AN ANALYSIS ON LEARNING OF VISUAL QUESTION ANSWERING USING MULTI-MEDIA COMPREHENSION ALGORITHM (MMCQA) IN NATURAL LANGUAGE PROCESSING

Dr S Venkata Lakshmi[1], M Therasa[2],Karthik Elangovan[3] , S Sharanyaa[4]

[1]Professor, Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, India.
[2]Associate Professor, Department of Computer Science and Engineering,Panimalar Institute of Technology, Chennai, India.
[3]Assistant Professor, Department of Computer Science and Engineering,Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, India.
[4]Assistant Professor, Department of Information and Technology,Panimalar Engineering College, Chennai, India[4].

Email:venkatalakshmis@skcet.ac.in[1], therasamic@gmail.com[2], auphdpapers@gmail.com[3], rnsharanyaa@gmail.com[4]

Abstract. Gaining knowledge of Vision and language is becoming a happening topic with in-depth research in Artificial Intelligence (AI). This AI, NLP, and computer vision have recently hit by issues like Graphical Question Answering (GQA). Here, we present the mission of the Multimedia Machine Modal Comprehension Question Answering Algorithm (MMCQA), focusing on addressing multimodal queries concerning words, figures, and pictures. Dataset addresses the questions and has around twelve thousand and odd lessons and more than thirty-six thousand multi-modal questions obtained from the science curriculum. Our study demonstrates that a significant part of queries need resolving of texts, figures, and cognitive analysis, denoting that the info in our report is way more complicated than the earlier studies and obvious question answering datasets. Lastly, we put forth a method based on dual-LSTM having spatial as well as temporal focus and prove that it is useful compared to other standard GQA methodologies via experiential studies.

1. Introduction

The Natural Language Processing (NLP) community is mainly concentrating on Question answering (QA) for a decade or more, and lately, it has gained acceptance in the computer vision community too.NLP [1] has numerous QA samples that could be grouped based on the information with which the queries addressed, and this consists of both structured and confined databases (e.g., Freebase) as well as unstructured and unbounded natural language form (NLF) (e.g., documents). A model somewhere in-between the above said types are being famous among the Machine Comprehension (MC), and here the data is not structured with limited paragraph size.QA in the vision community, also called GQA, is entirely in demand because of having a massive database of image-based QA. One way, GQA comes under the machine comprehension task, and here the queries are in NLF for the figures.

There is a huge demand for AI [2]. And its subordinates chat bots and robots are being in use for specific jobs. In such circumstances, GQA, on providing a picture, concentrates on addressing the queries and offering the best method for interacting efficiently with AI agents. The solutions for the questions raised in standard GQA given purely based on the image content only. One such being a GK related, e.g., the microphone that amplifies sound and here the former is a collective noun few such types of GQA given in Figure 1.

Figure 1.Sample of GQA vs. TQA

Graphical QA

How many images of cycling of man? 5 Times

Figure 2. Focus of GQA tasks on image reasoning

Image GQA

What is the color of Cat? White

Yet, in reality, we are required to fulfill our info requirements on the named components of the picture. Refer figure 2, for example,), the image in Figure 1, addressing similar queries needs global awareness on the named parts available in the image and also explore it further to look at this issue as GQA.

The issues interconnect two areas: Information Retrieval (IR) and NLP. The former is enhanced by fitting in NLP activities in a large scale-like, not depending on the field, thereby essentially possessing a grand exposure of language. This addition permits the choice of related paragraphs using language traits syntactic or even semantic level [14].

Global knowledge is naturally multi-faceted, extended across text documents, images, and videos. An AI device capable of addressing random queries on the universe has to update itself to understand such multi-faceted data resources. Hence, we put forth the idea of Multimedia Machine Modal Comprehension (MMC) [3], an extended part of the standard textual machine conception to multi-aspect information. In this model, it works to understand multiple aspects with a multifaceted query and naturally provide a solution with multiple features. But this contradicts the standard relying method where the plot is generally on a single facet (either language or vision).

For achieving the aim, we introduce the GQA dataset taken from the science curriculum (Figure 2). The textual and pictorial matter in middle school science, a slightly intricate incident was happening globally. GQA is a better place for experimenting with the MMCQA. It has around 1,076 lessons with78, 338 sentences, and 3,455 images. Every experience was having a bunch of queries that are addressable with the matter available in the syllabus. This GQA dataset contains26, 260 questions out of which12, 567 come with a picture and classified as training, validating, and testing at a lesson level.

In brief, our principal inputs have given below:

1. We present fresh ideas created exclusively for GQA that need spatial-temporal reasoning from pictures and videos for providing correct solutions.

2. We present a method based on a dual-LSTM [15] using a scheme focusing on resolving our GQA tasks.

3. We create awareness of the essential yet not much-researched issue of the GQA relating identified objects in a picture. Addressing queries based on them needs info on the universe and analyzing it, so we call this as knowledge-aware GQA.

4. We even offer activities of recent methodologies if used on the GQA dataset. With confidence, we say that GQA would lead to the emergence of different novel study are related to Computer Vision, NLP, and, more broadly, AI.

2. Related Works

QA systems have witnessed an enormous transition in the last few decades, which is unmatchable compared to the entire NLP. Here, we present earlier research on developing a QA system and its goal. The previous tool was introduced in 1959 (during The Conversation Machine), and numerous QA system was put forth from the 1960s, and the best among them was BASEBALL developed by [4].

A novel QA system with a unique methodology with a List queries for asking on various situations of a particular type of date.In2012 a QA Systems developed that got adapted to toil with thoughts; instead, the realities could add too. In 2014, Cabrio et al. developed QAKIS and FREITAS14 by Freit and Curry and created three QA system at Dima, INTUI3 in 2015, originally by HAKIMOV15 [5].

START QA system is the pioneer with respect to Web-based Question Answer system for Englishlanguage and during the same time came the Answer bus QA system(ODQA) which was developed for accepting queries in different languages (include the process of extracting answers from the local database to the WWW, allowing you to work with a larger number of queries). The following year, they introduced the QA system (ARANEA) [6] it was the first open-source web-based Factoid QA system that could be fully downloaded, and soon they developed a QA system AQUA, a classic automated QA system that specialized in natural language comprehension, physics knowledge, logical reasoning skills and advanced knowledge extraction techniques.

The integration of image representation with additional data obtained for GK based was proposed on the basis of image prediction for VQA. It enables addressing queries outside the picture, but the obtained info is parts of the plain text with no structural representations.

Author [7] used a clear understanding of their source description framework knowledge base for getting the solutions. Yet, it hugely dependent on the pre-set templates that limits its usage. “Fact-based VQA (FVQA) [8]”has proposed a methodology based on semantic analysis to support the facts retrieval. A similar score calculated for getting the best result. It is highly susceptible to misinterpretations due to meanings and homographs. An understanding-based method later formed in for FVQA, which is used for parameters mapping of realities to an embedding space and for understanding the image of the question pair, allowing its performance and applicability to be evaluated. So, Concatenation of featureson image-question-answer-facts recordsare taken into account [9].

3. Proposed MMCQA Models

Our methodology based on the latest method called MMCQA for the TQA problem. Here, initially, we brief MMCQA and its components that were employed in our MMCQA models. Followed by the introduction with respect to two categories of models for handling the open-ended MMCQA and the multiple-choicepattern MMCQA individually.

3.1. *Multiple-choice MMCQA model*

Multiple-choice MMCQA offers various pre-defined solution options apart from the figure and queries. The algorithm directed to choose the best matching answer from the given options. This can be directly corrected using the open-ended MMCQA model mentioned above by guessing thepossible answer thatmatches with the given exams. Yet, this method as no complete lead over the given data. Based on [10], that get the answer as input as well and show considerable development in output, we put forth a different model for multiple-choice MMCQA problem.

As shown in as presented in Figure 3, apart from the queries and the transformed image details, our methodology, too, receives a candid answer as input and computes the interface amid the answer and the scenario. With factual answer, the programmed features v0a and v1 are highly related to v0q and v1q. Or else, the structures may not be dependent. A MMCQA is accomplished on the concatenated facial appearance, i.e. Exp = X2Max (0, X1 [p0q; p1q; p0q; p1q]). After the first layer, we use the dropout with 0.5 probability value. The goal is to find out if the picture-question-answer is triple valid. Therefore, we follow the sigmoid function to modify the feature possible- the loss of binary logistics used to train the model.

In comparison, the open-ended MMCQA which selects the best answers as the best labels and rejects the rarest answers, directly encrypts the multiple choice MMCQA model candidates' answers – thereby covering the maximum number of options.

For similar type of answering expressions, such as “daytime”, sample knowledge can be obtained and informationfrom the similarityobserved by including an encoder, and not employing the experimental chooses. Besides, it evades the possibility of acknowledging them as a separate division and learns to differentiate them from the practice data [11].

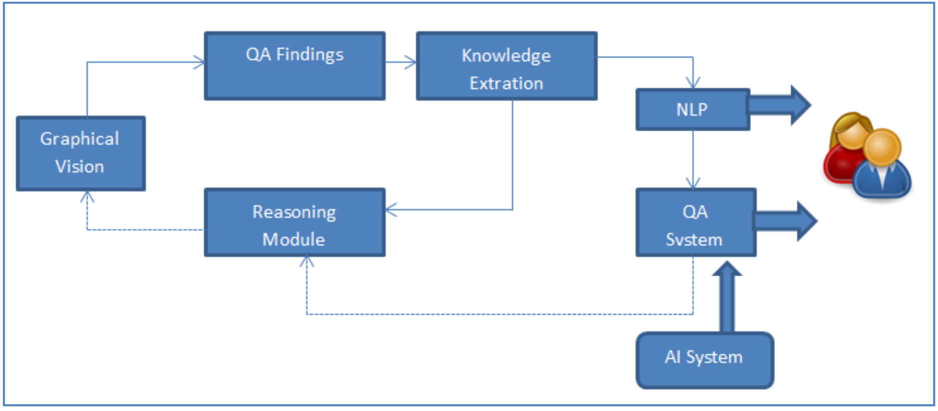


Figure 3– Multiple-choice VQA model

3.2. MMCQADataset

Our dataset has around 103,919 pair of QAs gathered from56, 720 animated GIFs [12]. We detail our latest jobs created for GQA [13] and show the info gathering procedure.

3.2.1. Task Definition

We present the types of the task types that work here. Of these, three are innovation and different for the video domain, including:

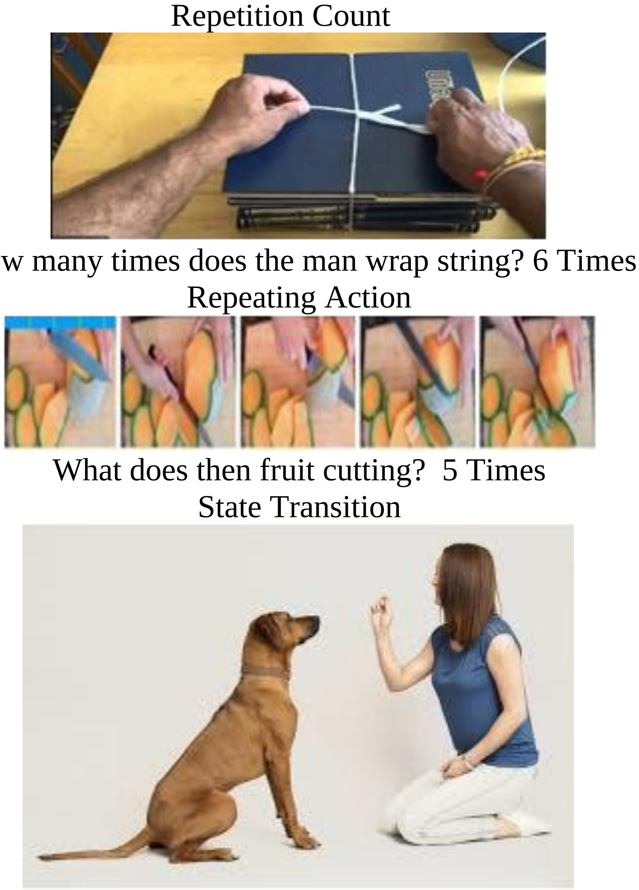
Number of Repetitions:this is really different from videos that calculate repetitive actions. We call this the open question of counting the steps recited.

Repeating Action: It went with the above mention task and called a multiple-choice question that finds an activity that is happening in a video.

State Transition: One more distinctive feature to videos is enquiring about transitions of particular states, which include facial expressions, actions, places, and functions of an object.

The tasks mentioned above need to study the several frames of a video (figure 4); we suggest all of them together by GQA. Apart from our GQA tasks, we recommend frame QA for highlighting the fact that the queries here could be addressed from one of the video frames. Based on the content available in the video, it could be a random frame or any specific frame of a video. To this end, we rely heavily on video titles given in the MMCQA database, and use the NLP – based method recommended for automatic generation QA pairs from titles. The query is called the open question about finding the best option given in the whole sentence (table 1).

Video QA Test



Repetition Count

How many times does the man wrap string? 6 Times

Repeating Action

What does then fruit cutting? 5 Times

State Transition

What does the dog on the left do? Sitting

Figure 4. Our MMC dataset tasks

Table 1. Templates used for creating video QA pairs

Task	Question	Answer
Repetition count	How many times does the	[#Repeat]
Repeating action	What does the [SUB] do [#Repeat] times?	[VERB] [OBJ]
State Transition	What does the [SUB] do before [Next state]?	[Previous State]
	What does the [SUB] do after [Previous state]?	[Next state]

3.2.2. The algorithm I: Generative Model for Images

```
Process Image Group (Colors)
Objects ←Empty array list
No. of objects=Length (colors)
For i in 0:No. Objects Do
While True Do
```

```
Xj~Unrestricted (0, Size of Image)
Yj~Unrestricted (0, Size of Image)
If (Xj, Yj) True Then
Break
Colorj=Image Colorsj
Add Object (Image Color ID=i, Image Position Center= (xi, yi), Image Shape=shapei) to objects
Image=Reduce (Objects)
Return   Image
```

3.2.3. Algorithm II: Generative Model for Questions

```
Process Question Group (Colors)
Cj=Unconditional (Length (Colors))
Tj=Unconditional (2)
STj=Unconditional (3)
TMPj=Unconditional (2)
Question =Question (Type=Tj, Subtype=STj, Template=TMPj, Color=Ci)
Return Question
```

4. Experiment Setup

4.1. Text Questions

Figure 5(a) (b) (c) illustrates the allocation of the query length in the dataset. It denotes that this one has lengthier queries than the GQA and GQA. Also, the spread of queries belonging to the entire W categories (what, where, when, who, why, how, and which) has shown. Fascinatingly, the later carried a decent count of queries. Additional studies show that a decent part of the queries marked down in the routine notepads is proofing in contrary to the interrogative statements. This is also one more reason behind the poor performance of baseline models in Section 5.

4.2. Diagram Questions

The figures in the queries of the QA database match the values in the queries of the MMCQA database, which is about the subject and problem. We present with the help of diagrammatic explanation using graphs for representation and also a hierarchical image of components and associations. We studied MMCQA and understood that there is perfect similarity amid the intricacy of the figure and the number of text boxes situated in the figure. The allocation of the text boxes for all the figures in queries of the QA dataset as a representative for the distribution of diagram intricacy is shown in Figure 5. That illustrates that the figures in the questions are intricate, and additional studies show that an in-depth analysis of them frequently needs to be done to address the queries.

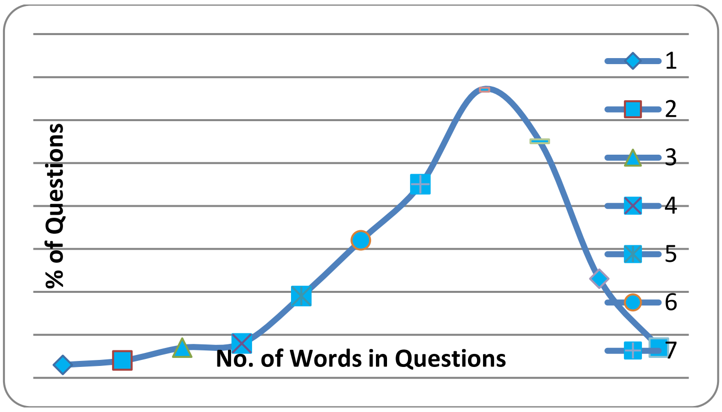


Figure 4. (a)

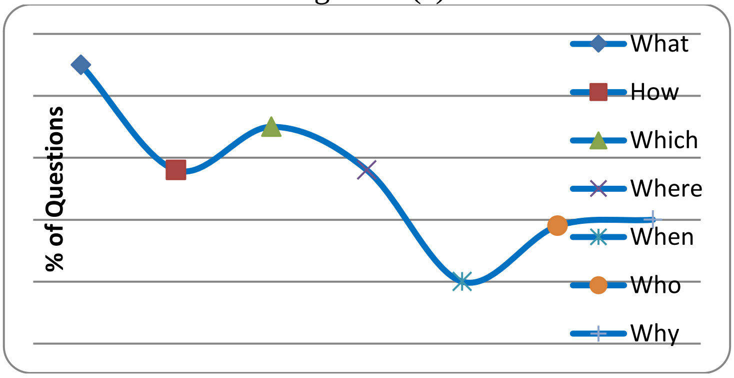


Figure 4. (b)

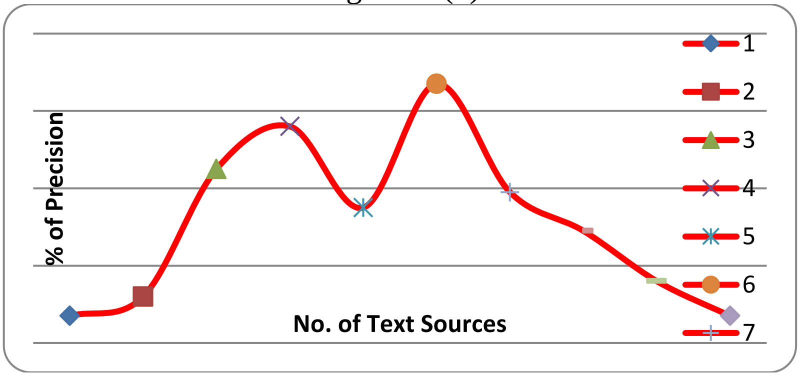


Figure 4 (c).

Figure 4.An analysis of questions in the GQA dataset

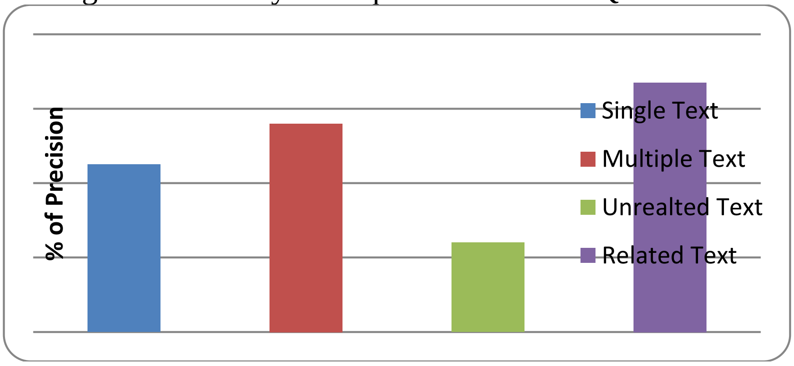
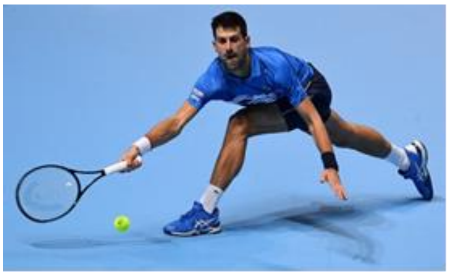


Figure 5. An analysis of GQA dataset

4.3. Results Analysis of MMCQA

For evaluating the precision of the model, we use the below method — the resulting forecast calculated to be exact when the string corresponds to the particular fact: the top-1 and top-3 precisions determined for every analyzed methodology. The mean result precision for all the five test splits is shown here as the total precision. Table 2 and Table 3show the complete accuracy of this methodology according to the supportive realities obtained using the outputs. Our method with fine-tuned MMC has got the highest top-1 precision, which is 0.7% more than the standard methods' o/p. Point to be noted that it has the top-3 Q-mapping accuracy of 91.97% that is 9% more than what we used. The Q mapping outputs have a straight impact on getting pertinent supportive data. With the above outputs, our approach beats the method more than6% on top- 1 and top-3 output precisions, and also shows improved performance compared to the ensemble method. The development of Q-mapping has been put forward for future research as we aim to suggest an improved approach to GQA issues by taking into account the complete natural language data given and resolving GQA as a read Comprehensible type.

4.3.1. Image Successful and failure cases of MMCQA dataset

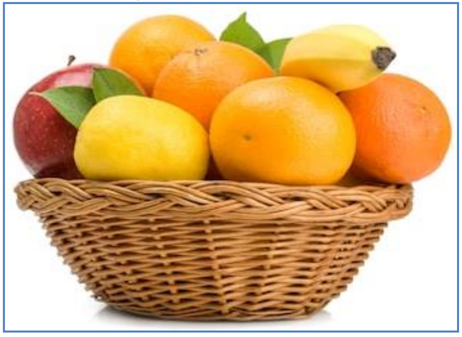


Question 1: In this image, which object is round in shape? (Which is the circular object in the picture?

Answer 1: A man is playing the game tennis. He holds the rocket with the tennis bat. The rocket is seeming to round in shape. A tennis ball is often yellow in colour. Tennis balls are spherical in shape. Tennis balls are moreover hollow.

Predicted Answer 1:Tennis ball

Resulting Answer 1: Tennis ball



Question 2:What sort of food can you see in this image? (What food items displayed in this picture?)

Answer 2: A bunch of fruits placed on the bowl with yellow banana and red apples. Fruit belongs to the food class. So, Apple, pear, banana belongs to the food class.

Predicted Answer 2: Apple

Resulting Answer 2: Fruits

Table 2 – The effect of maximum paragraph length basedon performance.

Max Length (Words)	**Overall Accuracy (%)**	**Inference Time (GQA /Sec)**	**Training time (GQA /Sec)**
300	42.19	156	48
400	43.18	145	47
500	43.19	135	39
600	44.48	125	29
1100	45.04	101	15

Table 3 – Experimental Results on MMCQA-GQA

Method	**What**	**Where**	**When**	**Who**	**Why**	**How**	**Overall**
Existing VGG Method	45.34	18.19	52.07	37.18	12.34	34.56	36.43
Proposed MMCQA –GQP Method	36.45	17.91	56.89	49.19	15.68	45.91	34.19

5. Conclusion

Here, we have tried to resolve GQA in the machine reading comprehension perception. On the contrary, analyzing unclear data from the picture, we suggest to characterizing picture data using natural language and alter GQA to word-based QA. This article presents a new work of MMCQA, which is an augmentation of MC and GQA. We introduce the GQA dataset as a platform for testing the MMCQA task. It contains around a thousand lessons with more than twenty-six thousand multi-faceted queries. Our analysis illustrates that the additions of the latest methodologies for MMCQA and GQA work so weak in this dataset, validating the tasks presented here. Later, we plan to design systems for addressing the MMCQA task in the TQA dataset. On suggesting an alternate method, we evade the joint-embedding of multi-faceted structures in a concealed space. By running multi-faceted fusion in the word-based combination field, semantic data is well-maintained and hence is highly precious for VQA. The outputs show that in spite of showing development in face recognition and QA, substantial additional studies were required for attaining excellent results on knowledge-aware VQA.

References

[1] Yash Srivastava et al, “Visual Question Answering using Deep Learning: A Survey and Performance Analysis”, arXiv:1909.01860, 2019.

[2] Peng Gao et al, “Question-Guided Hybrid Convolution for Visual Question Answering”, Proceedings of the European Conference on Computer Vision (ECCV), pp. 469-485, 2018.

[3] Zhou Yu et al, “Deep Modular Co-Attention Networks for Visual Question Answering”, Proceedings of the IEEE/CVF Conferenece on Computer Vision and Pattern Recogniton (CVPR), pp. 6281-6290, 2019.

[4] Huaizu Jiang et al, “In Defense of Grid Features for Visual Question Answering”, Proceedings of the IEEE/CVF Conferenece on Computer Vision and Pattern Recogniton (CVPR), pp. 10267-10276, 2020.

[5] Vatsal Goel, Mohit Chandak, Ashish Anand, PrithWijit Guha, “IQ-VQA:Intelligent

Visual Question Answering", arXiv:2007.04422, July 2020.

[6] Pranita P. Deshmukh, Rani S. Lande, "Convolution Neural Network based Review System for Automatic Past Event Search using Visual Question Answering", International Conferenece on Invention Computer Technology (ICICT), June 2020.

[7] Sudha Jha, Anirban Dey, Raghvendra Kumar, Vijendra Kumar-Solanki, "A Novel Approach on Visual Question Answering by Parameter Prediction Using Faster Region Based Convolutional Neural Network", IJIAMI, 2019.

[8] Zhou Zhao et al, "Multi-Turn Video Question Answering via Hierarchical Attention Context Reinforced Networks", IEEE Transaction on Image Processing, Vol. 28, Issue. 8, Aug 2019.

[9] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, Anton van den Hangel, "Image Captioning and Visual Based on Attributes and External Knowledge", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 40, Issue. 6, June2018.

[10] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, Anton van den Hangel, "Visual Question Answering: A Survey of methods and datasets", Computer Vision and Image Understanding, Vol. 163, October 2017.

[11] Yan Zhang, Jonathon Hare, Adam Prugel-Bennett, "Learning to Count Objects in Natural Image for Visual Question Answering", arXiv:1802.05766, Feb 2018.

[12] Kushal Kafle, Mohammed Yousefhussien, Christopher kanan, "Data Agumentation for Visual Question Answering", Proceedings of 10th International Natural Language Generation Conference, 2017.

[13] Peng Wang et al, "The VQA-Machine: Learning How to Use Existing Vision Algorithm to Answer New Questions", Peoceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1173-1182, 2017.

[14] Kushal Kafle, Christopher kanan, "An Analysis of Visual Question Answering Algorithms", Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1965-1973, 2017.

[15] Akshaya Kumar Gupta, "Survey of Visual Question Answering: Datasets and Techniques", arXiv:1705.03865, 2017.