VQD: Visual Query Detection in Natural Scenes

Manoj Acharya[1] **Karan Jariwala**[1] **Christopher Kanan**[1,2,3]
[1]Rochester Institute of Technology [2]PAIGE [3]Cornell Tech
{ma7583, kkj1811, kanan}@rit.edu

Abstract

We propose Visual Query Detection (VQD), a new visual grounding task. In VQD, a system is guided by natural language to localize a *variable* number of objects in an image. VQD is related to visual referring expression recognition, where the task is to localize only *one* object. We describe the first dataset for VQD and we propose baseline algorithms that demonstrate the difficulty of the task compared to referring expression recognition.

1 Introduction

In computer vision, object detection is the task of identifying all objects from a specific closed-set of pre-defined classes by putting a bounding box around each object present in an image, e.g., in the widely used COCO dataset there are 80 object categories and an algorithm must put a box around all instances of each object present in an image (Lin et al., 2014). Recent deep learning based models have significantly advanced the state-of-the-art in object detection (Ren et al., 2015b); however, many applications demand more nuanced detection of objects with specific attributes or objects in relation to each other. Here, we study goal-directed object detection, where the set of possible valid objects is far greater than in the typical object detection problem. Specifically, we introduce the Visual Query Detection (VQD) task (see Fig. 1). In VQD, a system is given a query in natural language and an image and it must produce 0–N boxes that satisfy that query. VQD has numerous applications, ranging from image retrieval to robotics.

VQD is related to the visual referring expression recognition (RER) task (Kazemzadeh et al., 2014); however, in RER every image has only a *single* correct box. In contrast, in VQD there could be no valid outputs for a query or multiple valid outputs, making the task both harder and more useful. As discussed later, existing RER datasets have multiple annotation problems and have significant language bias problems. VQD is also related to Visual Question Answering (VQA), where the task is to answer questions about images in natural language (Malinowski and Fritz, 2014; Antol et al., 2015). The key difference is that in VQD the algorithm must generate image bounding boxes that satisfy the query, making it less prone to the forms of bias that plague VQA datasets.

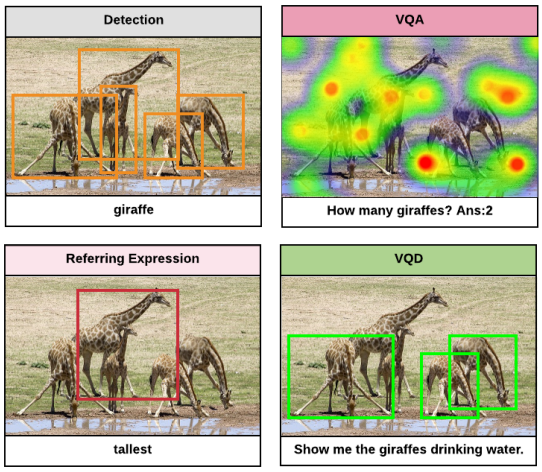


Figure 1: Unlike VQD, object detection cannot deal with attributes and relations. In VQA, often algorithms produce the right answers due to dataset bias without 'looking' at relevant image regions. RER datasets have short and often ambiguous prompts, and by having only a single box as an output, they make it easier to exploit dataset biases. VQD requires goal-directed object detection and outputting a variable number of boxes that answer a query.

We make the following contributions:

1. We describe the first dataset for VQD, which will be publicly released.
2. We evaluate multiple baselines on our VQD dataset.

Proceedings of NAACL-HLT 2019, pages 1955–1961
Minneapolis, Minnesota, June 2 - June 7, 2019.

2 Related work

Over the past few years, a large amount of work has been done at the intersection of computer vision and natural language understanding, including visual madlibs (Yu et al., 2015; Tommasi et al., 2018), captioning (Farhadi et al., 2010; Kulkarni et al., 2013; Johnson et al., 2016; Liu et al., 2018), and image retrieval (Wan et al., 2014; Li et al., 2016). For VQD, the most related tasks are VQA and RER, which we review in detail.

2.1 Visual Question Answering

VQA systems take in an image and open-ended natural language question and then generate a text-based answer (Antol et al., 2015; Goyal et al., 2017; Acharya et al., 2019; Kafle et al., 2018). Many VQA datasets have been created. However, initial datasets, e.g., VQAv1 (Antol et al., 2015) and COCO-QA (Ren et al., 2015a), exhibited significant language bias in which many questions could be answered correctly without looking at the image, e.g., for VQAv1 it was possible to achieve 50% accuracy using language alone (Kafle and Kanan, 2016). To address the bias issue, the VQAv2 dataset was created with a more balanced distribution for each possible answer to make algorithms analyze the image (Goyal et al., 2017), but it still had bias in the kinds of questions asked, with some questions being scarce, e.g., reasoning questions. Synthetic datasets such as the CLEVR dataset (Johnson et al., 2017) addressed this by being synthetically generated to emphasize hard reasoning questions that are rare in VQAv1 and VQAv2. The TDIUC dataset addresses bias using both synthetically generated and human gathered questions about natural images, with performance evaluated for 12 kinds of questions (Kafle and Kanan, 2017a). While the state-of-the-art has rapidly increased on both synthetic and natural image VQA datasets, many models do not generalize across datasets (Shrestha et al., 2019).

2.2 Referring Expression Recognition

Unlike VQA, RER algorithms must produce evidence to justify their outputs. A RER algorithm outputs a box around the image location matching the input string, making it easier to tell if an algorithm is behaving correctly. The RefCOCO and RefCOCO+ datasets for RER were collected from the two-player 'ReferIt' Game (Kazemzadeh et al., 2014). The first player is asked to describe an outlined object and the second player has to correctly localize it from player one's description. The test datasets are futher split into the 'testA' and 'testB' splits. The split 'testA' contains object categories sampled randomly to be close to the original data distribution, while 'testB' contains objects sampled from the most frequent object categories, excluding categories such as 'sky', 'sand', 'floor', etc. Since, there is a time limit on the game, the descriptions are short, e.g., 'guy in a yellow t-shirt,' 'pink,' etc.

Instead of playing a timed game, to create the RefCOCOg dataset for RER, one set of Amazon Mechanical Turk (AMT) users were asked to generate a description for a marked object in an image and other users marked the region corresponding to the description (Mao et al., 2016). This resulted in more descriptive prompts compared to RefCOCO and RefCOCO+.

The Visual7W dataset for VQA includes a 'pointing' task that is closely related to RER (Zhu et al., 2016). Pointing questions require choosing which box of the four *given* boxes correctly answered a query. Systems did not generate their own boxes, and there is always one correct box.

Cirik et al. (2018) showed that RER datasets suffer from biases caused by their dataset collection procedure. For RefCOCOg, they found that randomly permuting the word in the referring expression caused only about a 5% drop in performance, suggesting that instead of relying on language structure, systems may be using some hidden correlations in language. They further showed that an image only model that ignores the referring expression yielded a precision of 71.2% for top-2 best predictions. They also found that predicting the object category given the image region produced an accuracy of 84.2% for top-2 best predictions. By having 0–N boxes, VQD is harder for an image-only model to perform well.

3 The VQD 1.0 Dataset (VQDv1)

We created VQDv1, the first dataset for VQD. VQDv1 is created synthetically using annotations from Visual Genome (VG), COCO, and COCO Panoptic. While this limits variety, it helps combat some kinds of bias and serves as an initial version of the dataset. VQDv1 has three distinct query categories:

1. Object Presence (e.g., 'Show the dog in the image')
2. Color Reasoning (e.g., 'Which plate is white

Type	# Questions
Simple	391,628
Color	172,005
Positional	57,904
Total	621,537

Table 1: VQDv1 Query Types

Dataset	# Images	# Questions
RefCOCO	19,994	142,209
RefCOCO+	19,992	141,564
RefCOCOg	26,711	85,474
VQDv1	123,287	621,537

Table 2: VQDv1 compared to RER datasets.

in color?')

3. Positional Reasoning (e.g., 'Show the cylinder behind the girl in the picture')

The number of queries per type are given in the Table 1. The dataset statistics and example images and are shown in Fig. 2 and Fig. 3, respectively. We show statistics for VQDv1 compared to RER datasets in Table 2.

All images in VQDv1 are from COCO. The ground truth bounding box annotations are derived from the COCO Panoptic annotations dataset (Kirillov et al., 2018). The questions are generated using multiple templates for each question type, which is an approach that has been used in earlier work for VQA (Kafle and Kanan, 2017a; Kafle et al., 2017). The query objects and their attributes are extracted by integrating the annotations from images that have both COCO and VG annotations. COCO annotations are focused on objects, while VG also has attribute and relationship information, e.g., size, color, and actions for scene objects.

3.1 Object Presence

Object presence questions require an algorithm to determine all instances of an object in an image without any relationship or positional attributes, for example, 'Show me the horse in the image' or 'Where is the chair?' We use all of the COCO 'things' labels and half of the COCO 'stuff' labels to generate these questions, making this task test the same capabilities as conventional object detection. We filter some 'stuff' categories that do not have well defined bounding boxes such as 'water-other', 'floor-stone', etc. We use multiple templates to create variety, e.g., 'Show the <object> in the image', 'Where are the <object> in the picture?' etc.

3.2 Color Reasoning

Color questions test the presence of objects modified by color attributes, e.g., 'Show me the cat which is grey in color' or 'Which blanket is blue in color?' Since, COCO has only object annotations, color attributes are derived from VG's attribute annotations. We align every VG image annotation with COCO annotations to obtain (object, color) annotations for each bounding box. When multiple color attributes for an object are present, the object is assigned a single color from that attribute set.

3.3 Positional Reasoning

Positional reasoning questions test the location of objects with respect to other objects, e.g., 'Show the building behind horses', 'Which people are in front of the lighthouse?', and 'Show the rhino behind elephant.' We again use VG's relationship and attribute annotations to create these questions.

3.4 Generating Counter-Concept Questions

Counter-concept questions have no valid boxes as outputs, and we endeavor to create hard counter-concept questions for each category. We ask 'Show me the zebra' only if there is a similar animal present (e.g., a cow), which was done by using COCO's super-categories. Likewise, 'Show me the donut that is brown in color' is only asked if a brown donut does not exist in the image.

4 Experiments

Our experiments are designed to probe the behavior of models on VQD compared to RER datasets. To facilitate this, we created a variant of our VQDv1 dataset that had only a single correct bounding box.

To evaluate performance for the RER and '1 Obj' version of the VQDv1 dataset, systems only output a single bounding box during test time, so the **Precision@1** metric is used. For the '0-N Obj' version of the VQDv1 dataset, we use the standard PASCAL VOC metric $\mathbf{AP}^{IoU=.50}$ from object detection, which calculates the average precision across the dataset using an intersection over union (IoU) greater than 0.5 criteria for matching with the ground truth boxes.

4.1 Models Evaluated

We implemented and evaluated four models for VQD. All models are built on top of Faster R-CNN with a ResNet-101 backbone whose output bounding boxes pass through Non-Maximal Suppression

	RefCOCO			RefCOCO+			RefCOCOg		VQDv1	
	val	testA	testB	val	testA	testB	val	test	1 Obj.	0–N Obj.
DETECT	38.63	37.82	38.32	38.85	37.85	38.98	50.13	50.03	30.44	26.94
RANDOM	16.51	14.30	19.81	16.67	14.10	20.45	19.87	19.76	9.77	2.38
Query-Blind	33.95	37.28	31.58	34.06	37.34	32.46	39.79	23.34	23.34	6.80
Vision+Query	69.41	**75.52**	**65.28**	**59.83**	**65.21**	**53.02**	**62.52**	**62.06**	**37.55**	**31.03**
SLR	**69.48**	73.71	64.96	55.71	60.74	48.80	60.21	59.63	–	–
SLR	68.95	73.10	64.85	54.89	60.04	49.56	59.33	59.21	–	–

Table 3: Results on RER datasets and two versions of our VQD dataset. The '1 Obj' version is trained and evaluated on queries with only a single box, analogous to RER, and the 0–N version contains the entire VQD dataset. All models use the same object proposals and visual features.

with a threshold of 0.7. This acts as a region proposal generator that provides CNN features for each region.

The four models we evaluate are:

1. **DETECT:** A model that uses the full Faster R-CNN system to detect all trained COCO classes, and then outputs the boxes that have the same label as the first noun in the query.
2. **RANDOM:** Select one random Faster R-CNN proposal.
3. **Query-Blind:** A vision only model that does binary classification of each region proposal's CNN features using a 3 layer MultiLayer Perceptron(MLP) with 1024-unit hidden ReLU layers.
4. **Vision+Query (V+Q):** A model that does binary classification of each region proposal. The query features are obtained from the last hidden layer of a Gated Recurrent Unit (GRU) network, and then they are concatenated with the CNN features and fed into a 3 layer MLP with 1024-unit hidden ReLU layers.

The primary reason for providing VQDv1 (1 obj.) and the RER results is to put the benefits of the VQD task in context. To aid in this endeavor, we also include comparison results directly from the SLR models (Yu et al., 2017) for RER, which is a recent system for that task.

4.2 Training Details

The Query-Blind and Vision+Query models are trained with binary cross-entropy loss. We use a learning rate of 0.0004, and perform learning rate decay of 0.8 when the training loss plateaus continuously for five epochs. The best model is selected based on the validation loss after training for 50 epochs.

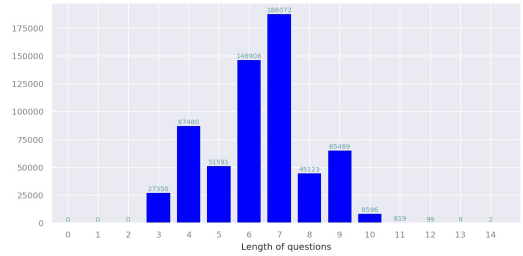
4.3 Results

Our main results are given in Table 3. Although simple, our Vision+Query model performs well across RER datasets, but it can also be applied to VQD tasks. As expected, RANDOM performs poorly on both VQDv1 datasets. DETECT beats RANDOM in the single object VQD setting by a large margin. Since, most of the questions in the RER datasets ask about common COCO categories, choosing one of those objects might be enough to get decent performance; however, DETECT performs poorly when evaluated under 0-N object settings in VQDv1. To handle queries in VQD, models must be able to understand the context and comprehend multiple objects in isolation.

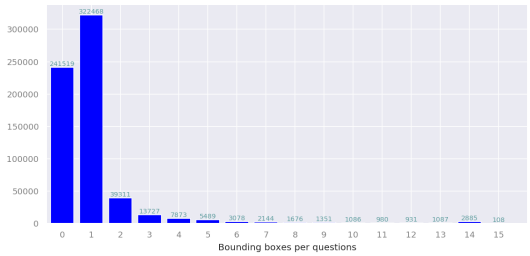
5 Conclusion

In this paper, we described our VQDv1 dataset as a test for visual grounding via goal-directed object detection. VQDv1 has both simple object presence and complex questions with 0–N bounding boxes. While VQDv1 contains only synthetically generated questions, this can help mitigate some forms of bias present in other VQA and RER datasets (Cirik et al., 2018; Kafle and Kanan, 2017b). While it would be expensive, a large, carefully filtered, and well designed human annotated VQD dataset is the next step toward advancing visual grounding research.

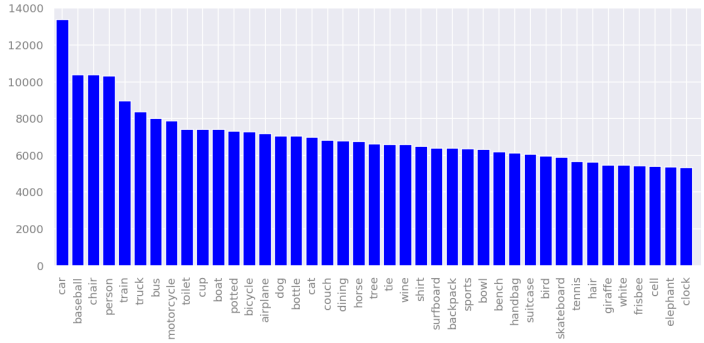
Compared to VQA, we argue that it is harder to be right for the *wrong* reasons in VQD because methods must generate bounding boxes. Compared to RER, we argue that it is harder to exploit bias in VQD since there are a variable number of boxes per image, making it considerably more difficult, as demonstrated by our experiments. We believe the VQD approach has considerable value and can be used to advance visual grounding research.



(a) Question length distribution.

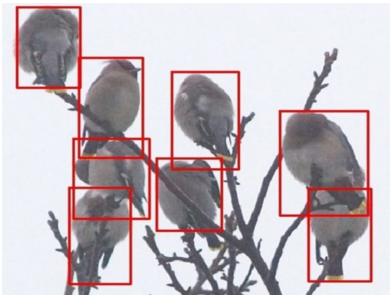


(b) Bounding boxes per question distribution.

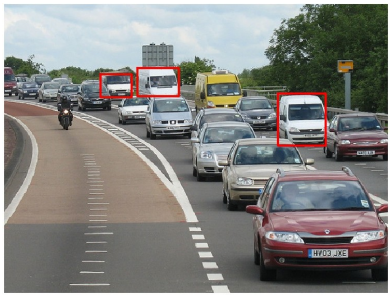


(c) Top-40 Object category distribution.

Figure 2: Distribution statistics for the VQD dataset.



(a) Where is the bird?



(b) Show me the van which is white in color.

(c) Which shirt is pink in color?

(d) Which glass is on the top of the head of the women?

(e) Show the lamp beside bed in the image.

(f) Where is the sink in the picture? Where is the toaster in the image?

Figure 3: Example query-detection pairs from the VQD dataset. Counter context questions that do not have a bounding box as an answer are generated in such a way that they are still relevant to the scene context. For example, in Fig. [3f] both questions pertain to the context 'kitchen'.

Acknowledgments

This work was supported in part by a gift from Adobe Research. The lab thanks NVIDIA for the donation of a GPU. We also thank fellow lab members Kushal Kafle and Tyler Hayes for their comments and useful discussions.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. TallyQA: Answering complex counting questions. In *AAAI*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *International Conference on Computer Vision (ICCV)*.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? *Association for Computational Linguistics (ACL)*.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision (ECCV)*.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Kushal Kafle and Christopher Kanan. 2017a. An analysis of visual question answering algorithms. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

Kushal Kafle and Christopher Kanan. 2017b. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding (CVIU)*.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *INLG-2017*, pages 198–202.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Empirical methods in natural language processing (EMNLP)*.

Alexander Kirillov, Kaiming He, Ross B. Girshick, Carsten Rother, and Piotr Dollár. 2018. Panoptic segmentation. *CoRR*.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees GM Snoek, and Alberto Del Bimbo. 2016. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Xiaoxiao Liu, Qingyang Xu, and Ning Wang. 2018. A survey on deep neural network-based image captioning. *The Visual Computer*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Neural Information Processing Systems (NeurIPS)*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. *Computer Vision and Pattern Recognition (CVPR)*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015a. Exploring models and data for image question answering. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015b. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2019. Answer them all! toward universal visual question answering models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tatiana Tommasi, Arun Mallya, Bryan Plummer, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2018. Combining multiple cues for visual madlibs question answering. *International Journal of Computer Vision (IJCV)*.

Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. 2014. Deep learning for content-based image retrieval: A comprehensive study. In *ACM international conference on Multimedia*.

Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. 2015. Visual madlibs: Fill in the blank description generation and question answering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speakerlistener-reinforcer model for referring expressions. In *Computer Vision and Pattern Recognition (CVPR)*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *IEEE conference on computer vision and pattern recognition (CVPR)*.