'Just because you are right, doesn't mean I am wrong': Overcoming a Bottleneck in the Development and Evaluation of Open-Ended Visual Question Answering (VQA) Tasks

Man Luo, Shailaja Keyur Sampat, Riley Tallman, Yankai Zeng, Manuha Vancha, Akarshan Sajja, Chitta Baral
Arizona State University, Tempe, AZ, USA
{mluo26,ssampa17,rtallman,yzeng55,mvancha,asajja,chitta}@asu.edu

Abstract

GQA (Hudson and Manning, 2019) is a dataset for real-world visual reasoning and compositional question answering. We found that many answers predicted by the best vision-language models on the GQA dataset do not match the ground-truth answer but still are semantically meaningful and correct in the given context. In fact, this is the case with most existing visual question answering (VQA) datasets where they assume only one ground-truth answer for each question. We propose Alternative Answer Sets (AAS) of ground-truth answers to address this limitation, which is created automatically using off-the-shelf NLP tools. We introduce a semantic metric based on AAS and modify top VQA solvers to support multiple plausible answers for a question. We implement this approach on the GQA dataset and show the performance improvements.

1 Introduction

One important style of visual question answering (VQA) task involves open-ended responses such as free-form answers or fill-in-the-blanks. The possibility of multiple correct answers and multi-word responses makes the evaluation of open-ended tasks harder, which has forced VQA datasets to restrict answers to be a single word or a short phrase. Despite enforcing these constraints, from our analysis of the GQA dataset (Hudson and Manning, 2019), we noticed that a significant portion of the visual questions have issues. For example, a question *"Who is holding the bat?"* has only one ground truth answer *"batter"* while other reasonable answers like *"batsman"*, *"hitter"* are not credited. We identified six different types of issues with the dataset and illustrated them in Table 1.

A large-scale human-study conducted by (Gurari and Grauman, 2017) on VQA (Antol et al., 2015) and VizWiz (Gurari et al., 2019) found that almost 50% questions in these datasets have multiple possible answers. datasets had similar observations. The above evidence suggests that it is unfair to penalize models if their predicted answer is correct in a given context but does not match the ground truth answer.

With this motivation, we leverage existing knowledge bases and word embeddings to generate Alternative Answer Sets (AAS) instead of considering visual questions to have fixed responses. Since initially obtained AAS are generated from multiple sources and observed to be noisy, we use textual entailment to verify semantic viability of plausible answers to make alternative answer sets more robust. We justify the correctness and quality of the generated AAS by human evaluation. We introduce a semantic metric based on AAS and train two vision-language models LXMERT (Tan and Bansal, 2019) and ViLBERT (Lu et al., 2019) on two datasets. The experimental results show that the AAS metric evaluates models' performances more reasonably than the old metric. Lastly, we incorporate AAS in the training phase and show that it further improves on the proposed metric. Figure 2 gives an overview of our work.

2 Related Works

We discuss related works from two aspects, dataset creation and evaluation.

Dataset Creation-Level Large-scale VQA datasets are often curated through crowd-sourcing, where open-ended ground-truths are determined by majority voting or annotator agreement. The subjectivity in crowd-sourced datasets is well-studied in human-computer interaction literature- (Gurari and Grauman, 2016, 2017; Yang et al., 2018) etc. Ray et al. (2018) suggested creating a semantically-grounded set of questions

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2766–2771
April 19 - 23, 2021.

Issue Type	Definition	%
[1] Synonym and Hypernym	Synonym or hypernym of the ground-truth can also be considered as a correct answer for a given question-image pair.	9.1
[2] Singular/Plural	Singular or plural of the ground-truth can also be considered as a correct answer for a given question-image pair.	1.0
[3] Ambiguous Objects	Question refers to an object but the image contains multiple such objects that can lead to different possible answers.	5.8
[4] Multiple Correct Answers	If a given image-question pair is not precise, annotators might have different opinion which leads to multiple correct answers	7.0
[5] Missing Object(s)	Object referred in the question is not clearly visible in image.	4.3
[6] Wrong Label	The ground-truth answer to a question-image pair is incorrect.	6.7

Table 1: Six types of issues observed in the GQA dataset, their definition and their distribution observed in manual review of 600 samples from testdev balanced split. For example of each issue type, refer Figure 1.

for consistent answer predictions. (Bhattacharya et al., 2019) analyzed VQA and VizWiz datasets to present 9-class taxonomy of visual questions that suffer from subjectivity and ambiguity. Our analysis on GQA partially overlaps with this study. GQA dataset only provides one ground truth for each question; thus, we propose AAS to extend answers by phrases with close semantic meaning as the ground-truth answer.

Evaluation-Level For open-ended VQA tasks, the standard accuracy metric can be too stringent as it requires a predicted answer to exactly match the ground-truth answer. To deal with different interpretations of words and multiple correct answers, (Malinowski and Fritz, 2014) defined a WUPS scoring from lexical databases with Wu-Palmer similarity (Wu and Palmer, 1994). (Abdelkarim et al., 2020) proposed a soft matching metric based on wordNet (Miller, 1998) and word2vec (Mikolov et al., 2013). Different from them, we incorporate more advanced NLP resources tools to generate answer sets and rely on textural entailment to validate semantics for robustness. We propose a new metric to evaluate a system's response.

3 Analysis of GQA Dataset

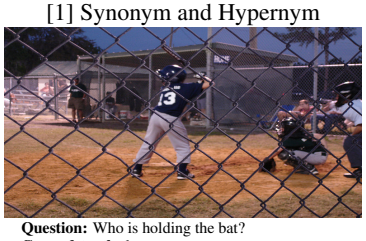
GQA is a dataset for real-world visual reasoning and compositional question answering. Instead of human annotation, answers to the questions in GQA are generated from the scene graphs of images. We found that automatic creation leads to flaws in the dataset; thus, we manually analyze 600 questions from the testdev balanced split of GQA dataset, and identify six issues shown in Table 1. Figure 1 shows examples of each type of issue.

These issues are caused by (not limited to) three reasons. First, the dataset assumes only one ground truth so that other answers with semantic closed meaning are ignored. We propose AAS to address this issue to some extent and describe AAS in the next section. Second, some questions referring to multiple objects cause ambiguous meaning. We leverage scene graphs to address this issue and found 2.92% and 2.94% ambiguous questions in balanced training split and balanced validation split, respectively. These ambiguous questions can be removed from the dataset. Third, there are incorrect scene graph detections so that some questions and/or labels do not match with the given images. We plan to address these issues in our future work.

4 Alternative Answer Set

To credit answers with semantically close meaning as the ground-truth, we propose a workflow that can be visualized from Figure 2. Each item in VQA dataset consists of <I, Q, GT>, where I is an image, Q is a question, and GT is a ground-truth answer. We define an Alternative Answer Set (AAS) as a collection of phrases $\{A_1, A_2, A_3, .., A_n\}$ such that A_i replaced with GT is still a valid answer to the given Image-Question pair. We construct AAS for each unique ground-truth automatically from two knowledge bases: Wordnet (Miller, 1998) and ConcpetNet (Liu and Singh, 2004), two word embeddings: BERT (Devlin et al., 2018) and counter-fitting (Mrkšić et al., 2016). We assign a semantic score to each alternative answer by textural entailment and introduce the AAS metric.

[1] Synonym and Hypernym

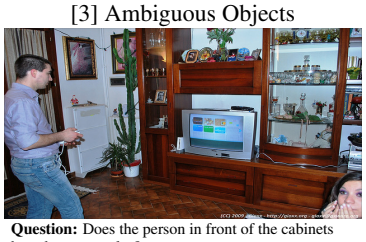


Question: Who is holding the bat?
Ground-truth: batter
Explanation: 'batsman' is a synonym of 'batter'
SU-AAS: batter, batsman, hitter, ballplayer, player

[2] Singular/Plural

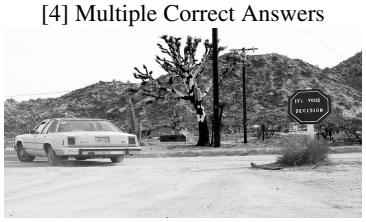
Question: Who is wearing the dress?
Ground-truth: women
Explanation: singular form 'woman' is also correct (there is only one woman in the picture)
SU-AAS: women, female, woman, people, adult female

[3] Ambiguous Objects



Question: Does the person in front of the cabinets have brunette color?
Ground-truth: Yes
Explanation: there are two people in front of the cabinet and it is not clear which person is being referred to
SU-AAS: yes

[4] Multiple Correct Answers



Question: Which place is it?
Ground-truth: road
Explanation: some person might answer 'road' and some might answer 'street'
SU-AAS: road, street, roadway, paved

[5] Missing Object(s)

Question: Does the marker look narrow?
Ground-truth: Yes
Explanation: the 'marker' is missing from the image
SU-AAS: yes

[6] Wrong Label

Question: Do the door and the logo have the same color?
Ground-truth: Yes
Explanation: the correct answer is 'no' as the door is white and logo is green.
SU-AAS: yes

Figure 1: Examples from GQA dataset for each issue type and SU-AAS i.e. AAS of ground-truth based on semantic union approach. SU-AAS can resolve Synonym and Hypernym, Singular/Plural, and Multiple Correct Answers for a given problem.

4.1 Semantic Union AAS

We take a union of four methods to find all alternative answers. For example, "stuffed animal" is semantic similar to "teddy bear", which appears in the AAS based on BERT but not in WordNet. However, the union might include phrases that we want to distinguish from the label like "man" is in the AAS of "woman" when using the BERT-based approach. For this reason, we employ the textural entailment technique to compute a semantic score of each alternative answer. For each label, we first obtain 50 sentences containing the ground-truth label from GQA dataset. We take each sentence as a premise, replace the label in this sentence with a phrase in its AAS as a hypothesis to generate an entailment score between 0-1. Specifically, we use publicly available RoBERTa (Liu et al., 2019) model trained on SNLI (Stanford Natural Language Inference) (Bowman et al., 2015) dataset for entailment computation. The semantic score of the alternative answer is the average of 50 entailment scores. If the semantic score is lower than the threshold of 0.5, then this alternative answer is thrown out. We choose 0.5 since it is the middle of 0 and 1.

Lastly, we sort the AAS by semantic score and keep the top K in the semantic union AAS, annotated by SU-AAS. We experiment with different values of K from 2 to 10, and decide K to be 6, a trade-off between accuracy and robustness. Note that the performance of textual entailment model is a contributing factor in obtaining quality AAS. Therefore, we recommend using the state-of-the-art entailment model when our proposed method is applied on other VQA datasets.

4.2 Evaluation Metric Based on AAS

We propose AAS metric and semantic score: given a question Q_i, an image I_i, the alternative answer set of GT_i denoted by S_{GT_i}, the prediction of model P_i is correct if and only if it is found in S_{GT_i}, and the score of P_i is $S_{GT_i}(P_i)$, where $S_{GT_i}(P_i)$ is the semantic score of P_i. Mathematically,

$$\text{Acc}(Q_i, I_i, S_{GT_i}, P_i) = \begin{cases} S_{GT_i}(P_i) & \text{if } P_i \in S_{GT_i} \\ 0 & \text{else} \end{cases}$$

5 Experiments

In this section, we first show that the performance of vision-language models on two datasets is improved based on the AAS metric. Then, we describe our experiment to incorporate AAS with one model on GQA dataset. Last, we verify the correctness of AAS by human evaluation.

5.1 Baseline Methods

We select two top Vision-and-Language models, ViLBERT (Lu et al., 2019) and LXMERT (Tan

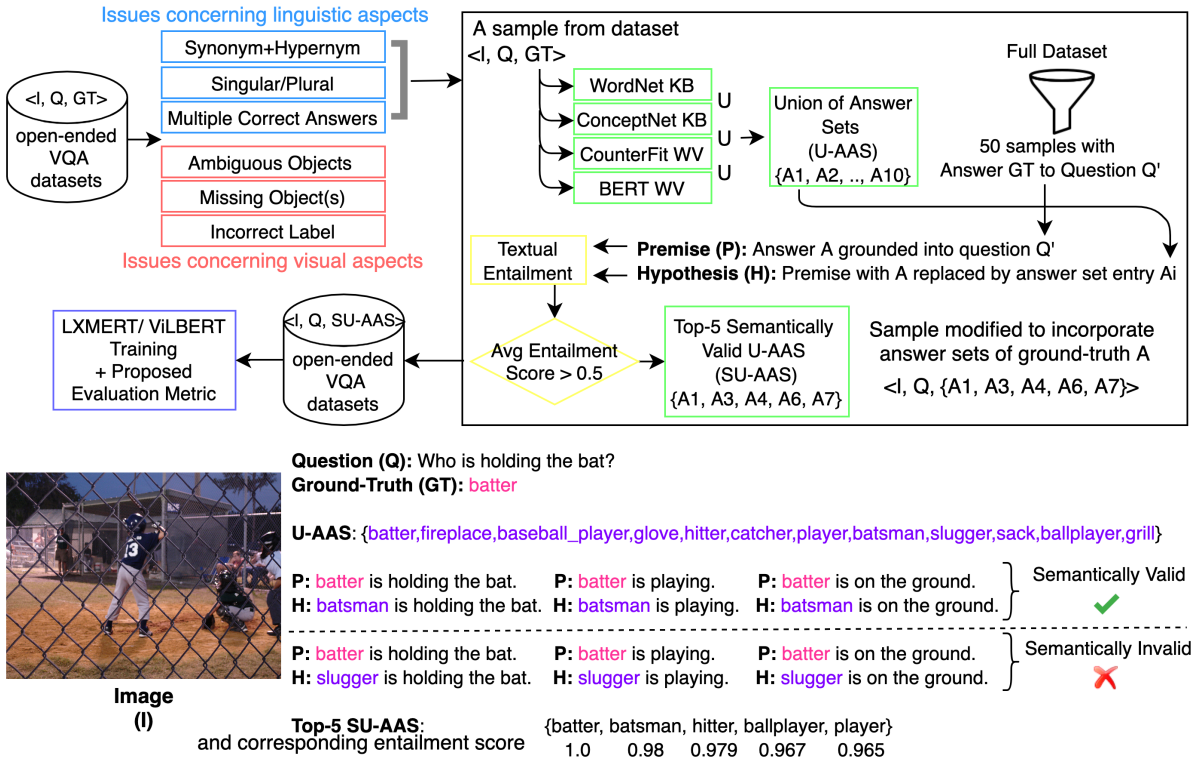


Figure 2: (top) The workflow for generating Alternative Answer Set (AAS) for VQA datasets (bottom) An example from GQA dataset showing semantically valid AAS for the answer 'batter' generated using above workflow

and Bansal, 2019) and evaluate their performances based on the AAS metric. From Table 2, we see that for the GQA dataset, LXMERT and ViLBERT have 4.49%, 4.26% improvements on union AAS metric separately. For VQA2.0 dataset, LXMERT and ViLBERT have 0.82%, 0.53% improvements on union AAS metric separately. It is expected that the improvement on VQA2.0 dataset is less than GQA since the former dataset already provides multiple correct answers. Figure 3 shows the impacts of the value K of Union AAS on the scores. From the figure, we see that when K increases from 2 to 6, the score gets increased significantly, and slightly when k increases from 6 to 9, but not increases more after K is 9. Since values 7 and 8 do not significantly improve the score, and the value 9 introduces noise, we take the top 6 as the SU-AAS.

5.2 Training with AAS

We incorporate SU-AAS of ground truth in training phase, so the model learns that more than one answer for a given example can be correct. We train LXMERT on GQA dataset with this objective.

Table 3 shows the results of LXMERT trained with AAS compared with the baseline. Not surprisingly, the performance evaluated on the original method drops because the model has a higher

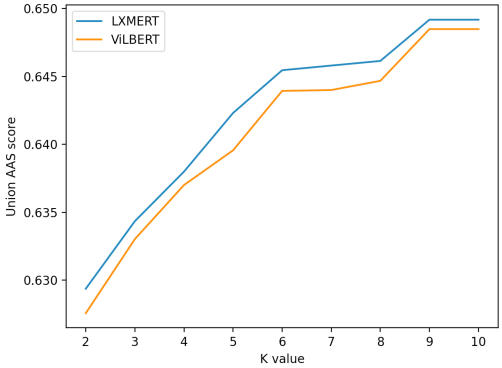


Figure 3: Union AAS score of different value of K

chance to predict answers in AAS, which are different from the ground truth, and thus the performance evaluated on SU-AAS metric increases.

Dataset	Exact Matching Accuracy		SU-AAS Accuracy	
	LXMERT	LXMERT$_{AAS}$	LXMERT	LXMERT$_{AAS}$
GQA(testdev)	**60.06**	59.02	64.55	**65.22**

Table 3: Incorporate AAS in the training phase of LXMERT (LMXERT$_{AAS}$) on GQA dataset.

5.3 Evaluation of AAS

To validate the correctness of AAS, we measure the correlation between human judgment and AAS.

Dataset	Model	Original Metric	WordNet	BERT	CounterFit	ConceptNet	Union
GQA	LXMERT	60.06	61.79	62.69	62.75	63.58	**64.55**
(testdev)	ViLBERT	60.13	61.90	62.69	62.74	63.67	**64.39**
VQA	LXMERT	69.98	70.21	70.54	70.33	70.52	**70.80**
(valid)	ViLBERT	77.65	77.82	78.10	77.93	78.06	**78.28**

Table 2: The evaluation of two models on GQA and VQA with original metric and AAS based metrics.

Specifically, for each label of GQA, we take the SU-AAS and ask three annotators to justify if alternative answers in AAS can replace the label. If the majority of annotators agree upon, we keep the answer in the AAS, remove otherwise. In this way, we collect the human-annotated AAS. We compare the human-annotated AAS with each automatically generated AAS. We take the intersection over union (IoU) score to evaluate the correlation between automatic approach and human annotation: a higher IoU score means stronger alignment.

Method	WordNet	BERT	CounterFit	ConceptNet	Union
IoU%	48.25	56.18	58.95	58.39	80.5

Table 4: The IoU scores between human annotations and AAS based on five approaches.

6 Discussion and Conclusion

To evaluate a model from a semantic point of view, we define an alternative answer set (AAS). We develop a workflow to automatically create robust AAS for ground truth answers in the dataset using Textual Entailment. Additionally, we did human verification to assess the quality of automatically generated AAS. The high agreement score indicates that entailment model is doing a careful job of filtering relevant answers. From experiments on two models and two VQA datasets, we show the effectiveness of AAS-based evaluation using our proposed metric.

AAS can be applied to other tasks, for example, machine translation. BLEU(Papineni et al., 2002) score used to evaluate machine translation models incorporates an average of n-gram precision but does not consider the synonymy. Therefore, METEOR (Banerjee and Lavie, 2005) was proposed to overcome this problem. However, METEOR only relies on the synset of WordNet to get the synonyms. Our proposed AAS has the advantage of both knowledge base and word embeddings, which would help better evaluate translation tasks.

Acknowledgments

We are thankful to Tejas Gokhale for useful discussions and feedback on this work. We also thank anonymous reviewers for their thoughtful feedback. This work is partially supported by the National Science Foundation grant IIS-1816039.

References

Sherif Abdelkarim, Panos Achlioptas, Jiaji Huang, Boyang Li, Kenneth Church, and Mohamed Elhoseiny. 2020. Long-tail visual relationship recognition with a visiolinguistic hubless loss. *arXiv preprint arXiv:2004.00436*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *IEEE ICCV*, pages 2425–2433.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. Why does a visual question have different answers?

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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

Danna Gurari and Kristen Grauman. 2016. Visual question: Predicting if a crowd will agree on the answer.

Danna Gurari and Kristen Grauman. 2017. Crowdverge: Predicting if people will agree on the answer to a visual question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3511–3522.

Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. 2019. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *IEEE CVPR*, pages 939–948.

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

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690.

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. 1998. *WordNet: An electronic lexical database*. MIT press.

Nikola Mrkšić, Diarmuid O Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. *arXiv preprint arXiv:1603.00892*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Arijit Ray, Giedrius T Burachas, Karan Sikka, Anirban Roy, Avi Ziskind, Yi Yao, and Ajay Divakaran. 2018. Make up your mind: Towards consistent answer predictions in vqa models. In *European Conference on Computer Vision (ECCV), Workshops*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, page 133–138, USA. Association for Computational Linguistics.

Chun-Ju Yang, Kristen Grauman, and Danna Gurari. 2018. Visual question answer diversity. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.