

# Visual Question Answering using Deep Learning: A Survey and Performance Analysis

Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, and Snehasis Mukherjee

Computer Vision Group,  
 Indian Institute of Information Technology, Sri City, Chittoor, Andhra Pradesh, India.  
 {srivastava.y15, murali.v15, srdubey,  
 snehasis.mukherjee}@iiits.in}

**Abstract.** The Visual Question Answering (VQA) task combines challenges for processing data with both Visual and Linguistic processing, to answer basic ‘common sense’ questions about given images. Given an image and a question in natural language, the VQA system tries to find the correct answer to it using visual elements of the image and inference gathered from textual questions. In this survey, we cover and discuss the recent datasets released in the VQA domain dealing with various types of question-formats and robustness of the machine-learning models. Next, we discuss about new deep learning models that have shown promising results over the VQA datasets. At the end, we present and discuss some of the results computed by us over the vanilla VQA model, Stacked Attention Network and the VQA Challenge 2017 winner model. We also provide the detailed analysis along with the challenges and future research directions.<sup>1</sup>

**Keywords:** Visual Question Answering · Artificial Intelligence · Human Computer Interaction · Deep Learning · CNN · LSTM.

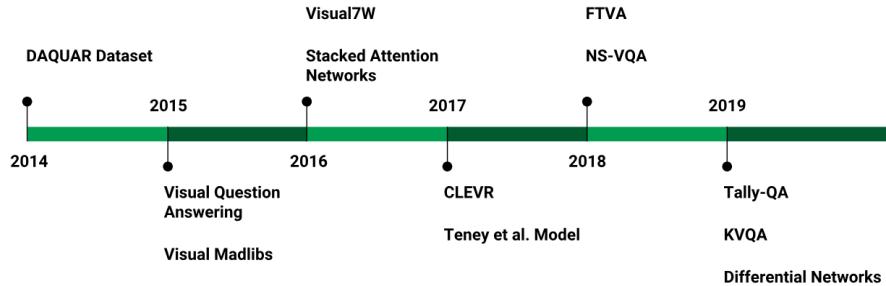
## 1 Introduction

Visual Question Answering (VQA) refers to a challenging task which lies at the intersection of image understanding and language processing. The VQA task has witnessed a significant progress the recent years by the machine intelligence community. The aim of VQA is to develop a system to answer specific questions about an input image. The answer could be in any of the following forms: a word, a phrase, binary answer, multiple choice answer, or a fill in the blank answer. Agarwal et al. [2] presented a novel way of combining computer vision and natural language processing concepts of to achieve **Visual Grounded Dialogue**, a system mimicking the human understanding of the environment with the use of visual observation and language understanding.

The advancements in the field of deep learning have certainly helped to develop systems for the task of Image Question Answering. Krizhevsky et al [14] proposed the AlexNet model, which created a revolution in the computer vision domain. The paper introduced the concept of Convolution Neural Networks (CNN) to the mainstream computer vision application. Later many authors have worked on CNN, which has resulted

---

<sup>1</sup> This paper is accepted in Fifth IAPR International Conference on Computer Vision and Image Processing (CVIP), 2020.



**Fig. 1.** Major Breakthrough Timeline in Visual Question Answering.

in robust, deep learning models like VGGNet [28], Inception [29], ResNet [6], and etc. Similarly, the recent advancements in natural language processing area based on deep learning have improved the text understanding performance as well. The first major algorithm in the context of text processing is considered to be the Recurrent Neural Networks (RNN) [21] which introduced the concept of prior context for time series based data. This architecture helped the growth of machine text understanding which gave new boundaries to machine translation, text classification and contextual understanding. Another major breakthrough in the domain was the introduction of Long-Short Term Memory (LSTM) architecture [7] which improvised over the RNN by introducing a context cell which stores the prior relevant information.

The vanilla VQA model [2] used a combination of VGGNet [28] and LSTM [7]. This model has been revised over the years, employing newer architectures and mathematical formulations as seen in Fig. 1. Along with this, many authors have worked on producing datasets for eliminating bias, strengthening the performance of the model by robust question-answer pairs which try to cover the various types of questions, testing the visual and language understanding of the system. Among the recent developments in the topic of VQA, Li et al. have used the context-aware knowledge aggregation to improve the VQA performance [15]. Yu et al. have performed the cross-modal knowledge reasoning in the network for obtaining a knowledge-driven VQA [35]. Chen et al. have improved the robustness of VQA approach by synthesizing the Counterfactual samples for training [3]. Li et al. have employed the attention based mechanism through transfer learning alongwith a cross-modal gating approach to improve the VQA performance [16]. Huang et al. [8] have utilized the graph based convolutional network to increase the encoding relational informatoin for VQA. The VQA has been also observed in other domains, such as VQA for remote sensing data [19] and medical VQA [37].

In this survey, first we cover major datasets published for validating the Visual Question Answering task, such as VQA dataset [2], DAQUAR [20], Visual7W [39] and most recent datasets up to 2019 include Tally-QA [1] and KVQA [26]. Next, we discuss the state-of-the-art architectures designed for the task of Visual Question Answering such as Vanilla VQA [2], Stacked Attention Networks [33] and Pythia v1.0 [10]. Next we present some of our computed results over the three architectures: vanilla VQA model

**Table 1.** Overview of VQA datasets described in this paper.

Dataset	# Images	# Questions	Question Type(s)	Venue	Model(s)	Accuracy
DAQUAR [20]	1449	12468	Object Identification	NIPS 2014	AutoSeg [5]	13.75%
VQA [2]	204721	614163	Combining vision, language and common-sense	ICCV 2015	CNN + LSTM	54.06%
Visual Madlibs [36]	10738	360001	Fill in the blanks	ICCV 2015	nCCA (bbox)	47.9%
Visual7W [39]	47300	2201154	7Ws, locating objects	CVPR 2016	LSTM + Attention	55.6%
CLEVR [12]	100000	853554	Synthetic question generation using relations	CVPR 2017	CNN + LSTM + Spatial Relationship	93%
Tally-QA [1]	165000	306907	Counting objects on varying complexities	AAAI 2019	RCN Network	71.8%
KVQA [26]	24602	183007	Questions based on Knowledge Graphs	AAAI 2019	MemNet	59.2%

[2], Stacked Attention Network (SAN) [33] and Teney et al. model [31]. Finally, we discuss the observations and future directions.

## 2 Datasets

The major VQA datasets are summarized in Table 1. We present the datasets below.

**DAQUAR:** DAQUAR stands for Dataset for Question Answering on Real World Images, released by Malinowski et al. [20]. It was the first dataset released for the IQA task. The images are taken from NYU-Depth V2 dataset [27]. The dataset is small with a total of 1449 images. The question bank includes 12468 question-answer pairs with 2483 unique questions. The questions have been generated by human annotations and confined within 9 question templates using annotations of the NYU-Depth dataset.

**VQA Dataset:** The Visual Question Answering (VQA) dataset [2] is one of the largest datasets collected from the MS-COCO [18] dataset. The VQA dataset contains at least 3 questions per image with 10 answers per question. The dataset contains 614,163 questions in the form of open-ended and multiple choice. In multiple choice questions, the answers can be classified as: 1) Correct Answer, 2) Plausible Answer, 3) Popular Answers and 4) Random Answers. Recently, VQA V2 dataset [2] is released with additional confusing images. The VQA sample images and questions are shown in Fig. 2.

**Visual Madlibs:** The Visual Madlibs dataset [36] presents a different form of template for the Image Question Answering task. One of the forms is the fill in the blanks type, where the system needs to supplement the words to complete the sentence and it mostly targets people, objects, appearances, activities and interactions. The Visual Madlibs samples are shown in Fig. 3.

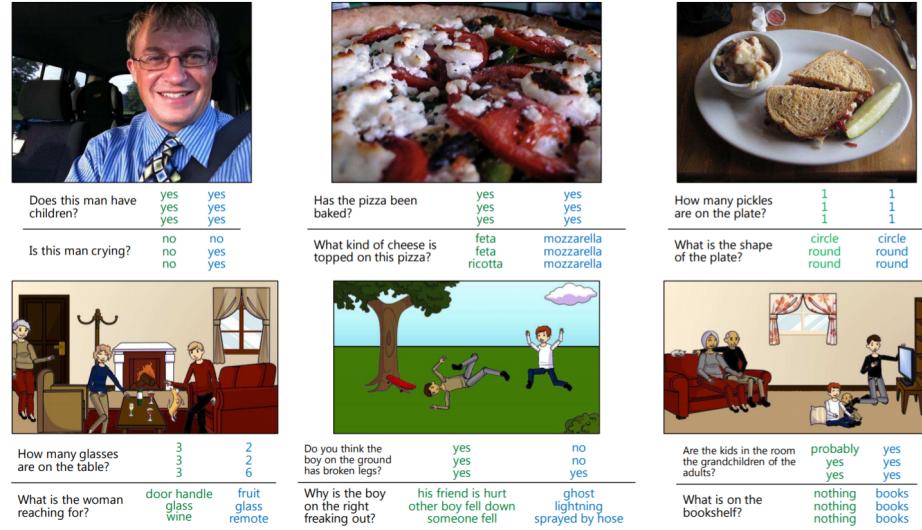


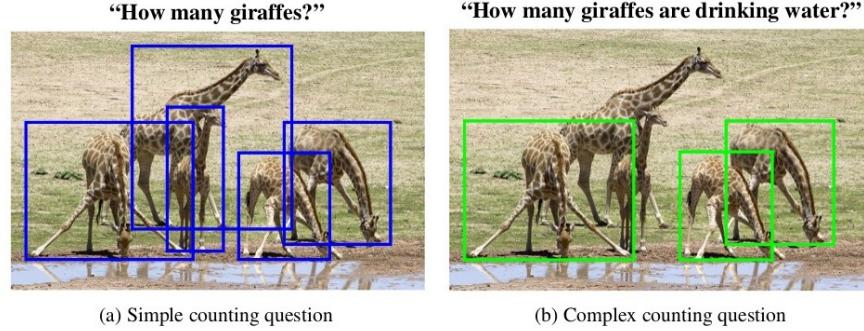
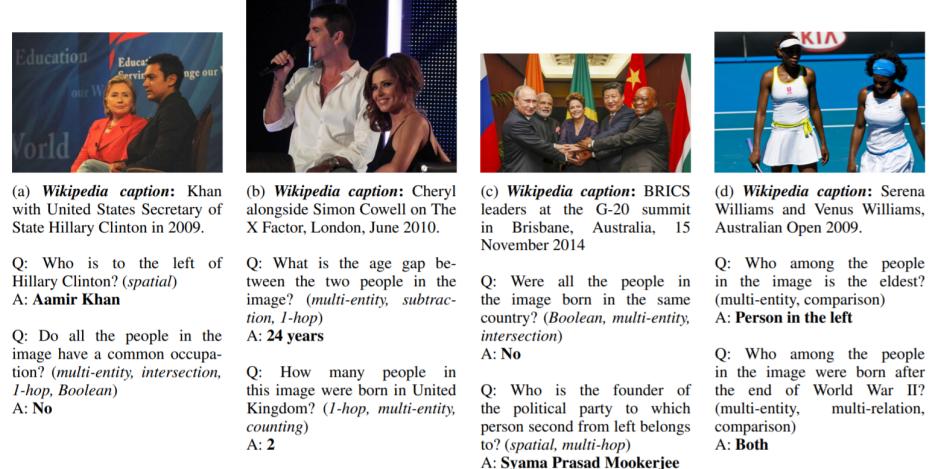
Fig. 2. Samples from VQA dataset [2].



Fig. 3. Samples from Madlibs dataset [36].

**Visual7W:** The Visual7W dataset [39] is also based on the MS-COCO dataset. It contains 47,300 COCO images with 327,939 question-answer pairs. The dataset also consists of 1,311,756 multiple choice questions and answers with 561,459 groundings. The dataset mainly deals with seven forms of questions (from where it derives its name): What, Where, When, Who, Why, How, and Which. It is majorly formed by two types of questions. The ‘telling’ questions are the ones which are text-based, giving a sort of description. The ‘pointing’ questions are the ones that begin with ‘Which,’ and have to be correctly identified by the bounding boxes among the group of plausible answers.

**CLEVR:** CLEVR [12] is a synthetic dataset to test the visual understanding of the VQA systems. The dataset is generated using three objects in each image, namely cylinder, sphere and cube. These objects are in two different sizes, two different materials and placed in eight different colors. The questions are also synthetically generated based on the objects placed in the image. The dataset also accompanies the ground-truth bounding boxes for each object in the image.

**Fig. 4.** Samples from Tally-QA dataset [1].**Fig. 5.** Samples from KVQA dataset [26].

**Tally-QA:** Very recently, in 2019, the Tally-QA [1] dataset is proposed which is the largest dataset of object counting in the open-ended task. The dataset includes both simple and complex question types which can be seen in Fig. 2. The dataset is quite large in numbers as well as it is 2.5 times the VQA dataset. The dataset contains 287,907 questions, 165,000 images and 19,000 complex questions. The Tally-QA samples are shown in Fig. 4.

**KVQA:** The recent interest in common-sense questions has led to the development of Knowledge based VQA dataset [26]. The dataset contains questions targeting various categories of nouns and also require world knowledge to arrive at a solution. Questions in this dataset require multi-entity, multi-relation, and multi-hop reasoning over large Knowledge Graphs (KG) to arrive at an answer. The dataset contains 24,000 images with 183,100 question-answer pairs employing around 18K proper nouns. The KVQA samples are shown in Fig. 5.

**Table 2.** Overview of Models described in this paper. The Pythia v0.1 is the best performing model over VQA dataset.

Model	Dataset(s)	Method	Accuracy	Venue
Vanilla VQA [2]	VQA [2]	CNN + LSTM	54.06 (VQA)	ICCV 2015
Stacked Attention Networks [33]	VQA [2], DAQAUR [20], COCO-QA [24]	Multiple Attention Layers	58.9 (VQA), 46.2 (DAQAUR), 61.6 (COCO-QA)	CVPR 2016
Teney et al. [31]	VQA [2]	Faster-RCNN + Glove Vectors	63.15 (VQA-v2)	CVPR 2018
Neural-Symbolic VQA [34]	CLEVR [12]	Symbolic Structure as Prior Knowledge	99.8 (CLEVR)	NIPS 2018
FVTA [17]	MemexQA [9], MovieQA [30]	Attention over Sequential Data	66.9 (MemexQA), 37.3 (MovieQA)	CVPR 2018
Pythia v1.0 [11]	VQA [2]	Teney et al. [31] + Deep Layers	72.27 (VQA-v2)	VQA Challenge 2018
Differential Networks [32]	VQA [2], TDIUC [13], COCO-QA [24]	Faster-RCNN, Differential Modules, GRU	68.59 (VQA-v2), 86.73 (TDIUC), 69.36 (COCO-QA)	AAAI 2019
GNN [38]	VisDial and VisDial-Q	Graph neural network	Recall: 48.95 (VisDial), 27.15 (VisDial-Q)	CVPR 2019

### 3 Deep Learning Based VQA Methods

The emergence of deep-learning architectures have led to the development of the VQA systems. We discuss the state-of-the-art methods with an overview in Table 2.

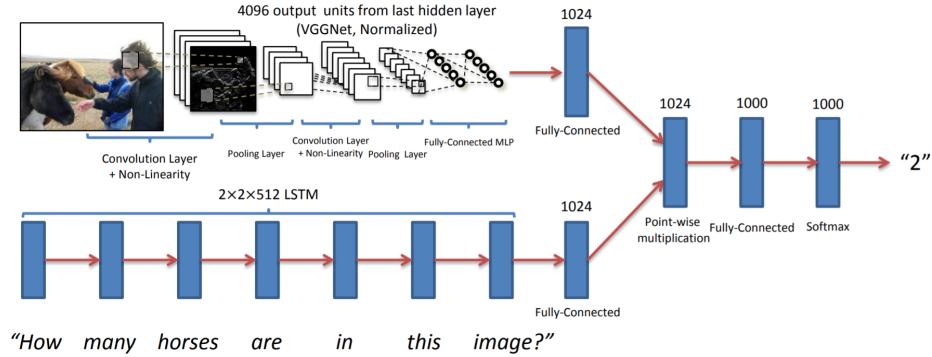
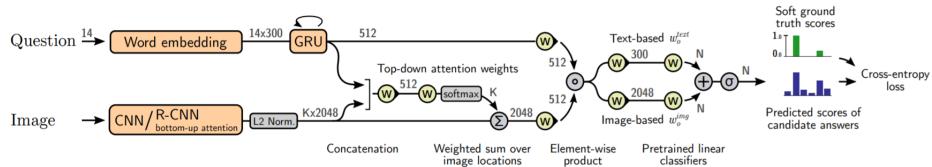
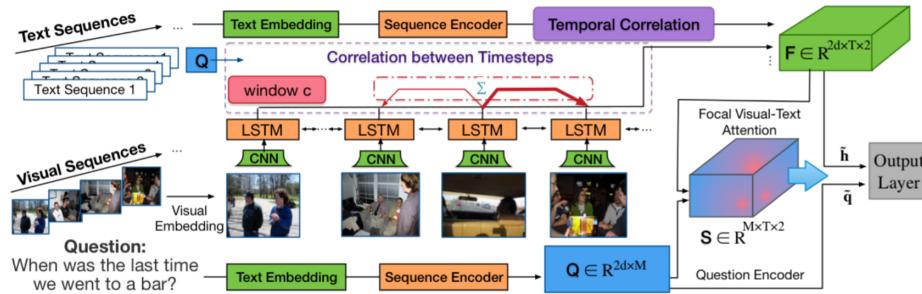
**Vanilla VQA [2]:** Considered as a benchmark for deep learning methods, the vanilla VQA model uses CNN for feature extraction and LSTM or Recurrent networks for language processing. These features are combined using element-wise operations to a common feature, which is used to classify to one of the answers as shown in Fig. 6.

**Stacked Attention Networks [33]:** This model introduced the attention using the softmax output of the intermediate question feature. The attention between the features are stacked which helps the model to focus on the important portion of the image.

**Teney et al. Model [31]:** Teney et al. introduced the use of object detection on VQA models and won the VQA Challenge 2017. The model helps in narrowing down the features and apply better attention to images. The model employs the use of R-CNN architecture and showed significant performance in accuracy over other architectures. This model is depicted in Fig. 7.

**Neural-Symbolic VQA [34]:** Specifically made for CLEVR dataset, this model leverages the question formation and image generation strategy of CLEVR. The images are converted to structured features and the question features are converted to their original root question strategy. This feature is used to filter out the required answer.

**Focal Visual Text Attention (FVTA) [17]:** This model combines the sequence of image features generated by the network, text features of the image (or probable an-

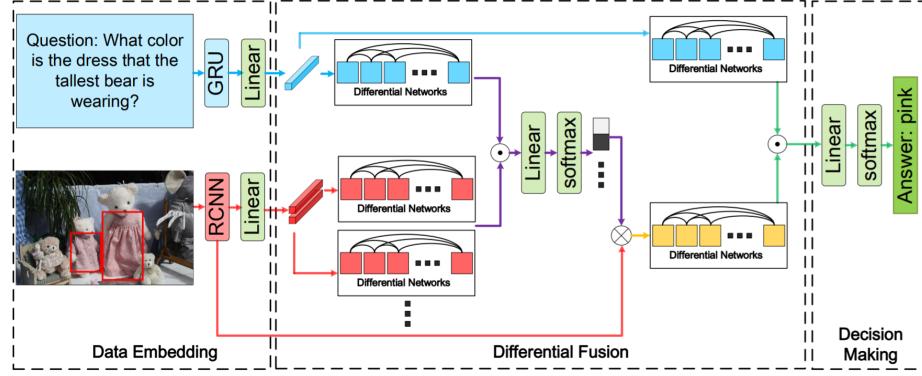
**Fig. 6.** Vanilla VQA Network Model [2].**Fig. 7.** Teney et al. VQA Model [31]**Fig. 8.** Focal Visual Text Attention Model [17]

swers) and the question. It applies the attention based on the both text components, and finally classifies the features to answer the question. This model is better suited for the VQA in videos which has more use cases than images. This model is shown in Fig. 8.

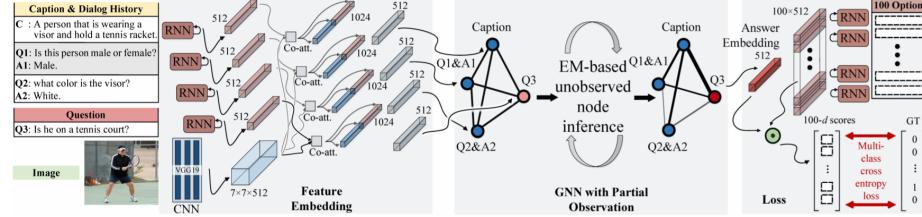
**Pythia v1.0 [11]:** Pythia v1.0 is the award winning architecture for VQA Challenge 2018<sup>2</sup>. The architecture is similar to Teney et al. [31] with reduced computations with element-wise multiplication, use of GloVe vectors [23], and ensemble of 30 models.

**Differential Networks [32]:** This model uses the differences between forward propagation steps to reduce the noise and to learn the interdependency between features. Image features are extracted using Faster-RCNN [25]. The differential modules [22] are

<sup>2</sup> <https://github.com/facebookresearch/pythia>



**Fig. 9.** Differential Networks Model [32].



**Fig. 10.** Differentiable Graph Neural Network [38].

used to refine the features in both text and images. GRU [4] is used for question feature extraction. Finally, it is combined with an attention module to classify the answers. The Differential Networks architecture is illustrated in Fig. 9.

**Differentiable Graph Neural Network (GNN) [38]:** Recently, Zheng et al. have discussed about a new way to model visual dialogs as structural graph and Markov Random Field. They have considered the dialog entities as the observed nodes with answer as a node with missing value. This model is illustrated in Fig. 10.

## 4 Experimental Results and Analysis

The reported results for different methods over different datasets are summarized in Table 1 and Table 2. It can be observed that VQA dataset is very commonly used by different methods to test the performance. Other datasets like Visual7W, Tally-QA and KVQA are also very challenging and recent datasets. It can be also seen that the Pythia v1.0 is one of the recent methods performing very well over VQA dataset. The Differential Network is the very recent method proposed for VQA task and shows very promising performance over different datasets.

As part of this survey, we also implemented different methods over different datasets and performed the experiments. We considered the following three models for our experiments, 1) the baseline Vanilla VQA model [2] which uses the VGG16 CNN ar-

**Table 3.** The accuracies obtained using Vanilla VQA [2], Stacked Attention Networks [33] and Teney et al. [31] models when trained on VQA [2] and Visual7W [39] datasets.

<b>Model Name</b>	<b>Accuracy</b>	
	<b>VQA Dataset</b>	<b>Visual7W Dataset</b>
CNN + LSTM	58.11	56.93
Stacked Attention Networks	60.49	61.67
Teney et al.	67.23	65.82

chitecture [28] and LSTMs [7], 2) the Stacked Attention Networks [33] architecture, and 3) the 2017 VQA challenge winner Teney et al. model [31]. We considered the widely adapted datasets such as standard VQA dataset [2] and Visual7W dataset [39] for the experiments. We used the Adam Optimizer for all models with Cross-Entropy loss function. Each model is trained for 100 epochs for each dataset.

The experimental results are presented in Table 3 in terms of the accuracy for three models over two datasets. In the experiments, we found that the Teney et al. [31] is the best performing model on both VQA and Visual7W Dataset. The accuracies obtained over the Teney et al. model are 67.23% and 65.82% over VQA and Visual7W datasets for the open-ended question-answering task, respectively. The above results re-affirmed that the Teney et al. model is the best performing model till 2018 which has been pushed by Pythia v1.0 [10], recently, where they have utilized the same model with more layers to boost the performance. The accuracy for VQA is quite low due to the nature of this problem. VQA is one of the hard problems of computer vision, where the network has to understand the semantics of images, questions and relation in feature space.

## 5 Conclusion

The Visual Question Answering has recently witnessed a great interest and development by the group of researchers and scientists from all around the world. The recent trends are observed in the area of developing more and more real life looking datasets by incorporating the real world type questions and answers. The recent trends are also seen in the area of development of sophisticated deep learning models by better utilizing the visual cues as well as textual cues by different means. The performance of the best model is still lagging and around 60-70% only. Thus, it is still an open problem to develop better deep learning models as well as more challenging datasets for VQA. Different strategies like object level details, segmentation masks, deeper models, sentiment of the question, etc. can be considered to develop the next generation VQA models.

## References

1. Acharya, M., Kafle, K., Kanan, C.: Tallyqa: Answering complex counting questions. arXiv preprint arXiv:1810.12440 (2018)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: IEEE ICCV, pp. 2425–2433 (2015)

3. Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y.: Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10,800–10,809 (2020)
4. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
5. Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from rgb-d images. In: IEEE CVPR, pp. 564–571 (2013)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR, pp. 770–778 (2016)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
8. Huang, Q., Wei, J., Cai, Y., Zheng, C., Chen, J., Leung, H.f., Li, Q.: Aligned dual channel graph convolutional network for visual question answering. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7166–7176 (2020)
9. Jiang, L., Liang, J., Cao, L., Kalantidis, Y., Farfade, S., Hauptmann, A.: Memexqa: Visual memex question answering. arXiv preprint arXiv:1708.01336 (2017)
10. Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D.: Pythia v0. 1: the winning entry to the vqa challenge 2018. arXiv preprint arXiv:1807.09956 (2018)
11. Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D.: Pythia v0. 1: the winning entry to the vqa challenge 2018. arXiv preprint arXiv:1807.09956 (2018)
12. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: IEEE CVPR, pp. 2901–2910 (2017)
13. Kafle, K., Kanan, C.: An analysis of visual question answering algorithms. In: ICCV (2017)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
15. Li, G., Wang, X., Zhu, W.: Boosting visual question answering with context-aware knowledge aggregation. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 1227–1235 (2020)
16. Li, W., Sun, J., Liu, G., Zhao, L., Fang, X.: Visual question answering with attention transfer and a cross-modal gating mechanism. Pattern Recognition Letters **133**, 334–340 (2020)
17. Liang, J., Jiang, L., Cao, L., Li, L.J., Hauptmann, A.G.: Focal visual-text attention for visual question answering. In: IEEE CVPR, pp. 6135–6143 (2018)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV, pp. 740–755 (2014)
19. Lobry, S., Marcos, D., Murray, J., Tuia, D.: Rsvqa: Visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing (2020)
20. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: NIPS, pp. 1682–1690 (2014)
21. Medsker, L.R., Jain, L.: Recurrent neural networks. Design and Applications **5** (2001)
22. Patro, B., Namboodiri, V.P.: Differential attention for visual question answering. In: IEEE CVPR, pp. 7680–7688 (2018)
23. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: EMNLP, pp. 1532–1543 (2014)
24. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: Advances in neural information processing systems, pp. 2953–2961 (2015)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS, pp. 91–99 (2015)
26. Shah, S., Mishra, A., Yadati, N., Talukdar, P.P.: Kvqa: Knowledge-aware visual question answering. In: AAAI (2019)

27. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV, pp. 746–760 (2012)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
29. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE CVPR, pp. 2818–2826 (2016)
30. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: IEEE CVPR, pp. 4631–4640 (2016)
31. Teney, D., Anderson, P., He, X., van den Hengel, A.: Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: IEEE CVPR, pp. 4223–4232 (2018)
32. Wu, C., Liu, J., Wang, X., Li, R.: Differential networks for visual question answering. AAAI 2019 (2019)
33. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: IEEE CVPR, pp. 21–29 (2016)
34. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.: Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In: NIPS, pp. 1031–1042 (2018)
35. Yu, J., Zhu, Z., Wang, Y., Zhang, W., Hu, Y., Tan, J.: Cross-modal knowledge reasoning for knowledge-based visual question answering. Pattern Recognition **108**, 107,563 (2020)
36. Yu, L., Park, E., Berg, A.C., Berg, T.L.: Visual madlibs: Fill in the blank description generation and question answering. In: IEEE ICCV, pp. 2461–2469 (2015)
37. Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M.: Medical visual question answering via conditional reasoning. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2345–2354 (2020)
38. Zheng, Z., Wang, W., Qi, S., Zhu, S.C.: Reasoning visual dialogs with structural and partial observations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6669–6678 (2019)
39. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: IEEE CVPR, pp. 4995–5004 (2016)