

VISUAL QUESTION ANSWERING

BHUPENDER SINGH

Abstract

— Visual Question and Answering (VQA) problems are attracting increasing interest from multiple research disciplines. Solving VQA problems requires techniques from both computer vision for understanding the visual contents of a presented image or video, as well as the ones from natural language processing for understanding semantics of the question and generating the answers. Regarding visual content modeling, most of existing VQA methods adopt the strategy of extracting global features from the image or video, which inevitably fails in capturing fine-grained information such as spatial configuration of multiple objects. Extracting features from auto-generated regions – as some region-based image recognition methods do – cannot essentially address this problem and may introduce some overwhelming irrelevant features with the question. In this work, we propose a novel Focused Dynamic Attention (FDA) model to provide better aligned image content representation with proposed questions. Being aware of the key words in the question, FDA employs off-the-shelf object detector to identify important regions and fuse the information from the regions and global features via an LSTM unit. Such question-driven representations are then combined with question representation and fed into a reasoning unit for generating the answers. Extensive evaluation on a large-scale benchmark dataset, VQA, clearly demonstrate the superior performance of FDA over well-established baselines

1 Introduction

Visual question answering (VQA) is an active research direction that lies in the intersection of computer vision, natural language processing, and machine learning. Even though with a very short history, it already has received great research attention from multiple communities. Generally, the VQA investigates a generalization of traditional QA problems where visual input (e.g., an image) is necessary to be considered. More concretely, VQA is about how to provide a correct answer to a human posed question concerning contents of one presented image or video. VQA is a quite challenging task and undoubtedly important for developing modern AI systems. The VQA problem can be regarded as a Visual Turing Test [1,2], and besides contributing to the advancement of the involved research areas, it has other important applications, such as blind person assistance and image retrieval. Coming up with solutions to this task requires natural language processing techniques for understanding the questions and generating the answers, as well as computer vision techniques for understanding contents of the concerned image. With help of these two core techniques, the computer can perform reasoning about the perceived contents and posed questions. Recently, VQA is advanced significantly by the development of machine learning methods (in particular the deep learning ones) that can learn proper representations of questions and images, align and fuse them in a joint question-image space and provide a direct mapping from this joint representation to a correct answer. For example, consider the following image-question pair: an image of an apple tree with a basket of apples next to

it, and a question “How many apples are in the basket?”. Answering this question requires VQA methods to first understand the semantics of the question, then locate the objects (apples) in the image, understand the relation between the image objects (which apples are in the basket), and finally count them and generate an answer with the correct number of apples. The first feasible solution to VQA problems was provided by Malinowski and Fritz in [2], where they used a semantic language parser and a Bayesian reasoning model, to understand the meaning of questions and to generate the proper answers. Malinowski and Fritz also constructed the first VQA benchmark dataset, named as DAQUAR, which contains 1,449 images and 12,468 questions generated by humans or automatically by following a template and extracting facts from a database [2]. Shortly after, Ren et al. [3] released the TORONTO-QA dataset, which contains many images (123,287) and questions (117,684), but the questions are automatically generated and thus can be answered without complex reasoning. Nevertheless, the release of the TORONTO-QA dataset was important since it provided enough data for deep learning models to be trained and evaluated on the VQA problem [4,5,3]. More recently, Antol et al. [6] published the currently largest VQA dataset. It consists of three human posed questions and ten answers given by different human subjects, for each one of the 204,721 images found in the Microsoft COCO dataset [7]. Answering the 614,163 questions requires complex reasoning, common sense, and real-world knowledge, making the VQA dataset suitable for a true Visual Turing Test. The VQA authors split the evaluation on their dataset on two tasks: an open-ended task, where the method should generate a natural language answer, and a multiple-choice task, where for each question the method should chose one of the 18 different answers. The current top performing methods [8,9,10] employ deep neural network model that predominantly uses the convolutional neural network (CNN) architecture [11,12,13,14] to extract image features and a Long Short-Term Memory (LSTM) [15] network to extract the representations for questions. The CNN and LSTM representation vectors are then usually fused by concatenation [16,3,5] or element-wise multiplication [17,18]. Other approaches additionally incorporate some kind of attention mechanism over the image features [18,19,20]. Properly modeling the image contents is one of the critical factors for solving VQA problems well. A common practice with existing VQA methods on modeling image contents is to extract global features for the overall image. However, only using global feature is arguably insufficient to capture all the necessary visual information and provide full understanding of image contents such as multiple objects, spatial configuration of the objects and informative background. This issue can be relieved to some extent by extracting features from object proposals – the image regions that possibly contain objects of interest. However, using features from all image regions [19,18] may provide too much noise or overwhelming information irrelevant to the question and thus hurt the overall VQA performance. In this work, we propose a question driven attention model that is able to automatically identify and focus on image regions relevant for the current question. We name our proposed model Focused Dynamic Attention (FDA) for Visual Question Answering. With the FDA model, computers can select and recognize the image regions in a well-aligned sequence with the key words containing in a given question. Recall the above VQA example. To answer the

question of “How many apples are in the basket?”, FDA would first localize the regions corresponding to the key words “apples” and “basket” (with the help of a generic object detector) and extract description features from these regions of interest. Then VQA compliments the features from selected image regions with a global image feature providing contextual information for the overall image and reconstruct a visual representation by encoding them with a Long Short-Term Memory (LSTM) unit. We evaluate and compare the performance of our proposed FDA model on two types of VQA tasks, i.e., the open-ended task and the multiple-choice task, on the VQA dataset – the largest VQA benchmark dataset. Extensive experiments demonstrate that FDA brings substantial performance improvement upon well-established baselines. The main contributions of this work can be summarized as follows: – We introduce a focused dynamic attention mechanism that learns to use the question word order to shift the focus from one image object, to another. – We describe a model that fuses local and global context visual features with textual features. – We perform an extensive evaluation, comparing to all existing methods, and achieve state-of-the-art accuracy on the open-ended, and on the multiple-choice VQA tasks. The rest of the paper is organized as follows. We review the current VQA models, and compare them to our model. We formulate the problem and explain our motivation. We describe our model and we evaluate and compare it with the current state-of-the-art models. We conclude our work.

The VQA Dataset

We have used MSCOCO dataset available from visualqa.com. It consists of 82,783 training images, 40,504 Validation images and 81,434 Testing images. The Question dataset consists of 60,000 training questions, 30,000 validation questions and 60,000 Testing Questions. The MS COCO dataset has images depicting diverse and complex scenes that are effective at eliciting compelling and diverse questions.

Splits. For real images, we follow the same train/val/test split strategy as the MC COCO dataset [32] (including testdev, test-standard, test-challenge, test-reserve). For the VQA challenge (see section 6), test-dev is used for debugging and validation experiments and allows for unlimited submission to the evaluation server. Test-standard is the ‘default’ test data for the VQA competition. When comparing to the state of the art (e.g., in papers), results should be reported on test-standard. Test-standard is also used to maintain a public leaderboard that is updated upon submission. Test-reserve is used to protect against possible overfitting. If there are substantial differences between a method’s scores on test-standard and test-reserve, this raises a red-flag and prompts further investigation. Results on test-reserve are not publicly revealed. Finally, test-challenge is used to determine the winners of the challenge. For abstract scenes, we created splits for standardization, separating the scenes into 20K/10K/20K for train/val/test splits, respectively. There are no sub splits (test-dev, test-standard, test-challenge, test-reserve) for abstract scenes.

Captions. The MS COCO dataset [32], [7] already contains five single-sentence captions for all images. We also collected five single-captions for all abstract scenes using the same user interface¹ for collection.

Questions. Collecting interesting, diverse, and well-posed questions is a significant challenge. Many simple questions may only require low-level computer vision knowledge, such as “What color is the cat?” or “How many chairs are present in the scene?”. However, we also want questions that require commonsense knowledge about the scene, such as “What sound does the pictured animal make?”. Importantly, questions should also require the image to correctly answer and not be answerable using just commonsense information, e.g., in Fig. 1, “What is the mustache made of?”. By having a wide variety of question types and difficulty, we may be able to measure the continual progress of both visual understanding and commonsense reasoning.

Related Work

VQA Efforts. Several recent papers have begun to study visual question answering [19], [36], [50], [3]. However, unlike our work, these are fairly restricted (sometimes synthetic) settings with small datasets. For instance, [36] only considers questions whose answers come from a predefined closed world of 16 basic colors or 894 object categories. [19] also considers questions generated from templates from a fixed vocabulary of objects, attributes, relationships between objects, etc. In contrast, our proposed task involves open-ended, free-form questions and answers provided by humans. Our goal is to increase the diversity of knowledge and kinds of reasoning needed to provide correct answers. Critical to achieving success on this more difficult and unconstrained task, our VQA dataset is two orders of magnitude larger than [19], [36] (>250,000 vs. 2,591 and 1,449 images respectively). The proposed VQA task has connections to other related work: [50] has studied joint parsing of videos and corresponding text to answer queries on two datasets containing 15 video clips each. [3] uses crowdsourced workers to answer questions about visual content asked by visually-impaired users. In concurrent work, [37] proposed combining an LSTM for the question with a CNN for the image to generate an answer. In their model, the LSTM question representation is conditioned on the CNN image features at each time step, and the final LSTM hidden state is used to sequentially decode the answer phrase. In contrast, the model developed in this paper explores “late fusion” – i.e., the LSTM question representation and the CNN image features are computed independently, fused via an element-wise multiplication, and then passed through fully connected layers to generate a SoftMax distribution over output answer classes. [34] generates abstract scenes to capture visual common sense relevant to answering (purely textual) fill-in the-blank and visual paraphrasing questions. [47] and [52] use visual information to assess the plausibility of common-sense assertions. [55] introduced a dataset of 10k images and prompted captions that describe specific aspects of a scene (e.g., individual objects, what will happen next). Concurrent with our work, [18] collected questions & answers in Chinese (later translated

to English by humans) for COCO images. [44] automatically generated four types of questions (object, count, color, location) using COCO captions.

Describing Visual Content. Related to VQA are the tasks of image tagging [11], [29], image captioning [30], [17], [40], [9], [16], [53], [12], [24], [38], [26] and video captioning [46], [21], where words or sentences are generated to describe visual content. While these tasks require both visual and semantic knowledge, captions can often be non-specific (e.g., observed by [53]). The questions in VQA require detailed specific information about the image for which generic image captions are of little use [3].

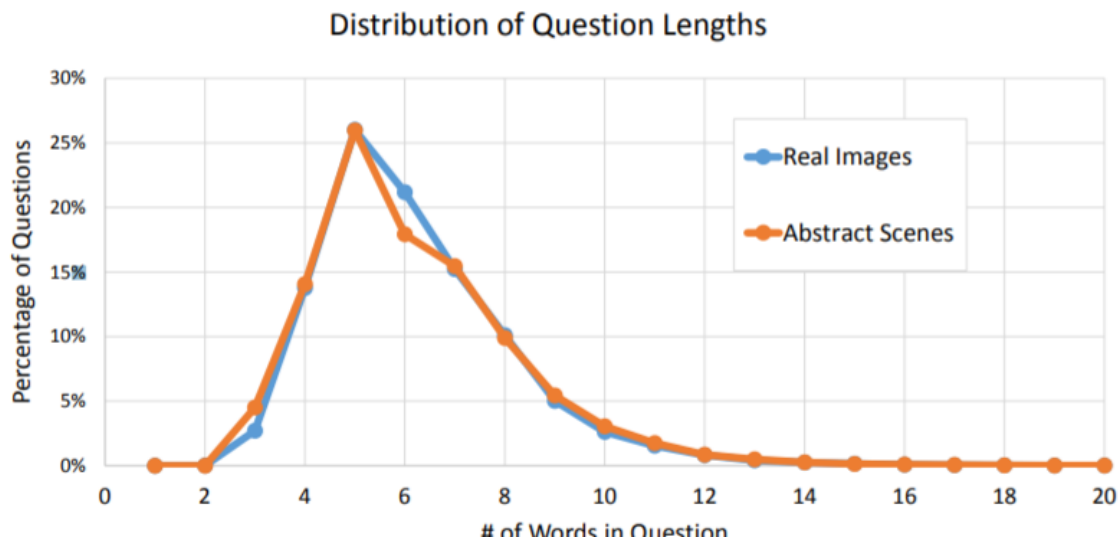
VQA DATASET ANALYSIS

In this section, we provide an analysis of the questions and answers in the VQA train dataset. To gain an understanding of the types of questions asked and answers provided, we visualize the distribution of question types and answers. We also explore how often the questions may be answered without the image using just commonsense information. Finally, we analyze whether the information contained in an image caption is sufficient to answer the questions.

Questions:

Types of Question. Given the structure of questions generated in the English language, we can cluster questions into different types based on the words that start the question. Fig. 3 shows the distribution of questions based on the first four words of the questions for both the real images (left) and abstract scenes (right). Interestingly, the distribution of questions is quite similar for both real images and abstract scenes. This helps demonstrate that the type of questions elicited by the abstract scenes is similar to those elicited by the real images. There exists a surprising variety of question types, including “What is. . .”, “Is there. . .”, “How many. . .”, and “Does the. . .”. Quantitatively, the percentage of questions for different types is shown in Table 3. Several example questions and answers are shown in Fig. 2. A particularly interesting type of question is the “What is. . .” questions, since they have a diverse set of possible answers. See the appendix for visualizations for “What is. . .” questions.

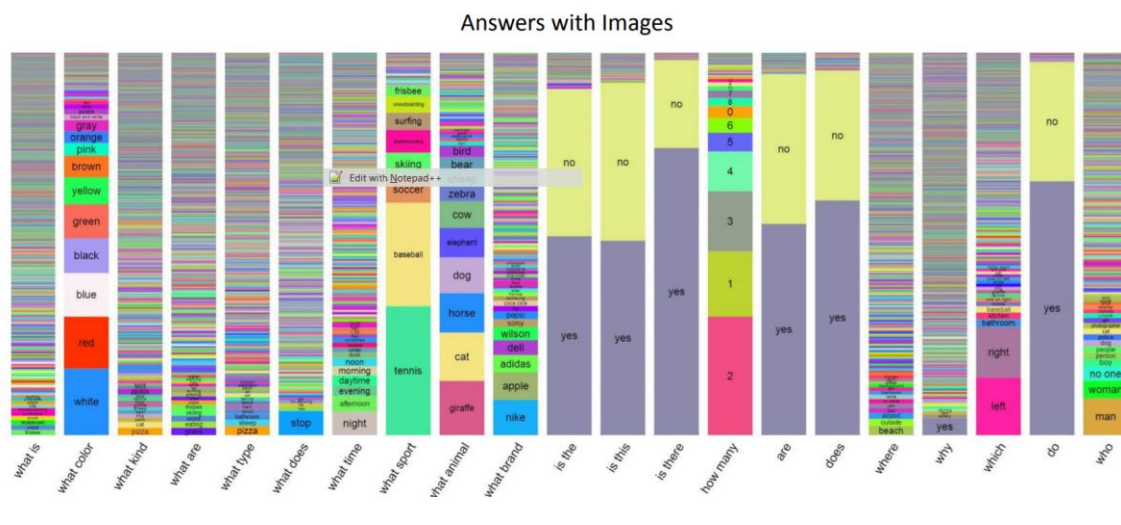
Lengths. Fig. 4 shows the distribution of question lengths. We see that most questions range from four to ten words.



Answers:

Typical Answers. Fig. 5 (top) shows the distribution of answers for several question types. We can see that a number of question types, such as “Is the . . .”, “Are . . .”, and “Does . . .” are typically answered using “yes” and “no” as answers. Other questions such as “What is. . .” and “What type. . .” have a rich diversity of responses. Other question types such as “What color. . .” or “Which. . .” have more specialized responses, such as colors, or “left” and “right”. See the appendix for a list of the most popular answers.

Lengths. Most answers consist of a single word, with the distribution of answers containing one, two, or three words, respectively being 89.32%, 6.91%, and 2.74% for real images and 90.51%, 5.89%, and 2.49% for abstract scenes. The brevity of answers is not surprising, since the questions tend to elicit specific information from the images. This is in contrast



with image captions that generically describe the entire image and hence tend to be longer. The brevity of our answers makes automatic evaluation feasible. While it may be tempting to believe the brevity of the answers makes the problem easier, recall that they are human-provided open-ended answers to open-ended questions. The questions typically require complex reasoning to arrive at these deceptively simple answers (see Fig. 2). There are currently 23,234 unique one-word answers in our dataset for real images and 3,770 for abstract scenes. ‘Yes/No’ and ‘Number’ Answers. Many questions are answered using either “yes” or “no” (or sometimes “maybe”) – 38.37% and 40.66% of the questions on real images and abstract scenes respectively. Among these ‘yes/no’ questions, there is a bias towards “yes” – 58.83% and 55.86% of ‘yes/no’ answers are “yes” for real images and abstract scenes. Question types such as “How many. . .” are answered using numbers – 12.31% and 14.48% of the questions on real images and abstract scenes are ‘number’ questions. “2” is the most popular answer among the ‘number’ questions, making up 26.04% of the ‘number’ answers for real images and 39.85% for abstract scenes. Subject Confidence. When the subjects answered the questions, we asked “Do you think you were able to answer the question correctly?”. Fig. 6 shows the distribution of responses. A majority of the answers were labeled as confident for both real images and abstract scenes. Inter-human Agreement. Does the self-judgment of confidence correspond to the answer agreement between subjects? Fig. 6 shows the percentage of questions in which (i) 7 or more, (ii) 3–7, or (iii) less than 3 subjects agree on the answers given their average confidence score (0 = not confident, 1 = confident). As expected, the agreement between subject.

Method Overview In this section, we briefly describe the motivation and give formal problem formulation. 3.1 Problem Formulation The visual question answering problem can be represented as predicting the best answer \hat{a} given an image I and a question q . Common practice [6,16,3,19] is to use the 1,000 most common answers in the training set and thus simplify the VQA task to a classification problem. The following equation represents the problem mathematically: $\hat{a} = \arg \max_{a \in \Omega} p(a|I, q; \theta)$ (1) where Ω is the set of all possible answers and θ are the model weights. 3.2 Motivation The baseline methods from [6] show only modest increase in accuracy when including the image features (4.98% for open-ended questions, and 2.42% for multiple-choice question). We believe that the image contains a lot more information and should increase the accuracy much more. Thus, we focus on improving the image features and design a visual attention mechanism, which learns to focus on the question related image regions. The proposed attention mechanism is loosely inspired on the human visual attention mechanism. Humans shift the focus from one image region to another, before understanding how the regions relate to each other and grasping the meaning of the whole image. Similarly, we feed our model image regions relevant for the question at hand, before showing the whole image.

Focused Dynamic Attention for VQA The FDA model is composed of question and image understanding components, attention mechanism, and a multimodal representation fusion network (Figure 1). In this section we describe them individually.

Question Understanding Following a common practice, our FDA model uses an LSTM network to encode the question in a vector representation [15,5,18,8]. The LSTM network learns to keep in its state the feature vectors of the important question words, and thus provides the question understanding component with a word attention mechanism.

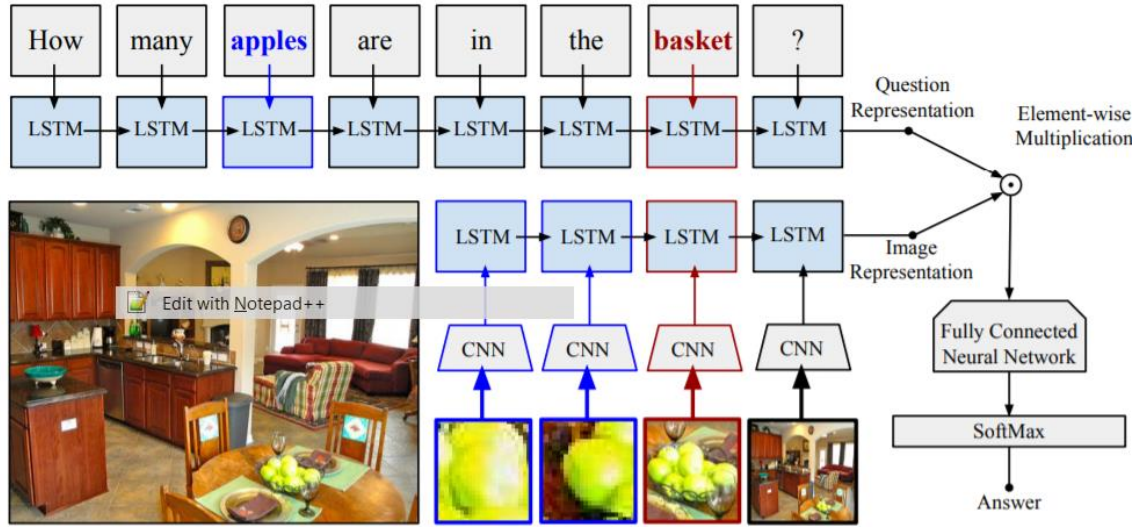


Fig. 1. Focused dynamic attention model diagram.

Understanding Following prior work [3,4,5], we use a pre-trained convolutional neural network (CNN) to extract image feature vectors. Specifically, we use the Deep Residual Networks model used in ILSVRC and COCO 2015 competitions, which won the 1 st places in: ImageNet classification, ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation [24]. We extract the weights of the layer immediately before the final SoftMax layer and regard them as visual features. We extract such features for the whole image (global visual features) and for the specific image regions (local visual features). However, contrary to the existing approaches, we employ an LSTM network to combine the local and global visual features into a joint representation.

Focused Dynamic Attention Mechanism We introduce a focused dynamic attention mechanism that learns to focus on image regions related to the question words. The attention mechanism works as follows. For each image object_i it uses word2vec word embeddings [26] to measure the similarity between the question words and the object label. Next, it selects objects with similarity score greater than 0.5 and extracts the feature vectors of the objects bounding boxes with a pre-trained ResNet model [24]. Following the question word order, it feeds the LSTM network with the corresponding object feature vectors. Finally, it feeds the LSTM network with the feature vector of the whole image and it uses the resulting LSTM state as a visual representation. Thus, the attention mechanism 1 During training we use the ground truth object bounding boxes and labels. At test time we use the

precomputed bounding boxes from [25] and classify them with [24] to obtain the object labels. A Focused Dynamic Attention Model for Visual Question Answering [7] enables the model to combine the local and global visual features into a single representation, necessary for answering complex visual questions. Figure 1 illustrates the focused dynamic attention mechanism with an example.

Multimodal Representation Fusion We regard the final state of the two LSTM networks as a question and image representation. We start fusing them into single representation by applying Tanh on the question representation and ReLU [2] on the image representation [3]. We proceed by doing an element-wise multiplication of the two vector representations and the resulting vector is fed to a fully-connected neural network. Finally, a SoftMax layer classifies the multimodal representation into one of the possible [4] answers.

Evaluation

In this section we detail the model implementation and compare our model against the current state-of-the-art methods.



					
Why does this male have his arms in this position?	balance for balance for balance	angry he's carrying bags hug	How many people are wearing an orange shirt?	3 3 3	1 3 3
Are the clouds high in the sky?	yes yes yes	no no yes	Is this a trained elephant?	yes yes yes	yes yes yes

Fig. 2. Representative examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for two images from the VQA dataset. Examples provided by [6].

Baseline

We compare our model against the baseline models provided by the VQA dataset authors [27], which currently achieve the best performance on the test-standard split for the multiple-choice task. The model, first described in [6], is a standard implementation of an LSTM+CNN VQA model. It uses an LSTM to encode the question and CNN features to encode the image. To answer a question, it multiplies the last LSTM state with the image CNN features and feeds the result into a SoftMax layer for classification into one of the 1,000 most common answers. The implementation in [27] uses a deeper two layer LSTM network for encoding the question, and normalized image CNN features, which showed crucial for achieving the state-of-the-art.

Model Implementation and Training details

We transform the question words into a vector form by multiplying one-hot vector representation with a word embedding matrix. The vocabulary size is 12,602 and the word embeddings are 300 dimensional. We feed a pre-trained ResNet network [24] and use the 2,048-dimensional weight vector of the layer before the last fully-connected layer. The word and image vectors are feed into two separate LSTM networks. The LSTM networks are standard implementation of one-layer LSTM network [15], with a 512 dimensional state vector. The final state of the question LSTM is passed through Tanh, while the final state of the image LSTM is passed through ReLU5 . We do element-wise multiplication on the resulting vectors, to obtain a multimodal representation vector, which is then fed to a fully-connected neural network.

Qualitative Results We qualitatively evaluate our model on a set of examples where complex reasoning and focusing on the relevant local visual features are needed for answering the question correctly. Figure 3 shows particularly difficult examples (the predominant image color is not the correct answer) of “What color” type of questions. But, by focusing on the question related image regions, the FDA model is still able to produce the correct answer. In Figure 4 we show examples where the model focuses on different regions from the same image, depending on the words in the question. Focusing on the right image region is crucial when answering unusual questions for an image (Row 1), questions about small image objects (Row 2), or when the most dominant image object partly occludes the question related region and can lead to a wrong answer (Row 3). Representative examples of questions that require image object identification. We can observe that the focused attention enables the model to answer complex questions (Row 1, left) and counting questions (Row 1, right). The question guided image object identification greatly simplifies the answering of questions like the ones shown in Row 2 and Row 3.

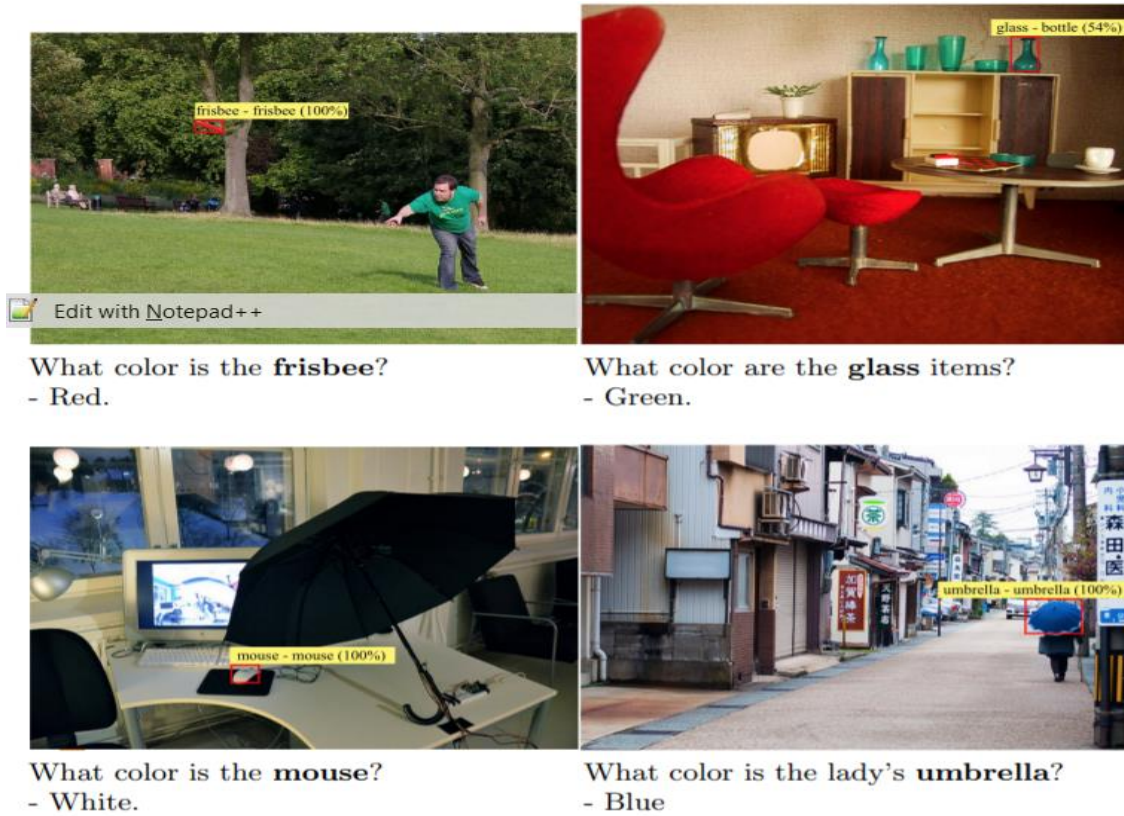


Fig. 3. Representative examples where focusing on the question related objects helps FDA answer “What color” type of questions. The question words in bold have been matched with an image region. The yellow region caption box contains the question word, followed by the region label, and in parenthesis their cosine similarity



Is this a birthday **cake**?

- Yes.

Is someone in all likelihood, a **zoo** fancier? - Yes.

with Notepad++



What **fruit** is by the sink?

- Apples.

Is there a **cookbook** in the picture?

- Yes.



What type of **vehicle** is pictured?

- Motorcycle.

Does the **elephant** have tusks?

- No.

Fig. 4. Representative examples where the model focuses on different regions from the same image, depending on the question. The question words in bold have been matched with an image region. The yellow region caption box contains the question word, followed by the region label, and in parenthesis their cosine similarity

Conclusion

In this work, we proposed a novel Focused Dynamic Attention (FDA) model to solve the challenging VQA problems. FDA is built upon a generic object-centric attention model for extracting question related visual features from an image as well as a stack of multiple LSTM layers for feature fusion. By only focusing on the identified regions specific for proposed questions, FDA was shown to be able to filter out overwhelming irrelevant information's from cluttered background or other regions, and thus substantially improved the quality of visual representations in the sense of answering proposed questions. By fusing cleaned regional representation, global context and question representation via LSTM layers, FDA provided significant performance improvement over baselines on the VQA benchmark datasets, for both the open-ended and multiple-choices VQA tasks. Excellent performance of FDA clearly demonstrates its stronger ability of modeling visual contents and also verifies paying more attention to visual part in VQA tasks could essentially improve the overall performance. In the future, we are going to further explore along this research line and investigate different attention methods for visual information selection as well as better reasoning model for interpreting the relation between visual contents and questions.

References

1. Geman, D., Geman, S., Hallonquist, N., Younes, L.: Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences* 112(12) (2015) 3618–3623
2. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: *Advances in Neural Information Processing Systems*. (2014) 1682–1690
3. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: *Advances in Neural Information Processing Systems*. (2015) 2935– 2943
4. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? dataset and methods for multilingual image question. In: *Advances in Neural Information Processing Systems*. (2015) 2287–2295
5. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: *Proceedings of the IEEE International Conference on Computer Vision*. (2015) 1–9
6. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: *The IEEE International Conference on Computer Vision (ICCV)*. (December 2015)

7. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll'ar, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision– ECCV 2014. Springer (2014) 740–755
8. Noh, H., Seo, P.H., Han, B.: Image question answering using convolutional neural network with dynamic parameter prediction. arXiv preprint arXiv:1511.05756 (2015)
9. Wu, Q., Wang, P., Shen, C., Hengel, A.v.d., Dick, A.: Ask me anything: Freeform visual question answering based on knowledge from external sources. arXiv preprint arXiv:1511.06973 (2015)
10. Ma, L., Lu, Z., Li, H.: Learning to answer questions from image using convolutional neural network. arXiv preprint arXiv:1506.00333 (2015)
11. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation 1(4) (1989) 541–551
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1–9
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8) (1997) 1735–1780
16. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)
17. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. arXiv preprint arXiv:1511.07394 (2015)
18. Chen, K., Wang, J., Chen, L.C., Gao, H., Xu, W., Nevatia, R.: Abc-cnn: An attention based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960 (2015)
19. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. arXiv preprint arXiv:1511.02274 (2015)
20. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. arXiv preprint arXiv:1511.05234 (2015)

21. Zitnick, C.L., Doll'ar, P.: Edge boxes: Locating object proposals from edges. In: Computer Vision–ECCV 2014. Springer (2014) 391–405
22. Jiang, A., Wang, F., Porikli, F., Li, Y.: Compositional memory for visual question answering. arXiv preprint arXiv:1511.05676 (2015)
23. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Learning to compose neural networks for question answering. arXiv preprint arXiv:1601.01705 (2016)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
25. Pont-Tuset, J., Arbel'aez, P., Barron, J., Marques, F., Malik, J.: Multiscale combinatorial grouping for image segmentation and object proposal generation. In: arXiv:1503.00848. (March 2015)
26. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. (2013) 3111–3119
27. Jiasen Lu, Xiao Lin, D.B., Parikh, D.: Deeper lstm and normalized cnn visual question answering model. https://github.com/VT-vision-lab/VQA_LSTM_CNN (2015)