill and Morgan C Benton. 2017. Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, pages 1541–1551.

Arijit Ray, Gordon Christie, Mohit Bansal, Dhruv Batra, and Devi Parikh. 2016. Question relevance in vqa: identifying non-visual and false-premise questions. *arXiv preprint arXiv:1606.06622*.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. *arXiv preprint arXiv:1902.05660*.

Manfred Stede. 1992. The search for robustness in natural language understanding. *Artificial Intelligence Review*, 6(4):383–414.

Alfred Tarski and Jan Tarski. 1994. *Introduction to Logic and to the Methodology of the Deductive Sciences*. 24. Oxford University Press on Demand.

Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. 2018. Fooling vision and language models despite localization and attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4951–4961.

Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2016. Automatic generation of grounded visual questions. *arXiv preprint arXiv:1612.06530*.