**Logical based Visual Question Answering (VQA) Model for Image Classification**

Dissertation Submitted by

**Mussa Ibn Ashraf Al Arafi**

**Student Id: 209139176**

Under Supervision of

**Dr. Swathi Ganesan**

In Partial Fulfilment for the Degree of

**Computer Science (MSc)**



# Declaration

I, Mussa Ibn Ashraf Al Arafi, declare that this description which has been submitted for assessment work by myself, this report has been written in my personal creative words and no portion of it has been formally submitted for a few other assessments. Slightly use of other author's work, whether that be quotes, concepts, images, figures, etc. Properly acknowledged that and itemized in the references section at the finish of this report.

# Acknowledgments

I would like to thanks my supervisor Dr. Swathi Ganesan for being so cooperative through the successful journey. The expert advice and direction remain precious for all the phases of this study. I would also wish to express my gratitude my supervisor for his comments and suggestions, which makes this study possible.

A very humble thanks to all indirectly supporting partner of my thesis i.e., my parents. I take this moment to thanks them for their encouragement and continuous support.

The thesis has been completed and written during my stay at Computer Science Department of the York St. John University. I would like thanks all the university stakeholders like librarian, lab engineers for their support to happen this moment.

09.12.2022

Mussa Al Arafi

# **Abstract**

Artificial Intelligence is a cutting-edge scheme to introduce domain like Computer Vision (CV) to perform computation on visualize information. Although, Natural Language Process (NLP) has also performing at its best when it comes process text related information from enormous amount of semantic data. Both domain (CV and NLP). These domains reaped a huge interest from the deep learning. We system gets some textual based questions about the images. Then the system has to compute the results from it. Amalgamation of deep learning and computer vision along with neural network makes humans capable of doing classification of the images. Being humans, we can understand the information exist in images i.e., objects, attributes, position of the object, relationship between the object etc. further, system encompasses the object recognition, object detection, attributes classification, scenes classification. Above all, scenarios can be complex if the nature of the question becomes spatial. VQA work with variety of applications. One of the most widely used domain for VQA is to facilitate the visually impaired and blind people. The most frequently used is to aid the blind and visually impaired people. Those people capable of getting information contain the image because a picture equal to thousands of words. VQA can facilitate in image recovery process without any meta information about the image. VQA system has several problems and captioning the task is one of them. The ideal technique is to have human judges evaluate the data, but this is time-consuming and costly. As a result, a number of automated evaluation techniques have been presented. BLEU, ROUGE, METEOR, and CIDEr are the most extensively utilized caption assessment techniques. Every caption related assessment metric were established foe evaluation purpose through machine translation through the exception for CIDER, that is generally developed to assess the description of the image. These kinds of indexes have their own certain group of interlined limitation. One of the most frequently use metric to assess image description was BLEUE i.e., referred to have similar kind of score for important alteration in different sentences from largely different kinds of semantics. In BLEU ratings for captions generated, machine captions were rated higher than human descriptions. The study proposed a VQA algorithm for larger dataset of images. Algorithm will evaluate the images in nuanced manner. The main ingredient of algorithm is that it will infer the results upon different mathematical rule of inferences. Then we Empirically investigate the algorithm in particularly scenario where it will aid the visual impaired person. During the research dataset will be used from (www.visualqa.org), and proposed will be applied to compute the results.

**Keywords:**

Visual Question Answering, Machine Learning, Image Classification.

# Table of Content

[Declaration 2](#_gjdgxs)

[Acknowledgment 2](#_30j0zll)

[Abstract 3](#_1fob9te)

[Table of Content 5](#_3znysh7)

[Chapter 1 Introduction 6](#_2et92p0)

[Chapter 2 Literature Review 16](#_tyjcwt)

[2.1](#_3dy6vkm) Image Classification Model 17

[Chapter 3 Methods 26](#_1t3h5sf)

[3.1](#_4d34og8) Research Objective 27

[3.2](#_2s8eyo1) Research Methodology 28

[Chapter 4 Results, data, findings 28](#_17dp8vu)

[4.1. Installation of Dataset 28](#_3rdcrjn)

[4.2. Making the dataset ready 29](#_26in1rg)

[4.3. Creating the Model 29](#_lnxbz9)

[4.4. Training Model 30](#_35nkun2)

[4.5. Analysis of Experiment 31](#_1ksv4uv)

[Chapter 6 Conclusion and recommendations 33](#_44sinio)

[Chapter 7 Bibliography 36](#_2jxsxqh)

[Appendices 38](#_z337ya)

# Chapter 1 Introduction

Answering a natural language query about a picture is what visual question answering (VQA) is all about. Many obstacles exist in VQA, including different technology-oriented language representation with different grounding. It also helps in recognition of semantics from the context and common-sense reasoning. VQA was able to perform specialized tasks i.e., calculating, reading and counting from given context. (Shih, Singh and Hoiem, 2016). As the computer vision research field progresses beyond "bucketed" identification and toward tackling multi-modal issues, language and vision related problems i.e., captioning the pictures and performing VQA. The captioning of images process and VQA has followed up the top down visual co-attention computation that can be exploited to tolerate dense image through proper understanding them and performing the fine-grained image process analysis with several time reasoning.

The (Jiasen Lu, Jianwei Yang, Dhruv, 2016) proposed different level of attention that includes top-down attention and bottom up attention for visual areas. The proposed system calculates attention at the level of objects and other prominent visual areas. This is a natural starting point for thinking about attention. Image captioning and visual question answering (VQA) are two problems at the intersection of computer vision and natural language processing that continue to drive a lot of research. To produce high-quality outcomes, both of these jobs frequently need fine-grained visual processing and even many levels of reasoning. As a result, visual attention mechanisms have become widely used in both picture captioning and visual quality assurance(Anderson *et al.*, 2018). These systems present the results-based network design through deep learning technology. The system has increased the throughput of the model by learning the most significant semantic of the images. Whereas, Top-down key signals defined by the present task (e.g., looking for something) can concentrate attention voluntarily, while bottom-up signals associated with unexpected, unusual, or salient stimuli can focus attention automatically in the human visual system. The attention-based processes can be retrieved by non-visual semantic or context specific tasks has defined to as "top-down" in this research, whereas solely visual feed-forward attention mechanisms are referred to as "bottom-up" in this paper. The top-down visual based attention processes are the most common in picture captioning and VQA. This happens due to the important co-ordination between different fields of artificial intelligence i.e., NLP and CV. In addition to that, VQA is also an emerging technology to explore new concept of deep learning, and attracting the research community to explore more relationship between VQA and Deep Learning.

Both the technology CV and NLP i.e. Natural Language Process and Computer Vision are collided in the VQA domain that is much more exciting and demanding for research community. The semantic information in picture captioning and video summarization is included in still images or video dynamics, and it simply has to be mined and presented in a human-consistent manner. In contrast, semantic information in the same media must be compared with the semantics inferred by a natural language inquiry in VQA, tripling the artificial intelligence-related work.

Recent research has examined coreference resolution tasks or developing referring phrases for a specific item in an image that would allow a human to distinguish which object is being referred to as challenges that are easier to assess than picture captioning at the intersection of vision and language. Despite being task-driven and specific, referring expressions frequently only include a few visual notions (such as colour and position). As we demonstrate, visual inquiries and their responses cause a broader variety of visual conceptions to arise.

Some recent VQA surveys have focused on the processes underpinning either image-related or verbal-related processing, as well as how to reliably fuse the transmitted information. The use of general-purpose datasets to analyze the building blocks of a VQA system is simply proposed; in reality, the majority of mentioned studies rely on them. VQA is now one of the most intriguing combined applications of Artificial Intelligence (AI) to Computer Vision (CV) and Natural Language Processing (NLP) (Ben-younes, Cord and Thome, 2017). Its goal is to develop systems that can respond to a variety of inquiries given in natural language and about any picture. To do this, a VQA system employs a variety of algorithms that take a picture and a natural language query as input and provide a natural language response as output. Except under rare circumstances, humans are inherently good at it, and AI aspires to replicate this capacity(Barra *et al.*, 2021). The function of NLP in tackling this multidisciplinary challenge is to comprehend the query and, of course, to generate a solution based on the CV results. In NLP, text-based Q&A has been explored for a longer time. The distinction with VQA is that both search and reasoning take into account the image's content. In its optimal form, VQA (Visual Questioning Answering) allows us to examine different reasoning models within the combined space of visual computation and computer-based language, and it may be used as a surrogate for the AI problem of scene interpretation. However, the majority of VQA benchmarks to far have been focused on basic counting schemes along with various visual features and different object recognition schemes for questions which doesn’t involve any thinking or knowledge above than what an image contains.

During recent years, the discipline of VQA (Visual Questioning Answering) has made incredible progress, hitting record numbers on common VQA datasets. VQA, as designed, is not only a fruitful environment for vision and language research but also a proxy for evaluating AI models for open-ended scene interpretation(Agrawal *et al.*, 2016). VQA in its optimal form would need not just visual recognition but also logical thinking and the incorporation of global information. However, current VQA datasets are mostly focused on recognition, and the majority of questions are about basic counting, colors, and other visual detection tasks, which do not need any logical thinking or external knowledge connection. The most challenging and intriguing inquiries, in theory, need understanding more than the question's scope or the information included in the photos. Over the last several years, one of the most prominent subjects in the computer vision field has been VQA (Visual Questioning Answering).

To merge textual and visual data, early attempts to VQA included recurrent networks and CNNs. By emphasizing visual regions that are important to the issue, attention-based models better help the model in answering queries. Deep neural networks use modular networks to make use of the language's compositional structure. Above mentioned methodologies are helpful and used in video processing. Whereas, there is no such activity was as such helpful to use external domain knowledge for computation. This can led towards a situation where we can conclude that images are no reflecting the entire semantics results that was queried in the question (Das *et al.*, 2016).

The duty of VQA was just recently introduced, and it immediately generated a great deal of attention and accelerated development. VQA is a challenging endeavor that was first made possible by the maturity attained in basic computer vision operations like image recognition. VQA is especially appealing since it, in its most full version, is an AI job. A goal called VQA (Visual Questioning and Answering) was put out to combine computer vision with natural language processing (NLP), to encourage research, and to push the limits of both disciplines. Computer vision, on the other hand, researches techniques for gathering, analysing, and comprehending pictures. Its main objective is to instruct robots in vision. NLP, on the other hand, focuses on facilitating natural language interactions between computers and people, which includes, among other things, teaching robots how to read. NLP and computer vision both fall under the umbrella of artificial intelligence, and they both use comparable machine learning-based techniques. They have, however, traditionally grown apart. In the last several decades, both studies have made great progress toward their own objectives, and the increasing expansion of both visual and textual data is driving a convergence of both fields' efforts. For instance, studies on captioning of images have developed effective techniques for combining learning from inputs of both text and images to create higher-level representations. When CNN (Convolutional Neural Network) technology trained on different object identification makes it more effective strategy using combining embedding word. The embedded words learned using CNN based on corpora text dataset that includes enormous amount of embedded words.

The area of VQA (Visual Questioning and Answering) is very young but has recently made significant strides. Another possible field for study that combines computer vision and natural language processing is VQA. Based on a picture and a question connected to it, a VQA system may determine the correct response. Although the process is straightforward for people, computers find it difficult.

VQA may be used in our daily lives. For instance, we may integrate automatic VQA (Visual Questioning and Answering) into the Chatbot platform to respond to queries and find information. In many real-world scenarios, such as customer assistance, advice, question-answering, conversation, and customer system administration, visual question-answering systems are required. It also has amazing possibilities for scenarios like bringing attention to important and priceless information in the immediate vicinity of displayers.

In the majority of current machine learning projects, data preprocessing is a crucial step. Cleaning a dataset increases the amount of data that can be extracted for model training, improving experimental outcomes.

Systems that can visually answer questions are crucial in AI applications for human life. Despite the fact that several other languages including English are also used for some recent research publications. Due to a lack of available data, there are no studies on visual question responding in Vietnamese. Because of this, we made the decision to conduct this research in order to provide a fresh dataset for testing Vietnamese visual question-answering systems. Based on the picture data source from MS COCO, this dataset was created. Although, when this VQA (Visual Questioning and Answering) model has been used as a practical example, the assessment result were perfomed on dataset using the cutting edge technologies like DLSTM (Deeper Long Short Term Memory), HCA (Hierarchical Co-Attention, BLSTM (Bidirectional Long & Short Term Memory). Mentioned study further highlighted the improvements made to VQA (Visual Questioning and Answering) model. This is because of the identification of the best performing model on the available dataset.

Many datasets have been suggested especially for VQA study. They at least include triples, which are made up of a picture, a question, and the response. There may also be additional annotations, including picture descriptions, image sections that support the answers, or multiple-choice candidate responses. The complexity of the datasets and the questions within them, as well as the amount of reasoning and non-visual information needed to deduce the answer, varies greatly. Malinowski and Fritz's idea was one of the earliest efforts at "open-world" visual question answering (2014). They provided details of a technique that combines semantic text parsing with picture segmentation in a Bayesian formulation that draws samples from training data's closest neighbours. The approach needs predicates that are human-defined, which are inherently dataset-specific and challenging to scale. Furthermore, it heavily depends on how accurate the picture segmentation technique and predicted image depth are.

Unpredictable text-based queries concerning pictures must be answered by an algorithm in open-ended visual question answering (VQA). VQA is a challenging computer vision challenge that calls for a system to be able to do a variety of jobs. Realizing VQA would mark a turning point in artificial intelligence and enhance human-computer interaction. To properly quantify improvement, however, a broad variety of competencies must be tested in VQA datasets.

When the DAQUAR dataset was made available in late 2014, VQA research really got going. Six significant VQA datasets, including DAQUAR, have been made available, and algorithms have advanced quickly. The top algorithms are presently close to 70% accurate on the most widely used dataset, "The VQA Dataset" (human performance is 83%). Even though these findings are encouraging, there are serious issues with biases in the present datasets. Furthermore, it is difficult to evaluate the capabilities of different algorithms since available datasets do not classify occurrences into useful categories. One approach may, for instance, do better at colour problems than at spatial reasoning ones. Due to the assessment measures utilised, an algorithm that excels in spatial reasoning will not be fairly rewarded because colour questions are far more frequent in the dataset.

The most popular kind of Visual Query Answering (VQA) involves showing the computer a picture and a text question about it. The right response, which is often a few words or a brief sentence, must subsequently be determined. Binary (yes/no) and multiple-choice situations, in which potential responses are presented, are examples of variations. A similar assignment is "fill in the blank," which requires you to add the missing word or words to an affirmation that describes a picture. These affirmations are basically inquiries that have been given a declarative form. The question to be answered does not become clear until run time, which is a key differentiator between VQA and other computer vision tasks. Traditional issues like segmentation or object recognition have a single question that an algorithm must answer, and only the input picture is subject to change. In contrast, the question's format and the sequence of steps necessary to respond to it are both unknown in VQA. This way, it more accurately captures the difficulty of interpreting broad images. The duty of answering textual questions is known as VQA, and the solution might be found in extensive knowledge bases or a specialised textual story (e.g., reading comprehension) (i.e information retrieval). The NLP community has long studied textual quality assurance, and VQA is its expansion to include extra visual supporting data. The additional difficulty is considerable since graphics often have greater resolution and are noisier than plain text. Additionally, pictures lack the grammatical structure and norms of language, and there is no direct substitute for NLP techniques like regular expression matching and syntactic parsers. Finally, although natural language already exhibits a greater degree of abstraction, photos capture more of the richness of the actual world. For instance, contrast the word "a red hat" with the many other ways that one may imagine it, many of which cannot be adequately expressed in a single line.

Visual question answering sometimes needs information that is not included in the picture, making it a far more difficult task than image captioning. This additional knowledge might be anything from common sense to exhaustive understanding of a particular aspect of the picture. This makes VQA a genuinely AI-complete work as it calls for multimodal knowledge that goes beyond a single subdomain. This eases the growing interest in VQA since it serves as a proxy for measuring our advancements toward AI systems that are capable of deep language and picture processing together with sophisticated reasoning. Note that picture captioning might theoretically be used to test image comprehension just as effectively. However, VQA offers the benefit of a simpler assessment criteria practically speaking. Usually, answers are simply a few sentences long. It is more challenging to compare extensive ground truth picture descriptions with expected ones. Despite research on enhanced assessment measures, this is still an unsolved research issue.

The "SHRDLU" system from 1972 (Winograd, 1972), which let users to use language to tell a computer to move different items about in a "blocks world," is one of the first examples of the integration of visual and language. Recently developed robotic conversational agents have a visual foundation as well. These works, however, were often constrained to certain fields or to particular linguistic forms. VQA, in contrast, focuses particularly on open-ended, free-form inquiries. The presence of established methods in computer vision and NLP as well as the accessibility of pertinent large-scale datasets are what are fueling the growing interest in VQA. As a result, during the last several years, a substantial amount of literature on VQA has been published. This survey's objective is to provide a thorough review of the subject, including models, datasets, and possible future prospects.

Both the natural language processing and computer vision communities are extremely interested in multimodal learning to connect vision and language. Image-text matching, visual captioning, visual grounding, and visual question answering are just a few of the vision-language activities that have greatly evolved. Since VQA needs both visual reasoning and fine-grained semantic knowledge of the picture and the question, it is more difficult than other multimodal learning tasks to anticipate the correct response.

Deep Neural Networks provide unique attention mechanism which has been successfully used to both the aforementioned multimodal activities as well as unimodal ones. Since it was initially proposed in VQA, the scheme of learning through visual based attention from image-based areas can input question which works as a de facto part of practically all VQA approaches. Both paying attention visually and learning to concentrate on the query's essential terms while reading text are crucial. According to recent research, learning different co-attention-based models which is simultaneously. The models present verbal and different visual information might facilitate very fine-grained depiction of provided images and different inquiries as well. This can lead towards more specific and accurate kind of prediction from images. However, these co-attention related models only pick up on the rough interactions of multimodal instances; they are unable to recognize the relationship between each picture region and each question phrase. This brings up a serious flaw with these co-attention theories.

Different kind of co-attention image models i.e., BAN & DCN have been suggested to represent rich kind of interactions among any visual area and any question word in order to address the issue of inadequate multimodal interactions. Understanding the relationship between the picture and the query is made easier by the dense co-attention mechanism. It's fascinating to see that both of these dense co-attention models have the potential to be combined to create deep co-attention models, which would allow for more sophisticated visual reasoning and maybe enhance VQA performance. But compared to the MFH model (Co-Attention model for Images) or their comparable shallow versions, these deep models don't perform appreciably better. The study aims to understand that these models lacks to presents concurrent modeling with rich sort of self-attention due to each model. This objective presents the bottleneck through out these co-attentions with deep learning models.

These systems assessed on performance metric which present apparently significant in the mind. The fundamental statistics based on any tradition VQA models presents through accuracy of these systems. Although these systems are very viable to generate multiple choice questions. These multiple option question provide very open-ended responses. These responses can be generating by asking the what, how, why type of the questions. The WUPS measure, developed by Malinowski and Fritz in 2014 and based on the WUP measure introduced by Wu and Palmer in 1994, is used to compute the semantic distance between an answer and the truth, which has a value between 0 and 1. WordNet is used to compute similarity between the phrases in the answer and the location of the ground truth in the semantic tree. WUPS, like practically all other semantic metrics, assigns very significant scores to ideas with completely unrelated meanings. To address this problem, the authors recommend that scores less than 0.9 be scaled down by a factor of 0.1. In many cases, WUPS is unquestionably superior to traditional accuracy.

We might have numerous ground truth solutions for each question instead of depending just on semantic measurements, as we saw with the VQA dataset. Then, for instance, if a particular response fits the more common response or at least one of the potential ground truth responses, we may say that the response is right. Because there is no agreed-upon answer to a yes-or-no question, the latter must be used with caution because any response would be valid.

Finally, the system evaluates different responses with different strategy to evaluate and assess the replied through human (Expert) judges. This kind of evaluation is very expensive because of the involvement of many resources. Although, different rules has been created to develop an accurate criteria. These criteria provide an ease for the judge to evaluate or assess the responses. For that purpose, a certain level of training has devised to those judges to improve the quality of the evaluations on every assessment level of the responses.

In recent years, there have been significant developments in a wide range of machine learning fields. Neural networks are increasingly being used to perform NLP problems including entity recognition, language form, and question answering as well as computer vision tasks like object detection and image process with greater speed and accuracy.

Visual question responding is a task that lately caught the interest of the AI community. This will examine the issue of visual question answering, several solutions to it, related problems, datasets, and incoming techniques.

Systems that use visual input to answer questions aim to provide accurate responses in natural language. The goal of this topic is to develop systems that can successfully communicate in natural language about an image and comprehend its contents in a manner similar to that of humans. This is a difficult undertaking since it calls for interaction and complementarity between image-based models and natural language models.

The challenge, which faces the Artificial General Intelligence problem, i.e., making computers as intelligent as people, has been universally acknowledged as AI-complete.

VQA is a relatively new area that requires knowledge of both text and vision. Since NLP and CV findings are being considerably improved by deep learning approaches, we may safely anticipate that VQA will become more accurate over the next years. The developed datasets and established metrics, like many other IA activities, have in some way influenced the research that has been conducted so far. The most effective approach to assess a VQA system is still up for debate, but it is extremely probable that new datasets and metrics will make it possible to further explore and hone the idea of quality.

# Chapter 2 Literature Review

VQA is a fundamental challenge in computer vision and natural language processing that demands a system to perform far more than task-specific algorithms like object recognition and detection. An algorithm that can answer arbitrary inquiries about pictures would be a watershed moment. VQA, in our opinion, should be an integral feature of any visual Turing test. Algorithm will work on mathematical model of rule of inference and experiment will be conducted by visually impaired people.

These approaches can only handle knowledge represented as subject-relation-object or visual concept-relation attributes triplets, and they rely on supervision for fact retrieval. Answering queries in our dataset, on the other hand, necessitates working with unstructured knowledge resources. It's a hot issue in deep learning research since it necessitates combining natural language processing and computer vision modules into a unified architecture. All attention models for VQA in the literature have so far concentrated on the challenge of determining "where to gaze" or visual attention. In this research, we suggest that determining "which words to listen to" or questioning attention is just as crucial as determining "which words to listen to." Consider the following inquiries: "How many horses are in this image?" and "How many horses can you perceive in this image?" They both signify the same thing, which is encapsulated in the first three words(Singh, Ying and Nutkiewicz, 2018). A computer that focuses on the first three words would be more resistant to linguistic variants that have no bearing on the meaning or response to the query. We address the problem of question attention as well as visual attention as a result of this discovery. VQA model is a significant refined and instance interpretation all the text-based contents and visual queries. Therefore, the development of co-attention based VQA model correlate with significant embedded words with key items of the image data. The images dataset contains all the content that is demanding by question to make it a success story (usecases). One of the most successful and effective attempt was generated through co-attention based vqa model which used a shallow model to implement deep co-attention based model. These deep models show minor improvement as compare to other shallow based co-attention model.

GNN (Graph Neural Network) is sub routine approach of deep learning model. It can apply on every kind of scenario either it could be structural or unistructural. Specially, in former mentioned scenario, there comes some significant topics referred as visual questioning answering model. VQA model has primary objective to gives answer from the available test images by understanding the semantic of the image. In GNN based VQA, the processing of the textual scenarios takes place using graph reasoning model It facilitate in identifying the critical issues from the textual or visual feature and merge them on a single space, so that the end user can made the decision about the accuracy. For same topic, its quite challenging to measure and compare the performance of different training and testing model. This is because of the variation in nature of the dataset.

## 2.1 Image Classification Model

To answer open-ended queries about pictures, Visual Question Answering (VQA) necessitates a model that can comprehend and reason about visual-linguistic notions. Object localization, attribute identification, activity classification, scene interpretation, reasoning, counting, and other skills are required to correctly answer these questions(Li *et al.*, 2019). Real-world photos were combined with crowd-sourced questions and responses in the first VQA datasets. It was anticipated that this would be an incredibly tough challenge and was offered as a type of Visual Turing Test to measure performance in computer vision. However, it became evident that many high-performing algorithms were relying on biases and surface connections rather than truly comprehending the visual information. In VQAv1, for example, answering “yes” to all yes/no questions resulted in a 70 percent accuracy rate on these questions. Later natural image VQA datasets sought to overcome this problem. By linking each question with complementary visuals and varied responses, VQAv2 decreases some kinds of linguistic bias(Chen *et al.*, 2020). TDIUC examines generalization to a variety of questions and replie“. W”en the train and test distributions diverge, CVQA checks concept compositionality, and VQACPv2 measures performance.

One of the main strategies towards full AI is Visual Question Answering (VQA), which involves answering natural language inquiries regarding visual input. VQA has gotten a lot of interest with the publication of many large-scale VQA datasets, and hundreds of models have been created as a result. However, because of the inherent annotation distortions in real-world picture datasets, today’s VQA models always depend too much on superficial language connections. For example, regardless of the X, a model replying “2” for all “how many X” inquiries can still get acceptable results(Jiang *et al.*, 2020). A diagnostic benchmark called VQA-CP has recently been developed to detangle bias factors and track the progress of VQA research. The train and test portions of the VQA-CP have purposefully differing question-answer distributions. When compared to other datasets, the performance of several cutting-edge VQA models suffers dramatically on VQA-CP. Deep learning breakthroughs have accelerated both computer vision and natural language processing (NLP). Image captioning, text-to-image synthesis, and visual question answering (VQA) are examples of interdisciplinary areas connecting language and vision that have gotten a lot of interest recently from both the vision and NLP communities(Jiang *et al.*, 2020). In the case of VQA, the aim (and the key problem) is to train a model that can comprehend multimodal input in a thorough and semantically aligned manner. The aim is to match visual aspects in the picture with the semantic meaning in the question to correctly answer the question, which is provided an image and a natural language question based on the image.

Opportunities to build and exploit algorithms for automated medical image interpretation are being investigated at a quicker rate as interest in artificial intelligence technology to enhance clinical decision-making and increase patient involvement grows. A “second opinion” supplied by an automated system can boost professionals’ confidence in evaluating difficult medical pictures. Furthermore, because patients may now access structured and unstructured data about their health through patient portals, there is a need to assist them in better comprehending their problems in terms of the data they have access to, including medical imaging.

In the computer vision and pattern recognition communities, optical character recognition (OCR) has a long history. In the beginning, OCR research was limited to handwritten digits and clear pictures of printed documents. OCR has recently taken on new forms, such as photoOCR, also known as scene text recognition, and unconstrained handwritten text recognition. All of these types of OCR problems have seen tremendous development. Nonetheless, several issues remain unsolved, such as text recognition with variable fonts and layout. They combine the problems of OCR with those of the VQA, and add a new job of reading and interpreting text appearing in the pictures to answer visual questions. Furthermore, by reading words in pictures, the research presents a revolutionary deep model for VQA. The study employs cutting-edge text block recognition and OCR modules, as well as a trainable visual question answering system, to achieve this goal(Mishra *et al.*, 2019). For example, basic CNN features for visual representation, text block coordinates and named-entity tags on OCRed text for textual representation, and bi-directional LSTM for question representation make up the baseline VQA system. To arrive at an accurate response, all of these representations are fed into a trainable feed-forward neural network.

Learning textual focus on the crucial words in the query is just as vital as visual attention. Recent research has demonstrated that learning co-attention for both visual and textual modalities at the same time can improve the fine-grained representation of the picture and inquiry, resulting in more accurate predictions. However, because these co-attention models only learn the coarse interactions of multimodal occurrences, they are unable to infer the association between each picture region and each question word. As a result, these co-attention models suffer a fundamental restriction. Two dense co-attention models, BAN and DCN, have been suggested to describe dense interactions between any visual region and any question word, to address the problem of inadequate multimodal interactions. To correctly answer questions, the dense co-attention mechanism aids comprehension of the image-question link. Both of these dense co-attention models may be cascaded in-depth, resulting in deep co-attention models that support more complicated visual reasoning, potentially boosting VQA performance. We propose two generic attention units, both of which are inspired by the Transformer model in machine translation: a self-attention (SA) unit that can model dense intra-modal interactions and a guided-attention (GA) unit that can represent dense inter-modal interactions (word-to-region). After that, we create several Modular Co-Attention (MCA) layers that may be cascaded in depth by combining the SA and GA units in a modular fashion(Abacha *et al.*, 2020). Finally, the research suggests a deep Modular Co-Attention Network (MCAN) with cascaded MCA levels.

Beyond recognizing the visual components of the picture, VQA also needs a thorough comprehension of the natural language question’s semantics. As a result, learning textual focus for the inquiry and visual attention for the image at the same time is required. The researchers presented a co-attention learning paradigm for learning picture and question attention alternatively. Yu and colleagues split the co-attention approach into two parts: self-attention for question embedding and question-conditioned attention for visual embedding. The researchers presented a multi-stage co-attention learning model for refining attention based on past attention recollection. These co-attention models, on the other hand, train distinct attention distributions for each modality, ignoring the complex interplay between each question word and each picture area(Yu *et al.*, 2019). Understanding fine-grained connections of multimodal features become a bottleneck as a result of this. Dense co-attention models, which establish the whole interaction between each question word and each visual region, have been proposed to overcome this issue. The dense co-attention models outperform the prior co-attention models with coarse interactions in terms of VQA performance. The MCA layer is a modular composition of the two fundamental attention units, namely the self-attention (SA) unit, and the guided-attention (GA) unit, as suggested by the scaled dot-product attention.

Researchers were able to handle multimodal challenges that included visual and textual modalities thanks to Deep Learning’s recent achievements in computer vision and natural language interpretation. Visual Question Answering (VQA) is one of these activities that is gaining popularity. The VQA job requires you to respond to a query concerning a picture. It necessitates not just a thorough comprehension of the visual scene and the inquiry, but also the ability to anchor textual notions in the image and effectively employ both modes. Solving the VQA task could have a huge impact on real-world applications like assisting visually impaired users in comprehending their physical and online environments, searching through large amounts of visual data using natural language interfaces, and even communicating with robots using more efficient and intuitive interfaces(Shrestha and Tech, 2019). Several significant real-image VQA datasets have lately become available. Each one focuses on various skills that a VQA model would need in real-world situations, such as fine-grained identification, object detection, counting, activity recognition, commonsense reasoning, and so on. On most of these benchmarks, current end-to-end VQA models outperform humans, and on one specific benchmark accounting for compositional reasoning, they even outperform humans. However, statistical regularities between response occurrences and specific patterns in the question are exploited by them. While they are supposed to combine information from both modalities, they frequently respond without taking the picture modality into account. When the majority of the bananas are yellow, a model does not need to learn the proper behavior to get high accuracy when answering questions regarding banana hue. It’s considerably simpler to learn from the statistical shortcut associating the terms what, color, and bananas with the most common answer yellow than it is to look at the image, locate a banana, and judge its hue. Unmoral models can be used to assess the number of statistical shortcuts from each modality. For example, across the test set, a question-only model trained on the widely available VQA v2 dataset(Marino *et al.*, 2019) predicts the correct answer around 40% of the time. Because their training and testing sets frequently have the same distribution, VQA models are not inhibited from using these statistical shortcuts from the question modality. When tested on a test set with differing statistical regularities, however, they frequently have a large decline inaccuracy. Unfortunately, when gathering real datasets, these statistical regularities are difficult to prevent. It is suggested that developing innovative ways to lessen the number of biases resulting from question modality is critical to acquiring better behaviors.

Parallel to these past dataset balancing efforts, a significant effort has been made to create VQA models to overcome dataset biases. The limitation of the aforementioned multimodal fusion models is that the global feature representation of an image may exclude data required to properly answer questions about local picture regions (such as "what is in the woman's left hand"). The attended picture properties are learned by adaptive learning for a given question, followed by multimodal feature fusion to produce the accurate prediction, and as a result, new approaches have integrated the visual attention mechanism into VQA. The study presented the Grounded VQA paradigm, which is a hand-designed architecture (GVQA). It divides the VQA problem into two parts: finding and detecting the visual areas required to answer the question and identifying the space of possible solutions based on a question-only branch. This method necessitates independent training of several sub-models. Our learning strategy, on the other hand, is comprehensive. Our technique is model-agnostic, whereas their sophisticated design is difficult to apply to multiple systems. We use a question-only branch during the training, but we delete it at the conclusion. Practical work is the most similar to ours in terms of methodology. In VQA models, the authors offer a learning technique to overcome linguistic priors. They begin by introducing an antagonistic question-only branch. It provides a question-only loss using the question encoding from the VQA model as input(Nguyen and Okatani, 2018). They utilize a gradient negation of this loss to prevent the question encoder from capturing undesired biases that the VQA model may exploit. They also suggest a loss based on the entropy difference between the VQA model and the output distributions of the question-only branch. Only the backpropagation of these two losses reaches the question encoder. Our learning technique, on the other hand, focuses on the whole VQA model parameters to more effectively limit the influence of undesired biases. Rather than depending on these two extra losses, the study utilizes the question-only branch to dynamically adjust the value of the classification loss in the VQA model to decrease bias learning.

We quickly go through earlier research on VQA, particularly those that include co-attention models. Answering Questions Visually (VQA). Over the last several years, interest in VQA has grown. The simplest VQA methods include the multimodal fusion of global characteristics. To predict the response, a multimodal fusion model first fuses the picture and the question as global characteristics. Some methods use a more complicated model to develop better multimodal fusion models using residual networks or LSTM networks to learn better question representations.

The aforementioned multimodal fusion models have the drawback that the global feature representation of an image could exclude information that is necessary to accurately respond to queries regarding local picture areas (such as "what is in the woman's left hand"). Therefore, new methods have integrated the visual attention mechanism into VQA by learning the attended picture characteristics via adaptive learning for a specific question, followed by multimodal feature fusion to provide the precise prediction. The question embeddings were projected into the visual space and a tunable convolutional kernel was developed by Chen et al. to explore the picture attention area. To learn the attention iteratively, it presents a stacked attention network. In order to combine the textual data from the questions with the visual information from the image's spatial grids, Different studies used several multimodal bilinear pooling approaches. In order to train the attention on candidate items rather than spatial grids, Anderson et al. developed a bottom-up and top-down attention mechanism.

**Co-Attention Model:** Beyond recognizing the visual components of the picture, VQA also needs to completely grasp the semantics of the natural language inquiry. Therefore, it is vital to train the textual attention for the inquiry and the visual attention for the picture concurrently. The study (Singh, Ying and Nutkiewicz, 2018) suggested a co-attention learning paradigm to alternatively train the picture attention and question attention. The study condensed the co-attention approach into two parts, self-attention for a question embedding and the question-conditioned attention for a visual embedding. The study (Ben-younes, Cord and Thome, 2017) suggested a multi-stage co-attention learning model to modify the attentions based on recall of prior attentions. However, these co-attention models train different attention distributions for each modality (image or question), and overlook the intense interaction between each question word and each picture area. This creates a barrier for interpreting fine grained connections of multimodal features. To solve this problem, dense co-attention models have been presented, which demonstrate the whole interaction between each question word and each visual area. Compared to the prior co-attention models with coarse interactions, the dense co-attention models give much higher VQA performance.

Visual question-answering has been the subject of many recent articles. These settings and datasets are rather constrained (and sometimes generated), in contrast to our work. For instance, only takes into account questions whose solutions fall into one of 894 item groups or the 16 fundamental colors. additionally takes into account queries produced by templates based on a set vocabulary of things, characteristics, connections between things, etc. Our suggested work, however, entails open-ended, free-form queries and human-provided responses. Our objective is to broaden the range of information and types of reasoning required to provide accurate solutions. crucial to succeed on this more challenging and unrestricted work. The proposed VQA job is connected to previous relevant work that has looked at the combined parsing of videos and the text that goes with them to reply to inquiries on two datasets that each comprise 15 video clips. utilizes employees from the public to respond to queries from visually impaired users concerning visual information. advocated using an LSTM for the question and a CNN for the picture in continuous processing to provide a response. The final LSTM hidden state is utilized to successively decode the response phrase in their model, which conditions the LSTM question representation on the CNN image characteristics at each time step. The model created in this study, however, experiments with "late fusion," where the CNN image features and the LSTM question representation are computed separately, combined using element-wise multiplication, and then passed through fully connected layers to produce a softmax distribution over output answer classes. creates abstract sceneries to capture common sense related to fill-in-the-blank and visual paraphrase queries that are just textual. and evaluate the veracity of arguments made using common sense using visual data. provided a collection of 10,000 photographs and asked for captions to explain certain details of a scene (e.g., individual objects, what will happen next). Concurrent with our effort, we gathered Chinese queries and responses for COCO pictures that were then translated to English by humans. Using COCO captions, four different question kinds were automatically generated: item, count, colour, and location.

**Text-based Q&A:** The NLP and text processing groups have researched the issue of text-based Q&A extensively. Sentence completion is another textual assignment that is similar (e.g., with multiple-choice answers). Techniques for VQA are inspired by these methods. The way questions are rooted in the text is one major issue. For instance, QA-pairs and synthesized textual descriptions based on a simulation of people and items in predetermined places. VQA is innately visual, requiring a comprehension of both text (questions) and imagery (images). Since humans are the ones who come up with the questions, having common sense knowledge and using complicated reasoning are much more crucial.

**Describing Visual Content:** The activities of image tagging, image captioning, and video captioning are related to VQA since they involve creating words or phrases to describe visual content. While these activities need both semantic and visual expertise, captions are often vague. Generic picture descriptions are of limited help in VQA questions since they call for comprehensive, particular information about the image.

**Other Vision + Language Tasks:** Recent research has looked at tasks that are simpler to assess than picture captioning at the junction of vision and language, for coreference resolution or creating referring phrases for a specific item in an image that would enable a human to recognize which object is being referred to. Although task-driven and concrete, referencing expressions often only capture a small number of visual concepts (such as color and position). As we show, visual inquiries and their solutions lead to the emergence of a greater diversity of visual conceptions.

There exist different dataset and algorithms to validate VQA model. None of the algorithm has shown accuracy more than 66%(Kafle and Kanan, 2017). They key element to gain such accuracy was using the attention of the participant. All the method were lack to include compositional reasoning in their attributes. This led towards ambiguity in the participants minds. There exists a compositional reasoning framework with over 90% of accuracy to answer visual question in subsequent manners, referred neural module network(Kafle and Kanan, 2017). Each sub task performed a well-defined work in sub-network. Combining this framework, we proposed an algorithm that meet the limitation and can work better on larger dataset with more accurate results. Referred framework was improved with mathematical rules to propose algorithm.

It is thought to be an ambitious objective to build a system that can automatically answer questions from random photos. Recently, significant progress has been achieved in resolving the image-based question and answer issue thanks to the creation of contemporary machine learning techniques and the implementation of a number of relevant research.

# Chapter 3 Methods

For this experimental model, we selected the dataset (Goyal *et al.*, 2019). This dataset contains 204k images with 614k question. The dataset was evaluated for language biasness based on three tuple relation (I, Q, A). I for image, Q for Question and A for Answer. The data was collected from Amazon Mechanical Turk workers. We have reported the statistics based upon frozen calculation model. Frozen is multi-layers language model with billions of parameters. This study presented the question template that ask from the participants in table 1.

| Question | What is the color of\_\_\_\_\_\_\_? |
| --- | --- |
| Demonstration level question | Floor area?  Plow on which cat lay down?  Short’s that wear by the child?  Sign of the business? |
| General Question | Fence besides the man?  Statue beside the man? |
|  | Near building? |
| Question | Why? |
| Demonstration level question | Colour of the ground surface differ from another.  Can is sitting under an umbrella?  Car is tow down. |
| General Question | Boys having entertainment?  Element trucks in two colours? |
| Question | Which |
| Demonstration level question | Utensil is placed over the table?  Way the train is going?  Hands grips the rocket?  Operating system (OS) is used? |
| General Question | Side of the room, television is placed. |
| Question | How many |
| Demonstration level question | Unopened paper roll is contained by the picture.  Engine an aeroplane has.  Door is available?  Cabinets are installed?  People wearing jeans? |
| Question | Does this |
|  | Positive type question:  Eatable looks burnt?  Sound like noisy around?  Ships have an engine?  Negative type question:  Fruit, change it colour type?  Animal produce dairy products?  Fast food item (pizza) looks hot? |
| General question | Seems to be a happy occasion?  Yes/No  Person on his skis?  Yes/No  Transportation depends on gasoline. |

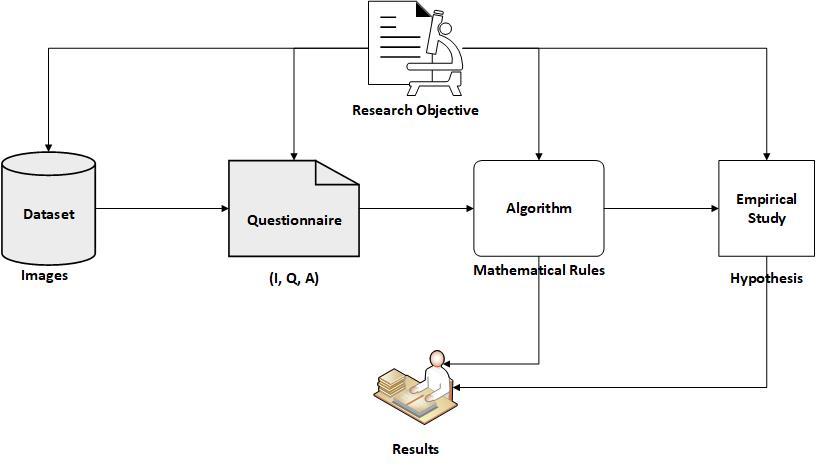
## 3.1 Research Objective

The research objectives are mentioned below:

* To proposed a VQA algorithm for larger dataset of images.
* To evaluate the algorithm images in nuanced manner. With different mathematical rule of inferences.
* To empirically investigate the algorithm in particularly scenario (Amazon Case Study) where it will aid the visual impaired person.

## 3.2 Research Methodology

Flow of the research methodology has been explained below fig 3.1.



*Figure 3.1 Research Methodology*

# Chapter 4 Results, data, findings

This chapter has five sections in which clear demonstration have been made using the VQA model implementation with dataset.

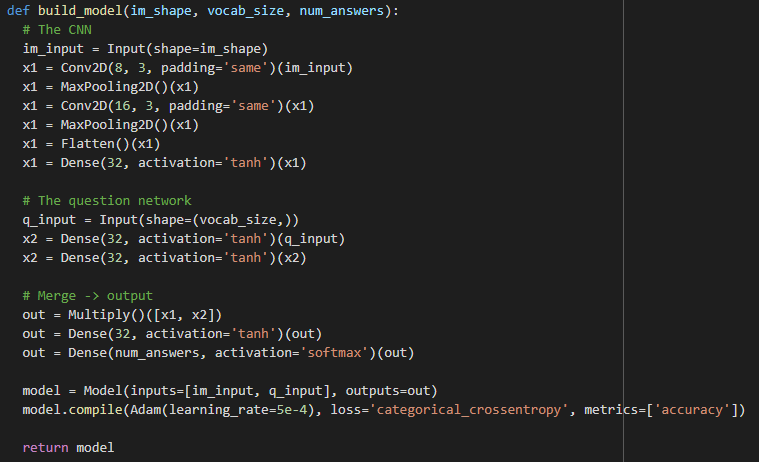
## 4.1. Installation of Dataset

First, we need to install easy-VQA dataset which we are going to use in our experiment. The mentioned dataset contains set of images and repository of questions to train the algorithm. In this dataset, we have 4000 images and more than 38000 questions for training the algorithms. The experiment use 1000 images and more than 9000 questions for testing purpose. The set of expected answers includes thirteen answers to evaluate the training dataset. More than twenty-eight thousand training questions and over seven thousand testing questions have the answer in simple yes/no format. The first step called as Installation, in which dataset begins by running the query “pip install easy-vqa”. This dataset is highly accurate to train the algorithm. It works faster than any other dataset regardless from the hardware incompetence.

## 4.2. Making the dataset ready

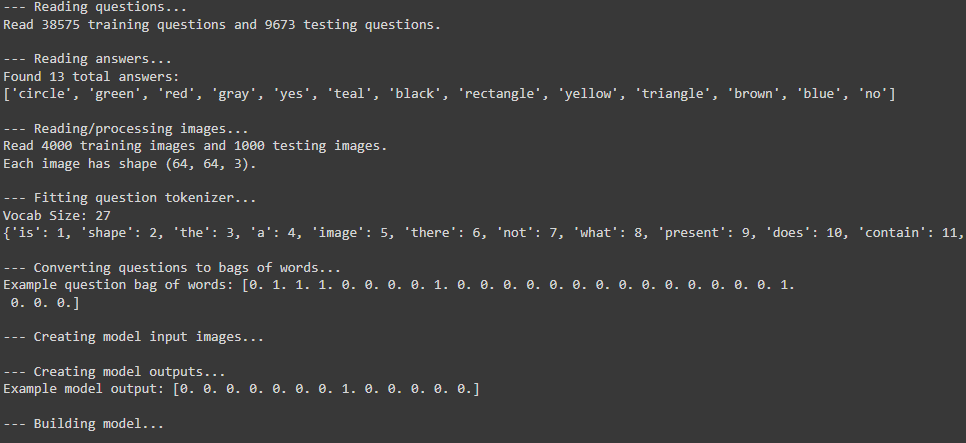
The second step helps to prepare dataset ready for experimental analysis. The step reads out the questions and their actual available answers. It also includes reading the images of the dataset and save them into there set directory. These sub task helps to set the training and testing data.

## 4.3. Creating the Model

The third step include creating the model for Visual Question Answering which is the simplest one because of the dataset, presented in figure 1. The first of the code present the Convolutional Neural Network (CNN) based model for images. CNN model works as grid like protocol. Any kind of digital image is a collection of binary representation of given visualized data. The pixels of the image in CNN arrange in grid like structure. CNN architecture consists of three pre-defined layers i.e., convolutional, pooling and connected one. The first layer is considered as the fundamental unit of CNN as it took all the load of the computations. Actually, it performs a cross product between two different Metrix. First metric includes all kinds of parameters that involves learning, also referred as kernel Metrix. The next Metrix include the receptive fields. The kernel Metrix has smaller size as compare to other Metrix but its depth becomes higher when the cross product took more data. The conclude the kernel is made up of three-color scheme channels, typically referred as RGB. The next layer referred as pooling layer in CNN. The primary concern of this layer is to replace the value-added output from the dedicated network at pre-defined locations. The concept reduces the computational load which directly minimize the spatial size of the program. The last layer is known as connected layer in which neuron are interconnected with its succeeding and proceeding layer. In the code snippet figure 1.0 presents that we use connected layer for combining the two-prediction model to get answer from the images. In combining the prediction model, we define our building model in a function which takes three parameters. Input takes the images in predefined shape a size and output merge the two models using layers of CNN. 

## 4.4. Training Model

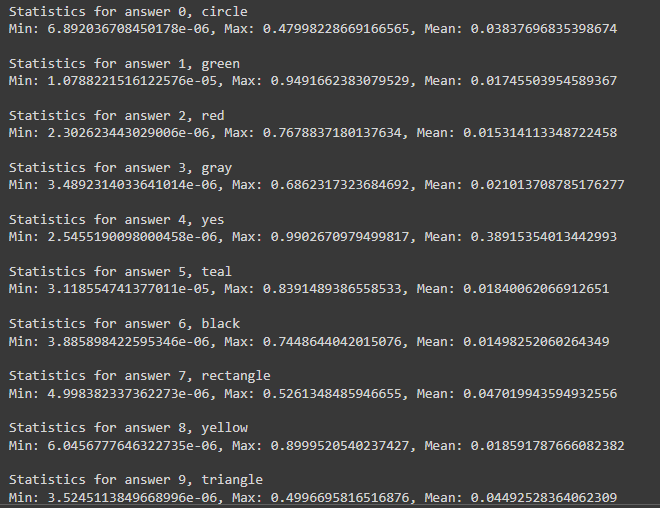
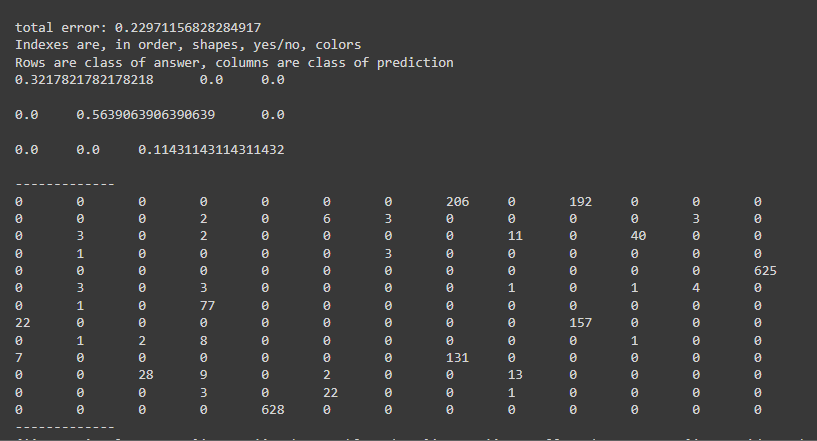
After building the model, next significant step is to train it on particular parameters. These parameters are the chunks that you want to see in your answers. Training involves reading the question properly, reading the images properly and the answers of the visual questions as well. The figure 2 demonstrate the output of training model in which shows our model take more thirty-eight thousand questions for training purpose and more nine thousand answers for testing purpose. It took four thousand images with size of 64x64. Model is trained over image tokenization technique. The training involves combining the CV and NLP based techniques for high interpretation of the presented scenes. Our model use the easyvqa and CNN build model fro training purpose. Then training model will be evaluated through experiment with propose algorithm.



*Figure 2 Training Model*

## 4.5. Analysis of Experiment

The following part present the experimental analysis for our model. The analysis was mapped with three parameters i.e., presenting the accuracy of the system against each kind of question from the dataset and the other one is to present the error while in last parameter our model present the wrong answers. For analysis, we use matplot library and argument parser library. The output is presented in figure 4 where the output was shown with respect to statistical manner against each answer.

While Figure 4 present the output in which error are shown on terminal. This presents that 0.23 percent error our model has produced. 

# Chapter 6 Conclusion and recommendations

An application of computer vision and natural language processing is called visual question answering (VQA). Recent developments in the intriguing and challenging subject of visual question answering (VQA) have brought together computer vision (CV) and natural language processing (NLP). Simply mined and presented in a human-consistent way, the semantic information needed for picture captioning and video summarizing is already contained in still pictures or video dynamics. Contrarily, the effort associated with artificial intelligence is tripled since semantic data from the same media must be compared with the semantics deduced from a natural language inquiry in VQA. Deep learning showed a great deal of interest in these fields. Our system receives some text-based inquiries regarding the photographs. The system must then compute the outcomes from it. Deep learning, computer vision, and neural networks combined with deep learning enable people to classify pictures. Humans are capable of comprehending the information included in pictures, such as objects, properties, positions of things, relationships between objects, etc. Additionally, the system includes scene categorization, attribute classification, object recognition, and object detection. Above all, if the topic takes on a geographical dimension, situations might get complicated. There are numerous potentials uses for VQA. The image is accessible to those who can access information since a picture is worth hundreds of words. Without the picture meta data, image recovery via VQA is also an option. One of the issues with the VQA system is the process of captioning. The best method is to use human judges to assess the data, but this takes time and money. Several automated evaluation methods have been introduced as a result. The most popular caption assessment methods are BLEU, ROUGE, METEOR, and CIDEr. With the exception of CIDEr, which was created expressly for evaluating image descriptions, all caption assessment measures were first devised for machine translation evaluation. Each of these indices has a unique set of restrictions. It is well known that major changes in sentence structure with essentially changing semantic content result in the same score for the most used metric, BLEU. Machine captions received higher evaluations than human descriptions in BLEU ratings for produced captions. A VQA technique was suggested in the study for bigger picture datasets. The photographs will be evaluated by the algorithm in a subtle way. The essential component of an algorithm is how it will infer the outcomes from various mathematical inference rules. The algorithm is then empirically investigated, specifically in situations where it will benefit a person with vision impairment. Dataset from (www.visualqa.org) will be utilized during the research, and the suggested method will be used to compute the findings. As per our research objective, we have utilized easy-VQA dataset and enables its significant to work with larger dataset. The second objective was to evaluate using mathematical inference, where we have used CNN based mathematical model for prediction and the third objective was to evaluate it using experiment where we performed experimental analysis. The experimental analysis describes that algorithm has very minimal error rate and work correctly with large dataset.

**Future Expectations**

Less bias is required in upcoming datasets. In this work, we have frequently addressed the issue of bias in existing VQA datasets and highlighted the kind of issues these biases lead to when trying to accurately evaluate a VQA algorithm. This will be challenging to accomplish in open-ended real-world VQA without thoroughly training the people who create the questions. In addition to bias in the questions, bias in the images used for computer vision datasets has long been a concern in VQA. Future datasets for benchmarking should be bigger and less biassed, and should also have more detailed analysis. Some types of questions are much easier than others, either due to bias or because existing algorithms are particularly good at answering those types of questions, such as object recognition questions, in all of the publicly available datasets that use evaluation metrics that treat every question equally. Certain datasets, like COCO-QA, have separated VQA questions into separate categories. For COCO-QA, these categories include colour, counting (number), position, and object. Standard accuracy should be replaced, in our opinion, by mean per-question type performance, which would give no one question an equivalent weight when assessing performance. This would go a long way toward ensuring that a VQA algorithm must perform well at a wide variety of question types to perform well overall. Otherwise, a system that performed exceptionally well at "Why" questions but slightly worse than another model at questions that were more prevalent would not be fairly evaluated. Each question would need to be given a category in order to accomplish this. This endeavour, in our opinion, would considerably increase the significance of benchmark findings. The results of each question type's scores could be compared between algorithms to determine which types of questions they do best on.

An agent in an interactive environment may have a VQA assignment that requires it to respond to queries from humans. For illustration, the agent might be asked, "Are there any apples in the fridge?" The proposed Interactive Question Answering presents a difficult issue. In order to complete this task, the agent must be able to move around in the environment, understand it, interact with it, and plan and carry out a sequence of activities. An intelligent agent that is created to solve the difficulty of interactive question answering may include a visual question and answer system.

# Chapter 7 Bibliography

Abacha, A. Ben *et al.* (2020) ‘Overview of the VQA-Med Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain’, *CEUR Workshop Proceedings*, 2696, pp. 22–25.

Agrawal, A. *et al.* (2016) ‘VQA : Visual Question Answering’, pp. 1–25.

Anderson, P. *et al.* (2018) ‘Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086. doi: 10.1109/CVPR.2018.00636.

Barra, S. *et al.* (2021) ‘Visual question answering: Which investigated applications?’, *Pattern Recognition Letters*, 151, pp. 325–331. doi: 10.1016/j.patrec.2021.09.008.

Ben-younes, H., Cord, M. and Thome, N. (2017) ‘MUTAN: Multimodal Tucker Fusion for VQA’, *Iccv*, pp. 2612–2620.

Chen, L. *et al.* (2020) ‘Counterfactual samples synthesizing for robust visual question answering’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 10797–10806. doi: 10.1109/CVPR42600.2020.01081.

Das, A. *et al.* (2016) ‘Human attention in visual question answering: Do humans and deep networks look at the same regions?’, *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 932–937. doi: 10.18653/v1/d16-1092.

Goyal, Y. *et al.* (2019) ‘Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering’, *International Journal of Computer Vision*, 127(4), pp. 398–414. doi: 10.1007/s11263-018-1116-0.

Jiang, H. *et al.* (2020) ‘In Defense of Grid Features for Visual Question Answering’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 10264–10273. doi: 10.1109/CVPR42600.2020.01028.

Jiasen Lu, Jianwei Yang, Dhruv, D. P. (2016) ‘Hierarchical Question-Image Co-Attention for Visual Question Answering’, *30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain*, 90(9), pp. 343–348.

Li, L. *et al.* (2019) ‘Relation-aware graph attention network for visual question answering’, *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October, pp. 10312–10321. doi: 10.1109/ICCV.2019.01041.

Marino, K. *et al.* (2019) ‘OK-VQA: A visual question answering benchmark requiring external knowledge’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June, pp. 3190–3199. doi: 10.1109/CVPR.2019.00331.

Mishra, A. *et al.* (2019) ‘OCR-VQA: Visual question answering by reading text in images’, *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pp. 947–952. doi: 10.1109/ICDAR.2019.00156.

Nguyen, D. K. and Okatani, T. (2018) ‘Improved Fusion of Visual and Language Representations by Dense Symmetric Co-attention for Visual Question Answering’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6087–6096. doi: 10.1109/CVPR.2018.00637.

Shih, K. J., Singh, S. and Hoiem, D. (2016) ‘Where to look: Focus regions for visual question answering’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, pp. 4613–4621. doi: 10.1109/CVPR.2016.499.

Shrestha, R. and Tech, C. (2019) ‘Answer Them All ! Toward Universal Visual Question Answering Models Rochester Institute of Technology’, *Cvpr*, pp. 10464–10473.

Singh, J., Ying, V. and Nutkiewicz, A. (2018) ‘Attention on Attention: Architectures for Visual Question Answering (VQA)’, pp. 1–7.

Yu, Z. *et al.* (2019) ‘Deep modular co-attention networks for visual question answering’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June, pp. 6274–6283. doi: 10.1109/CVPR.2019.00644.

# Appendices

