# VQAR: Review on Information Retrieval Techniques based on Computer Vision and Natural Language Processing

**Conference Paper** · March 2019

DOI: 10.1109/ICCMC.2019.8819803



**2 authors:**

Shivangi Modi
Sardar Vallabhbhai National Institute of Technology
**4** PUBLICATIONS **19** CITATIONS

SEE PROFILE

Dhatri Pandya
Sarvajanik College of Engineering and Technology
**7** PUBLICATIONS **51** CITATIONS

SEE PROFILE

Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019)
IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4

# VQAR: Review on Information Retrieval Techniques based on Computer Vision and Natural Language Processing

Ms. ShivangiModi
PG Scholar
Computer Engineering Department
SCET, Surat, India
smodi0123@gmail.com

Prof. DhatriPandya
Assistant Professor
Computer Engineering Department
SCET, Surat, India
dhatri.pandya@scet.ac.in

***Abstract*—Recently Computer vision and Natural language processing paradigm contains enormous research progress in their respective areas. Despite the progress in both areas, still it remains as a challenging task for machines to extract image semantics and then communicate this extracted information with the desired users. These problems will be solved by Visual Question Answering (VQA) system by connecting both computer vision and natural language processing paradigms. In VQA, system is presented with an image and textual question related to that image. The system will generate the answer by processing on both image and textual features. Answer generated by VQA is in one word, phrase or in sentence. Various datasets are available for training and evaluating VQA system which contains real or abstract images and question-answer pairs related to the semantics available in the image. VQA is being used in many areas such as for blind and visually impaired users, robotics, art gallery and many more areas. This paper discusses VQA techniques, VQA datasets and highlights the parametric evaluation of these techniques along with generic issues in VQA system.**

***Keywords*—visual question answering, computer vision, natural language processing, attention model, joint embedding, compositional model, external knowledge base mechanism.**

## 1. INTRODUCTION

Since vision is the core constituent of the intelligence, Computer vision plays a significant role for a computers to receive and analyze visual data on its own and then make decision about images and videos [1][2]. For an example various computer vision application areas such as face identification, object detection has been solved with quite good accuracy [3][4].Computer vision applications have become everywhere with products such as Google's new product Goggles worked with the concept of object detection, Facebook's facial recognition task is also worked with the concept of face identification [5]. Similarly natural language processing field also received a lot of attention due to its wide variety of applications like robotics, language translation, text summarization and many more[1][2]. Despite the progress in both areas, it is still become challenging task for machines to extract semantics information from images and communicate this extracted information to humans [1][3]. As this task requires understanding of an image and communicate this extracted semantic information with machine as a humanin natural language form require vision and language knowledge to infer the correct answer [1-4].This problem will be solved by Visual Question Answering (VQA) techniques [3][4].

VQA is a system which will predict the answer of a given question related to an image [4-6]. This system provides insights into the relationships between two important sources of information from computer vision and natural language processing fields [1-4]. In VQA, system has an image and textual question as an input and generates textual answer as an output [5]. VQA is a challenging task because system need to process both image and textual question and then after system is capable to give answer. Sample example of VQA system is shown in Fig.1.

Q: What is the color of the ball?
A: yellow

Q: How many giraffes?
A: 2

Fig.1. Real Image [12]

VQA has attracted much consideration due to its various usages in real day to day life such for blind or visually impaired users, image retrieval, robotics, art gallery and many more areas [6].For example, blind user captures the image and use VQA system to get knowledge about visual information [15].

There are two different types of assessment format in visual question answering [3][6]. One is open-ended type of evaluation format and another is multi-choice type evaluation format [3][22]. In open-ended type, VQA system is provided with input image and textual question. System will generate

answer by processing on both image and textual question. Input image is either natural scene image or cartoon scene image [25]. In multi-choice type format, system is provided with input image, textual question and answer options. These options will be correct, plausible, popular or random. In open-ended type of visual question answering, questions are organized into primarily twelve different categories such as object presence, subordinate Object recognition, counting, color attributes, other attributes, activity recognition, sport recognition, positional reasoning, scene classification, object utilities, sentiment understanding and absurd [7]. VQA methods are divided into joint embedding, attention, compositional and external knowledge basedmodels which are used for implementing VQA system [5].

This paper aims to make survey of VQA models and techniques. This paper is organized as follows: In section 2, VQA system framework, VQA models and their different VQA techniques are discussed. Section 3 discusses parametric evaluation of VQA techniques. Major issues of existing VQA techniques are presented in section 4.

## 2. Related Work

This section gives a detailed description of various visual question answering models; their techniques and parametric evaluation of various VQA techniques.

### 2.1Framework of VQA System

General framework of the VQA system is shown in Fig. 2. VQA framework includes system input, computer vision task, natural language processing task and answer generation module. Initially VQA system is presented with an image and textual question as an input. Then after through computer vision and natural language processing task system is processing on both input image and textual question and generate the visual and textual representation. After generating both image features and question encoding vector system is combined both output vectors. Generated output vectoris going to appropriate VQA model and VQA model is predict and generate the answer of a given question accordingly semantic present in the input image.

To generate visual representation of an input image different VQA techniques are using different CNN (Convolutional Neural Network) like AlexNet [27], VggNet [24], GoogleNet [28], ResNet [29] etc. CNN takes an image as an input and extract semantic features of an image [23]. Initial layers of CNN is extracting the lower level features of an image such as edges, lines, corners, brightness etc. and later on layers of CNN is extracting the whole object. To generate the textual representation of an input question different VQA techniques are using RNN (Recurrent Neural Network or LSTM (Long Short Term Memory). To encode the features of an input question VQA system requires several natural language processing task such as tokenization [32], word embedding [8][24] etc. Tokenization operation is performed on an input textual question and generates the tokens. These tokens are then passed into word embedding technique. Word embedding converts the tokens or words or texts into numberform. There are various word embedding techniques such as CBOW (Continuous bag of Word) [8], Skip-gram model [8], GloVe [24] and many more. Then after each and every vector of real value numbers are passing into LSTM network and generate the question encoding vector.Generated image feature matrix and question encoded vectors are combined with each other through one of the different operations like concatenation, multiplication, addition etc.

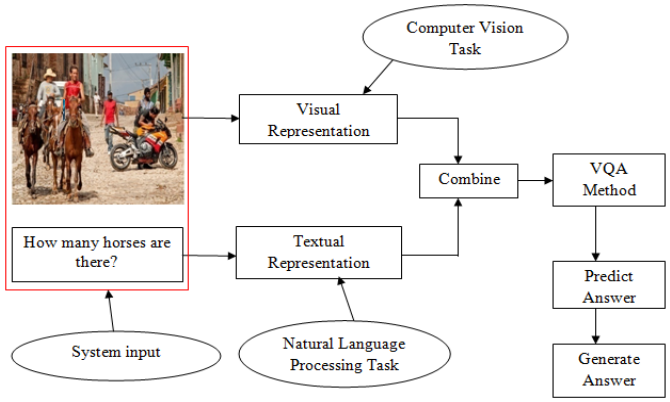


Fig.2. General framework of VQA system [3][4]

### 2.2 VQA Methods

Visual question answering is able to answer free form open-ended type of question through different VQA methods. VQA methods are divided into four models which are shown in Fig.3.

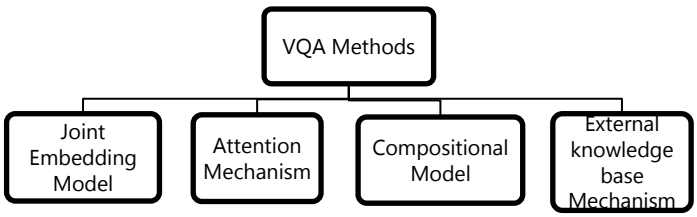


Fig.3. Different VQA Methods [3]

#### 2.2.1 Joint Embedding Model

H. gao et al. [8] proposed joint embedding model for implementing VQA. In this model, image and textual question take as input. Then afterimage and textual question features are to be extracted through different deep learning and NLP techniques. After getting both features now these both feature vectors are jointly embedded into common feature space. Then after this combined feature vector feed into classifier. Finally classifier predicts the answer of a given question. The main part of this model is that it is focuses on the global features of an image. Example of this model is shown in Fig.4. Question asked related to image is like `What's in the background?' and Fig. 4 imagecontains all the global features like two man,

mountain, stick, skateboardand all other small features. For embedding the image and textual question nowadays deep neural network based mechanism is used.

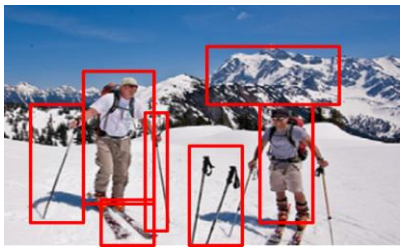

Fig.4. Real image with global features [18]

Q: What's in the background?
A: mountain

### 2.2.2 Attention Mechanism

Z. Yang et al. [9] proposed an attention mechanism which is extension of joint embedding model. In joint embedding model, VQA system is focuses on all the global features of an image rather than focuses on only question specific features of an image. Extracting the global features alone may difficult to understand question specific semantic information of an image. Therefore this limitation of the joint embedding model is solving by the attention mechanism. Attention model will focuses on question specific region of an image rather than all the global featuresof an image. Fig.5 is the example of this model for a VQA system answering is `What is the color of umbrella?' then system will focuses on the question specific umbrella region of an image rather other image regions. There are so many different ways to apply attention. One approach to apply attention to this representation is by suppressing or improving the features at various spatial areas. Utilizing the question features with these local image features like only umbrella region, a weighting factor for each lattice area can be figured that decides the spatial area's significance to the question, which would then be able to be utilized to process attention-weighted image features.

Fig.5. Real image with attention [5]

Q: What is the color of umbrella?
A: red

### 2.2.3 Compositional Model

R. Hu et al. [12] proposed compositionalmodel. This model is useful when questions require multi-step reasoning to answer properly. Example of this model isshown in Fig.6. Visual question is `What is to the left of the dog?' then this model first finding dog, and then after recognize the object which is left of dog. Two compositional systems have been proposed for VQA that attempt to handle solving VQA in a series of sub-steps. First framework is Neural Module Network (NMN) and second structure is Recurrent Answering Units (RAU). The NMNstructure utilizes external question parsers to find the subtask in the question whereas RAU is prepared end-to-end and sub-tasks can be implicitly learned.

Fig.6.Real image [19]

Q: What color is she wearing?
( describe[color]
find[wear] )
A: white

### 2.2.4 External Knowledge base Mechanism

Q. Wu et al. [16] proposed external knowledge base mode. This model is useful when some common sense or additional background knowledge type of questions require some external knowledge sources to answer properly. For example, questions like 'Which transportation way in this image is cheaper than taxi?' can require some external knowledge source to answer. One way of external knowledge source is supporting fact. Supporting fact is like 'bus is cheaper than taxi'. Due to this additional knowledge sources now system is able to answer the question properly. There are so many other ways to provide external knowledge sources to system.

Fig.7. Image with supporting fact [11]

Q: Which transportation way in this image is cheaper than taxi?
**Supporting fact**: bus is cheaper than taxi
A: bus

### 2.3 VQA Techniques

There are so many various techniques are available in visual question answering. These all techniques are based on VQA four methods. In this section we are discussing about all four methods joint embedding, attention based, compositional, external knowledge based and hybrid model based techniques.

### 2.3.1 mQA Model based Technique

H.gao et al. [8] paper presentsmQA model based technique which is based on joint embedding method to answer a

question related to image. In this technique answer would be in single word, phrase or in a sentence. This technique contains four major components. First component is question representation. For that they use LSTM to extract features of question. Second component is visual representation. For that they use CNN to extract visual features. This component is pre-trained on ImageNet classification task. Most important is third component which is another LSTM component. This component main aim is to generate answer in sentence form. This LSTM component encodes current word and previous words of answer intodense representations. The fourth component is fusing component. This component fuses the information from the first three componentsto predict the next word in the answer. mQA model train and evaluate on Freestyle Multilingual ImageQuestion Answering (FM-IQA) dataset. This dataset contain Chinese question answer pairs and their English translations.This dataset have action, object recognition, positions and interactions among objects, recognition based on commonsense and visual content type of questions.

#### 2.3.2 Ask Your Neurons Technique

M. Malinowski et al. [17] paper presents joint embedding based method for implementing visual question answering. They evaluate their approach on DAQUAR and VQA datasets. They used deep learning method CNN for image information encoding. Encoded image information and question together fed into LSTM network. Combination of CNN with an LSTM predicts answers in a multiple words or a single word.

#### 2.3.3 Stacked Attention Network (SAN) Technique

Z. Yang et al. [9] paper presents SAN technique which is based on attention mechanism. This technique performing multi-step reasoning through multiple attention layers for implementing visual question answering. SAN technique consists of three major components. Among all three components, first component of the SAN technique is to extract image features through VGGNet CNN. Second component is question model that uses a CNN or LSTM models for question features extraction. Third component is stacked attention model. This model locates image regions that are relevant to the question for answer prediction through multiple attention layers. Finally combine image features from last attention layer and last query vector to predict the answer. This technique will predict answer in single word.

#### 2.3.4 ABC-CNN Technique

K. Chen et al. [10] paper presents anattention based configurable convolutional neural network technique. This technique is focuses on the information region of an image which is based on question guided attention. This technique contains four major components. First component is image feature extraction part. This component useVGG-19 deep-CNN to extract the image features. Second component is question understanding part. This component use LSTM model to encode question features. Third component is attention extraction part. This component set the convolutional kernels which will map the question features with the image features and to generate question guided attention map. Fourth component is answer generation part. This component use the classifier which is based on question guided attention map and generates answer. Answer would be generated in single word.

#### 2.3.5 Graph based Technique

W. Brown et al. [14] paper presents a graph based approach for implementing visual question answering. This technique is use graph learner module for makingagraph of an input image which is related to question.Due to creating a question specific graph now system is able to answer high level reasoning type of questions such as semantic and spatial representation type of questions. This technique contains four major components. First component is question encoder which embedding the question through word embedding and recurrent neural network. Second component is object detector. This object detector performs on image and gets bounding box coordinates and feature vectors of an image. Third component is graph learner module which combines both image and question feature vectors and generates adjacency matrix of the image objects and a given question. Fourth component is spatial graph convolution which focuses on the objects, object relationship which are relevant to the question. Finally perform max pool and element wise product operations to get the final answer of an input question which is related to input image.

#### 2.3.6 Ask Me Anything (AMA) Technique

Q. Wu et al. [16] paper presents AMA model which handles general knowledge based questions from external sources. They combine an automatically generated description of an image with external knowledge base to provide an answer of a general question answer pairs. There are three major components in this model. First component is extracting, encoding and merging. First task of this component is to predict set of attributes of the image through CNN. Based on attributes image captioning model generate series of captions. Second task of this component is to extract relevant information from DBpedia knowledge base based on top-5 attributes detected in the image. Then that information is encoded into fixed length feature representation through Doc2Vec. Second component is VQA model with multiple inputs. In this component encoded attributes, captions and KB information are taken as single input and fed into LSTM to interpret question and generate answer.

#### 2.3.7 FVQA: Fact based Technique

P. Wang et al. [11] paper presents fact based visual question answering task. This technique is handle the common sense and basic factual knowledge type of questions through explicit

reasoning. In this paper, explicit reasoning is provided through supporting facts. For adding supporting facts with question answer pairs they introduce new dataset with name FVQA dataset. Supporting facts are linked with visual concepts. There are three types of visual concepts: object, scene and action. Knowledge about each visual concept is extracted from knowledge bases which are DBpedia, ConceptNet and WebChild. After constructing knowledge base they perform question-query mapping throughRNN or LSTM. For retrieving the correct supporting fact they used query; which is performed over entire knowledge base. LSTM or SVM classifier is used to selects most relevant supporting fact among list of pairs returned by query.

#### 2.3.8 End-to-End Module Network Technique

R. Hu et al. [12] paper presents an end to end module network without the help of parsers. This model is used to handle compositional reasoning type of visual question answering. To solve compositional reasoning problem, they breaks down problem into set of neural modules. Then after layout policy is implemented with RNN to predict layout expression for each question. This expression is passed to the network builder to dynamically predict instance specific neural network for questions and applies to the input image to get the answer.

#### 2.3.9 R-VQA Technique

P. Lu et al. [13] paper presents a novel semantic attention model framework for visual question answering. This framework is useful to learn visual relation facts as semantic knowledge in images. They build Relational-VQA (R-VQA) dataset which is based on large-scale Visual Genome dataset. Each data instance in the R-VQA dataset composed of image, question, relation fact and answer. This technique consists of three major components. First component is context aware visual attention module. This component is used to extract image feature representation. Second component is fact-aware semantic attention module. In this module, output of the first component is fed into second component to select related relation facts. These relation facts are generated by relation detector from image and question. Third Component is joint knowledge embedding learning. This component simultaneously merges final visual and semantic attention representation to learn visual and semantic knowledge.

## 3. PARAMETRIC EVALUATION OF VQA TECHNIQUES

This section shows comparison of previously discussed VQA techniques. Table-1 shows comparison of VQA techniques based on parameters are as follows. VQA models parameter indicates that technique use which model to predict the answer of a given question related to image. Datasets parameter used to determine that different VQA technique used which dataset for training and evaluation purpose. Types of questions parameter represent that which types of questions handled by that particular technique. Key characteristics parameter shows main features of the techniques and scope of improvement parameter indicate the improvement or future work of technique.

TABLE-I: PARAMETRIC EVALUATION OF VQA TECHNIQUES

| Technique-Year | VQA model | Datasets | Type of questions | Image feature extraction | Question feature extraction | Image input size | Scope of improvement |
|---|---|---|---|---|---|---|---|
| Are You Talking to a Machine? [8] - 2015 | Joint embedding model | FM-IQA [8] | Action, object recognition, positions, recognition based on commonsense | CNN-GoogleNet [28] | LSTM [30] | 224*224 | - Small object detection |
| Ask Your Neurons [17] - 2015 | Joint embedding model | DAQUAR [3][4] | Identifying object, color and count type of questions for only indoor scene images | CNN [26] | LSTM [30] | 227*227 | - Small objects recognition<br>- Handle spatial reasoning type of questions |
| Stacked Attention Networks (SAN) [9]-2016 | Attention mechanism | DAQUAR, COCO-QA and VQA [3][4] | Object, color, count and location | CNN-VGGNet [24] | CNN / LSTM [24][30] | 224*224 | - Predict answer in sentence |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ABC-CNN [10] -2016 | Attention mechanism | DAQUAR, Toronto COCO-QA and VQA [3] | Object, color, count and location | CNN- VGG-19 [24] | LSTM [30] | 224*224 | - Predict answer in sentence |
| Learning Conditioned Graph Structures [14]-2018 | Attention mechanism | VQAv2 [4][6] | Object, color, count | Object detector | RNN | 227*227 | - Time specific answer<br>- Improve performance of "Number" questions |
| Ask Me Anything [16]-2016 | External knowledge base model | DBpedia knowledge base [16] | Scene or Knowledge based reasoning | CNN-VGGNet-16 [24] | LSTM [30] | 224*224 | - Precise image caption model |
| FVQA [11]-2017 | External knowledge base model | FVQA [3][11] | Object, scene and action | Fast R-CNN | RNN / LSTM [30] | 224*224 | - Handle more numbers of visual reasoning type of questions |
| Learning to Reason: End-to-End Module Networks [12]-2018 | Attention + Compositional model | CLEVR [21] | Object recognition, position, shape | Heuristic rule-based semantic parsing | Stanford Dependency parser | 224*224 | - Predict answer in multiple word |
| R-VQA [13]-2018 | Attention + External knowledge base model | R-VQA [13] | Object, color, activity, position, scene, relation, commonsense | CNN - ResNet-152 [29] | LSTM [30] | 224*224 | -Predict answer in multiple word |

## 4. MAJOR ISSUES

Visual question answering system contains many open research issues due to its wide variety of applications and its broader area of research. After studding all the visual question answering methods and techniques we have identified certain issues in visual question answering.

**A. Single word answer [9][10]:** Recent visual question answering system generates the answer in single word. Few of the VQA techniques generate the answer in multiple words. However these few techniques are not generating answer in sentence form or in proper human understandable form. Some of the examples which fall in this issue are shown below. In below first example, some of the cows are in black and some of the cows are in brown color. However the current VQA system generates only 'brown' answer. Proper human understandable answer is like 'some cows are in brown and some cows are in black color'.


Fig.8.Different types of cows and sheep [20]

Q: What is the color of cow? Q: What is different between two sheep?
Generated A: brown Generated A: black

**B. Time specific answer [14]:** Recent visual question answering techniques are enables to answer the image specific questions like 'What time is it?', 'What time does the clockshow?' and many more time specific questions. Some of the examples of which fall in this issue are shown below figure.


Fig.9.Evening scene and Clock image [12]

Q: What time is it?
Predicted answer: evening

Q: What time does the clock show?
Predicted answer: 8:26

**C. Number specific visual question answer [14]:** Recent visual question answering techniques are enables to answer the image specific questions like 'What is the number of table?', 'What is the bus number?' and many more number specific visual questions. Some of the examples of which fall in this issue are shown below.

Fig.10.Real images [14]

Q: What is the number of bus? Q: What is the number of table?

Wrong A: 23 Wrong A: 4

**D. Commonsense reasoning [16]:** Existing VQA system is not able to handle all types of commonsense reasoning type of questions like 'is this child is boy or girl?', 'What is the age of the man?'

**E. Handle limited number of knowledge base reasoning type of questions [11]:** Existing VQA system is handle knowledge base reasoning type of questions with the help of additional and background knowledge provided to the system in the form of external sources or supporting facts. Therefore they are handling only those questions whose background information is given.

**F. Not handle too small object detection and recognition [8][17]:** Existing VQA system is able to answer object detection and also identified those objects. However object is too small at that time existing system is not able to detect and identified those objects.

## 5. Conclusion and Future Work

In this paper we studied about Visual Question Answering (VQA) methods and their different techniques for implementing VQA. There are primarily four approaches available in VQA. These four approaches are Joint embedding, Attention based, Compositional and External knowledge based methods. We studied these four approaches and their techniques in detail. While studying these all techniques we have identified several issues in existing visual question answering system. These all techniques are based on computer vision and natural language processing domains. These all techniques are used to predict the answer of a given question related to an image. VQA techniques are useful in many applications such as for blind or visually impaired users, interact with robot, providing information to a spectator at an art gallery and many more.

The major challenge in visual question answering system is to develop a more efficient technique which will predictthe answer in subjective form and handle time specific question answer pairs.

## References

[1] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]," in IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48-57, May 2014.

[2] R. S. Dudhabaware and M. S. Madankar, "Review on natural language processing tasks for text documents," Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference, pp. 1-5, 2014.

[3] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick and A. van den Hengel, "Visual question answering: A survey of methods and datasets", Computer Vision and Image Understanding, vol. 163, pp. 21-40, 2017.

[4] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges", Computer Vision and Image Understanding, vol. 163, pp. 3-20, 2017.

[5] S Antol, A Agrawal, J Lu and M Mitchell, "Vqa: Visual question answering" In IEEE International Conference on Computer Vision (ICCV), pp. 2425-2433, 2015.

[6] AAgrawal, J Lu, S Antol, M Mitchell and CL Zitnick, "Vqa: Visual question answering", In International Journal of Computer Vision, vol. 123, pp. 4-31, 2017.

[7] K Kafle and C Kanan, "An Analysis of Visual Question Answering Algorithms," In Computer Vision and Pattern Recognition, 2017.

[8] H Gao, J Mao, J Zhou and Z Huang -"Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering", Computer Vision and Pattern Recognition, 2015.

[9] Z Yang, X He, J Gao, L Deng and A Smola -"Stacked Attention Networks for Image Question Answering", Computer Vision and Pattern Recognition, 2016.

[10] K Chen, J Wang, LC Chen, H Gao and W Xu -"ABC-CNN: An attention based convolutional neural network for visual question answering," Computer Vision and Pattern Recognition, 2016.

[11] P. Wang, Q. Wu, C. Shen, A. Dick and A. Hengel, "FVQA: Fact-based Visual Question Answering", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-1, 2017.

[12] S Aditya, Y Yang and C Baral – "Explicit Reasoning over End-to-End Neural Architectures for Visual Question Answering", Computer Vision and Pattern Recognition, 2018.

[13] P Lu, L Ji, B Li, W Zhang and N Duan -"R-VQA: Learning Visual Relation Facts with Semantic Attention for Visual Question Answering", Computer Vision and Pattern Recognition, 2018.

[14] W Norcliffe-Brown, E Vafeais and S Parisot -"Learning Conditioned Graph Structures for Interpretable Visual Question Answering," Computer Vision and Pattern Recognition, 2018.

[15] R Hu, J Andreas and M Rohrbach - "Learning to reason: End-to-end module networks for visual question answering", Computer Vision Foundation, 2017.

[16] Q Wu, P Wang, C Shen and A Dick -"Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources", Computer Vision and Pattern Recognition, 2016.

[17] M Malinowski, M Rohrbach and M Fritz –"Ask Your Neurons: A Neural-based Approach to Answering Questions about Images", Computer Vision and Pattern Recognition, 2015.

[18]Machine Learning [Online], Available: https://machinelearningmastery.com/how-to-caption-photos-with-deep-learning/, [Accessed on 30th November 2018].

[19] J Andreas, M Rohrbach, T Darrell and D Klein - "Learning to compose neural networks for question answering", Computer Vision and Pattern Recognition, 2016.

[20]R Krishna, Y Zhu, O Groth and J Johnson - "Visual Genome: Connecting Language and Vision Using Crowd sourced Dense Image Annotations", Computer Vision and Pattern Recognition, 2016.

[21] J Johnson and B Hariharan– "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning", Computer Vision Foundation, 2017.

[22] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. Zitnick, D. Batra and D. Parikh-"VQA: Visual Question Answering",Computation and Language (cs.CL),2016.

[23] Convolution Neural Networks for Visual Recognition [Online], Available: http://cs231n.github.io/classification/, [Accessed on 30th November 2018].

[24]K. Simonyan and A. Zisserman–"Very Deep Convolutional Networks for Large-Scale Image Recognition", Computer Vision and Pattern Recognition (cs.CV), 2014.

[25] Visual Question Answering [Online], Available: http://visualqa.org/evaluation.html, [Accessed on 18th October 2018].

[26] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould and L. Zhang -"Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering", Computer Vision and Pattern Recognition (cs.CV), 2017.

[27]A. Krizhevsky, I. Sutskever and Geoffrey E. Hinton-"ImageNet Classification with Deep Convolutional Neural Networks", ACM Digital Library, 2012.

[28] C. Szegedy, V.Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna-"Rethinking the Inception Architecture for Computer Vision", Computer Vision and Pattern Recognition (cs.CV), 2015.

[29] K.He, Xi.Zhang, S. Ren and J. Sun -"Deep Residual Learning for Image Recognition", Computer Vision and Pattern Recognition (cs.CV), 2015.

[30] Sundermeyer, M., Schlüter, R., and H. Ney – "LSTM neural networks for language modeling", In Thirteenth Annual Conference of the International Speech Communication Association, 2012.