IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Survey On Visual Question Answering

[1]Varalakshmi Iyer, [2]Dr. Darshak G. Thakore

[1]M. Tech Student, [2]Head of the Department

[1]Department of Computer Engineering,

[1]Birla Vishvakarma Mahavidyalaya, Vallabh Vidyanagar, Gujarat, India.

Abstract: Visual question answering (VQA) is a multi-disciplinary task. The main aim of VQA system is to provide natural language answer to an open-ended question about a given image. This task involves both image understanding and natural language processing. In order to provide answer to the question, VQA system performs image classification, object detection and reasoning over the image. In most of the works, feature extraction of the image and the question is done in order to provide appropriate results. Convolution neural network (CNN) and Recurrent Neural Network (RNN) are utilized in a VQA system. This paper provides a brief survey of various research carried out so far in this field. This paper also discusses the existing datasets, feature extraction methodologies and evaluation metrics used in this system.

Index Terms **- Visual Question Answering, Convolutional Neural Network, Recurrent Neural Network, Image classification, Feature Extraction.**

I. INTRODUCTION

The latest advancement in the field of Artificial Intelligence (AI) has been making the lives of people much easier. Today, most of the human work is carried out by robots. The introduction of Visual Question Answering task has given birth to a renewed excitement in the field of multi-disciplinary AI research problems. Visual Question Answering is one of the most interesting tasks that has attracted the attention of many researchers. It is a free- form and open-ended AI task. Visual Question Answering (VQA) is a multi-disciplinary AI research problem that requires both image understanding and natural language processing. This research problem introduction is a breakthrough towards introducing more “AI complete” tasks. "AI complete" tasks require multiple domain knowledge beyond a single sub domain knowledge and have a well-defined evaluation metric.

Visual Question Answering can be defined as a system that takes an image and a free-form, open-ended natural language question about the image as input and provides a natural language answer as the output. This natural language answering process requires various capabilities such as object recognition (“How many people are there?”), activity recognition (“Are they playing?”), scene recognition (“What is the weather in the image?”), attribute classification (“What is the shape of box in the image?”), knowledge-based reasoning and common-sense reasoning to answer questions that require additional information other than that available from the image.

A basic VQA system involves the extraction of image features and question features through Convolution Neural Network (CNN) and Recurrent Neural Network (RNN) and then combining those features to generate an answer. Most of the VQA systems treat answer generation as a classification problem. Currently, the VQA systems are trained well to answer counting-based and object detection-based questions. Attention-based VQA systems were introduced to provide accurate answers. These systems only focus on a specific part of the image to answer the natural language question asked about the image. But these systems are far behind when reasoning and knowledge-based answering of the question is required. Also, the same model provides different answers to different users. A possible reason behind this is the irrelevance of question to the image present. Most of the VQA systems when asked irrelevant questions, provide an answer based on their trained features. Question relevance and knowledge-based VQA are possible areas where there is a need for development.

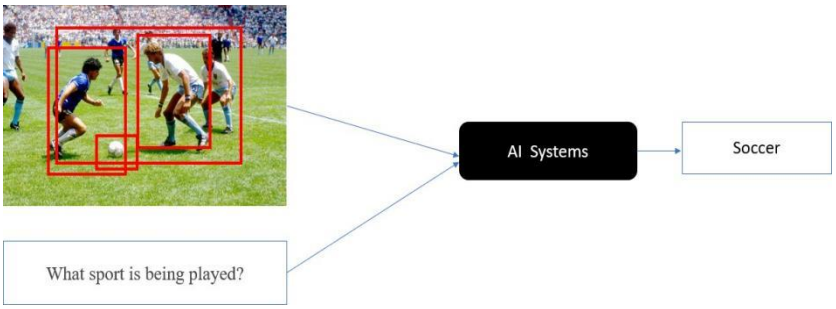

Fig. 1. Visual Question Answering System

Figure 1 provides a general idea of the Visual Question Answering System. Here, as shown in Figure 1, the VQA system obtains an image and an open-ended question related to the image as inputs and provides a natural language answer as output.

Many datasets have been introduced in the past few years to solve the VQA research problems. Some of the widely used datasets such as DAQUAR [23], VQA v1 and v2 [14], VizWiz [22], etc. have become prominent. VQA is a challenging task. It involves the proper extraction of both the image and text features. Much research work has been published in this area in the last few years.

II. Literature Review

Computer Vision and Natural Language Processing have gained a lot of recognition and development in recent years. With the growing interest in the field of computer vision, there is increasing research work in the area of image understanding. Similarly, there is progress in Natural Language Processing side. Visual Question Answering is a combination of these two research areas. VQA can be termed as a multi–disciplinary field in which the system requires to gain image understanding as well as language understanding to provide appropriate results. The open-ended natural language output depends on the input image as well as the input question.

Malinowski et al. [20] proposed the first VQA approach. The authors proposed to process questions through semantic parsing and obtain answers through Bayesian Reasoning. They built their dataset on top of the NYU-Depth V2 dataset. After collecting the images, the authors created question-answer pairs for the NYU dataset. The VQA dataset created from the NYU-Depth V2 dataset was very small in size.

A general VQA approach was proposed by Parikh et al. [14]. In this approach, the authors proposed to extract the image features using CNN and encode questions using Long Short Term Memory (LSTM). They treated VQA as a clas- sification system in which the extracted image and question features were combined and passed through a multi-layer perceptron to obtain the results. To fuse the question and the image features, Hadamard Product [56] of their vectors was used. Some of the common methods of combining image and question features include concatenation, elementwise product or elementwise sum. In the early-stage for combining question and image features these linear pooling methods were utilized. Later on, some of the bilinear pooling methods were proposed such as MCB [40], MLB [41], MFB [42] and MLPB [43] that have been shown to be much more effective than the linear pooling methods.

Malinowski et al. [44] have also proposed a CNN-LSTM based approach. In this approach, the author proposed to combine CNN features with LSTM features into an end-to-end architecture to predict the correct answer for the given question and the image. A large number of algorithms have been proposed in this research area. All these algorithms consist of image feature extraction, question feature extraction, and an algorithm to combine these features to obtain appropriate results.

Image feature extraction was performed using CNN algorithms. Various CNN architectures such as VGGNet [18], ResNet [46], Faster-RCNN [47], GoogleNet [45], etc. were utilized for this task. Similarly, for the question embedding task, RNN (LSTM and Gated Recurrent Unit (GRU)) was preferred by most of the researchers. Some of the researchers such as Antol et al. [20] utilized the idea of creating a Bag of Words(BoW) with the top 1000 words in the questions from the dataset. They have also considered strong co-relation between the words that start a question and the answer. Kafle et al. [48] proposed "answer type prediction". In this approach, they took advantage of the fact that the type of the answer can be predicted based on the question. This made the VQA problem a multiple choice one. In this research work, ResNet was used to extract the image features and skip-through vectors were used to represent the question feature.

Most of these research works published in the field of VQA make use of global image features to answer questions. This may result in limiting the capacity of systems to answer questions about the local image regions. For example, the question, "What is located on the right of the chair?" requires a system to identify a chair and look on the right of the chair to answer this question correctly. As a result, recent VQA approaches have introduced the visual attention mechanism into them by learning attention-based image features for a given question and then performing multi-modal feature integration to obtain accurate answers [1].

The attention mechanism has already been utilized in the task of image captioning, object detection, etc. Visual attention focuses on "where to look" to provide accurate results. Chen et al. [49] introduced an attention mechanism based on deep learning architecture for VQA. They proposed an Attention- based Configurable Convolutional Neural Network (ABC- CNN). Since different questions focus on different regions of the image, the attention mechanism implemented in their research work was question-guided. ABC-CNN [49] determined the attention regions by finding the corresponding visual features in the visual feature maps using a "configurable convolution" operation. Yang et al. [50] proposed a "stacked attention network". In this research work, the authors proposed a multi-layer Stacked Attention Network (SAN) in which they query an image multiple times to obtain appropriate answers. Here, the image features were extracted through VGGNet architecture and the question features through LSTM/GRU. These features were then passed through a single-layer neural network. To generate the attention distribution over the regions of an image, a softmax

function was used. Depending upon the attention distribution, the weighted sum of the image vectors was calculated from each region and then combined with the question vector to enhance and update the query for obtaining a relevant answer.

In the work by Gao et al. [2], the authors proposed "question-led object attention for visual question answering". In this research work, the image and question features were passed through a single-layer perceptron and then a softmax function was applied to it to generate the attention distribution. Here, instead of using LSTM or GRU to encode a question, the authors used a convolution unit with a local "receptive field" and "shared weights" to capture the semantics information of the question. In this work, the answer was treated as a multi-class classification problem. The authors observed that the model gave lower results for number-based questions. The possible cause of this according to the authors was the consideration of only objects present in the image while performing question-led object attention without considering the relationship between those objects.

Along with the visual content, VQA also needs to under-stand the semantics of the question. Therefore, it was later proposed that along with visual attention, textual attention also needs to be studied. Lu et al. [13] proposed co-attention- based framework. In this work, the authors proposed that "what words to listen" is equally important to "where to look" [13]. The model proposed by authors performed question feature extraction hierarchically (word level, phrase level, and question-level) using a novel one-dimensional convolution neural network. Using this model, the authors also presented two approaches of co-attention mechanism i.e., parallel co-attention and alternative co-attention.

In the work by Peng et al. [11] the authors proposed a "word to region attention model for visual question answering". Here, the authors challenged other attention-based models [50], by proposing that only some important keywords in the query question were involved in identifying the relevant image regions to answer the question. Compared to previous work [13], here the authors proposed to generate a "word map" to highlight core words and extract important information from the question. They utilized Faster- RCNN with Resnet to extract image features and RNNs to represent question features. In this work, the authors used local object regions of the image instead global image regions. The authors observed that this model provided good accuracy when the questions were straight-forward and about an object in the image, but performed poorly for complicated questions or questions having lots of nouns. In the work by Nam et al. [52] the authors proposed "dual attention networks". According to the authors, that DANs (Dual Attention Network) focuses on specific regions in the image and text. To capture essential information from the visual and textual features, it performs multiple reasoning steps.

In the work by Shi et al. [53] the authors proposed "question type guided attention in VQA". Here, the model used the information "question type" to guide the visual feature extraction process. This reduced the search space for answers. Inspired by their work, Yang et al. [12] proposed a co-attention mechanism with a question type-based combined model.

In the work by Yu et al. [1], the authors proposed a "deep-modular co-attention networks-based" model to perform visual question answering. Here, the authors focused on the dense interaction between the question word and image regions [1]. To understand the relationships among these question and image features, the authors proposed a dense co-attention model. Here the authors designed two general attention units: a self-attention (SA) unit and a guided-attention (GA) unit. The SA unit helped model the dense interaction between word- to-word or region-to-region. The GA unit helped model the dense interactions between word-to-region. After that, by the modular composition of the SA and GA units, the authors obtained different Modular Co-Attention (MCA) layers, which were cascaded in depth. Finally, they proposed a deep Modular Co-Attention Network (MCAN) which consisted of cascaded MCA layers [1].

VQA systems are mostly trained to answer questions that are present in the images. Some of the open-ended questions asked by humans often require external knowledge to answer them. This gave a rise to External Knowledge-based Visual Question Answering.

Wu et al. [54] introduced external knowledge-based Visual Question answering. In this research work, the authors combined the image description with external knowledge to provide an answer to a question about the image. The external knowledge base that the authors utilized in this research work was DBpedia [23]. First, a CNN was used to obtain attributes of the image. Image captions were later generated based on these attributes using a state of art-based image captioning model. Then, these detected attributes were used to extract relevant information from the KB (Knowledge Base). Later the captions, attributes, and the KB information were passed into an LSTM which was trained to obtain appropriate ground- truth answers.

Wang et al. [15] claimed that the conventional LSTM based approach for answer generation does not provide an appropriate explanation of how they arrived at a particular answer. So, the authors proposed the "Ahab" approach. This was based on reasoning about the content of the images. To provide appropriate answers to the questions, the relevant information from the images was obtained. This relevant information included image scenes, objects in the image, and image attributes. They were extracted using Faster-RCNN, VGGNet- 16, etc. Later this information was linked with appropriate external information. This information was then stored in RDF (Resource Description Framework) triples format. The question was first converted into a form that can be used to query the RDBMS. The questions were parsed by a set of regex and then connected to appropriate slot phrases. To obtain an answer to the question, equivalence between the slot phrases and entities in the graph was used. Similarly, they also introduced FVQA [17] dataset which primarily contained image-question-answer- supporting facts tuples.

Another work by, Schwing et al. [5] was developed on the FVQA dataset. In contrast to the previous work by [15] and [17], here the authors tried to eliminate the errors due to synonyms, homographs, and incorrect prediction of visual concept type and answer type [51].

In the work by Shah et al. [3], the authors proposed "knowledge aware visual question answering". In the proposed methodology the authors first performed visual entity linking which for them was face identification. The identification of the face was then followed by linking the face data with Wiki- data. After this, the visual question answering task involved obtaining relevant facts from KG (Knowledge Graphs), reasoning over them, and learning to answer questions. The authors used the memory network as one of their baselines for KVQA. First visual entity linking was done, then fetching facts from KGs, then each knowledge and the spatial fact was fed to a Bi- LSTM (Bi-Directional LSTM) to get corresponding memory embeddings. Then, at last, the sum of output representation 'o' and the question 'q' were fed to a multi-layer perceptron. Presently, the KVQA dataset is limited to persons.

In the work by Mishra et al. [55], the authors proposed a VQA system that can read. They published the OCR-VQA dataset. They proposed a novel methodology of answering the question by reading text in images.

Along with this some of the researchers also explored other areas in VQA. In the paper "Why Does a Visual Question Have Different Answers? "by Bhattacharya et al. [4] the authors presented reasons why a VQA system might answer different answers to different users. They listed 6 different reasons why a visual question is unanswerable. Question relevance is also one of the important issues in the field of Visual Question Answering.

In the work by Toor et al. [5], the authors proposed Question action relevance and editing for visual question answering. The authors here focused on "action-verb". In this work, the authors obtained the action-verb using a pre-trained image caption model (NeuralTalk2 [59]) that utilized the fundamental (Res-Net50 CNN) as a visual feature extractor. They proposed to identify irrelevant questions and edit them. In this work, the authors worked on action-based question editing. In another work by Ray et al. [6], the authors proposed identification of a question as visual or non-visual by applying LSTM and Rule- based methodology on POS (Part of Speech) tags associated with the questions. Later the question relevance identification was done based on the object present in the image.

Along with these developments in the field of VQA, Acharya et al. [7] proposed "Tally VQA" to answer complex counting questions. In this work, the authors performed question feature extraction using –one-layer GRU. For the foreground patches, the authors used Faster R-CNN. For the background patches, the authors extracted ResNet-152 features from the entire image before the last pooling layer and then applied average pooling over these features.The outputs of these networks were then concatenated and passed to one hidden layer with 1024 units and ReLU activation. The authors have pointed out the need for an intelligent pruning method. VQA systems are becoming very prominent these days as they are helpful in the field of image retrieval. They are helpful in assisting blind people to understand the environment around them. Many field-specific VQA models have also been developed such as DocVQA [32] for document-based images, RecipeQA [33] for cooking recipe-based images, Visual Ques- tion Answering for Cultural Heritage [34] and VQA-Med [35] for medical images.

III. Dataset and Evaluation

Several datasets have been published for the visual question answering system. These datasets contain images, questions associated with the image, and correct answers to those questions. These datasets also contain some additional annotations associated with the image.

DAQUAR [20] which stands for Dataset for Question Answering on Real-world images is the very first dataset published in this field. This is the smallest dataset in this research area. It consists of approximately 1449 images (795 training and 654 testing images). It takes images from the NYU Depth V2 dataset [61]. All the images in this dataset are indoor images. It consists of only 6795 training and 5673 testings QA pairs based on images [26].

Fig. 2. Sample Image from NYU-Depth v2 Dataset [61]

Image 5: Question: How many chairs are on the right side of the table in the image5? Correct Answer:3.
Figure 2 provides a sample of the image and question- answer pair are obtained from the DAQUAR [20] dataset.

Another dataset that was released in the year 2015 was COCO-QA [27]. The images in the dataset were obtained from MS-COCO dataset [21]. COCO-QA dataset consists of object, number, color, and location-related questions. The maximum length of the questions in this dataset is 55. This dataset is larger than the DAQUAR [20] dataset. It contains 123287 images, 78736 training questions, and 38948 testing questions. Figure 3 shows the sample image of the COCO-QA dataset.

Fig. 3. Sample Image from COCO-QA dataset [60]

VQA v1 and v2 [14] is one of the largest datasets in the field of visual question answering. This dataset contains open-ended questions about images. It consists of around 265,016 images (COCO and abstract scenes), at least three questions per image, 10 ground-truth answers per question, and 3 plausible answers per question. There are 82000 training images, 40000 validation images and about 81000 testing images. V2 dataset was introduced to remove biases in the dataset. Figure 4 shows a sample image of the VQA dataset.

Fig. 4. Sample Image from VQA dataset [62]

Visual7W [22] dataset is generated using images from the MS-COCO dataset [21]. The Visual7W dataset is a part of the Visual Genome project [30]. Visual Genome contains 1.7 million QA pairs of the 7W question types, which offers the largest visual QA collection to date for training models. The QA pairs in Visual7W are a subset of the 1.7 million QA pairs from Visual Genome [20]. The dataset contains 47300 images. Totally, it has 327,939 QA pairs, along with 1,311,756 human- generated multiple-choices and 561,459 object groundings from 36,579 different categories [29]. This dataset contains questions of the form Who, What, Which, When, Why, How and Where. They also provide complete groundingannotations that link the object mentioned in the QA sentences to their bounding boxes in the images and therefore introduce a new QA type with image regions as the visually grounded answers [29]. Figure 5 shows a sample image from Visual7W dataset.

Fig. 5. Sample image from Visual7W dataset [63]

Some other Visual Question Answering datasets were also published to accomplish specific tasks. FigureQA [31] dataset is specifically designed for research related to graphical plots and figures. These images consist of pie charts, line plots, dot- line plots, histograms, etc. The dataset is generated at a very large scale. Its training set contains 100, 000 images and 1.3 million questions. The validation and test sets each contain 20, 000 images with more than 250, 000 questions [31]. The questions present in the dataset are relational. Figure 6 shows a sample image from the FigureQA dataset.

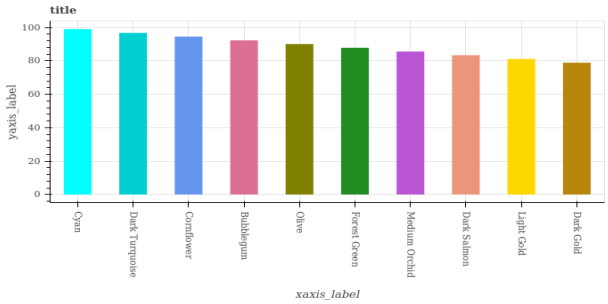

Fig. 6. Sample image from FigureQA dataset [64]

Similarly, many such VQA datasets published in recent times are dedicated towards specific research area such as DocVQA [32] for document-based images, RecipeQA [33] for cooking recipe-based images, Visual Question Answering for Cultural Heritage [34] and VQA-Med [35] for medical images. CLEVR [37] is a dataset to test the Visual understanding of the VQA system. The dataset consists of three types of object shapes (cube, sphere and cylinder) in both small and large sizes. This dataset is used to analyze a group of VQA models and identify their weaknesses. Most of the counting questions in the VQA dataset perform simple counting which mostly requires object detection. So, recently in 2019 TallyQA [7] was introduced. In this dataset, the authors studied algorithms to answer complex counting questions that involved relationships between objects, identification of attributes, reasoning over those objects and more [7]. The authors believe this dataset to be the world's largest dataset for open-ended counting- based questions. It includes both simple and complex counting questions. It contains 287,907 questions, 165,000 images and 19,000 complex questions [36]. Figure 7 shows an example presented by the authors of the TallyQA dataset.

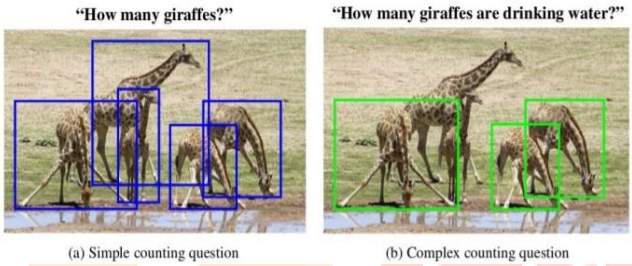

Fig. 7. Image showing the comparison between simple and complex questions [65]

In conventional VQA systems, the questions asked to the VQA systems were answered purely based on the content of the images. Recently there has been growing interest in the development of common-sense-based and external knowledge- based VQA systems. KVQA [3] is the first dataset for the task of World -Knowledge Enabled VQA systems. The questions in this dataset are based on various categories of nouns and also requires external knowledge to obtain an answer to the question [36]. KVQA [3] dataset consists of around 183K question-answer pairs. These pairs involve more than 18K named entities and 24K images [3]. Figure 8 shows sample images questions and captions present in the dataset.

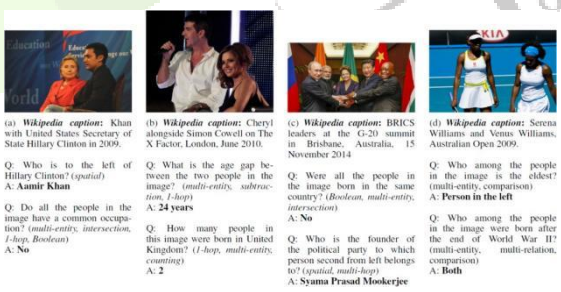

Fig. 8. Sample images and questions from KVQA dataset [66]

Visual Question Answering system provides various evaluation metrics to evaluate the generated answers. For evaluation of VQA systems Simple accuracy can also be utilized. In a simple accuracy system, the answer is considered accurate if the generated answer matches exactly with the ground truth. Since the answer needs to match exactly with the proposed ground truth, this cannot handle incomplete or less incorrect answers. For example, if the ground truth answer is bat and the system responds with bats, then it would be treated in the same manner if the responded answer were ball. Due to this, some other evaluation metrics have also been introduced for VQA systems. Another Evaluation metrics that is utilized by DAQUAR [20] and COCO-QA [27] dataset is WUPS (Wu- Palmer similarity) [58]. WUPS evaluation metrics measures how much a predicted word differs from the ground truth word based on the differences in their semantic meaning. WUPS assigns a value between 0 to 1 based on the similarity between the predicted word and the ground truth word. In WUPS certain issues may arise such as some of the answers may be lexically similar but have a different meaning. For example, the name of the animals. WUPS works properly for single-word answers.

VQA dataset [14] consists of Consensus Metrics to evaluate the Visual Question Answering system. In this evaluation metrics method, every question has 10 ground truth answers. Equation 1 shows the formula for calculating the accuracy of the answer generated using the Consensus metrics [38].

$$\text{Accuracy} = \min(\#\text{humans that said that ans}/3, 1) \quad (1)$$

For the answer to be consistent with the human answers, the machine-generated answers are averaged over all 10 choose 9 sets of human annotators [38]. Using this metric, if the algorithm agrees with the answer of three or more annotators then a full score is given to that question [26]. In DAQUAR [20] dataset consensus, an average of five human-annotated ground truths were collected. The consensus-based accuracy calculation was proposed in two ways, which they called average consensus and min consensus. In average consensus, the final score is given to the more popular answer provided by the annotators. In min consensus, the answer should agree with at least one of the annotator [39]. VQA-med [35] dataset creators have suggested BLEU (bilingual evaluation understudy) [57] metric to capture the similarity between a machine-generated answer and the ground truth answer.

IV. Conclusion

Visual question answering is a technique to generate natural language answers to the open-ended question about a given image. This paper presents a comprehensive review of the ongoing research and works done in the field of visual question answering. The number of works in this field is constantly increasing day by day due to the interesting features of this research area. Most of the work in this field has been done on visual-based question answering. There are various new methodologies in this field such as attention-based VQA, knowledge and common-sense enabled VQA system as well as answering unanswerable questions and question relevance. This paper also provides a comprehensive survey of various datasets that are utilized so far in the realization of visual question answering. Various evaluation metrics have also been discussed. We believe that the ongoing and future work in this research area will benefit the VQA task.

V. Acknowledgment

I express my sincere thanks and gratitude to my guides Dr. Darshak G. Thakore and Dr. Mayur M. Vegad who guided me through the various stages of this research work. I am thankful for their constant support and guidance in identification of appropriate area of research work.

I am also thankful to all my colleagues who have given me support and assistance during this work. I am thankful to my parents and friends for their constant support and motivation. Last but not the least I am thankful to all the people who have directly or indirectly helped me realize this work.

References

[1] Yu, Zhou; Yu, Jun; Cui, Yuhao; Tao, Dacheng; Tian, Qi, "Deep Modular Co-Attention Networks for Visual Question Answering", Proceedings of the IEEE conference on computer vision and pattern recognition 2019.

[2] Lianli Gao, Liangfu Cao, Xing Xu, Jie Shao, Jingkuan Song, "Question- Led object attention for visual question answering", Neurocomputing Volume 391, 28 May 2020, Pages 227-233.

[3] Shah, S., Mishra, A., Yadati, N., Talukdar, P. P. (2019)," KVQA: Knowledge-Aware Visual Question Answering", Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 8876-8884.

[4] Nilavra Bhattacharya, Qing Li, Danna Gurari, "Why Does a Visual Question Have Different Answers?", The IEEE International Conference on Computer Vision (ICCV) 2019

[5] Andeep S. Toor, Harry Wechsler, and Michele Nappi. 2019 "Question action relevance and editing for visual question answering", Multimedia Tools Appl. 78, 3 (February 2019), 2921–2935.

[6] Arijit Ray, Gordon Christie, Mohit Bansal, Dhruv Batra, Devi Parikh, "Question Relevance in VQA: Identifying Non-Visual and False-Premise Questions", arXiv:1606.06622

[7] Acharya, M., Kafle, K., Kanan, C. (2019, July), "TallyQA: Answering complex counting questions", In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 8076-8084).

[8] Ernest Davis, "Unanswerable Questions About Images and Texts", Department of Computer Science, New York University, New York, NY, United States, URL: https://www.frontiersin.org/articles/10.3389/frai.2020.00051/fullB2

[9] Lianli Gao, Liangfu Cao, Xing Xu, Jie Shao, Jingkuan Song, "Question Led object attention for visual question answering", Neurocomputing, Volume 391, 2020, Pages 227-233, ISSN 0925-2312, doi: https://doi.org/10.1016/j.neucom.2018.11.102.

[10] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang, "Bottom-Up and Top-Down At- tention for Image Captioning and Visual Question Answering", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[11] Liang Peng, Yang Yang, Yi Bin, Ning Xie, Fumin Shen, Yanli Ji, Xing Xu, "Word-to-region attention network for visual ques- tion answering", Multimed Tools Appl 78, 3843–3858 (2019), doi: https://doi.org/10.1007/s11042-018-6389-3

[12] C. Yang, M. Jiang, B. Jiang, W. Zhou and K. Li, "Co-Attention Network with Question Type for Visual Question Answering," in IEEE Access, vol. 7, pp. 40771-40781, 2019, doi: 10.1109/ACCESS.2019.2908035.

[13] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016," Hierarchical question-image co-attention for visual question answering", In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 289–297.

[14] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, "VQA: Visual Question Answering", International Conference on Computer Vision (ICCV), 2015.

[15] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, Anton van den Hengel, "Explicit Knowledge-based Reasoning for Visual Question Answering", Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)

[16] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, Anirban Chakraborty, "OCR-VQA: Visual Question Answering by Reading Text in Images", ICDAR, 2019.

[17] Peng Wang; Qi Wu; Chunhua Shen; Anthony Dick; Anton van den Hengel, "FVQA: Fact-Based Visual Question Answering", IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 40, Issue: 10, Oct. 1 2018)

[18] Karen Simonyan, Andrew Zisserman, "Very Deep Convolutional Networks for Large Scale Image Recognition", arXiv:1409.1556.
[19] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham." VizWiz Grand Challenge: Answering Visual Questions from Blind People" IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[20] Malinowski, Mateusz and Fritz, Mario, "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input", Advances in Neural Information Processing Systems 27, Pages, 1682—1690, 2014.
[21] MSCOCO dataset URL: https://cocodataset.org/home
[22] Yuke Zhu, Oliver Groth, Michael Bernstein, Li Fei-Fei, Visual7W: Grounded Question Answering in Images, 2015, arXiv:1511.03416
[23] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives., "DBpedia: A nucleus for a web of open data.", Springer, 2007.
[24] Rupal Agarwal, University of South Florida "State Classification of Cooking Images Using VGG19 Network", 2019
[25] Introduction to Visual Question Answering: Datasets, Approaches and Evaluation Retrieved from https://tryolabs.com/blog/2018/03/01/introduction-to-visual-questionanswering/.
[26] Sunny Katiyar, M. S. Wakode, "A Survey on Visual Questioning Answering: Datasets, Approaches and Models", International Journal of Scientific Technology Research Volume 9, Issue 01, January 2020.
[27] Mengye Ren, Ryan Kiros, and Richard S. Zemel, "Exploring models and data for image question answering", In Proceedings of the 28th International Conference on Neural Information Processing Systems, Volume 2 (NIPS'15). MIT Press, Cambridge, MA, USA, 2953–2961, 2015.
[28] Gupta, A. K., "Survey of Visual Question Answering: Datasets and Techniques", arXiv e-prints, 2017.
[29] Visual7W Toolkit retrieved from https://github.com/yukezhu/visual7wtoolkit
[30] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowd sourced dense image annotations. In arXiv preprint arxiv:1602.07332, 2016.
[31] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, Yoshua Bengio, "FigureQA: An Annotated Figure Dataset for Visual Reasoning", arXiv preprint arXiv:1710.07300, 2017.
[32] Manmatha, C.V. Jawahar "DocVQA: A Dataset for VQA on Document Images", arXiv preprint arXiv:2007.00398, 2020.
[33] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, "RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes", arXiv:1809.00812, 2018.
[34] Pietro Bongini, Federico Becattini, Andrew D. Bagdanov, Alberto Del Bimbo, "Visual Question Answering for Cultural Heritage", arXiv:2003.09853, 2020.
[35] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, Henning Muller, "VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019", CLEF 2019.
[36] Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, Snehasis Mukherjee, "Visual Question Answering using Deep Learning: A Survey and Performance Analysis", arXiv:1909.01860, 2020.
[37] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li FeiFei, C. Lawrence Zitnick, Ross Girshick," CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[38] VQA Evaluation Metrics retrieved from https://visualqa.org/evaluation.html.
[39] Kushal Kafle, Christopher Kanan, "Visual Question Answering: Datasets, Algorithms, and Future Challenges", arXiv:1610.01465, 2017.
[40] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal Compact Bilinear Pooling for Visual Question Answering +and Visual Grounding," in Proc. EMNLP, 2016, pp. 457- 468.
[41] J. H. Kim, K. W. On, W. Lim, J. W. Ha, and B. T. Zhang, "Hadamard Product for Low-rank Bilinear Pooling," in Proc. ICMR, 2016.
[42] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering," in Proc. ICCV, 2017, pp. 1839-1848.
[43] M. Lao, Y. Guo, H. Wang, and X. Zhang, "Multimodal Local Perception Bilinear Pooling for Visual Question Answering," IEEE Access, vol. 6, pp. 57923-57932, Oct. 2018.
[44] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in The IEEE International Conference on Computer Vision (ICCV), 2015.
[45] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Rabinovich A" Going deeper with convolutions", In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1–9,2015.
[46] He K, Zhang X, Ren S, Sun J," Deep residual learning for image recognition", In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778, 2016.
[47] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", arXiv:1506.01497, 2016.
[48] K. Kafle and C. Kanan, "Answer-type prediction for visual question answering," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[49] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, Ram Nevatia, "ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering", arXiv:1511.05960, 2016.
[50] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[51] Medhini Narasimhan, Alexander G. Schwing, "Straight to the Facts: Learning Knowledge Base Retrieval for Factual Visual Question Answering", arXiv:1809.01124, 2018.
[52] H. Nam, J. Ha, and J. Kim, "Dual Attention Networks for Multimodal Reasoning and Matching," in Proc. CVPR, 2017, pp. 2156-2164.

[53] Y. Shi, T. Furlanello, S. Zha, and A. Anandkumar, "Question Type Guided Attention in Visual Question Answering," in Proc. ECCV, 2018, pp. 158-175.
[54] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel, "Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources", arXiv:1511.06973, 2016.
[55] A. Mishra, S. Shekhar, A. K. Singh and A. Chakraborty, "OCR-VQA: Visual Question Answering by Reading Text in Images," 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 2019, pp. 947-952, doi: 10.1109/ICDAR.2019.00156.
[56] Million, Elizabeth (April 12, 2007). "The Hadamard Product". buzzard.ups.edu.
[57] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a method for automatic evaluation of machine translation", in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02), Association for Computational Linguistics, USA, 311–318, 2002.
[58] Z. Wu and M. Palmer, "Verbs semantics and lexical selection", in Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL '94). Association for Computational Linguistics, USA, 133–138, 1994.
[59] NeuralTalk2-Image Captioning Model Retrieved from: https://github.com/karpathy/neuraltalk2neuraltalk2.
[60] COCO-QA dataset Available At: http://www.cs.toronto.edu/ mren/research/imageqa/data/cocoqa/.
[61] NYU Depth Dataset V2 Available at : https://cs.nyu.edu/ silberman/datasets/nyu depth v2.html.
[62] VQA dataset Available at: https://visualqa.org/download.html
[63] Visual7W dataset Available at: http://ai.stanford.edu/ yukez/visual7w/
[64] Figure QA dataset Available at: https://www.microsoft.com/en-us/research/project/figureqadataset/!download.
[65] TallyQA: Answering Complex Counting Questions Retrieved from: https://www.manojacharya.com/tallyqa
[66] KVQA dataset Available at: http://malllabiisc.github.io/resources/kvqa/