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

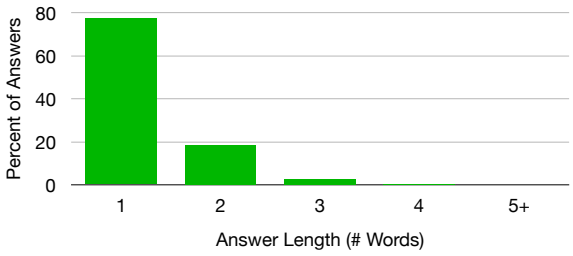


Figure 1: **Answer length distribution.** Histogram of the answer lengths in the dataset.

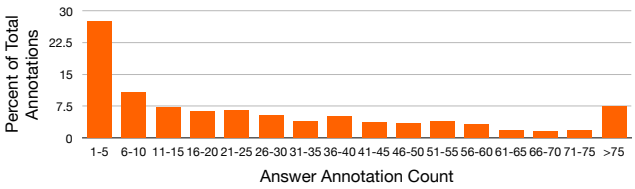


Figure 2: **Answer frequency distribution.** Histogram of the frequency of answers in the dataset. All 5 answers for each question are considered to compute the histogram. This shows for instance that answers that appear between 6 and 10 times in the dataset make up about 10% of all answers.

A. More Dataset Statistics

In Figure 1 we show the distribution of the lengths of the answers in the dataset. In Figure 2, we show the distribution of answer frequency for each of the unique answers in the dataset.

In Figure 3 we show the most common and the highest "relative frequency" question words and answers in each category.

Some other relevant dataset statistics for OK-VQA can be found in Table 1.

Number of unique answers	14,454
Test set covered by top 2000 answers	88.5057%
Number of unique questions	12,591
Number of unique question words	7,178

Table 1: More OK-VQA dataset statistics.

B. ArticleNet Details

B.1. Article Collection

Extracting the articles is composed of three steps: collecting possible search queries, using the Wikipedia search API to get the top retrieved article for each query and extracting a small subset of each article that is most relevant for the query.

To perform the search query step, first we need to come up with possible queries for each question. We extract all non-stop words (i.e. remove "the", "a", "what", etc.) from the question itself. Next we extract visual entities from the images by taking the top classifications from trained classifiers. We take the top classifications from an object classifier (trained on ImageNet [12]), a place classifier (trained on Places2 [14]) and an object detector [4, 3] (trained on COCO dataset [11]). Figure 4 shows some example images and their corresponding questions and classifications and shows some example queries that can be generated for that question.

Once the query words are selected, we compute possible queries. We choose every query word by itself and every two word combination of the query words as possible queries. We then retrieve the top article for each query from the Wikipedia search. Using the retrieval model from [2] to achieve a consistent snapshot, we retrieve the raw text.

Finally, we use the original query and the retrieved Wikipedia article to extract the most relevant sentences from the article for the query. Essentially, we perform another step of retrieval. The sentence priority is determined by three hierarchical metrics: (1) the number of unique query words in the sentence, (2) the total number of query words in the sentence, counting repeats, (3) the order of the sentence in the article. The priority is determined by factor

Knowledge Category	Most common question words	Highest relative frequency question words	Most common answers	Highest relatively frequency answers
Overall	what, the, is, this, of	N/A	ski, bake, surf, skateboard, cook	N/A
1. Vehicles and Transportation	what, the, is, this, of	bus, train, truck, buses, jet	stop, fly, passenger, transport, motorcycle	jet, double decker, take off, coal, freight
2. Brands, Companies and Companies	what, the, is, this, of	measuring, founder, advertisements, poster, mobile	dell, sony, samsung, computer, coca cola	ebay, logitech, gift shop, flickr, sprint
3. Objects, Material and Clothing	what, the, is, of, this	scissors, toilets, disk, teddy, sharp	bear, teddy bear, clothing, brick, flower	sew, wrench, quilt, teddy, bib
4. Sports and Recreation	what, the, is, this, of	tennis, players, player, baseball, bat	ski, surf, tennis, skateboard, snowboard	umpire, serve, catcher, ollie, pitcher
5. Cooking and Food	what', is, the, this, of	dish, sandwich, meal, cook, pizza	bake, carrot, banana, orange, fry	donut, fork, meal, potato, vitamin c
6. Geography, History, Language and Culture	what, is, this, the, of	denomination, nation, festival, century, monument	chinese, church, washington dc, lighthouse, england	prom, spire, illinois, past, bern
7. People and Everyday Life	what, the, is, this, of	expressing, emotions, haircut, sunburned, punk	sleep, happy, bathroom, living room, relax	hello, overall, twice, get married, cross leg
8. Plants and Animals	what, is, this, the, of	animals, wild, cows, habitat, elephants	cat, rose, africa, elephant, grass	herbivore, zebra, herd, giraffe, ivory
9. Science and Technology	what, the, is, this, of	indoor, mechanical, technology, voltage, connect	computer, laptop, window, phone, cell phone	surgery, earlier, 1758, thumb, alan turing
10. Weather and Climate	what, the, is, of, this	weather, clouds, forming, sunrise, windy	rain, winter, rainy, sunny, snow	stormy, noah, chilly, murky, oasis

Figure 3: For each category we show the question words and answers that have the highest frequency and relative frequency across our knowledge categories (i.e. frequency in category divided by overall frequency).

(1). If two sentences tie on this metric, we use metric (2) as a tie breaker, and similarly we use metric (3) to break ties for metric (2).

After these steps, we have our final "article" for each query consisting of the title, and T most relevant sentences (in our case $T = 5$). In our experiments, we retrieve on the order of 100 articles for each question at this step.

B.2. ArticleNet Overview

Once the Wikipedia articles have been retrieved, the next step is to filter and encode them for use in VQA. Simple encodings such as an average word2vec encoding, or with skip-thought [10] are not suitable for encoding long articles. Hence, we train an encoding specific to our data and useful for our eventual task. Taking inspiration from the fact that a hidden layer of a network trained on ImageNet is a good representation for images, we train a network on the retrieved articles on a proxy task to get a good representation. Specifically, we train ArticleNet to predict whether and where the ground truth answers appear in the article and each sentence. This also gives a way to narrow down the hundreds of articles for each question-image pair to a handful for the final VQA training.

For each of the Wikipedia articles, each word and series of words in the sentence are compared to the ground truth answer for that question to see if they match (using Porter stemming). Hence, a label l_{art} is obtained if the answer appears in the article, and also a label l_{title} and l_{sent_i} if the answer appears in the title or sentence i, and a label l_{word_j} for each word in the title and sentence.

The architecture of the ArticleNet is shown in Figure 5. The inputs to the network are the question Q, the visual features V taken from an ImageNet trained ResNet152 [6], the title of the Wikipedia article, and the T sentences of the article (retrieved by the method explained in the previous section). From these inputs, it predicts whether the answer is in the title a_{title}, any of the sentences a_{sent} or the entire article a_{art}. The hidden states of this network are used later in the VQA pipeline to encode the sentences.

After training, the network is evaluated on the articles for

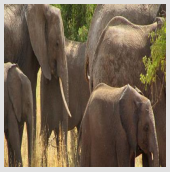
Image	Question	ImageNet Classification	Places Classifications	COCO Classifications	Example Queries
	What musical instrument used to be made with parts from these animals?	African elephant, Loxodonta africana | tusker | Indian elephant, Elephas maximus	watering hole | natural history museum	elephant	elephant parts instrument elephant elephant musical instrument animals parts musical elephant
	Which of these fruits is highest in potassium?	banana | butternut squash | spaghetti squash	orchard | bowling alley	banana | apple	fruits potassium banana potassium banana potassium apple potassium highest potassium
	These are types of what?	broccoli | cauliflower | hotdog, hot dog, red hot	fabric store | aquarium	broccoli | carrot	type broccoli broccoli carrot type carrot type carrot broccoli
	What type of hairstyle?	racket, racquet | volleyball | tennis ball	volleyball court/outdoor | baseball field	person | sports ball | tennis racket	type hairstyle hairstyle tennis person hairstyle hairstyle person tennis person
	What company or organization made this plane?	warplane, military plane | aircraft carrier, carrier, flattop, attack aircraft carrier | wing	hangar/outdoor | runway	person | car | airplane	organization plane military organization warplane military runway airplane organization plane company

Figure 4: **Example generated queries.** For some example questions, we show the image, question and top classification results from the trained models. In the rightmost column, we show some example queries that can be constructed for each question.

each question, the sentences that have the highest prediction score a_{sent} are used in our VQA training.

B.3. ArticleNet Performance

We rank each sentence during evaluation by the sentence score a_{sent}, and then plot on average how many sentences should be retrieved to find one including the answer. We compute the same curve for words where the ranking is based on the word score a_{w_i} multiplied by the sentence score a_{sent}. Product of these scores results in a higher retrieval than a_{w_i} by itself. These results show that ArticleNet is able to retrieve relevant sentences and words from the articles with reasonable accuracy. The plots that show Recall for top K sentences or words are shown in Figure 6.

C. MUTAN+AN and BAN+AN Details

We provide more details for the MUTAN+AN and BAN+AN models in this section. The MUTAN model is the Multimodal Tucker Fusion (MUTAN) model [1]. Specifically, we use the attention version of MUTAN, and choose the parameters to match the single best performing model of [1].

The BAN model is the single model version of Bilinear Attention Networks [8]. We use the single model version, and we use faster-rcnn features trained on COCO train (to avoid overlap with our test set). For both BAN and MUTAN, we use the top 2000 answers in train as our answer vocabulary.

We incorporate hidden states of ArticleNet for the top retrieved sentences into MUTAN and BAN. During VQA training and testing, we take the hidden states for the top N_{art} predicted sentences (ignoring duplicate sentences),

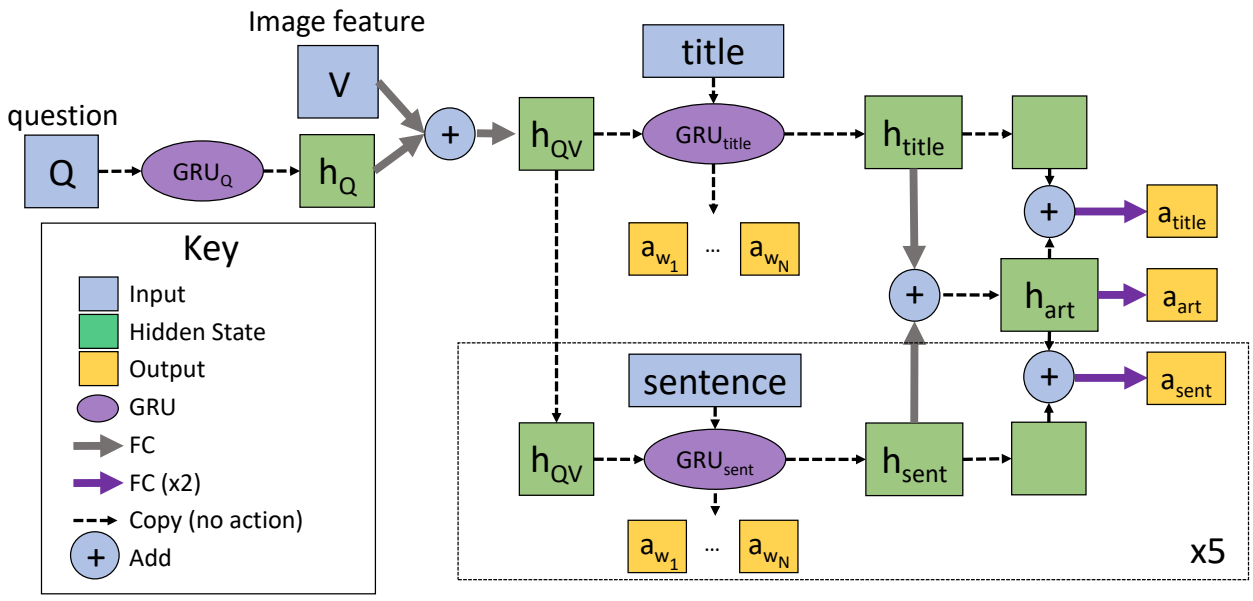


Figure 5: **ArticleNet architecture.** ArticleNet takes in the question Q and visual features V. All modules within the dotted line box share weights. The output of the GRUs is used to classify each word as the answer or not a_{w_i}. The final GRU hidden states h_{title} and h_{sent} are put through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

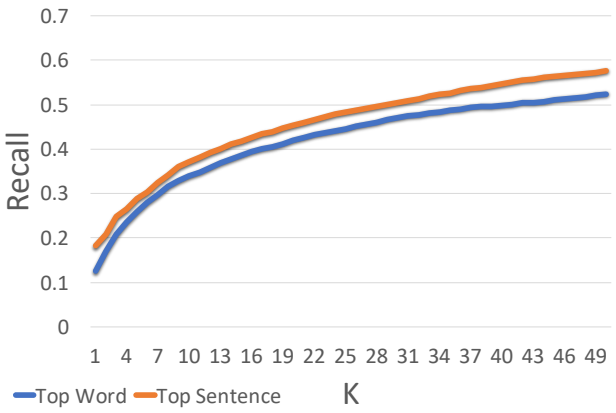


Figure 6: Retrieval@K curve for words and sentences.

and feed them in the memory in an end-to-end memory network [13].

We use the visual features V and encoded question Q passed through a hidden layer as the key to the memory network. To incorporate the memory network into the VQA system, we concatenate the output of the memory network to the hidden layer of the MUTAN after the attention MUTAN fusion and before the final MUTAN fusion. For the BAN model, we feed the output of the question embedding as the key to the memory network, and concatenate the output of the memory network to BAN right before the final classification layers.

D. Training and Model Details

D.1. ArticleNet

The question is encoded using a pre-trained skip-thought [10] encoder. All fully connected layers (except at output layers) have batch normalization [7] and ReLU activations. All output layers have Sigmoid before the final output. We train ArticleNet for $10,000$ iterations with a batch size of 64 using ADAM [9] with a learning rate of 10^{-4}, and using a balanced training set of "positive" and "negative" articles, meaning that with equal probability, an input article will contain the answer somewhere.

D.2. VQA Models

The MUTAN models as well as the MLP and Q-Only models were trained for 500 epochs. All use batch size of 128 using ADAM [9] with learning rate 10^{-4}.

The BAN models were trained for 200 epochs. We found that setting γ (number of glimpses) to 2 and the hidden feature size to 512 yielded much better performance on our dataset than the default parameter options used for VQAv2 [5].

We choose N_{art} to be 20, number of hops in the memory network as 2, and the hidden size of the memory network as 300.

E. Additional Dataset Examples

In Figures 7, 8, 9 we provide additional examples of Outside Knowledge VQA (OK-VQA) .

References

[1] Hedi Ben-younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, 2017. 3

[2] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. *arXiv*, 2017. 1

[3] Xinlei Chen and Abhinav Gupta. An implementation of faster rcnn with study for region sampling. *arXiv*, 2017. 1

[4] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 1

[5] 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. 4

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[7] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 4

[8] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear Attention Networks. *arXiv preprint arXiv:1805.07932*, 2018. 3

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 4

[10] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *NIPS*, 2015. 2, 4

[11] 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. 1

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1

[13] Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *NIPS*, 2015. 4

[14] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 1

Q: How old do you have to be in Canada to do this?

A: 18

Q: What is the hairstyle of the blond called?

A: pony tail, ponytail

Q: When was this piece of sporting equipment invented?

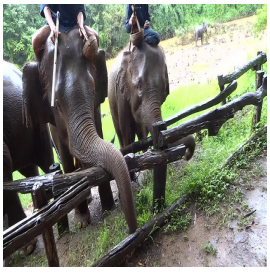
A: 1926

Q: What toy store used a mascot similar to these animals for many years?

A: Toys R Us

Q: In what country would you find this bus?

A: United Kingdom, England



Q: What religion holds these animals as sacred?

A: Hindu, Hinduism

Q: What creates those type of clouds?

A: water vapor

Q: What movie is the elephant from?

A: Horton Hears a Who

Q: Which Beatles song features a road like this?

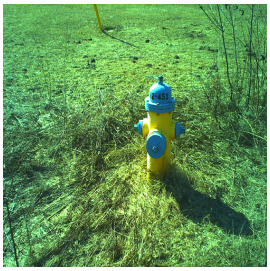
A: Abbey Road

Q: Is this recreational or a skiing competition?

A: competition

Q: Is this flora or fauna?

A: flora



Q: Who invented this device?

A: Frederick Graff

Q: What healthy oil is this dish a source of?

A: omega 3

Q: Which region of the united states is well known for the white object?

A: Texas, western

Q: What kind of event can be celebrated with these cakes?

A: Easter

Q: What makes the vegetable shown here unhealthy?

A: frying, grease

Figure 7: **Dataset examples.** Some more sample questions of OK-VQA.

Q: What century is this?

A: 20th

Q: When is best to use this toy?

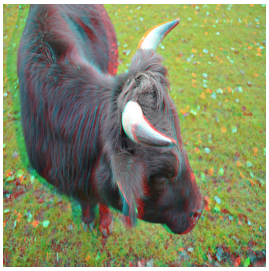
A: when its windy, windy days

Q: What is the process called that produces the red area on the chair?

A: rust, oxidization

Q: Which of these streets is famous for theater?

A: Broadway



Q: If wearing proper glasses what might this picture do?

A: pop out, 3d

Q: Who is the owner of this building?

A: Pope, Catholic Church

Q: Where is the monument located?

A: washington dc

Q: Can you guess the celebration where the people are enjoying?

A: fourth of July, 4th of July

Q: What does the thing in the sky need for it to be aimed by the user?

A: string

Q: What level of baseball is this?

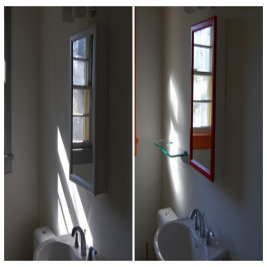
A: minor league, minor

Q: What is he jumping off of?

A: ramp, halfpipe

Q: What city does this player play for?

A: Oakland



Q: Behind the mirrors shown here there is likely what kind of cabinet?

A: medicine

Q: What animal is this device intended for?

A: human

Q: What is usually in the big blue metal box?

A: newpapers

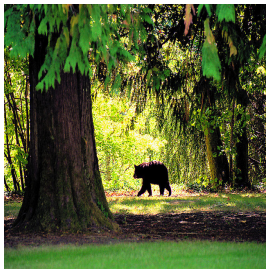
Q: Where are the two outer fruits primarily grown?

A: South America

Figure 8: **Dataset examples.** Some more sample questions of OK-VQA.

Q: Note the most largest device in the photo what is apple's line of these devices called?

A: Macbook



Q: What is the name of the popular confectionery named after the animal in this picture?

A: gummi bear

Q: What state in the United States gets the largest amount of what is landing on this umbrella?

A: Washington

Q: Although this is a British plane which colors shown here are also famously associated with the us?

A: red white and blue

Q: When was this sport invented?

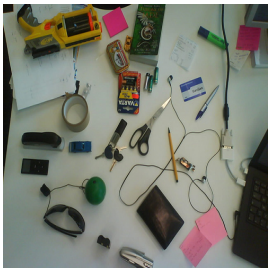
A: 1968

Q: How is Toyota involved with his activity?

A: sponsor, advertising

Q: What phylum does this animal belong to?

A: chordate, chordata



Q: What usually contains these objects?

A: drawer, junk drawer

Q: What channel is likely broadcasting this game?

A: ESPN

Q: Which Greek god is associated with this type of scene?

A Poseidon

Q: What is the term for a craftsman that specializes in producing this object?

A: clockmaker

Q: What is the boy in red attempting to do?

A: steal base

Q: What sound does this animal make?

A: moo

Q: When was the first stadium of this type built?

A: 1909, 1910

Q: What profession would need to be called if a fix is needed?

A: plumber

Q: Is this a team or individual sport?

A: individual

Figure 9: **Dataset examples.** Some more sample questions of OK-VQA.