‘Just because you are right, doesn’t mean I am wrong’: Overcome a bottleneck in development and evaluation of Open-Ended VQA tasks

Anonymous EMNLP submission

Abstract

Visual question answering (VQA) systems aim at responding to natural language questions about visual content with a valid answer. Despite the agreement or majority voting among crowd-workers, a significant portion of visual questions have been observed to be subjective and/or ambiguous. Previous work has analyzed many VQA examples from popular datasets and found that people provide multiple different answers in about half of the questions. This makes the evaluation of open-ended VQA tasks far more challenging. To address this challenge, we propose Alternative Answer Sets (AAS) for such visual questions curated using existing NLP tools and techniques. We then modify best VQA solvers to support multiple plausible answers for a visual question and show the performance improvement over the GQA and VQA datasets.

1 Introduction

In recent years, a large body of visual question answering (VQA) datasets have been proposed and compiled to evaluate the ability of AI systems to understand images by asking questions in natural language. VQA datasets have demonstrated two major question-answering (QA) styles. One style is modelling QA as a classification problem with multiple-choice or identifying relational tuples where output space is mutually exclusive. Another style uses open-ended Q such as free-form answers or fill-in-the-blank.

The possibility of multiple correct answers and multi-word responses makes evaluating open-ended tasks harder, which has forced VQA datasets to restrict answers to be a single word or very short phrase. Despite enforcing these constraints, based on our analysis of the GQA dataset (Hudson and Manning, 2019), we noticed that a significant portion of visual questions suffer from problems of subjectivity and ambiguity, as per examples provided in Figure 1. A large-scale human-study conducted by (Gurari and Grauman, 2017) on VQA (Antol et al., 2015) and VizWiz (Gurari et al., 2019) datasets had a similar observation, where they found almost 50% questions with muliple possible answers. Both of the above evidences suggest that ‘just because crowd-workers have agreed upon a particular ground-truth answer, it is unfair to penalize other humans or AI models based on their subjectivity’.

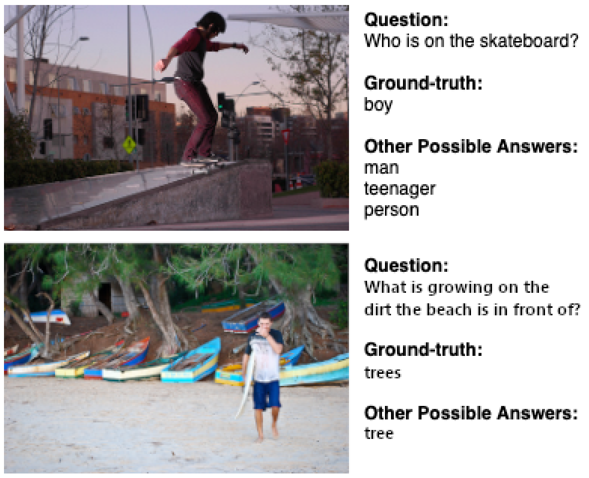

Figure 1: Examples from the GQA dataset with multiple correct answers

With this motivation, we leverage a combination of 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 combined from multiple sources and observed to be noisy, we use textual entailment to verify semantic viability of plausible answers in a given context to make alternative answer sets more robust. We then modify training objective and evaluation metric for pre-trained vision-language models- LXMERT (Tan and Bansal, 2019) and ViLBERT (Lu et al., 2019b) to incorporate AAS. Finally, we benchmark performance of modified models over GQA (Hudson and Manning, 2019) and VQA-Real Images (Antol et al., 2015) which demonstrates performance improvement by 5% and 1% respectively. We believe that this work will advance development of VQA models that can address subjectivity from a lingusitic point of view.

2 Related Works

Several works have attempted to address challenges related to open-ended VQA tasks which we categorize into 3 levels;

Dataset Creation-Level. Large-scale VQA datasets are often curated through crowd-sourcing, where open ended ground-truth answers are determined by majority voting or annotator agreement. The subjectivity in crowd-sourced datasets has been well-studied in human-computer interaction literature- (Gurari and Grauman, 2016), (Gurari and Grauman, 2017), (Yang et al., 2018) etc., which has been of interest to computer vision researchers in recent years. (Ray et al., 2018) suggested to create a semantically-grounded set of questions which leads to consistent answer predictions. (Bhattacharya et al., 2019) have conducted detailed analyses of the VQA (Goyal et al., 2017) and VizWiz (Bigham et al., 2010) datasets, and proposed a 9-class taxonomy of visual questions that might suffer from subjectivity and ambiguity. Our proposed AAS based method overcomes three taxonomies specific to subjectivity of text.

Model-Level. Several works have attempted to reduce the output space in open-ended tasks through question categorization (Mishra et al., 2020), by generating plausible answers (Bakhshandeh et al., 2016) or incorporating answer-type predictions (Kafle and Kanan, 2016) as mechanisms to combat ambiguity. Contrary, we are in favor of developing models that can handle subjectivity rather than limiting the kind of questions one can ask to a VQA system, which is a more realistic manifestation of real world natural language. (Hu et al., 2018) proposed learning of answer embeddings along with the image+question embeddings and learn best parameterization to maximize the likelihood of correct answer.

Evaluation-Level. For open-ended VQA task, use of standard accuracy metric can be too stringent as algorithm's predicted answer must exactly match the ground truth answer. To deal with different interpretations of words and multiple possible 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 and rely on sentence entailment validate semantics for robustness. However, the best way to evaluate open-ended VQA tasks remains the topic of ongoing debate and active research in AI community. Considerable work needs to be done to develop better approaches for measuring semantic similarity and handling multi-word answers in open-ended tasks.

3 Proposed Evaluation Method

Due to human bias and annotation inconsistency, flaws in the VQA dataset are well known. Like in (Bhattacharya et al., 2019), we categorize six issues of GQA dataset; the details can be found the Appendix A.1.

To combat the flaws present in VQA, we propose a semantic way to evaluate a model's accuracy. Each item in a VQA dataset consists of <I, Q, GT>, where I is an image, Q is a question and GT is a ground-truth answer(s). 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 following knowledge bases and word embeddings;

3.1 Alternative Answer Sets (AAS) Generation

WordNet. Wordnet (Miller, 1998) is a large lexical database for English language which groups distinct concepts based on their semantic and lexical relations in a network like structure. We particularly focus on Synonyms and immediate Hypernyms of the labels to generate AAS.

ConceptNet. ConcpetNet (Liu and Singh, 2004) is a database of terms and relations with a total of 34 types of relationships. We use relationships "Synonym", "IsA" and "FormOf" to obtain the phrase's synonyms, hypernyms, hyponyms, and plural forms to create AAS.

Counter-Fitted Word Vectors. Word vector methods that derive representations from co-occurrence of words from similar contexts are unable to distinguish between semantic similarity and conceptual association of words. To overcome this limitation, a counter-fitting (Mrkšić et al., 2016) method is proposed, which injects antonymy and synonymy constraints into vector space representations. We use counter-fitted embeddings with a cosine similarity threshold of 0.6 (empirically derived) to generate alternative answer sets.

BERT. Proposed by (Devlin et al., 2018), BERT is a language model generated through pre-training deep transformers in a bidirectional fashion, which achieved state-of-art results over many NLP tasks. We use contextual embeddings of phrases from BERT to extract the top 15 most similar words to the given ground-truth answer using cosine similarity for answer set expansion.

Filtered Union. Finally, we take a direct union of all four methods, and include the original label and use textual entailment to filter out irrelevant terms.

By aggregating the previous methods we hope to form a more robust set and find all possible alternative answers. However, the AAS of a label might include phrases that we want to distinguish from the label, like ”man” is in the AAS of ”woman” when using BERT-based approach. For this reason we employ a sentence entailment technique to filter incorrect terms. Specifically, we take a sentence containing the label as a premise, and then take the same sentence but replace the label with any phrase in AAS as hypothesis. If the entailment score is lower than threshold 0.5 (empirically derived), then this phrase is thrown out. Lastly, each term is sorted by its entailment score and only the top 5 are kept in the final AAS.[1] The complete algorithm can be found in Appendix A.3. Examples of different AAS based approach is shown in Appendix A.2.

3.2 Evaluation Metric Based on AAS

The accuracy based on extract matching is that given a question Q_i, an image I_i, and a ground truth label GT_i, the prediction of model P_i is correct if and only if it is exactly the same as GT_i. Instead of exact matching, we propose a new metric based on AAS: given a question Q_i, an image I_i, the alternative answer set of GT_i denoted by $\mathrm{S}_{\mathrm{GT}_i}$, the prediction of model P'_i is correct if and only if it is found in $\mathrm{S}_{\mathrm{GT}_i}$. The mathematical expression is,

$$\mathrm{Acc}(\mathrm{Q}_i, \mathrm{I}_i, \mathrm{S}_{\mathrm{GT}_i}, \mathrm{P}'_i) = \begin{cases} 1 & \text{if } \mathrm{P}'_i \in \mathrm{S}_{\mathrm{GT}_i} \\ 0 & \text{else} \end{cases} \quad (1)$$

[1]Some ground truths have less than 5 alternative answer sets.

4 Experiments

In this section, we experiment with two datasets GQA(Hudson and Manning, 2019) and VQA v2.0 (Goyal et al., 2017) and pick top performing models on these two datasets and benchmark their performance. Then we finetune two models to incorporate AAS and compare with benchmark.

4.1 Baseline Methods

ViLBERT. Vision-and-Language BERT (ViLBERT) (Lu et al., 2019a) utilized two transformer-based mechanisms (one language only single-modal and one cross-modal transformer) pretrained over Conceptual Captions (Sharma et al., 2018). In (Lu et al., 2019b), they train ViLBERT with 12 different tasks and demonstrate that multi-task training objective outperforms single task training.

LXMERT. Proposed by (Tan and Bansal, 2019), LXMERT incorporates two single-modality transformers for vision and language respectively and one cross-modal transformer. It was pretrained with large amounts of image-sentence pairs via five diverse pretraining tasks based on popular captioning and VQA datasets like COCO-Caption, VG Caption, VGQA, VQA and GQA.

4.2 Training with AAS

Instead of only using provided ground-truth, we extend ground-truth with its AAS, so the model learns that more than one answer for a given example is correct. We train LXMert on both GQA and VQA with this objective. More training setting details can be found in Appendix A.4.

GQA. Firstly,we extract 1842 unique labels from training and validation sets, and we generate the AAS of each ground truth label based on union approach. Then during training, instead of only using GT as label, we use the AAS of GT as labels to train LXMert with binary cross entropy.

VQA. Similarly, we find 3129 unique ground truths from training and validation set and create an AAS for each. Different from GQA, VQA provides

Dataset	Model	Original Metric	WordNet	BERT	CounterFit	ConceptNet	Union
GQA	LXMERT	60.06	62.08	62.95	63.03	64.31	**64.45**
(testdev)	ViLBERT	60.13	62.24	62.99	63.0	**64.43**	64.18
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 1: The evaluation of two models on GQA and VQA with original metric and AAS based metrics.

a set of labels with different scores (confidences) for each question. Inside of the AAS of each label, we pair matching alternative answers with the same score of that label. [2] We use the extended labels to train LXMert with binary cross entropy.

4.3 Results

From Table 1, the AAS-based metrics show improvement over both datasets compared to original accuracy, with GQA increasing by at least 2% and VQA by at most 0.82%.[3] LXMERT and ViLBERT show consistent improvements by AAS-based metrics.

Table 2 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 higher chance to predict answers in AAS which are different than the ground truth, and thus the performance evaluated on AAS-based metric increased.

Dataset	**Metric$_{old}$**		**Metric$_{new}$**	
	LXMERT	LXMERT$_{AAS}$	LXMERT	LXMERT$_{AAS}$
GQA(testdev)	**60.06**	51.53	64.45	**65.13**
VQA(valid)	**69.98**	53.74	70.80	**71.59**

Table 2: Incorporate AAS with LXMERT (LMXERT$_{AAS}$) and compare the results of LXMERT and LXMERT$_{AAS}$ on original metric (Metric$_{old}$) and union-based metric (Metric$_{new}$).

5 Analysis and Discussion

Analysis. From Table 3, we see that although WordNet provides more alternative answers, many of them are outside of the candidate answer set, thus the overlap between AAS and the candidate answer sets is lowest. Since both LXMERT and ViLBERT predict answer from a candidate answer set, wordNet based AAS show least improvement. We conclude that larger alternative answer set does not indicate more improvement. Both BERT and CounterFit are based on cosine similarity of vectors and sentence entailment filtering to generate AAS, therefore they show almost equal improvement on both datasets. ConceptNet-based AAS has close improvement to union-based approach demonstrating the semantic robustness of ConceptNet.

AAS	**GQA**		**VQA**	
	Avg Len	Overlap	Avg Len	Overlap
WordNet	3.62	0.41	3.18	0.40
BERT	2.23	0.64	2.12	0.64
CounterFit	1.84	0.72	1.54	0.72
ConceptNet	2.27	0.79	2.01	0.78
Union	3.67	0.78	3.31	0.79

Table 3: The average length (Avg Len) of AAS of different approaches, and the overlap ratio (Overlap) of the aas with the ground truth.

Discussion. AAS-based metrics show more improvement in GQA than VQA 2.0. We have two insights. VQA has counting questions where the answers are numbers, and in this case, the AAS-based metric has little effect when the model is based on classification and there is no alternative of numbers in the candidate answer set. Second, the AAS-based metric shows more impacts when the question falls into semantic issues, including singular/plural, synonym/hypernym, and the answers are words or short phrases.

6 Conclusion

To address the human annotation mistake and individual subjectivity, we define alternative answer set and automatically create robust AAS for ground truths in dataset. Based on AAS, we propose a semantic metric to evaluate VQA system's performance . By the experiments on two models and two VQA datasets, we show the effectiveness of AAS-based evaluation.

[2]If one phrase happens in AAS of multiple ground truth, we take the lowest score.

[3]In equation 1, for GQA, we credit models with score 1; for VQA, we credit models with the soft score.

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.

Omid Bakhshandeh, Trung Bui, Zhe Lin, and Walter Chang. 2016. Proposing plausible answers for open-ended visual question answering.

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

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

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*.

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.

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.

Hexiang Hu, Wei-Lun Chao, and Fei Sha. 2018. Learning answer embeddings for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*.

Kushal Kafle and Christopher Kanan. 2016. Answer-type prediction for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

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

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2019b. 12-in-1: Multi-task vision and language representation learning. *arXiv preprint arXiv:1912.02315*.

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.

Aakansha Mishra, Ashish Anand, and Prithwijit Guha. 2020. Cq-vqa: Visual question answering on categorized questions.

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*.

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*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

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*.

A Appendix

A.1 Six issues existing in GQA dataset

We conduct a detail analysis on GQA dataset and identify six issues from human annotations. We manually analyze the first 600 questions from test-dev balanced questions,

Issue Type	Notation	Percentage%
Ambiguous Question	multiple possible references of the object in the question	10.3
Multiple Correct Answers	more than one answers to the question	12.6
Missing Object(s) in Image	the object is invisble in the image	4.3
Synonym and hypernym	synonyms or hypernyms of labels can answer the question	9.1
Wrong Label	the ground truth is incorrect	5.8
Singular/Plural	the singular or plural of the groud truth is correct	1.0

Table 4: General issues present in the GQA dataset identified through manual review.

A.2 Examples of AAS for GQA Ground Truth

AAS	Ground Truth		
	beneath	shops	teddy bear
WordNet	to a lower place below, beneath	retail stores, shops,store, outlet,shop	teddy bears
BERT	underneath thin	shop, stores buildings,	stuffed bears tuffed bear stuffed animals
CounterFit	below, under underneath, bottom	shop, stores,store, outlet,shop	teddy bear
ConceptNet	below, under underneath,	shop, store,	stuffed animal teddy bears bears
Union	to a lower place below, beneath underneath	retail store, shops,stores, shop	stuffed animal stuffed bear stuffed bears teddy bears

Table 5: Different Alternative Answer Sets of three ground truth labels in GQA.

A.3 Textual Entailment Algorithm

To make the AAS more robust, we rely on textural entaiment approach to filer not good alternative answers found by four approaches. Algorithm 1 shows the process.

A.4 Experiment Details

A.4.1 Training

For GQA, we use balanced training set to train models. We use the default training setting of LXMERT and ViLBERT for both GQA and VQA tasks.

Algorithm 1: The textural entailment algorithm used to filter incorrect answers from an AAS

Result: Filtered AAS of a ground truth
a ground truth gt;
a list of sentence containing gt, S;
a list of candidate alternative answer AAS;
a threshold $\theta = 0.5$;
an empty set $L = \{\}$;
for aa *in* AAS **do**
  a initial $score = 0$;
  for S_i *in* S **do**
    get S_i' by replacing gt in S_i with aa ;
    call textural entailment system with (S, S');
    get $prob$ of entailment;
    $score += prob$;
  end
  if $score/len(S) > \theta$ **then**
    add aa to L with $score/len(S)$
  else
    aa is not good;
  end
end
sort L by score;
get top five;

LXMERT. In GQA, we fine tune LXMERT in 4 epochs with learning rate 1e-5. We use the validation set to save the best model. In VQA, the learning rate is 5e-5. In both training, we use batch size 32.

ViLBERT. We use the pretrained model provided by (Lu et al., 2019b). In both GQA and VQA, we fine tune ViLBERT in 20 epochs with learning rate 4e-5, batch size 32.

A.4.2 Testing

For GQA, We use the testdev set provided by LXMERT which includes 12579 questions from 398 images. For VQA, we use the validation set provided by LMXERT with includes 25994 questions from 5000 images.

A.5 Examples from GQA with Possible Disagreements or Multiple Correct Answer Possibility

Figure 2: Example Questions (Q) and Ground-truth (GT) answers from GQA with multiple correct answers.
(a) Q: What animal is standing in front of the giraffes? | GT: zebras | Some giraffes are facing some humans standing by a cart and other giraffes are facing the zebras so “human”, “humans”, “zebra”, or “zebras” are acceptable answers
(b) Q: Who is on the skateboard? | GT: boy | Hyponyms/Hypernyms such as “teenager”, “person”, and “male” are acceptable
(c) Q: What is growing on the dirt the beach is in front of? | GT: trees | There are trees growing, and there is also a tree growing, so “tree” is acceptable
(d) Q: What do you think is on the couch? | GT: pillow | Alternative responses “pillows”, “throw pillow”, and “throw pillows” are acceptable
(e) Q1: What food is brown? | GT: pancake | The bacon is also brown so “bacon” is acceptable
(e) Q2: What is the fluffy food called? | GT: eggs | The pancake also looks fluffy, and one could also answer “scrambled eggs”
(f) Q: Who is walking next to the boy on the left of the picture? | GT: girl | Alternative possibilities “woman” and “women” are acceptable, as well as “girls”, “female”, or “females”