International Journal of Computer Vision (2019) 127:398–414
https://doi.org/10.1007/s11263-018-1116-0

Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering

Yash Goyal[1] · Tejas Khot[2] · Aishwarya Agrawal[1] · Douglas Summers-Stay[3] · Dhruv Batra[1,4] · Devi Parikh[1,4]

Received: 14 February 2018 / Accepted: 21 August 2018 / Published online: 11 September 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

The problem of visual question answering (VQA) is of significant importance both as a challenging research question and for the rich set of applications it enables. In this context, however, inherent structure in our world and bias in our language tend to be a simpler signal for learning than visual modalities, resulting in VQA models that ignore visual information, leading to an inflated sense of their capability. We propose to counter these language priors for the task of VQA and make vision (the V in VQA) matter! Specifically, we *balance* the popular VQA dataset (Antol et al., in: ICCV, 2015) by collecting complementary images such that every question in our balanced dataset is associated with not just a single image, but rather a *pair of similar images* that result in two different answers to the question. Our dataset is by construction more balanced than the original VQA dataset and has approximately *twice* the number of image-question pairs. Our complete balanced dataset is publicly available at http://visualqa.org/ as part of the 2nd iteration of the VQA Dataset and Challenge (VQA v2.0). We further benchmark a number of state-of-art VQA models on our balanced dataset. All models perform significantly worse on our balanced dataset, suggesting that these models have indeed learned to exploit language priors. This finding provides the first concrete empirical evidence for what seems to be a qualitative sense among practitioners. We also present interesting insights from analysis of the participant entries in VQA Challenge 2017, organized by us on the proposed VQA v2.0 dataset. The results of the challenge were announced in the 2nd VQA Challenge Workshop at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017. Finally, our data collection protocol for identifying complementary images enables us to develop a novel interpretable model, which in addition to providing an answer to the given (image, question) pair, also provides a counter-example based explanation. Specifically, it identifies an image that is similar to the original image, but it believes has a different answer to the same question. This can help in building trust for machines among their users.

Keywords Visual question answering · VQA · VQA challenge

1 Introduction

Language and vision problems such as image captioning (Fang et al. 2015; Chen and Zitnick 2015; Donahue et al. 2015; Karpathy and Fei-Fei 2015; Vinyals et al. 2015; Kiros et al. 2015; Mao et al. 2014) and visual question answering (VQA) (Antol et al. 2015; Malinowski and Fritz 2014; Malinowski et al. 2015; Gao et al. 2015; Ren et al. 2015) have gained popularity in recent years as the computer vision research community is progressing beyond

Communicated by Svetlana Lazebnik.

Yash Goyal, Tejas Khot, and Aishwarya Agrawal: Part of this work was done while at Virginia Tech, Blacksburg, VA, USA.

✉ Yash Goyal
ygoyal@gatech.edu

Tejas Khot
tkhot@andrew.cmu.edu

Aishwarya Agrawal
aishwarya@gatech.edu

Douglas Summers-Stay
douglas.a.summers-stay.civ@mail.mil

Dhruv Batra
dbatra@gatech.edu

Devi Parikh
parikh@gatech.edu

[1] Georgia Tech, Atlanta, GA, USA

[2] Carnegie Mellon University, Pittsburgh, PA, USA

[3] Army Research Lab, Adelphi, MD, USA

[4] Facebook AI Research, Menlo Park, CA, USA

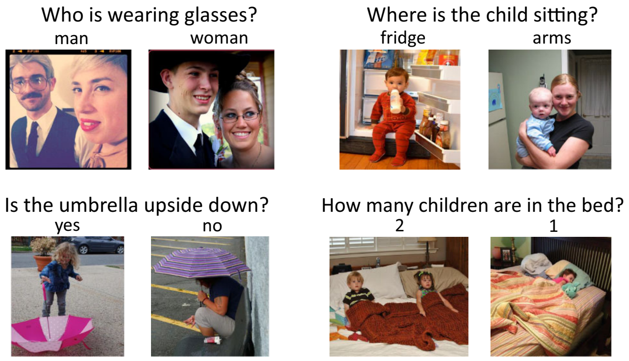

Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

Fig. 1 Examples from our balanced VQA v2.0 dataset

"bucketed" recognition and towards solving multi-modal problems.

The complex compositional structure of language makes problems at the intersection of vision and language challenging. But recent works (Devlin et al. 2015; Zhang et al. 2016; Zhou et al. 2015; Jabri et al. 2016; Kafle and Kanan 2016b; Agrawal et al. 2016) have pointed out that language also provides a strong prior that can result in good superficial performance, without the underlying models truly understanding the visual content.

This phenomenon has been observed in image captioning (Devlin et al. 2015) as well as visual question answering (Zhang et al. 2016; Zhou et al. 2015; Jabri et al. 2016; Kafle and Kanan 2016b; Agrawal et al. 2016). For instance, in the VQA (Antol et al. 2015) dataset, the most common sport answer "tennis" is the correct answer for 41% of the questions starting with "What sport is", and "2" is the correct answer for 39% of the questions starting with "How many". Moreover, (Zhang et al. 2016) points out a particular 'visual priming bias' in the VQA dataset—specifically, subjects saw an image while asking questions about it. Thus, people only ask the question "Is there a clock tower in the picture?" on images actually containing clock towers. As one particularly perverse example—for questions in the VQA dataset starting with the n-gram "Do you see a …", blindly answering "yes" without reading the rest of the question or looking at the associated image results in a VQA accuracy of 87%!

Such language priors are not specific to VQA dataset from (Antol et al. 2015), but are also present in other VQA datasets. For example, in another popular VQA dataset Visual7W (Zhu et al. 2016), a question-only baseline achieves an accuracy of 46.2%, while the baseline question + image model achieves 52.1%. So, models are able to achieve good accuracy using only language information and without even looking at the image, and visual information is only making slight relative improvement compared to the question-only baseline.

Hence, these language priors can give a false impression that machines are making progress towards the goal of understanding images correctly when they are only exploiting language priors to achieve high accuracy. This can hinder progress in pushing state of art in the computer vision aspects of multi-modal AI (Torralba and Efros 2011; Zhang et al. 2016).

In this work, we propose to counter these language biases and elevate the role of image understanding in VQA. In order to accomplish this goal, we collect a balanced VQA dataset with significantly reduced language biases. Specifically, we create a balanced VQA dataset in the following way—given an (image, question, answer) triplet (I, Q, A) from the VQA dataset, we ask a human subject to identify an image I' that is similar to I but results in the answer to the question Q to become A' (which is different from A). Examples from our balanced dataset are shown in Fig. 1. More random examples can be seen in Fig. 2 and on the project website.[1]

Our hypothesis is that this balanced dataset will force VQA models to focus on visual information. After all, when a question Q has two different answers (A and A') for two different images (I and I' respectively), the only way to know the right answer is by looking at the image. Language-only models have simply no basis for differentiating between the two cases—(Q, I) and (Q, I'), and by construction must get one wrong. We believe that this construction will also prevent language+vision models from achieving high accuracy by exploiting language priors, enabling VQA evaluation protocols to more accurately reflect progress in image understanding.

Our balanced VQA dataset is also particularly difficult because the picked complementary image I' is close to the original image I in the semantic (fc7) space of VGGNet (Simonyan and Zisserman 2015) features. Therefore, VQA models will need to understand the subtle differences between the two images to predict the answers to both the images correctly.

Note that simply ensuring that the answer distribution $P(A)$ is uniform across the dataset would not accomplish the goal of alleviating language biases discussed above. This is because language models exploit the correlation between question n-grams and the answers, e.g. questions starting with "Is there a clock" has the answer "yes" 98% of the time, and questions starting with "Is the man standing" has the answer "no" 69% of the time. What we need is not just higher entropy in $P(A)$ across the dataset, but higher entropy in $P(A|Q)$ so that image I must play a role in determining A. This motivates our balancing on a per-question level.

Our complete balanced dataset contains approximately *1.1 Million* (image, question) pairs—almost *double* the size of the VQA (Antol et al. 2015) dataset—with approximately *13 Million* associated answers on the ∼200 k images from COCO (Lin et al. 2014). We believe this balanced VQA dataset is a better dataset to benchmark VQA approaches,

[1] http://visualqa.org/.

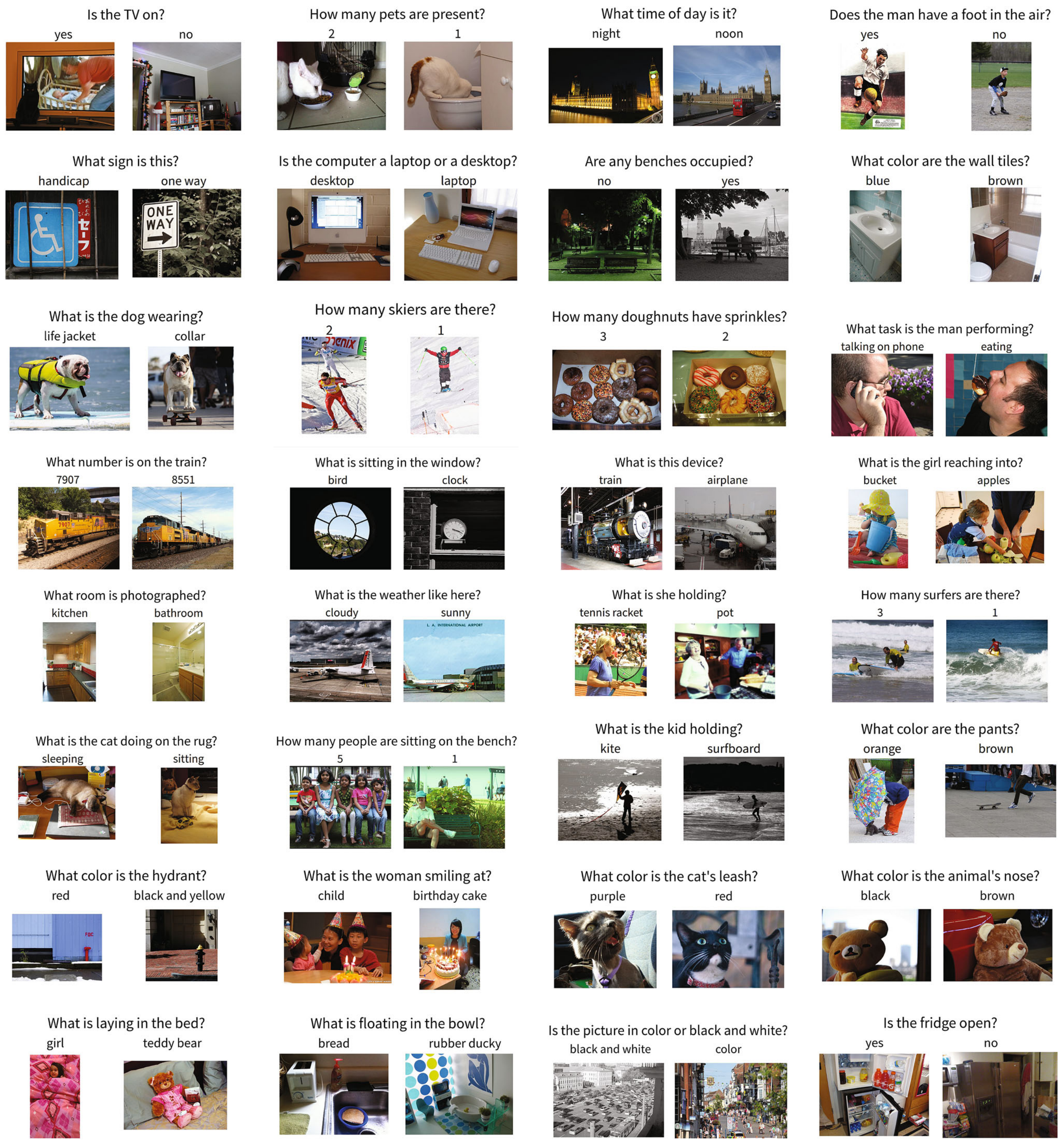

Is the TV on?
yes
no
How many pets are present?
2
1
What time of day is it?
night
noon
Does the man have a foot in the air?
yes
no
What sign is this?
handicap
one way
Is the computer a laptop or a desktop?
desktop
laptop
Are any benches occupied?
no
yes
What color are the wall tiles?
blue
brown
What is the dog wearing?
life jacket
collar
How many skiers are there?
2
1
How many doughnuts have sprinkles?
3
2
What task is the man performing?
talking on phone
eating
What number is on the train?
7907
8551
What is sitting in the window?
bird
clock
What is this device?
train
airplane
What is the girl reaching into?
bucket
apples
What room is photographed?
kitchen
bathroom
What is the weather like here?
cloudy
sunny
What is she holding?
tennis racket
pot
How many surfers are there?
3
1
What is the cat doing on the rug?
sleeping
sitting
How many people are sitting on the bench?
5
1
What is the kid holding?
kite
surfboard
What color are the pants?
orange
brown
What color is the hydrant?
red
black and yellow
What is the woman smiling at?
child
birthday cake
What color is the cat's leash?
purple
red
What color is the animal's nose?
black
brown
What is laying in the bed?
girl
teddy bear
What is floating in the bowl?
bread
rubber ducky
Is the picture in color or black and white?
black and white
color
Is the fridge open?
yes
no

Fig. 2 Random examples from our proposed balanced VQA v2.0 dataset. Each question has two similar images with different answers to the question

and is publicly available for download on the project website.

Finally, our data collection protocol enables us to develop a counter-example based explanation modality. We propose a novel model that not only answers questions about images, but also 'explains' its answer to an image-question pair by providing "hard negatives" i.e., examples of images that it believes are similar to the image at hand, but it believes have different answers to the question. Such an explanation modality will allow users of the VQA model to establish greater trust in the model and identify its oncoming failures.

Our main contributions are as follows:

1. We balance the existing VQA v1.0 dataset (Antol et al. 2015) by collecting complementary images such that almost every question in our balanced dataset is associated with not just a single image, but rather a pair of similar images that result in two different answers to the question. The result is a more balanced VQA v2.0 dataset, which is also approximately twice the size of the VQA v1.0 dataset.
2. We evaluate state-of-art VQA models (with publicly available code) on our balanced VQA v2.0 dataset, and show that models trained on the existing 'unbalanced' VQA v1.0 dataset perform poorly on our new balanced VQA v2.0 dataset. This finding confirms our hypothesis that these models have been exploiting language priors in the existing VQA v1.0 dataset to achieve higher accuracy.
3. We organized the VQA Challenge 2017 on the proposed VQA v2.0 dataset. The results were announced in the 2nd VQA Challenge Workshop at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017. In this work, we analyze the challenge entries for various factors such as statistical significance, sensitivity to subtle changes in images, compositionality, effect of priors, etc., and present interesting insights into the results.
4. Finally, our data collection protocol for identifying complementary scenes enables us to develop a novel interpretable model, which in addition to answering questions about images, also provides a counter-example based explanation—it retrieves images that it believes are similar to the original image but have different answers to the question. Such explanations can help in building trust for machines among their users.

2 Related Work

2.1 Visual Question Answering

A number of recent works have proposed visual question answering datasets (Antol et al. 2015; Krishna et al. 2016; Malinowski and Fritz 2014; Ren et al. 2015; Gao et al. 2015; Yu et al. 2015; Tapaswi et al. 2016; Shin et al. 2016) and models (Fukui et al. 2016; Lu et al. 2016; Andreas et al. 2016; Xiong et al. 2016; Lu et al. 2015; Malinowski et al. 2015; Zhang et al. 2016; Yang et al. 2016; Xu and Saenko 2016; Wang et al. 2015; Shih et al. 2016; Kim et al. 2016; Noh and Han 2016; Ilievski et al. 2016; Wu et al. 2016; Saito et al. 2016; Kafle and Kanan 2016a). Our work builds on top of the VQA dataset from (Antol et al. 2015), which is one of the most widely used VQA datasets. We reduce the language biases present in this popular dataset, resulting in a dataset that is more balanced and about twice the size of the VQA dataset. We benchmark one 'baseline' VQA model (Lu et al. 2015), one attention-based VQA model (Lu et al. 2016), and the winning model from the VQA Real Open Ended Challenge 2016 (Fukui et al. 2016) on our balanced VQA dataset, and compare them to a language-only model.

2.2 Data Balancing and Augmentation

At a high level, our work may be viewed as constructing a more rigorous evaluation protocol by collecting 'hard negatives'. In that spirit, it is similar to the work of Hodosh and Hockenmaier (2016), who created a binary forced-choice image captioning task, where a machine must choose to caption an image with one of two similar captions. To compare, (Hodosh and Hockenmaier 2016) implemented hand-designed rules to create two similar captions for images, while we create a novel annotation interface to collect two similar images for questions in VQA.

Perhaps the most relevant to our work is that of Zhang et al. (2016), who study this goal of balancing VQA in a fairly restricted setting—binary (yes/no) questions on abstract scenes made from clipart [part of the VQA abstract scenes dataset (Antol et al. 2015)]. Using clipart allows Zhang et al. to ask human annotators to "change the clipart scene such that the answer to the question changes". Unfortunately, such fine-grained editing of image content is simply not possible in real images. The novelty of our work over Zhang et al. is the proposed complementary image data collection interface, application to real images, extension to *all* questions (not just binary ones), benchmarking of state-of-art VQA models on the balanced dataset, and finally the novel VQA model with counter-example based explanations.

2.3 Models with Explanation

A number of recent works have proposed mechanisms for generating 'explanations' (Hendricks et al. 2016; Selvaraju et al. 2016; Zhou et al. 2015; Goyal et al. 2016; Ribeiro et al. 2016) for the predictions made by deep learning models, which are typically 'black-box' and non-interpretable. (Hendricks et al. 2016) generates a natural language explanation (sentence) for image categories. (Selvaraju et al. 2016; Zhou et al. 2015; Goyal et al. 2016; Ribeiro et al. 2016) provide 'visual explanations' or spatial maps overlaid on images to highlight the regions that the model focused on while making its predictions. In this work, we introduce a third explanation modality: counter-examples, instances the the model believes are close to but not belonging to the category predicted by the model.

Closest to our counter-example explanation work is the work by Berg and Belhumeur (2013) for fine-grained bird classification. They identify features which best show the

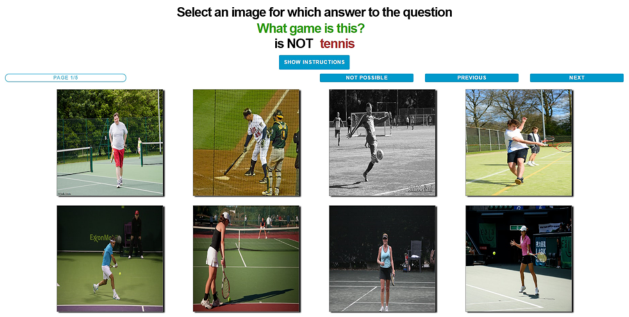

Select an image for which answer to the question
What game is this?
is NOT tennis

Fig. 3 A snapshot of our Amazon Mechanical Turk (AMT) interface to collect complementary images

difference between two similar classes, exemplify this difference using example images and identify the regions in these images to show the distinguishing features. Other example-based explanation works include (Doersch et al. 2012) which finds most distinctive visual elements for each class represented by clusters of image patches and (Koh and Liang 2017) which identifies training examples most responsible for a prediction.

3 Dataset

We build on top of the VQA dataset introduced by Antol et al. (2015). VQA real images dataset contains just over 204K images from COCO (Lin et al. 2014), 614 K free-form natural language questions (3 questions per image), and over 6 million free-form (but concise) answers (10 answers per question). While this dataset has spurred significant progress in VQA domain, as discussed earlier, it has strong language biases.

Our key idea to counter this language bias is the following—for every (image, question, answer) triplet (I, Q, A) in the VQA dataset, our goal is to identify an image I' that is similar to I, but results in the answer to the question Q to become A' (which is different from A). We built an annotation interface (shown in Fig. 3) to collect such complementary images on Amazon Mechanical Turk (AMT). AMT workers are shown 24 nearest-neighbor images of I, the question Q, and the answer A, and asked to pick an image I' from the list of 24 images for which Q "makes sense" and the answer to Q is *not* A.

To capture "question makes sense", we explained to the workers (and conducted qualification tests to make sure that they understood) that any premise assumed in the question must hold true for the image they select. For instance, the question "What is the woman doing?" assumes that a woman is present and can be seen in the image. It does not make sense to ask this question on an image without a woman visible in it.

We compute the 24 nearest neighbors by first representing each image with the activations from the penultimate ('fc7') layer of a deep Convolutional Neural Network (CNN)—in particular VGGNet (Simonyan and Zisserman 2015)—and then using ℓ_2-distances to compute neighbors.

After the complementary images are collected, we conduct a second round of data annotation to collect answers on these new images. Specifically, we show the picked image I' with the question Q to 10 new AMT workers, and collect 10 ground truth answers (similar to (Antol et al. 2015)). The most common answer among the 10 is the new answer A'.

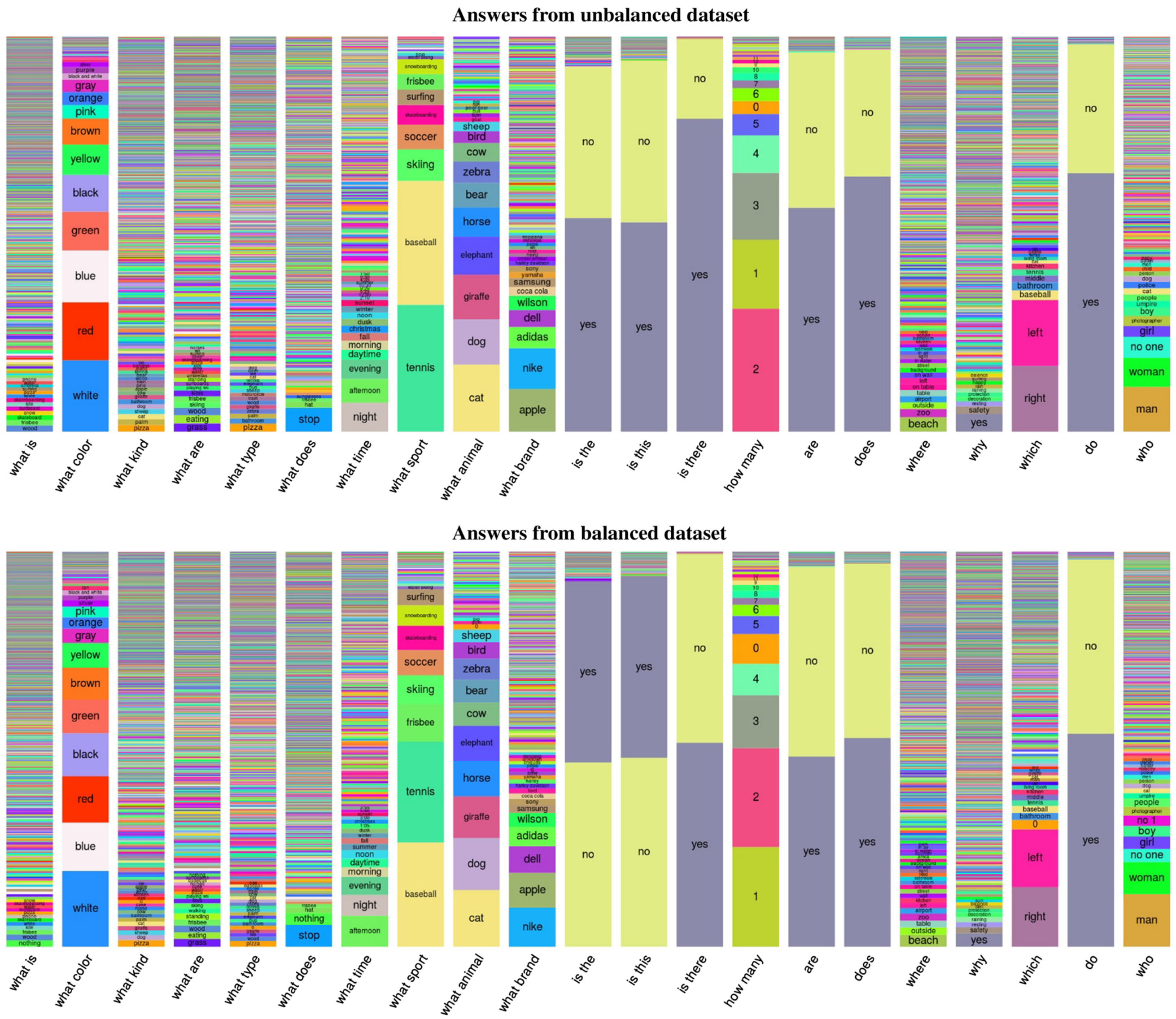
This two-stage data collection process finally results in pairs of complementary images I and I' that are semantically similar, but have different answers A and A' respectively to the same question Q. Since I and I' are semantically similar, a VQA model will have to understand the subtle differences between I and I' to provide the right answer to both images. Example complementary images are shown in Figs. 1, 2, and on the project website.

Note that sometimes it may not be *possible* to pick one of the 24 neighbors as a complementary image. This is because either (1) the question does not make sense for any of the 24 images (e.g. the question is 'what is the woman doing?' and none of the neighboring images contain a woman), or (2) the question is applicable to some neighboring images, but the answer to the question is still A (same as the original image I). In such cases, our data collection interface allowed AMT workers to select "not possible".

We analyzed the data annotated with "not possible" selection by AMT workers and found that this typically happens when (1) the object being talked about in the question is too small in the original image and thus the nearest neighbor images, while globally similar, do not necessarily contain the object resulting in the question not making sense, or (2) when the concept in the question is rare (e.g., when workers are asked to pick an image such that the answer to the question "What color is the banana?" is NOT "yellow").

In total, such "not possible" selections make up 22% of all the questions in the VQA dataset. We believe that a more sophisticated interface that allowed workers to scroll through many more than 24 neighboring images could possibly reduce this fraction. But, (1) it will likely still not be 0 (there may be no image in COCO where the answer to "is the woman flying?" is NOT "no"), and (2) the task would be significantly more cumbersome for workers, making the data collection significantly more expensive.

We collected complementary images and the corresponding new answers for all of train, val and test splits of the VQA dataset. AMT workers picked "not possible" for approximately 135 K total questions. In total, we collected approximately 195 K complementary images for train, 93 K complementary images for val, and 191 K complementary images for test set. In addition, we augment the test set with ~18K additional (question, image) pairs to provide additional means to detect anomalous trends on the test data.


Answers from unbalanced dataset

Answers from balanced dataset

Fig. 4 Distribution of answers per question type for a random sample of 60 K questions from the (unbalanced) VQA v1.0 dataset (Antol et al. 2015) (top) and from our proposed balanced VQA v2.0 dataset (bottom)

Hence, our complete balanced dataset contains more than 443 K train, 214 K val and 453 K test (question, image) pairs. Following VQA v1.0 dataset (Antol et al. 2015), we divide our test set into 4 splits: test-dev, test-standard, test-challenge and test-reserve. For more details, please refer to (Antol et al. 2015). Our complete balanced dataset is publicly available for download.

We use the publicly released VQA evaluation script in our experiments. The evaluation metric uses 10 ground-truth answers for each question to compute VQA accuracies. As described above, we collected 10 answers for every complementary image and its corresponding question to be consistent with the VQA dataset (Antol et al. 2015). Note that while unlikely, it is possible that the majority vote of the 10 new answers may not match the intended answer of the person picking the image either due to inter-human disagreement, or if the worker selecting the complementary image simply made a mistake. We find this to be the case—i.e., A to be the same as A'—for about 9% of our questions.

Figure 4 compares the distribution of answers per question-type in our balanced VQA v2.0 dataset with the (unbalanced) VQA v1.0 dataset (Antol et al. 2015). We notice several interesting trends. First, binary questions (e.g. “is the”, “is this”, “is there”, “are”, “does”) have a *significantly* more balanced distribution over “yes” and “no” answers in our balanced dataset compared to unbalanced VQA dataset. “baseball” is now slightly more popular than “tennis” under “what sport”, and more importantly, overall “baseball” and “tennis” dominate less in the answer distribution. Several other sports like “frisbee”, “skiing”, “soccer”, “skateboarding", "snowboard" and "surfing" are more visible in the answer distribution in the balanced dataset, suggesting that it contains heavier tails. Similar trends can be seen across the board with colors, animals, numbers, etc. Quantitatively, we find that the entropy of answer distributions averaged across various question types (weighted by frequency of question types) increases by 56% after balancing, confirming the heavier tails in the answer distribution.

As the statistics show, while our balanced dataset is not perfectly balanced, it is *significantly* more balanced than the original VQA v1.0 dataset. The resultant impact of this balancing on performance of state-of-the-art VQA models is discussed in the next section.

Table 1 Performance of VQA models when trained/tested on unbalanced/balanced VQA datasets

Approach	UU	UB	B_{half}B	BB
Prior	27.38	24.04	24.04	24.04
Language-only	48.21	41.40	41.47	43.01
d-LSTM+n-I (Lu et al. 2015)	54.40	47.56	49.23	51.62
NMN (Andreas et al. 2016)	54.83	47.97	49.52	51.62
HieCoAtt (Lu et al. 2016)	57.09	50.31	51.88	54.57
MCB (Fukui et al. 2016)	60.36	54.22	56.08	59.14

UB stands for training on **U**nbalanced VQA v1.0 train and testing on **B**alanced VQA v2.0 val datasets. UU, B_{half}B and BB are defined analogously

4 Benchmarking Existing VQA Models

Our first approach to training a VQA model that emphasizes the visual information over language-priors-alone is to re-train the existing state-of-art VQA models [with code publicly available (Lu et al. 2015; Andreas et al. 2016; Lu et al. 2016; Fukui et al. 2016)] on our new balanced VQA dataset. Our hypothesis is that simply training a model to answer questions correctly on our balanced dataset will already encourage the model to focus more on the visual signal, since the language signal alone has been impoverished. We experiment with the following models:

Deeper LSTM Question + norm Image (d-LSTM+n-I) (Lu et al. 2015) This was the VQA model introduced in Antol et al. (2015) together with the dataset. It uses a CNN embedding of the image, a Long-Short Term Memory (LSTM) embedding of the question, combines these two embeddings via a point-wise multiplication, followed by a multi-layer perceptron classifier to predict a probability distribution over 1000 most frequent answers in the training dataset.

Neural Module Networks (NMN) (Andreas et al. 2016) This is a compositional VQA model which dynamically initiates a different network architecture for each test example based on the linguistic substructure of the question using neural "modules", which are specialized for subtasks such as recognizing dogs, classifying colors, etc.

Hierarchical Co-attention (HieCoAtt) (Lu et al. 2016) This is an attention-based VQA model that 'co-attends' to both the image and the question to predict an answer. Specifically, it models the question (and consequently the image via the co-attention mechanism) in a hierarchical fashion: at the word-level, phrase-level and entire question-level. These levels are combined recursively to produce a distribution over the 1000 most frequent answers.

Multimodal Compact Bilinear Pooling (MCB) (Fukui et al. 2016) This is the winning entry on the real images track of the VQA Challenge 2016. This model uses a multimodal compact bilinear pooling mechanism to attend over image features and combine the attended image features with language features. These combined features are then passed through a fully-connected layer to predict a probability distribution over the 3000 most frequent answers. It should be noted that MCB uses image features from a more powerful CNN architecture ResNet (He et al. 2016) while the previous three models use image features from VGGNet (Simonyan and Zisserman 2015).

Baselines To put the accuracies of these models in perspective, we compare to the following baselines: *Prior:* Predicting the most common answer in the training set, for all test questions. The most common answer is "yes" in both the unbalanced and balanced sets. *Language-only:* This language-only baseline has a similar architecture as Deeper LSTM Question + norm Image (Lu et al. 2015) except that it only accepts the question as input and does not utilize any visual information. Comparing VQA models to language-only ablations quantifies to what extent VQA models have succeeded in leveraging the image to answer the questions.

The results are shown in Table 1 when models are trained on train set and evaluated on val set. For fair comparison of accuracies with original (unbalanced) dataset (VQA v1.0), we create a balanced train set which is of similar size as VQA v1.0 train set (referred to as B_{half} in table). For benchmarking, we also report results using the full balanced train set.

We see that the current state-of-art VQA models trained on (unbalanced) VQA v1.0 train set perform significantly worse when evaluated on our balanced VQA v2.0 val set, compared to evaluating on the unbalanced VQA v1.0 val set (i.e., comparing UB to UU respectively in the table). This finding confirms our hypothesis that existing models have learned severe language biases present in the train set, resulting in a reduced ability to answer questions correctly when the same question has different answers on different images. When these models are trained on our balanced VQA v2.0 train set, their performance improves (compare UB to B_{half}B in the table). Further, when models are trained on complete VQA v2.0 train set (~twice the size of VQA v1.0 train set), the accuracy improves by 2–3% (compare B_{half}B to BB).

This increase in accuracy suggests that current VQA models are data starved, and would benefit from even larger VQA datasets.

As the absolute numbers in the table suggest, there is significant room for improvement in building visual understanding models that can extract detailed information from images and leverage this information to answer free-form natural language questions about images accurately. As expected from the construction of this balanced dataset, the question-only approach performs *significantly* worse on the balanced dataset compared to the unbalanced dataset, again confirming the language-bias in the VQA v1.0 dataset, and its successful alleviation (though not elimination) in our proposed balanced VQA v2.0 dataset.

Note that in addition to the lack of language bias, visual reasoning is also challenging on the balanced dataset since there are pairs of images very similar to each other in image representations learned by CNNs, but with different answers to the same question. To be successful, VQA models need to understand the subtle differences in these images.

The paired construction of our dataset allows us to analyze the performance of VQA models in unique ways. Given the prediction of a VQA model, we can count the number of questions where *both* complementary images (I, I') received correct answer predictions for the corresponding question Q, or both received identical (correct or incorrect) answer predictions, or both received different answer predictions. For the HieCoAtt (Lu et al. 2016) model, when trained on the unbalanced VQA v1.0 dataset, 13.5% of the pairs were answered correctly, 59.9% of the pairs had identical predictions, and 40.1% of the pairs had different predictions. In comparison, when trained on balanced VQA v2.0 dataset, the same model answered 17.7% of the pairs correctly, a 4.2% increase in performance! Moreover, it predicts identical answers for 10.5% fewer pairs (49.4%). This shows that by training on balanced dataset, this VQA model has learned to tell the difference between two otherwise similar images. However, significant room for improvement remains. The VQA model still can not tell the difference between two images that have a noticeable difference—a difference enough to result in the two images having different ground truth answers for the same question asked by humans.

To benchmark models on VQA v2.0 dataset, we also train these models on VQA v2.0 train+val and report results on VQA v2.0 test-standard in Table 2. Papers reporting results on VQA v2.0 dataset are suggested to report test-standard accuracies and compare their methods' accuracies with accuracies reported in Table 2.

4.1 Analysis of Accuracies for Different Answer Types

We further analyze the accuracy breakdown over answer types for Multimodal Compact Bilinear Pooling (MCB) (Fukui et al. 2016), Hierarchical Co-attention (HieCoAtt) (Lu et al. 2016) and Neural Module Networks (NMN) (Andreas et al. 2016) models.

The results are shown in Table 3. First, we immediately notice that the accuracy for the answer-type "yes/no" drops significantly from UU to UB (∼ 10.8% for MCB, ∼ 12.4% for HieCoAtt and ∼ 12.2% for NMN). This suggests that these VQA models are really exploiting language biases for "yes/no" type questions, which leads to high accuracy on unbalanced val set because the unbalanced val set also contains these biases. But performance drops significantly when tested on the balanced val set which has significantly reduced biases.

Second, we note that for all three state-of-art VQA models, the largest source of improvement from UB to $\mathrm{B_{half}B}$ is the "yes/no" answer-type (∼ 4.5% for MCB, ∼ 3.3% for HieCoAtt and ∼ 4.2% for NMN) and the "number" answer-type (∼ 3% for MCB, ∼ 2% for HieCoAtt and and ∼ 2.5% for NMN).

This trend is particularly interesting since the "yes/no" and "number" answer-types are the ones where existing approaches have shown minimal improvements. For instance, in the results announced at the VQA Real Open Ended Challenge 2016, the accuracy gap between the top-4 approaches is a mere 0.15% in "yes/no" answer-type category (and a gap of 3.48% among the top-10 approaches). Similarly, "number" answer-type accuracies only vary by 1.51% and 2.64% respectively. The primary differences between current generation of state-of-art approaches seem to come from the "other" answer-type where accuracies vary by 7.03% and 10.58% among the top-4 and top-10 entries.

This finding suggests that language priors present in the unbalanced VQA dataset (particularly in the "yes/no" and "number" answer-type questions) lead to similar accuracies for all state-of-art VQA models, rendering vastly different models virtually indistinguishable from each other (in terms of their accuracies for these answer-types). Benchmarking these different VQA models on our balanced dataset (with reduced language priors) may finally allow us to distinguish between 'good' models (ones that encode the 'right' inductive biases for this task, such as attention-based or compositional models) from others that are simply high-capacity models tuning themselves to the biases in the dataset.

5 VQA Challenge 2017

Following VQA v1.0, we have not released the test set annotations publicly. We have set up an evaluation server on EvalAI[2] where researchers can upload their models' predictions and evaluate their performance. To encourage,

[2] https://evalai.cloudcv.org/.

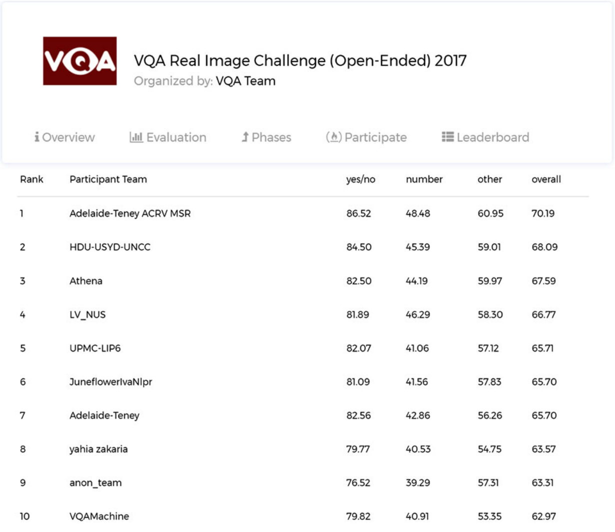
Table 2 Performance of VQA models when trained on VQA v2.0 train+val and tested on VQA v2.0 test-standard dataset

Approach	All	Yes/no	Number	Other
Prior	25.98	61.20	00.36	01.17
Language-only	44.26	67.01	31.55	27.37
d-LSTM+n-I (Lu et al. 2015)	54.22	73.46	35.18	41.83
MCB (Fukui et al. 2016)	62.27	78.82	38.28	53.36

Table 3 Accuracy breakdown over answer types for MCB (Fukui et al. 2016), HieCoAtt (Lu et al. 2016) and NMN (Andreas et al. 2016) models when trained/tested on unbalanced/balanced VQA datasets

Approach	Ans type	UU	UB	B_{half}B	BB
MCB (Fukui et al. 2016)	Yes/no	81.20	70.40	74.89	77.37
	Number	34.80	31.61	34.69	36.66
	Other	51.19	47.90	47.43	51.23
	All	60.36	54.22	56.08	59.14
HieCoAtt (Lu et al. 2016)	Yes/no	79.99	67.62	70.93	71.80
	Number	34.83	32.12	34.07	36.53
	Other	45.55	41.96	42.11	46.25
	All	57.09	50.31	51.88	54.57
NMN (Andreas et al. 2016)	Yes/no	80.39	68.21	72.45	73.38
	Number	33.45	30.00	32.53	33.23
	Other	41.07	37.33	36.57	39.93
	All	54.83	47.97	49.52	51.62

UB stands for training on **U**nbalanced VQA v1.0 train and testing on **B**alanced VQA v2.0 val datasets. UU, B_{half}B and BB are defined analogously



VQA Real Image Challenge (Open-Ended) 2017

Organized by: VQA Team

Rank	Participant Team	yes/no	number	other	overall
1	Adelaide-Teney ACRV MSR	86.52	48.48	60.95	70.19
2	HDU-USYD-UNCC	84.50	45.39	59.01	68.09
3	Athena	82.50	44.19	59.97	67.59
4	LV_NUS	81.89	46.29	58.30	66.77
5	UPMC-LIP6	82.07	41.06	57.12	65.71
6	JuneflowerIvaNlpr	81.09	41.56	57.83	65.70
7	Adelaide-Teney	82.56	42.86	56.26	65.70
8	yahia zakaria	79.77	40.53	54.75	63.57
9	anon_team	76.52	39.29	57.31	63.31
10	VQAMachine	79.82	40.91	53.35	62.97

Fig. 5 A snapshot of leaderboard of VQA Challenge 2017

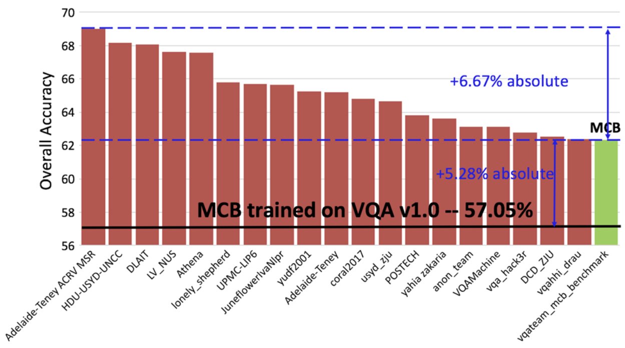


Fig. 6 Challenge accuracies for top-20 teams, including a single best model from VQA Challenge 2016—MCB (Fukui et al. 2016) trained on VQA v2.0 dataset (green bar) added to the challenge by us for comparison. We also show the accuracy of MCB model when trained on VQA v1.0 dataset (black horizontal line) for reference (Color figure online)

systematically track, and benchmark research in this area, we organized the VQA Challenge 2017 on the proposed VQA v2.0 dataset. The results were announced in the 2nd VQA Challenge Workshop[3] at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017. For more details, please see the challenge page.[4] A screenshot of the leaderboard has been shown in Fig. 5. We further analyze the challenge entries and present interesting insights into the results below.

[3] http://visualqa.org/workshop_2017.html.

[4] http://visualqa.org/challenge_2017.html.

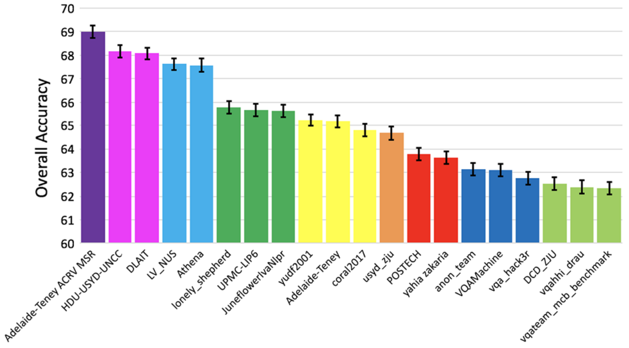


Fig. 7 Challenge accuracies for top-20 teams, where statistically similar teams have been grouped together (shown in same color) (Color figure online)

5.1 Analysis

5.1.1 Overall progress from VQA Challenge 2016.

Figure 6 shows the challenge accuracies for top-20 teams which participated in the 2017 challenge. The green bar corresponds to the MCB (Fukui et al. 2016) model[5]—the single best model of the challenge winning entry from VQA Challenge 2016, benchmarked by us on VQA v2 after retraining. As we can see, there is a significant improvement in the performance compared to the MCB model, with an absolute improvement of about 7%. It is worth noting that 19 teams in the VQA Challenge 2017 outperformed this MCB benchmark. Note that the performance of the MCB model trained on VQA v1 is only 57.05% (denoted by black solid line). So, training on balanced data improves the performance of the model by 5.28%.

5.1.2 Statistical Significance of the Results

In order to determine whether performance of teams are statistically significantly different from one another, we bootstrapped samples from predictions 5000 times and report statistical significance at 95% confidence. Figure 7 shows grouping of teams where teams which are statistically similar to each other have been grouped together (shown in same color). We can see that the challenge winner is statistically significantly different from everyone else. The team at the 2nd rank is statistically similar to 3rd team and is better than rest of the teams. And so on.

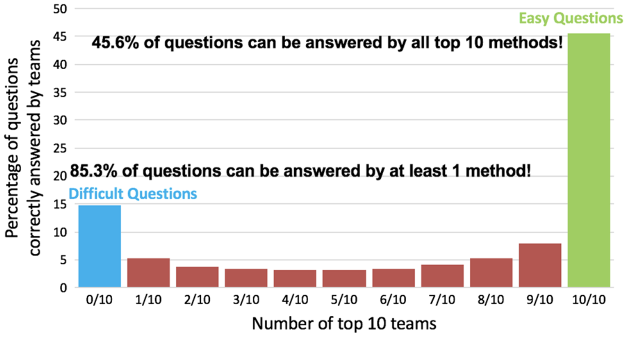


Fig. 8 Distribution of test set based on how many teams out of top-10 are able to correctly answer the question. The first bin (shown in blue) shows the percentage of questions which none of the top 10 teams answered correctly. The last bin (shown in green) shows the percentage of questions which all of the top 10 teams answered correctly (Color figure online)

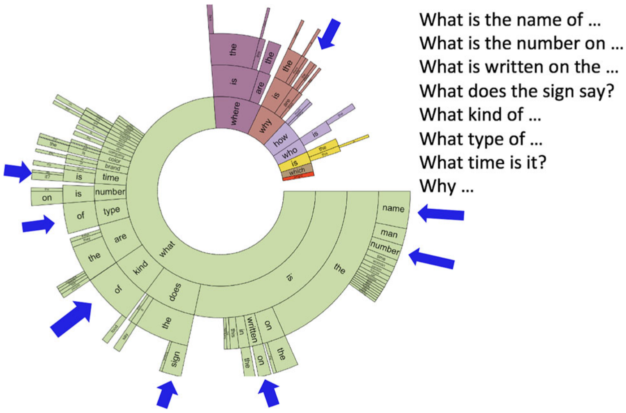


Fig. 9 Visualization of difficult questions whose ground-truth answers are not among the top-1000 ground-truth answers in the training set. Hence, the models which use classification over these top-1000 answers can not answer these questions correctly. (Best viewed after zooming in.)

5.1.3 Easy and Difficult Questions

We also analyze if there are certain questions which none of the top 10 teams could answer, and if there are certain questions which all top 10 teams were able to answer. In Fig. 8, the x-axis shows the number of teams out of top 10 that were able to correctly answer certain questions, and the y-axis shows the percentage of questions they could correctly answer. The first bin (shown in blue) shows the percentage of questions which none of the top 10 teams answered correctly, hence these are 'difficult' questions. This means that 85.3% of questions could be answered correctly by at least one method out of top 10. The last bin (shown in green) shows the percentage of questions which all of the top 10 teams got right, hence these are 'easy' questions.

Note that the set of difficult questions also includes those questions whose ground-truth answers are not in top K most

[5] Note that this entry is a single model and does not use pretrained word embeddings and data augmentation unlike the winning entry in VQA Challenge 2016 which was an ensemble of 7 such MCB models, and was trained with pretrained Glove (Pennington et al. 2014) embeddings and data augmentation from Visual Genome dataset (Krishna et al. 2016). These three factors lead to a 2–3% increase in performance.

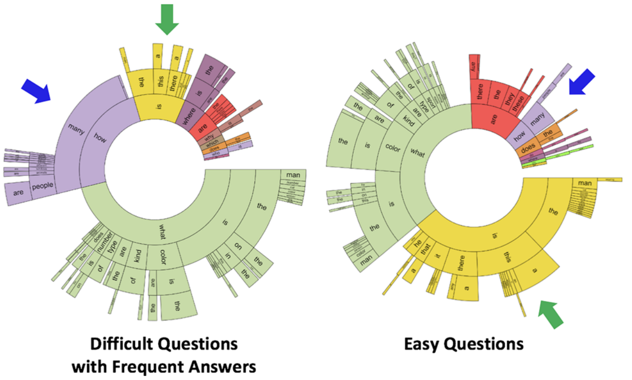


Fig. 10 Left: Visualization of difficult questions whose ground-truth answers are among the top-1000 ground-truth answers in the training set. Right: Easy questions which are correctly answered by all top-10 teams. (Best viewed after zooming in.)

common answers from the training set. Hence, methods using classification over top K (= 1000, typically) answers, would not be able to get those questions correct. We found that half of these difficult questions are the questions whose answers are not in top 1000 answers. The visualization in Fig. 9 shows such questions. The innermost arc contains the first word of the question, the next arc contains the 2nd word and so on. It is interesting to note that most of these questions are OCR questions, some are fine grained recognition and other are open-ended such as "Why ... ". These trends are similar to those from the VQA Challenge 2016, showing that progress towards models that can answer such questions still remains to be made. We also qualitatively examine the difficult questions whose answers are in the top 1000 answers, and see how they differ from easy questions. In Fig. 10, we observe that difficult questions contain more counting questions while easy questions contain more binary questions.

5.1.4 Rankings Based on Answer Types

We further analyze the teams' performance based on different answer types—"yes/no", "number" and "other". Interestingly, we observe that the team at 4th rank outperforms all other teams for "number" questions. This is probably because this entry is an ensemble of various models, one of which is trained only on "number" questions. The winner team performs the best for "yes/no" and "other" questions.

5.1.5 Are Models Sensitive to Subtle Changes in Images?

Recall that the VQA v2.0 dataset has 2 images with the same question but different answers and these complementary images are similar to each other. Such a dataset allows us to probe if the models are sensitive to subtle changes in images. We analyze the following: (1) if the model produces *different* predictions for the complementary images given the question, and (2) if the model produces *accurate* predictions for (both) the complementary images.

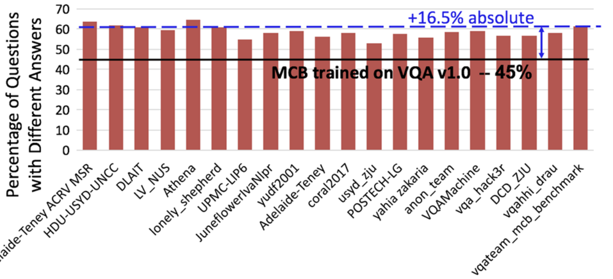


Fig. 11 Percentage of questions for which the model's predictions are different for complementary images. For an MCB model trained on VQA v1.0, the performance on this metric is 45%, as shown by black horizontal line

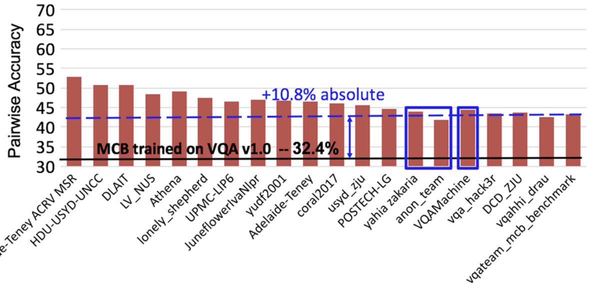


Fig. 12 Percentage of complementary pairs for which the model's predictions are accurate. For an MCB model trained on VQA v1.0, the performance on this metric is 32.4%, as shown by black horizontal line

In Fig. 11, we show the percentage of questions for which the model's predictions are different for complementary images. Most teams predict different answers for about 60% of the pairs. There is not much variation across the teams. Comparing MCB model's predictions with the same model trained on VQA v1.0 dataset (for which the predictions are different for 45% of the complementary pairs), we observe an improvement of 16.5% on this metric. Clearly, training models on the balanced set makes them more sensitive to subtle changes in images.

To check if the model predicts accurate answers for the complementary images, we use a stricter accuracy metric proposed in (Zhang et al. 2016). Under this metric, if the model predicts correct answers for both images, it gets one point. Otherwise, it gets zero points. The results have been shown in Fig. 12. We can see that the top teams get about 50% of the pairs correct. More interestingly, certain teams change their ranks on this metric (highlighted in blue), e.g. the team 'VQAMachine' outperforms the two teams that have higher overall accuracies. Comparing the accuracy of the MCB model under this pairwise metric, we observe that training on VQA v2.0 as compared to VQA v1.0 leads to a gain of about 11%.

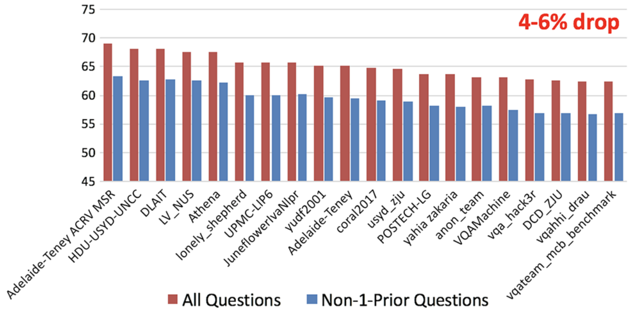


Fig. 13 Accuracy of top-20 teams on all test questions (shown in red) and on 'Non-1-Prior' questions, i.e. test questions whose ground-truth answers are not in top-1 most common answers for the corresponding question type in training set (shown in blue) (Color figure online)

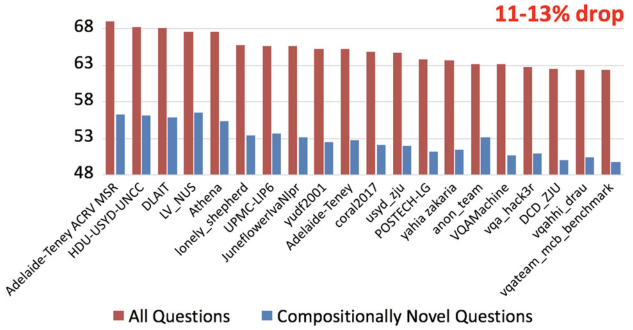


Fig. 14 Accuracy of top-20 teams on all test questions (shown in red) and on compositionally novel test questions, i.e. those (question, answer) pairs which are unseen in the training set but all the concepts in these (question, answer) pairs have been seen in the training set (shown in blue) (Color figure online)

5.1.6 Are VQA Models Driven by Priors?

It has been shown by recent studies (e.g. Zhang et al. 2016; Agrawal et al. 2016, 2017) that today's VQA models are heavily driven by superficial correlations in training data. So it is interesting to ask—how much are the VQA models driven by priors in training data? In order to answer this question, we compute the accuracy over those questions whose answers are not popular answers for the question n-gram in the training data. Therefore, inspired by Agrawal et al. (2018), we create 2 subsets of the VQA v2.0 test set by filtering out the test questions whose ground-truth answers are among top-1 or top-2 answers for the corresponding question type (e.g., "how much ... ", "what are the people ... ", etc.) in the training set. For more details, please refer to (Agrawal et al. 2018). In Fig. 13, we show (in blue) the accuracies of all entries for those test questions whose ground-truth answers are not in top-1 most common answers for the corresponding question type in training set. When compared to the accuracies on all test questions (shown in red), we can see that the performance drops by 4–6%. Extending this setting to evaluate the models on those test questions whose ground-truth answers are not among the top-2 most common answers per question type in training set, we observe a drop of 14–16% in performance as compared to the accuracies on all test questions. We also observe that the relative ranks for some of the teams change for both these subsets of the test set. Hence, we conclude that all of the challenge entries are significantly driven by priors; some more than others.

5.1.7 Are VQA Models Compositional?

We further evaluate how good these models are in answering test questions which are compositionally novel compared to the training set. We define compositionally novel test questions as those (question, answer) pairs which are unseen in the training set but all the concepts in these (question, answer) pairs have been seen in the training set. For example, given the QA pairs ("What color is the plate?", "green") and ("What color are stop lights?", "red") in the training set, a test instance ("What is the color of the plate?", "red") is compositionally novel. Following (Agrawal et al. 2017), we created a subset of the VQA v2.0 test set only containing compositionally novel test questions relative to the training set. In Fig. 14, we plot the accuracies of all teams on this subset of the test set (shown in blue). Compared to the accuracies on the complete test set, we observe a drop of 11–13% in performance for all models. Again, we see that the rankings for some teams change on this compositionally novel test set. Hence we conclude that all challenge entries are poor at dealing with compositionality in VQA, and the ability to deal with compositionality is not directly correlated with VQA v2 test set accuracy.

5.1.8 Trends in VQA v2.0 as Compared to v1.0.

Finally, we analyze the trends in VQA Challenge 2017 and compare them to those in VQA Challenge 2016. In VQA Challenge 2016, the "yes/no" accuracy was saturated (e.g., difference between "yes/no" accuracies of top-4 teams was only 0.15%), there was moderate difference in accuracy of "number" questions (the corresponding difference was 1.51%), and most discriminating were the "other" questions (the difference was 7.03%). On the other hand, in VQA Challenge 2017, both "yes/no" and "number" questions were crucial in determining the winner team. The corresponding differences among top-4 teams are 3.47% and 3.19% respectively, which are significantly better than those in VQA Challenge 2016. For "other" questions, the difference is 0.75%. Hence, VQA v2.0 dataset provides the opportunity to make better progress in "yes/no" and "number" questions which were previously saturated in VQA v1.0 dataset.

5.2 Take-Aways

Based on the analysis of challenge entries, (oral and poster) presentations by challenge participants at the VQA Workshop at CVPR 2017 and descriptions of challenge entries,[6] we provide some promising directions below which tend to be useful while building VQA models. We believe these directions might be helpful in promoting future research in VQA.

- A ResNet (He et al. 2016) model trained for image classification is a better image feature extractor (i.e., improves the performance) for the task of VQA than a VGG (Simonyan and Zisserman 2015) model trained for image classification.
- Using pretrained GloVe (Pennington et al. 2014) word embeddings and then fine-tuning with VQA loss provides better question features (i.e., improves the performance) than learning word embeddings from scratch for VQA task.
- Using additional (Image, Question, Answer) data from other VQA datasets such as Visual Genome (Krishna et al. 2016) helps in improving the performance.
- Using attention over object bounding boxes in the image is better than attention over equally spaced grids (Anderson et al. 2018).
- Various bilinear pooling approaches such as compact (Fukui et al. 2016), low-rank (Kim et al. 2017), factorized (Yu et al. 2017), etc. for combining visual and language features are helpful.
- Co-attention over images and questions i.e., guiding attention over question words using image features and attention over image regions using question features, benefits both image and question attention mechanisms individually.
- Supervision for attention over images seems to improve the attention mechanism as well as VQA performance.
- Training separate modules for different types of questions such as number and yes/no questions also seems to improve the performance of these questions.
- (Teney et al. 2018) presents various other tips and tricks which work well for the task of VQA, for instance: sigmoid outputs (instead of softmax), soft training targets, gated tanh activations, answer embeddings initialized using GloVe and Google Images, large mini-batches, and smart shuffling of training data.

5.3 Discussion

As we have seen in Sect. 5.1, current VQA models are still significantly driven by priors. In this subsection, we provide a brief discussion on some ways to tackle this problem and to build more generic models. We also briefly summarize some recent works which have taken first steps in tackling these problems. In our views, there are three possible ways:

1. *Remove Biases from the Dataset* Our balanced VQA v2.0 dataset is an effort in this direction. However, we balance each question individually. So, certain kinds of biases such as answers given a question type may still exist, however significantly reduced compared to the VQA v1.0 dataset as shown in Fig. 4.
2. *Adding Inductive Biases in the Models* to prevent them from relying on biases in the training data and to encourage visual grounding. For example, GVQA model proposed in Agrawal et al. (2018) explicitly disentangles the recognition of visual concepts present in the image from the identification of plausible answer space for a given question. As shown in Agrawal et al. (2018), such models are less prone to exploiting the priors in the dataset. Another such example is the inductive bias in modular networks (e.g., NMN Andreas et al. 2016, N2NMN Hu et al. 2017) which learn various submodules for different subtasks (such as recognizing dogs, classifying colors, etc.), and dynamically combine these submodules to instantiate a different network architecture for each test question based on the questions linguistic substructure.
3. *Better Evaluation Protocols* Another way to encourage researchers to develop more general models is to have evaluation protocols that explicitly reward grounding and generality. For example, a new split of VQA dataset called VQA under Changing Priors (VQA-CP) proposed in Agrawal et al. (2018) stress tests VQA models for visual grounding by having different distributions of answers given the question type in train and test. Similarly, there are compositionally novel splits of VQA dataset (C-VQA Agrawal et al. 2017) and the CLEVR dataset (Johnson et al. 2017) to test VQA models for compositionality. In terms of accuracy metric, (Kafle and Kanan 2017) proposed various accuracy metrics for VQA to analyze performance of VQA models on rare answers.

6 Counter-Example Explanations

We propose a new explanation modality: counter-examples. We propose a model that when asked a question about an image, not only provides an answer, but also provides example images that are similar to the input image but the model believes have different answers to the input question. This would instill trust in the user that the model does in fact ‘understand’ the concept being asked about. For instance, for a question “What color is the fire-hydrant?” a VQA model may be perceived as more trustworthy if in addition to saying

[6] http://www.visualqa.org/roe_2017.html.

"red", it also adds "unlike this" and shows an example image containing a fire-hydrant that is not red.[7]

6.1 Model

Concretely, at test time, our "negative explanation" or "counter-example explanation" model functions in two steps. In the first step, similar to a conventional VQA model, it takes in an (image, question) pair (Q, I) as input and predicts an answer A_{pred}. In the second step, it uses this predicted answer A_{pred} along with the question Q to retrieve an image that is similar to I but has a different answer than A_{pred} to the question Q. To ensure similarity, the model picks one of K nearest neighbor images of I, $I_{NN} = \{I_1, I_2, ..., I_K\}$ as the counter-example.

How may we find these "negative explanations"? One way of picking the counter-example from I_{NN} is to follow the classical "hard negative mining" strategy popular in computer vision. Specifically, simply pick the image that has the lowest $P(A_{pred}|Q, I_i)$ where $i \in 1, 2, ..., K$. We compare to this strong baseline. While this ensures that $P(A_{pred}|Q, I_i)$ is low for I_i, it does not ensure that the Q "makes sense" for I_i. Thus, when trying to find a negative explanation for "Q: What is the woman doing? A: Playing tennis", this "hard negative mining" strategy might pick an image without a woman in it, which would make for a confusing and non-meaningful explanation to show to a user, if the goal is to convince them that the model has understood the question. One could add a component of question relevance (Ray et al. 2016) to identify better counter-examples.

Instead, we take advantage of our balanced data collection mechanism to directly train for identifying a good counter-example. Note that the I' picked by humans is a good counter-example, by definition. Q is relevant to I' (since workers were asked to ensure it was), I' has a different answer A' than A (the original answer), and I' is similar to I. Thus, we have supervised training data where I' is a counter-example from I_{NN} $(K = 24)$ for question Q and answer A. We train a model that learns to provide negative or counter-example explanations from this supervised data.

To summarize, during test time, our model does two things: first it answers the question (similar to a conventional VQA model), and second, it explains its answer via a counter-example. For the first step, it is given as input an image I and a question Q, and it outputs a predicted answer A_{pred}. For the second (explaining) step, it is given as input the question Q, an answer to be explained A,[8] *and* a set I_{NN} from which the model has to identify the counter-example. At training time, the model is given image I, the question Q, and the corresponding ground-truth answer A to learn to answer questions. It is also given Q, A, I' (human-picked), I_{NN} $(I' \in I_{NN})$ to learn to explain.

Our model architecture contains two heads on top of a shared base 'trunk'—one head for answering the question and the other head for providing an explanation. Specifically, our model consists of three major components:

1. Shared Base: The first component of our model is learning representations of images and questions. It is a 2-channel network that takes in an image CNN embedding as input in one branch, question LSTM embedding as input in another branch, and combines the two embeddings by a point-wise multiplication. This gives us a joint QI embedding, similar to the model in Lu et al. (2015). The second and third components—the answering model and the explaining model—take in this joint QI embedding as input, and therefore can be considered as two heads over this first shared component. A total of 25 images—the original image I and 24 candidate images $\{I_1, I_2, ..., I_{24}\}$ are passed through this shared component of the network.

2. Answering Head: The second component is learning to answer questions. Similar to (Lu et al. 2015), it consists of a fully-connected layer fed into a softmax that predicts the probability distribution over answers given the QI embedding. Only the QI embedding corresponding to the original image I is passed through this component and result in a cross-entropy loss.

3. Explaining Head: The third component is learning to explain an answer A via a counter-example image. It is a 2-channel network which linearly transforms the joint QI embedding (output from the first component) and the answer to be explained A (provided as input)[9] into a common embedding space. It computes an inner product of these 2 embeddings resulting in a scalar number for each image in I_{NN} (also provided as input, from which a counter-example is to be picked). These K inner-product values for K candidate images are then passed through a fully connected layer to generate K scores $S(I_i)$, where $i \in \{1, 2, ..., K\}$. The K candidate images $\{I_1, I_2, ..., I_K\}$ are then sorted according to

[7] It could easily also convey what color it thinks the fire-hydrant is in the counter-example. We will explore this in future work.

[8] In practice, this answer to be explained would be the answer predicted by the first step A_{pred}. However, we only have access to negative explanation annotations from humans for the ground-truth answer

Footnote 8 continued
A to the question. Providing A to the explanation module also helps in evaluating the two steps of answering and explaining separately.

[9] Note that in theory, one *could* provide A_{pred} as input during training instead of A. After all, this matches the expected use case scenario at test time. However, this alternate setup (where A_{pred} is provided as input instead of A) leads to a peculiar and unnatural explanation training goal—specifically, the explanation head will *still be* learning to explain A since that is the answer for which we collected negative explanation human annotations. It is simply unnatural to build that model that answers a question with A_{pred} but learn to explain a different answer A! Note that this is an interesting scenario where the current push towards "end-to-end" training for everything breaks down.

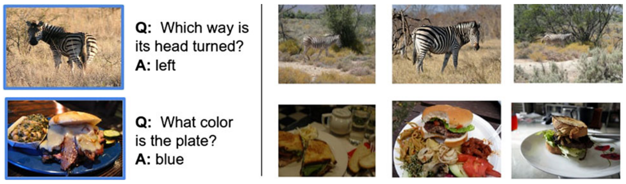


Fig. 15 Three counter-example or negative explanations (right three columns) generated by our model, along with the input image (left), the input question Q and the predicted answer A

Table 4 Negative or counter-example explanation performance of our model compared to strong baselines

	Recall@1	Recall@5	Mean
Random	4.22	20.79	12.51
Distance	9.64	42.84	8.95
VQA (Antol et al. 2015)	4.53	21.65	12.42
Ours	**12.23**	**46.70**	**8.12**

these scores $S(I_i)$ as being most to least likely of being good counter-examples or negative explanations. This component is trained with pairwise hinge ranking losses that encourage $S(I') - S(I_i) > M - \epsilon, \quad I_i \in \{I_1, I_2, ..., I_K\} \setminus \{I'\}$, i.e. the score of the human picked image I' is encouraged to be higher than all other candidate images by a desired margin of M (a hyperparameter) and a slack of ϵ. This is of course the classical 'constraint form' of the pairwise hinge ranking loss, and we minimize the standard expression $\max\Big(0, M - \big(S(I') - S(I_i)\big)\Big)$. The combined loss function for the shared component is

$$\mathcal{L} = -\log P(A|I, Q) + \lambda \sum_i \max\Big(0, M - \big(S(I') - S(I_i)\big)\Big) \tag{1}$$

where, the first term is the cross-entropy loss (for training the answering module) on (I, Q), the second term is the sum of pairwise hinge losses that encourage the explaining model to give high score to image I' (picked by humans) than other I_is in I_{NN}, and λ is the trade-off weight parameter between the two losses.

6.2 Results

Figure 15 shows qualitative examples of negative explanations produced by our model. We see the original image I, the question asked Q, the answer A_{pred} predicted by the VQA head in our model, and top three negative explanations produced by the explanation head. We see that most of these explanations are sensible and reasonable—the images are similar to I but with answers that are different from those predicted for I.

For quantitative evaluation, we compare our model with a number of baselines: *Random:* Sorting the candidate images in I_{NN} randomly. That is, a random image from I_{NN} is picked as the most likely counter-example. *Distance:* Sorting the candidate images in increasing order of their distance from the original image I. That is, the image from I_{NN} most similar to I is picked as the most likely counter-example. *VQA Model:* Using a VQA model's probability for the predicted answer to sort the candidate images in *ascending* order of $P(A|Q, I_i)$. That is, the image from I_{NN} *least likely* to have A as the answer to Q is picked as the *most likely* counter-example.

Note that while I'—the image picked by humans—is a good counter-example, it is not necessarily the unique (or even the "best") counter-example. Humans were simply asked to pick any image where Q makes sense and the answer is not A. There was no natural criteria to convey to humans to pick the "best" one—it is not clear what "best" would mean in the first place. To provide robustness to this potential ambiguity in the counter-example chosen by humans, we evaluate our approach using the following metrics: (1) recall@k i.e., how often the human picked I' is among the top-k in the sorted list of I_is in I_{NN} our model produces (higher is better), and (2) mean rank of human picked image (lower is better).

In Table 4, we can see that our explanation model significantly outperforms the random baseline, as well as the VQA (Antol et al. 2015) model. Interestingly, the strongest baseline is Distance. While our approach (statistically significantly) outperforms it, it is clear that identifying an image that is a counter-example to I from among I's nearest neighbors is a challenging task. Again, this suggests that visual understanding models that can extract meaningful details from images still remain elusive.

7 Conclusion

To summarize, in this paper we address the strong language priors for the task of visual question answering and elevate the role of image understanding required to be successful on this task. We develop a novel data-collection interface to 'balance' the popular VQA dataset (Antol et al. 2015) by collecting 'complementary' images. For every question in the dataset, we have two complementary images that look similar, but have different answers to the question.

This effort results in VQA v2.0 dataset that is not only more balanced than the original VQA v1.0 dataset by construction, but also is about twice the size. We find both qualitatively and quantitatively that the 'tails' of the answer distribution are heavier in this balanced dataset, which reduces the strong language priors that may be exploited by models. Our complete balanced dataset is publicly available at http://visualqa.org/ as part of the 2nd iteration of the visual question answering Dataset and Challenge (VQA v2.0).

We benchmark a number of (near) state-of-art VQA models on our balanced dataset and find that testing them on this balanced dataset results in a significant drop in performance, confirming our hypothesis that these models had indeed exploited language biases. We also present interesting insights from analysis of the participant entries in VQA Challenge 2017, organized on VQA v2.0 dataset.

Finally, our framework around complementary images enables us to develop a novel explainable model—when asked a question about an image, our model not only returns an answer, but also produces a list of similar images that it considers 'counter-examples', i.e. where the answer is not the same as the predicted response. Producing such explanations may enable a user to build a better mental model of what the system considers a response to mean, and ultimately build trust.

Acknowledgements We thank Anitha Kannan for helpful discussions. This work was funded in part by NSF CAREER awards to DP and DB, an ONR YIP award to DP, ONR Grant N00014-14-1-0679 to DB, a Sloan Fellowship to DP, ARO YIP awards to DB and DP, an Allen Distinguished Investigator award to DP from the Paul G. Allen Family Foundation, ICTAS Junior Faculty awards to DB and DP, Google Faculty Research Awards to DP and DB, Amazon Academic Research Awards to DP and DB, AWS in Education Research Grant to DB, and NVIDIA GPU donations to DB. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. In *EMNLP*.

Agrawal, A., Batra, D., Parikh, D., & Kembhavi, A. (2018). Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*.

Agrawal, A., Kembhavi, A., Batra, D., & Parikh, D. (2017). C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. arXiv preprint arXiv:1704.08243.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., et al. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016). Neural module networks. In *CVPR*.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., et al. (2015). VQA: Visual question answering. In *ICCV*.

Berg, T., & Belhumeur, P. N. (2013). How do you tell a blackbird from a crow? In *ICCV*.

Chen, X., & Zitnick, C. L. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*.

Devlin, J., Gupta, S., Girshick, R. B., Mitchell, M., & Zitnick, C. L. (2015). Exploring nearest neighbor approaches for image captioning. arXiv preprint arXiv:1505.04467.

Doersch, C., Singh, S., Gupta, A., Sivic, J., & Efros, A. A. (2012). What makes Paris look like Paris? *ACM Transactions on Graphics (SIGGRAPH)*, *31*(4), 101:1–101:9.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., et al. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

Fang, H., Gupta, S., Iandola, F. N., Srivastava, R., Deng, L., Dollár, P., et al. (2015). From captions to visual concepts and back. In *CVPR*.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

Gao, H., Mao, J., Zhou, J., Huang, Z., & Yuille, A. (2015). Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*.

Goyal, Y., Mohapatra, A., Parikh, D., & Batra, D. (2016). Towards transparent AI systems: Interpreting visual question answering models. In *ICML workshop on visualization for deep learning*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. In *ECCV*.

Hodosh, M., & Hockenmaier, J. (2016). Focused evaluation for image description with binary forced-choice tasks. In *Workshop on vision and language, annual meeting of the association for computational linguistics*.

Hu, R., Andreas, J., Rohrbach, M., Darrell, T., & Saenko, K. (2017). Learning to reason: End-to-end module networks for visual question answering. In *ICCV*.

Ilievski, I., Yan, S., & Feng, J. (2016). A focused dynamic attention model for visual question answering. arXiv preprint arXiv:1604.01485.

Jabri, A., Joulin, A., & van der Maaten, L. (2016). Revisiting visual question answering baselines. In *ECCV*.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Kafle, K., & Kanan, C. (2016a). Answer-type prediction for visual question answering. In *CVPR*.

Kafle, K., & Kanan, C. (2016b). Visual question answering: Datasets, algorithms, and future challenges. arXiv preprint arXiv:1610.01465.

Kafle, K., & Kanan, C. (2017). An analysis of visual question answering algorithms. In *ICCV*.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kim, J. H., Lee, S. W., Kwak, D. H., Heo, M. O., Kim, J., Ha, J. W., et al. (2016). Multimodal residual learning for visual QA. In *NIPS*.

Kim, J. H., On, K. W., Lim, W., Kim, J., Ha, J. W., & Zhang, B. T. (2017). Hadamard product for low-rank bilinear pooling. In *ICLR*.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2015). Unifying visual-semantic embeddings with multimodal neural language models. In *TACL*.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *ICML*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., et al. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common objects in context. In *ECCV*.

Lu, J., Lin, X., Batra, D., & Parikh, D. (2015). Deeper LSTM and normalized CNN visual question answering model. https://github.com/VT-vision-lab/VQA_LSTM_CNN. Accessed 1 Sep 2017.

Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In *NIPS*.

Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*.

Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*.

Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L. (2014). Explain images with multimodal recurrent neural networks. In *NIPS*.

Noh, H., & Han, B. (2016). Training recurrent answering units with joint loss minimization for vqa. arXiv preprint arXiv:1606.03647.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*.

Ray, A., Christie, G., Bansal, M., Batra, D., & Parikh, D. (2016). Question relevance in VQA: Identifying non-visual and false-premise questions. In *EMNLP*.

Ren, M., Kiros, R., & Zemel, R. (2015). Exploring models and data for image question answering. In *NIPS*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Knowledge discovery and data mining (KDD)*.

Saito, K., Shin, A., Ushiku, Y., & Harada, T. (2016). Dualnet: Domain-invariant network for visual question answering. arXiv preprint arXiv:1606.06108.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-CAM: Why did you say that? Visual explanations from deep networks via gradient-based localization. arXiv preprint arXiv:1610.02391.

Shih, K. J., Singh, S., & Hoiem, D. (2016). Where to look: Focus regions for visual question answering. In *CVPR*.

Shin, A., Ushiku, Y., & Harada, T. (2016). The color of the cat is gray: 1 Million full-sentences visual question answering (FSVQA). arXiv preprint arXiv:1609.06657.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., & Fidler, S. (2016). MovieQA: Understanding stories in movies through question-answering. In *CVPR*.

Teney, D., Anderson, P., He, X., & van den Hengel, A. (2018). Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *CVPR*.

Torralba, A., & Efros, A. (2011). Unbiased look at dataset bias. In *CVPR*.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *CVPR*.

Wang, P., Wu, Q., Shen, C., van den Hengel, A., & Dick, A. R. (2015). Explicit knowledge-based reasoning for visual question answering. arXiv preprint arXiv:1511.02570.

Wu, Q., Wang, P., Shen, C., van den Hengel, A., & Dick, A. R. (2016). Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*.

Xiong, C., Merity, S., & Socher, R. (2016). Dynamic memory networks for visual and textual question answering. In *ICML*.

Xu, H., & Saenko, K. (2016). Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*.

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In *CVPR*.

Yu, L., Park, E., Berg, A. C., & Berg, T. L. (2015). Visual madlibs: Fill-in-the-blank description generation and question answering. In *ICCV*.

Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*.

Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., & Parikh, D. (2016). Yin and Yang: Balancing and answering binary visual questions. In *CVPR*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Learning deep features for discriminative localization. In *CVPR*.

Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., & Fergus, R. (2015). Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167.

Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In *CVPR*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.