OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge

Kenneth Marino*[1], Mohammad Rastegari[2], Ali Farhadi[2,3] and Roozbeh Mottaghi[2]

[1]Carnegie Mellon University
[2]PRIOR @ Allen Institute for AI
[3]University of Washington

Abstract

Visual Question Answering (VQA) in its ideal form lets us study reasoning in the joint space of vision and language and serves as a proxy for the AI task of scene understanding. However, most VQA benchmarks to date are focused on questions such as simple counting, visual attributes, and object detection that do not require reasoning or knowledge beyond what is in the image. In this paper, we address the task of knowledge-based visual question answering and provide a benchmark, called OK-VQA, where the image content is not sufficient to answer the questions, encouraging methods that rely on external knowledge resources. Our new dataset includes more than 14,000 questions that require external knowledge to answer. We show that the performance of the state-of-the-art VQA models degrades drastically in this new setting. Our analysis shows that our knowledge-based VQA task is diverse, difficult, and large compared to previous knowledge-based VQA datasets. We hope that this dataset enables researchers to open up new avenues for research in this domain.

1. Introduction

The field of Visual Question Answering (VQA) has made amazing strides in recent years, achieving record numbers on standard VQA datasets [20, 4, 11, 17]. As originally conceived, VQA is not only a fertile ground for vision and language research, but is also a proxy to evaluate AI models for the task of open-ended scene understanding. In its ideal form, VQA would require not only visual recognition, but also logical reasoning and incorporating knowledge about the world. However, current VQA datasets (e.g., [3, 47]) are focused mainly on recognition, and most questions are about simple counting, colors, and other visual detection tasks, so do not require much logical reasoning or association with external knowledge. The most difficult and interesting questions ideally require knowing more than what the question entails or what information is contained in the images.

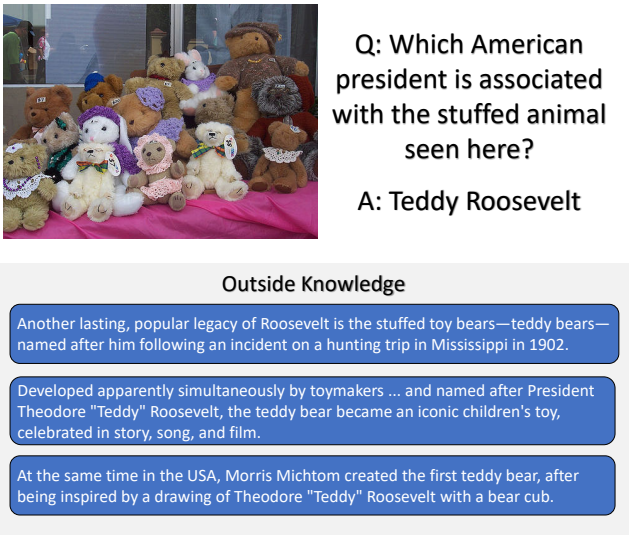

Figure 1: We propose a novel dataset for visual question answering, where the questions require external knowledge resources to be answered. In this example, the visual content of the image is not sufficient to answer the question. A set of facts about teddy bears makes the connection between teddy bear and the American president, which enables answering the question.

Consider the question in Figure 1, which asks about the relation between the teddy bear and an American president. The information in the image here is not complete for answering the question. We need to link the image content

*Work done during internship at Allen Institute for AI

to external knowledge sources, such as the sentences at the bottom of the figure taken from Wikipedia. Given the question, image, and Wikipedia sentences, there is now enough information to answer the question: Teddy Roosevelt!

More recent research has started to look at how to incorporate knowledge-based methods into VQA [29, 30, 36, 37]. These methods have investigated incorporating knowledge bases and retrieval methods into VQA datasets with a set of associated facts for each question. In this work, we go one step forward and design a VQA dataset which requires VQA to perform reasoning using unstructured knowledge.

To enable research in this exciting direction, we introduce a novel dataset, named Outside Knowledge VQA (OK-VQA), which includes only questions that require external resources for answering them. On our dataset, we can start to evaluate the reasoning capabilities of models in scenarios where the answer cannot be obtained by only looking at the image. Answering OK-VQA questions is a challenging task since, in addition to understanding the question and the image, the model needs to: (1) learn what knowledge is necessary to answer the questions, (2) determine what query to do to retrieve the necessary knowledge from an outside source of knowledge, and (3) incorporate the knowledge from its original representation to answer the question.

The OK-VQA dataset consists of more than 14,000 questions that cover a variety of knowledge categories such as science & technology, history, and sports. We provide category breakdowns of our dataset, as well as other relevant statistics to examine its properties. We also analyze state-of-the-art models and show their performance degrades on this new dataset. Furthermore, we provide results for a set of baseline approaches that are based on simple knowledge retrieval. Our dataset is diverse, difficult, and to date the largest VQA dataset focused on knowledge-based VQA in natural images.

Our contributions are: (a) we introduce the OK-VQA dataset, which includes only questions that require external resources to answer; (b) we benchmark some state-of-the-art VQA models on our new dataset and show the performance of these models degrades drastically; (c) we propose a set of baselines that exploit unstructured knowledge.

2. Related Work

Visual Question Answering (VQA). Visual question answering (VQA) has been one of the most popular topics in the computer vision community over the past few years. Early approaches to VQA combined recurrent networks with CNNs to integrate textual and visual data [27, 1]. Attention-based models [11, 25, 39, 40, 41, 47] better guide the model in answering the questions by highlighting image regions that are relevant to the question. Modular networks [2, 15, 19] leverage the compositional nature of the language in deep neural networks. These approaches have been extended to the video domain as well [16, 28, 35]. Recently, [13, 9] address the problem of question answering in an interactive environment. None of these approaches, however, is designed for leveraging external knowledge so they cannot handle the cases that the image does not represent the full knowledge to answer the question.

The problem of using external knowledge for answering questions has been tackled by [38, 36, 37, 23, 30, 29]. These methods only handle the knowledge that is represented by subject-relation-object or visual concept-relation-attribute triplets, and rely on supervision to do the retrieval of facts. In contrast, answering questions in our dataset requires handling unstructured knowledge resources.

VQA datasets. In the past few years several datasets have been proposed for visual question answering [26, 3, 12, 44, 31, 47, 34, 21, 18, 37]. The DAQUAR dataset [26] includes template-based and natural questions for a set of indoor scenes. [3] proposed the VQA dataset, which is two orders of magnitude larger than DAQUAR and includes more diverse images and less constrained answers. FM-IQA [12] is another dataset that includes multi-lingual questions and answers. Visual Madlibs [44], constructs fill-in-the-blank templates for natural language descriptions. COCO-QA [31] is constructed automatically by converting image descriptions to questions. The idea of Visual 7W [47] is to provide object-level grounding for question-answer pairs as opposed to image-level associations between images and QA pairs. Visual Genome [21] provides dense annotations for image regions, attributes, relationships, etc. and provide free-form and region-based QA pairs for each image. MovieQA [34] is a movie-based QA dataset, where the QAs are based on information in the video clips, subtitles, scripts, etc. CLEVR [18] is a synthetic VQA dataset that mainly targets visual reasoning abilities. In contrast to all these datasets, we focus on questions that cannot be answered by the information in the associated image and require external knowledge to be answered.

Most similar to our dataset is FVQA [37]. While that work also tackles the difficult problem of creating a VQA dataset requiring outside knowledge, their method annotates questions by selecting a fact (a knowledge triplet such as “dog is mammal") from a fixed knowledge base. While this dataset is still quite useful for testing methods’ ability to incorporate a knowledge base into a VQA system, our dataset tests methods’ ability to retrieve relevant facts from the web, from a database, or some other source of knowledge that was not used to create the questions. Another issue is that triplets are not sufficient to represent general knowledge.

Building knowledge bases & Knowledge-based reasoning. Several knowledge bases have been created using visual data or for visual reasoning tasks [46, 8, 10, 32, 49, 48]. These knowledge bases are potentially helpful resources for answering questions in our dataset. Knowledge-based

Figure 2: **Dataset examples.** Some example questions and their corresponding images and answers have been shown. We show one example question for each knowledge category.

question answering has received much more attention in the NLP community (e.g., [5, 43, 42, 6, 33, 22, 7]).

3. OK-VQA Dataset

In this section we explain how we collect a dataset which better measures performance of VQA systems requiring external knowledge. The common VQA datasets such as [3, 14] do not require much knowledge to answer a large majority of the questions. The dataset mostly contains questions such as "How many apples are there?", "What animal is this?", and "What color is the bowl?". While these are perfectly reasonable tasks for open-ended visual recognition, they do not test our algorithms' ability to reason about a scene or draw on information outside of the image. Thus, for our goal of combining visual recognition with information extraction from sources outside the image, we would not be able to evaluate knowledge-based systems as most questions do not require outside knowledge.

To see this specifically, we examine the "age annotations" that are provided for 10,000 questions in the VQA dataset [1]. For each question and image pair, an MTurk worker was asked how old someone would need to be to answer the question. While this is not a perfect metric, it is a reasonable approximation of the difficulty of a question and how much a person would have to know to answer a question. The analysis shows that more than 78% of the questions can be answered by people who are 10 years old or younger. This suggests that very little background knowledge is actually required to answer the vast majority of these questions.

Given that current VQA datasets do not test exactly what we are looking for, we collect a new dataset. We use random images from the COCO dataset [24], using the original 80k-40k training and validation splits for our train and test splits. The visual complexity of these images compared to other datasets make them ideal for labeling knowledge-based questions.

In the first round of labeling, we asked MTurk workers to write a question given an image. Similar to [3], we prompt users to come up with questions to fool a "smart robot." We also ask in the instructions that the question should be related to the image content. In addition, we prompt users not to ask what is in an image, or how many of something there is, and specify that the question should require some outside knowledge. In a second round of labeling, we asked 5 different MTurk workers to label each question-image pair with an answer.

Although this prompt yielded many high-quality questions, it also yielded a lot of low quality questions, for example, ones that asked basic questions such as counting, did not require looking at the image, or were nonsensical. To ensure that the dataset asked these difficult knowledge-requiring questions, the MTurk provided questions were manually filtered to get only questions requiring knowledge. From a pool of 86,700 questions, we filtered down to 34,921 questions.

One more factor to consider was the potential bias in the dataset. As discussed in many works, including [14],

	Number of questions	Number of images	Knowledge based?	Goal	Answer type	Avg. A length	Avg. Q length
DAQUAR [26]	12,468	1,449	✘	visual: counts, colors, objects	Open	1.1	11.5
Visual Madlibs [44]	360,001	10,738	✘	visual: scene, objects, person	FITB/MC	2.8	4.9
Visual 7W [47]	327,939	47,300	✘	visual: object-grounded questions	MC	2.0	6.9
VQA (v2) [14]	1.1M	200K	✘	visual understanding	Open/MC	1.2	6.1
MovieQA [34]	14,944	408V	✘	text+visual story comprehension	MC	5.3	9.3
CLEVR [18]	999,968	100,000	✘	logical reasoning	Open	1.0	18.4
KB-VQA [36]	2,402	700	✓	visual reasoning with given KB	Open	2.0	6.8
FVQA [37]	5,826	2,190	✓	visual reasoning with given KB	Open	1.2	9.5
OK-VQA (ours)	14,055	14,031	✓	visual reasoning with open knowledge	Open	1.3	8.1

Table 1: **Comparison of various visual QA datasets.** We compare OK-VQA with some other VQA datasets. The bottom three rows correspond to knowledge-based VQA datasets. A length: answer length; Q length: question length; MC: multiple choice; FITB: fill in the blanks; KB: knowledge base.

the VQAv1 dataset had a lot of bias. Famously, questions beginning with "Is there a ..." had a very strong bias towards "Yes." Similarly, in our unfiltered dataset, there were a lot of questions with a bias towards certain answers. For instance, in a lot of images where there is snowfall, the question would ask "What season is it?" Although there were other images (such as ones with deciduous trees with multi-colored leaves) with different answers, there was a clear bias towards "winter." To alleviate this problem, for train and test, we removed questions so that the answer distribution was uniform; specifically, we removed questions if there were more than 5 instances of that answer as the most common answer. This had the effect of removing a lot of the answer bias. It also had the effect of making the dataset harder by limiting the number of times VQA algorithms would see questions with a particular answer, making outside information more important. We also removed questions which had no inter-annotator agreement on the answer. Performing this filtering brought us down to 9,009 questions in train and 5,046 questions in test for a total of 14,055 questions.

Figure 2 shows some of the collected questions, images, and answers from our dataset. More will be provided in the supplementary material. You can see that these questions require at least one piece of background knowledge to answer. For instance, in the bottom left question, the system needs to recognize that the image is of a christian church and know that those churches hold religious services on Sundays. That latter piece of knowledge should be obtained from external knowledge resources, and it cannot be inferred from the image and question alone.

4. Dataset Statistics

In this section, we explore the statistical properties of our dataset, and compare to other visual question answering datasets to show that our dataset is diverse, difficult, and, to the best of our knowledge, the largest VQA dataset specifically targeted for knowledge-based VQA on natural scenes.

Knowledge category. Requiring knowledge for VQA is a good start, but there are many different types of knowledge that humans have about the world that could come into play. There is common-sense knowledge: water is wet, couches are found in living rooms. There is geographical knowledge: the Eiffel Tower is in Paris, scientific knowledge: humans have 23 chromosomes, and historical knowledge: George Washington is the first U.S. president. To get a better understanding of the kinds of knowledge our dataset requires, we asked five MTurk workers to annotate each question as belonging to one of ten categories of knowledge that we specified: Vehicles and Transportation; Brands, Companies and Products; Objects, Materials and Clothing; Sports and Recreation; Cooking and Food; Geography, History, Language and Culture; People and Everyday Life, Plants and Animals; Science and Technology; and Weather and Climate. If no one category had a plurality of workers, it was categorized as "Other". This also ensured that the final category labels are mutually exclusive. We show the distribution of questions across categories in Figure 3.

Comparison with other VQA datasets. In Table 1 we look at a number of other visual question answering datasets and compare them to our dataset in a number of different ways. In the top section, we look at a number of datasets which do not explicitly try to include a knowledge component including the ubiquitous VQAv2 dataset [14], the first version of which was one of the first datasets to investigate visual question answering. Compared to these datasets, we have a comparable number of questions to DAQUAR [26] as well as MovieQA [34], and many more questions than knowledge-based datasets KB-VQA [36] and FVQA [37]. We have fewer questions compared to CLEVR [18] where the images, questions and answers are automatically generated, as well compared to more large-scale human annotated visual datasets such as VQAv2 [14], and Visual

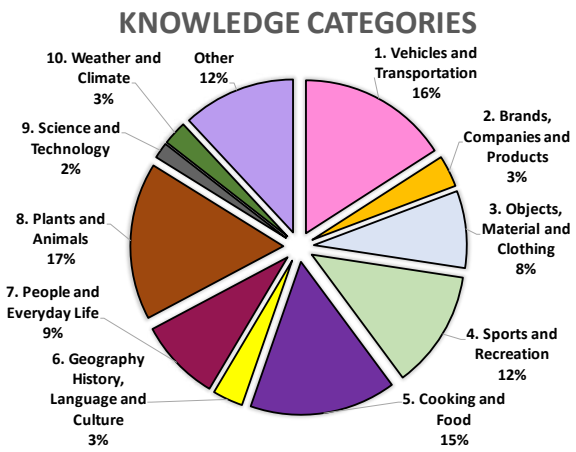

Figure 3: **Breakdown of questions in terms of knowledge categories.** We show the percentage of questions falling into each of our 10 knowledge categories.

Madlibs [44]. Since we manually filtered our dataset to avoid the pitfalls of other datasets and to ensure our questions are knowledge-based and because we filtered down common answers to emphasize the long tail of answers, our dataset is more time-intensive and expensive to collect. We trade off size in this case for knowledge and difficulty.

We can see from the average question lengths and average answer lengths that our questions and answers are about comparable to KB-VQA [36] and FVQA [37] and longer than the other VQA datasets with the exception of DAQUAR and CLEVR (which are partially and fully automated from templates respectively). This makes sense since we would expect knowledge-based questions to be longer as they are typically not able to be as short as common questions in other datasets such as "How many objects are in the image?" or "What color is the couch?".

Question statistics. We also collected statistics for our dataset by looking at the number of questions, and by looking at which were most frequent for each knowledge category. OK-VQA has 12,591 unique questions out of 14,055 total, and 7,178 unique question words. This indicates that we get a variety of different questions and answers in our dataset. We also looked at the variety of images in our dataset. As we stated earlier, our images come from the COCO image dataset, so our dataset contains the same basic distribution of images. However, we only use a subset of COCO images, so we want to see if we still get a wide distribution of images. For this, we ran a Places2 [45] classifier on our images and looked at the top-1 scene class for each image and compared that to COCO overall. Out of 365 scenes, our dataset contains all but 5 classes: hunting lodge, mansion, movie theater, ruin and volcano. These classes appear infrequently in the overall COCO dataset (10, 22, 28, 37 and 25 times respectively), so overall, we still captured quite a lot of the variation in scenes.

Knowledge Category	Highest relative frequency question words	Highest relatively frequency answers
1. Vehicles and Transportation	bus, train, truck, buses, jet	jet, double decker, take off, coal, freight
2. Brands, Companies and Companies	measuring, founder, advertisements, poster, mobile	ebay, logitech, gift shop, flickr, sprint
3. Objects, Material and Clothing	scissors, toilets, disk, teddy, sharp	sew, wrench, quilt, teddy, bib
4. Sports and Recreation	tennis, players, player, baseball, bat	umpire, serve, catcher, ollie, pitcher
5. Cooking and Food	dish, sandwich, meal, cook, pizza	donut, fork, meal, potato, vitamin c
6. Geography, History, Language and Culture	denomination, nation, festival, century, monument	prom, spire, illinois, past, bern
7. People and Everyday Life	expressing, emotions, haircut, sunburned, punk	hello, overall, twice, get married, cross leg
8. Plants and Animals	animals, wild, cows, habitat, elephants	herbivore, zebra, herd, giraffe, ivory
9. Science and Technology	indoor, mechanical, technology, voltage, connect	surgery, earlier, 1758, thumb, alan turing
10. Weather and Climate	weather, clouds, forming, sunrise, windy	stormy, noah, chilly, murky, oasis

Figure 4: For each category we show the question words and answers that have the highest relative frequency across our knowledge categories (i.e. frequency in category divided by overall frequency).

Finally, we show in Figure 4 the question words and answers in each category that are the most "unique" to get a better idea of what types of questions we have in each of these categories. We calculate these for each knowledge category by looking at the number of appearances within the category over the total number in the dataset to see which question words and answers had the highest relative frequency in their category. When looking at the question words, we see words specific to categories such as bus in Vehicles and Transportation, sandwich in Cooking and Food, and clouds in Weather and Climate. We also see that the answers are also extremely related to each category, such as herbivore in Plants and Animals, and umpire in Sports and Recreation. In the supplemental material, we also show the most common question words and answers.

5. Benchmarking

In this section, we evaluate current state-of-the-art VQA approaches and provide results for some baselines, includ-

Method	OK-VQA	VT	BCP	OMC	SR	CF	GHLC	PEL	PA	ST	WC	Other
Q-Only	14.93	14.64	14.19	11.78	15.94	16.92	11.91	14.02	14.28	19.76	25.74	13.51
MLP	20.67	21.33	15.81	17.76	24.69	21.81	11.91	17.15	21.33	19.29	29.92	19.81
ArticleNet (AN)	5.28	4.48	0.93	5.09	5.11	5.69	6.24	3.13	6.95	5.00	9.92	5.33
BAN [20]	25.17	23.79	17.67	22.43	30.58	27.90	**25.96**	20.33	25.60	20.95	**40.16**	22.46
MUTAN [4]	26.41	25.36	18.95	24.02	33.23	27.73	17.59	20.09	**30.44**	20.48	39.38	22.46
BAN + AN	25.61	24.45	19.88	21.59	30.79	29.12	20.57	21.54	26.42	**27.14**	38.29	22.16
MUTAN + AN	**27.84**	**25.56**	**23.95**	**26.87**	**33.44**	**29.94**	20.71	**25.05**	29.70	24.76	39.84	**23.62**
BAN/AN oracle	27.59	26.35	18.26	24.35	33.12	30.46	28.51	21.54	28.79	24.52	41.4	25.07
MUTAN/AN oracle	28.47	27.28	19.53	25.28	35.13	30.53	21.56	21.68	32.16	24.76	41.4	24.85

Table 2: **Benchmark results on OK-VQA.** We show the results for the full OK-VQA dataset and for each knowledge category: Vehicles and Transportation (VT); Brands, Companies and Products (BCP); Objects, Material and Clothing (OMC); Sports and Recreation (SR); Cooking and Food (CF); Geography, History, Language and Culture (GHLC); People and Everyday Life (PEL); Plants and Animals (PA); Science and Technology (ST); Weather and Climate (WC); and Other.

ing knowledge-based ones.

MUTAN [4]: Multimodal Tucker Fusion (MUTAN) model [4], a recent state-of-the-art tensor-based method for VQA. Specifically, we use the attention version of MUTAN, and choose the parameters to match the single best performing model of [4].

BAN [20]: Bilinear Attention Networks for VQA. A recent state-of-the art VQA method that uses a co-attention mechanism between the question features and the bottom-up detection features of the image. We modify some hyperparameters to improve performance on our dataset (see supplemental material).

MLP: The MLP has 3 hidden layers with ReLU activations and hidden size 2048 that concatenates the image and question features after a skip-thought GRU after one fully connected layer each. Like MUTAN, it uses fc7 features from ResNet-152.

Q-Only: The same model as MLP, but only takes the question features.

ArticleNet (AN): We consider a simple knowledge-based baseline that we refer to as ArticleNet. The idea is to retrieve some articles from Wikipedia for each question-image pair and then train a network to find the answer in the retrieved articles.

Retrieving articles is composed of three steps. First, we collect possible search queries for each question-image pair. We come up with all possible queries for each question by combining words from the question and words that are identified by pre-trained image and scene classifiers. Second, we use the Wikipedia search API to get the top retrieved article for each query. Third, for each query and article, we extract a small subset of each article that is most relevant for the query by selecting the sentences within the article that best correspond to our query based on the frequency of those query words in the sentence.

Once the sentences have been retrieved, the next step is to filter and encode them for use in VQA. Specifically, we train ArticleNet to predict whether and where the ground truth answers appear in the article and in each sentence. The architecture is shown in Figure 5. To find the answer to a question, we pick the top scoring word among the retrieved sentences. More specifically, we take the highest value of $a_{w_i}.a_{sent}$, where a_{w_i} is the score for the word being the answer and a_{sent} is the score for the sentence including the answer.

For a more detailed description of ArticleNet see the supplementary material.

MUTAN + AN: We augment MUTAN with the top sentence hidden states (h_{sent} in Figure 5) from ArticleNet (AN). During VQA training and testing, we take the top predicted sentences (ignoring duplicate sentences), and feed them in the memory of an end-to-end memory network [33]. The output of the memory network is concatenated with the output of the first MUTAN fusion layer.

BAN + AN: Similarly, we incorporate the ArticleNet hidden states into BAN and incorporate it into VQA pipeline with another memory network. We concatenate output of the memory network with the BAN hidden state right before the final classification network. See the supplemental material for details.

MUTAN/AN oracle: As an upper bound check, and to see potentially how much VQA models could benefit from the knowledge retrieved using ArticleNet, we also provide results on an oracle, which simply takes the raw ArticleNet and MUTAN predictions, taking the best answer (comparing to ground truth) from either.

BAN/AN oracle: Similar to the MUTAN/AN oracle, but we take the best answer from the raw ArticleNet and BAN instead, again taking the best answer for each question.

Benchmark results. We report the results using the common VQA evaluation metric [3], but use each of our answer annotations twice, since we have 5 answer annotations versus 10 in [3]. We also stem the answers using Porter stemming to consolidate answers that are identical except for pluralization and conjugation as in [37]. We also show the breakdowns for each of our knowledge categories. The

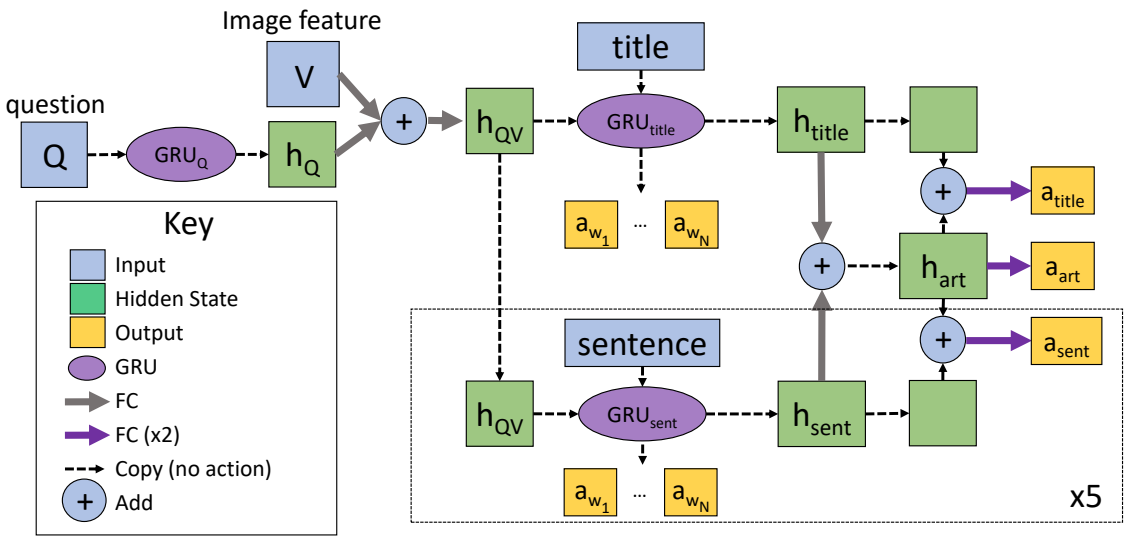

Figure 5: **ArticleNet architecture.** ArticleNet takes in the question Q and visual features V. All modules within the dotted line box share weights. The output of the GRUs is used to classify each word as the answer or not a_{w_i}. The final GRU hidden states h_{title} and h_{sent} are put through fully connected layers to predict if the answer is in the sentence a_{sent} or title a_{title}, and then are combined together and used to classify if the answer is in the article a_{art}.

results are reported in Table 2.

The first observation is that no method gets close to numbers on standard VQA dataset such as VQA [14] (where the best real open-ended result for the 2018 competition is 72.41). Moreover, state-of-the-art models such as MUTAN [4] and BAN [20], which are specifically designed for VQA to learn high-level associations between the image and question, get far worse numbers on our dataset. This suggests that OK-VQA cannot be solved simply by coming up with a clever model, but actually requires methods that incorporate information from outside the image.

It is interesting to note that although the performance of the raw ArticleNet is low, it provides improvement when combined with the state-of-the-art models (MUTAN + AN and BAN + AN). From the oracle numbers, we can see that the knowledge retrieved by ArticleNet provides complementary information to the state-of-the-art VQA models. These oracles are optimistic upper bounds using ArticleNet, but they show that smarter knowledge-retrieval approaches could have stronger performance on our dataset. Note that ArticleNet is not directly trained on VQA and can only predict answers within the articles it has retrieved. So the relatively low performance on VQA is not surprising.

Looking at the category breakdowns, we see that ArticleNet is particularly helpful for brands, science, and cooking categories, perhaps suggesting that these categories are better represented in Wikipedia. It should be noted that the major portion of our dataset requires knowledge outside Wikipedia such as commonsense or visual knowledge.

The Q-Only baseline performs significantly worse than the other VQA baselines, suggesting that visual features are indeed necessary and our procedure for reducing answer bias was effective.

Method	VQA score on OK-VQA
ResNet152	26.41
ResNet50	24.74
ResNet18	23.64
Q-Only	14.93

Table 3: Results on OK-VQA with different visual features.

Visual feature ablation. We also want to demonstrate the difficulty of the dataset from the perspective of visual features, so we show MUTAN results using different ResNet architectures. The previously reported result for MUTAN is based on ResNet152. We also show the results using extracted features from ResNet50 and ResNet18 in Table 3. From this table it can be seen that going from ResNet50 to ResNet152 features only has a marginal improvement, and similarly going from ResNet18 to ResNet50. However, going from ResNet18 to no image (Q-Only) causes a large drop in performance. This suggests that our dataset is indeed visually grounded, but better image features do not hugely improve the results, suggesting the difficulty lies in the retrieving the relevant knowledge and reasoning required to answer the questions.

Scale ablation. Finally, we investigate the degree to which the size of our dataset relates to its difficulty as opposed to the nature of the questions themselves. We first take a ran-

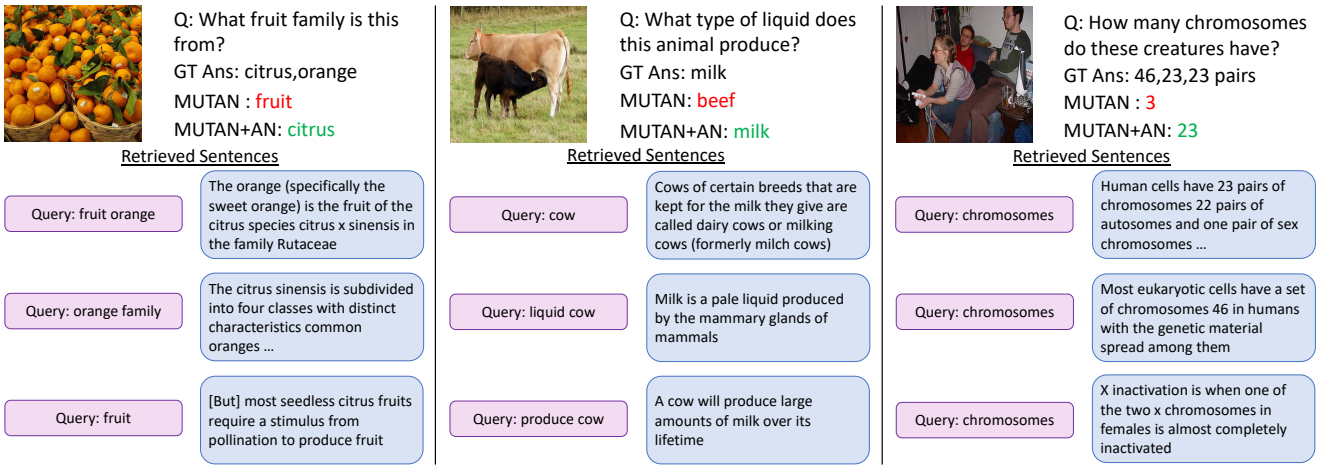

Figure 6: **Qualitative results.** We show the result of MUTAN+AN compared to the MUTAN baseline answer and the ground truth answer ('GT Ans'). We show the query words that were used by ArticleNet (pink boxes) and the corresponding most relevant sentences (blue boxes).

dom subdivision of our training set and train MUTAN on progressively smaller subsets of the training data and evaluate on our original test set. Figure 7 shows the results.

Qualitative examples. We show some qualitative examples in Figure 6 to see how outside knowledge helps VQA systems in a few examples. We compare MUTAN+AN method with MUTAN. The left example asks what "fruit family" the fruit in the image (oranges) comes from. We see that two sentences that directly contain the information that oranges are citrus fruits are retrieved —"The orange ... is a fruit of the citrus species" and "The citrus sinensis is subdivided into four classes [including] common oranges".

The middle example asks what liquid the animal (cow) produces. The first and third sentences tell us that cows produce milk, and the second sentence tells us that milk is a liquid. This gives the combined MUTAN+AN method enough information to correctly answer milk.

The example on the right asks how many chromosomes humans have. It is somewhat ambiguous whether it means how many individual chromosomes or how many pairs, so workers labeled both as answers. The retrieved articles are helpful here, retrieving two different articles referring to 23 pairs of chromosomes and 46 chromosomes total. The combined MUTAN+AN method correctly answers 23, while MUTAN guesses 3.

6. Conclusion

We address the task of knowledge-based visual question answering. We introduce a novel benchmark called OK-VQA for this task. Unlike the common VQA benchmarks, the information provided in the question and the corresponding images of OK-VQA is not sufficient to answer the questions, and answering the questions requires reasoning on external knowledge resources. We show that the performance of state-of-the-art VQA models significantly drops on OK-VQA. We analyze the properties and statistics of the dataset and show that background knowledge can improve results on our dataset. Our experimental evaluations show that the proposed benchmark is quite challenging and that there is a large room for improvement.

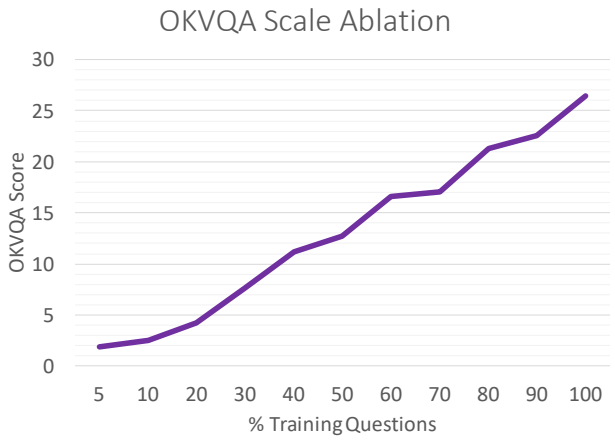

Figure 7: Results on OK-VQA using different sizes of the training set.

Acknowledgements: We would like to thank everyone who took time to review this work and provide helpful comments. This work is in part supported by NSF IIS-165205, NSF IIS-1637479, NSF IIS-1703166, Sloan Fellowship, NVIDIA Artificial Intelligence Lab, and Allen Institute for artificial intelligence. Thanks to Aishwarya Agrawal, Gunnar Sigurdsson, Victoria Donley, Achal Dave, and Eric Kolve who provided valuable assistance, advice and feedback. Kenneth Marino is supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *IJCV*, 2017. 2, 3

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, 2016. 2

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, 2015. 1, 2, 3, 6

[4] Hedi Ben-younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, 2017. 1, 6, 7

[5] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, 2013. 3

[6] Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. In *EMNLP*, 2014. 3

[7] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. *arXiv*, 2017. 3

[8] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013. 2

[9] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. *arXiv*, 2017. 2

[10] Santosh Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014. 2

[11] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016. 1, 2

[12] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *NIPS*, 2015. 2

[13] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. IQA: Visual question answering in interactive environments. *arXiv*, 2017. 2

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 3, 4, 7

[15] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, 2017. 2

[16] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017. 2

[17] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 1

[18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 2, 4

[19] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, 2017. 2

[20] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear Attention Networks. *arXiv preprint arXiv:1805.07932*, 2018. 1, 6, 7

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 2

[22] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*, 2016. 3

[23] Guohao Li, Hang Su, and Wenwu Zhu. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. *arXiv*, 2017. 2

[24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 3

[25] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016. 2

[26] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014. 2, 4

[27] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 2

[28] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. Marioqa: Answering questions by watching gameplay videos. In *ICCV*, 2017. 2

[29] Medhini Narasimhan, Svetlana Lazebnik, and Alexander G Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *NIPS*, 2018. 2

[30] Medhini Narasimhan and Alexander G. Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *ECCV*, 2018. 2

[31] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015. 2

[32] Fereshteh Sadeghi, Santosh K Divvala, and Ali Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *CVPR*, 2015. 2

[33] Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *NIPS*, 2015. 3, 6

[34] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016. 2, 4

[35] Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 2014. 2

[36] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*, 2017. 2, 4, 5

[37] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick. Fvqa: Fact-based visual question answering. *TPAMI*, 2017. 2, 4, 5, 6

[38] Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, 2016. 2

[39] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016. 2

[40] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016. 2

[41] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 2

[42] Xuchen Yao and Benjamin Van Durme. Information extraction over structured data: Question answering with freebase. In *ACL*, 2014. 3

[43] Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *ACL-IJCNLP*, 2015. 3

[44] Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *ICCV*, 2015. 2, 4, 5

[45] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 5

[46] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, 2014. 2

[47] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016. 1, 2, 4

[48] Yuke Zhu, Joseph J. Lim, and Li Fei-Fei. Knowledge acquisition for visual question answering via iterative querying. In *CVPR*, 2017. 2

[49] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base for visual question answering. *arXiv*, 2015. 2