Humanlike Visual Q&A System

Capstone Project Report

End-Semester Evaluation

Submitted by:

(101803207)	Ishan Manuja
(101803209)	Aadarsh Gupta
(101803222)	Sameep Singh
(101803246)	Harchetan Singh Chahal

BE Third Year- CoE

CPG No: 70

Under the Mentorship of

Dr. Jasmeet Singh

Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology

December 2021

ABSTRACT

In recent years, significant progress has been made in the areas of computer vision, object detection, and natural language processing (NLP). Artificial Intelligence (AI) systems, such as question and answer models, use NLP to provide "extensive" capabilities to a machine. Such a machine can answer natural language questions about any part of the structured text. An extension of this system is the integration of NLP with computer vision to accomplish the task of Visual Question Answering (VQA), which builds a system that can answer natural language questions about images. Several systems have been proposed for VQA that use in-depth learning structures and learning algorithms.

This project introduces a VQA system that provides in-depth insight into images using Deep Convulsive Neural Networks (CNNs) that detect image features.

More specifically, feature embedding from the output layer of the VGG19 model is used for this purpose. Our system acquires complex logical capabilities and natural language comprehension so that it can understand the question correctly and give an appropriate answer. The Inferior model is used to obtain sentence-level embedding to extract features from the query. Various structures have been proposed to merge image and language models. Our system achieves results comparable to the baseline system in the VQA dataset.

DECLARATION

We hereby declare that the design principles and working prototype model of the project entitled Humanlike Visual Q&A system is an authentic record of our own work carried out in the Computer Science and Engineering Department, TIET, Patiala, under the guidance of Dr. Jasmeet Singh during 6th,7th semester (2021).

Date:

Roll No.	Name	Signature
101803207	Ishan Manuja	Ishan
101803209	Aadarsh Gupta	Aadarsh Gupta
101803222	Sameep Singh	
101803246	Harchetan Singh Chahal	Harchetan

Counter Signed By:

Mentor:

Jasmeet Singh

Dr. Jasmeet Singh

Assistant Professor,

Computer Science & Engineering Department

TIET, Patiala

ACKNOWLEDGEMENT

We would like to express our thanks to our mentor Dr. Jasmeet Singh. He has been of great help in our venture, and an indispensable resource of technical knowledge. He is truly an amazing mentor to have.

We are also thankful to Dr. Maninder Singh Head, Computer Science and Engineering Department, entire faculty and staff of Computer Science and Engineering Department, and also our friends who devoted their valuable time and helped us in all possible ways towards successful completion of this project. We thank all those who have contributed either directly or indirectly towards this project.

Lastly, we would also like to thank our families for their unyielding love and encouragement. They always wanted the best for us and we admire their determination and sacrifice.

Date: 16-12-2021

Roll No.	Name	Signature
101803207	Ishan Manuja	Ishan
101803209	Aadarsh Gupta	Aadarsh Gupta
101803222	Sameep Singh	Sameep
101803246	Harchetan Singh Chahal	Harchetan

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBR	**ABBREVIATIONS**
ROI	Region of interest
OCR	Optical Character Recognition
VQA	Visual Question Answering
T5	Text-To-Text Transfer Transformer
R-CNN	Recurrent Convolutional Neural Network

CHAPTER 1 - INTRODUCTION

1.1 Project Overview

Recent advances in computer vision and deep learning research have made enormous strides in many computer-focused tasks such as image classification, object detection, and functional recognition. Given enough data, Deep Convolutional Neural Networks (CNNs) can compete with human capabilities to perform image classification. However, the scope of these issues is limited and does not require a thorough understanding of the images. As human beings we can identify the objects in the image understand the spatial location of these objects, their properties and relationships to each other and also describe the purpose of each object given the surrounding context. We can ask arbitrary questions about images and communicate information obtained from them. Until recently, developing a computer vision system that could answer arbitrary natural language questions about images was considered an ambitious, but difficult goal. However, since 2014, there has been huge progress in developing systems with these capabilities. Visual Question Answering (VQA) is a computer vision task.

A text-based question about a film is asked and he must answer. Questions come unilaterally and have many sub-problems in computer vision, e.g.,

- Object Recognition - What's in the picture?
- Object Detection - Are there cats in the picture?
- Class Feature Classification - What color is the cat?
- Class Visual Classification - What is Sunshine?
- Count - How many cats are in the picture?

In addition to these, very complex questions such as the spatial relationship between objects (what is between the cat and the bed?) And the trivia reasoning questions (why is the girl crying?) can be asked. A solid VQA system requires the ability to solve a wide range of classical computer vision tasks, as well as the ability to reason from images. VQA has many potential applications. It helps blind and visually impaired people access information about images on the web and in the real world.

For example, when a blind user is scrolling through their social media feed, the caption system describes the image and can use VQA to query the image to learn more about the user view. In general, VQA can be used to enhance human-computer interaction as a natural way of questioning visual content. A VQA without the use of image meta-data or tags, the system can also be used for image retrieval. For example, to find all the pictures taken in the background of rain, we ask 'Does it rain?' From all the images in the dataset. Beyond applications, VQA is an important fundamental research issue. Since a good VQA system should be able to solve most computer vision problems, it is considered a part of the Turing test for image comprehension. The Visual Turing test rigorously assesses whether a computer vision system is capable of human-level semantic analysis of images. Passing this test requires a system that can perform many different visual tasks. VQA can be considered a kind of visual Turing test that also requires the ability to understand questions, but does not require more advanced natural language processing. If an algorithm works better or better on arbitrary questions about images, computer vision is largely solved compared to humans. But, this is only true if the benchmarks and evaluation tools are sufficient to make such bold arguments.

The newer studies on machine reading and comprehension have focused on textual comprehension, but have not yet reached the level of human understanding of visual layout and real-world document content. In our project, we will introduce a new visual machine reading comprehension project called HumanLike Visual Q&A System, in which a machine can answer a question in natural language using text from a picture, given a question and document image. The goal is to create intelligent agents who can answer questions. To achieve this goal, much attention has been paid to machine reading comprehension (MRC), a challenge that enables the machine to read and understand natural language texts so that it can answer questions. MRC capability is important for customers if they can be hired by automated assistants on e-commerce websites such as customer-service chatbots or assistive systems for reading professional literature. Here, most real-world documents are presented in plain non-text formats (e.g., HTML and PDF). However, current studies in MRC focus almost entirely on text-level comprehension, while neglecting visual layout and document content (text display, tables, charts, etc.).

The Visual Question Answer (VQA) on images containing a few words has recently been studied as a challenging task, at the intersection of vision and language comprehension. However, these learning tasks do not focus on document comprehension. Compared to current visual query (VQA) datasets containing text in images, human-like visual Q&A systems focus more on developing natural language comprehension and generation capabilities. Our dataset contains 30,000+ pairs of questions and short answers to 10,000+ document images taken from multiple domains of web pages. We are working with the latest model to expand the current sequence-to-sequence model with a large text corporation trained to take into account visual training layout and document content. Current methods combine visual information and textual features in one place. This is a step towards using large amounts of text and natural language processing methods to solve the VQA problem. Due to its numerous potential applications in autonomous agents and virtual assistants, it is attracting a large amount of interest in the computer vision and natural language processing (NLP) communities. To some extent, VQA is closely related to the text query answer (TQA, also known as machine reading comprehension),which asks the machine to answer questions based on a given paragraph of text. This may seem even more challenging due to the additional visual aids information.

1.1.1 Technical terminology

Visual Question Answering (VQA) system- It is different from simple question and answering system. Unlike simple question answering system where only context and question is given, in VQA context is made from visual layout or document image and answer is generated.

Region of interest(ROI)- an area of an image defined for further analysis or processing. Here ROI means segmenting into paragraph, data, sub data, title, etc.

Optical Character Recognition(OCR)- is a process of recognizing text inside images and converting it into an electronic form. This is used for data extraction from printed or written text from a scanned document or image file and then converting the text into a machine-readable form to be used for data processing like editing or searching.

1.1.2 Problem statement

Recent studies on machine reading comprehension have focused on text-level understanding but have not yet reached the level of human understanding of the visual layout and content of real-world documents. Machine Reading Comprehension is one of the key problems in Natural Language Understanding, where the task is to read and comprehend a given text passage, and then answer questions based on it.

1.1.3 Goal

Given a document and a question relevant to document we need to generate answer. The objectives of the project are:

- Analysing different existing techniques for VQA.
- Extracting different features to feed as input embedding.
- Building the encoder decoder model.
- Designing framework for effective interaction with the user.

1.1.4 Solution

The system should answer a question similar to humans in the following aspects:

- Learning the visual and textual knowledge from the inputs (image and question respectively).
- Combining the two data streams.
- Using this advanced knowledge to generate the answer.

1.2 Need Analysis

The long-term ambitious goal of artificial intelligence is to create intelligent agents who can answer questions along with the public. To achieve this goal, Machine Reading Comprehension (MRC) allows a machine to read and understand natural language lessons so that it can answer questions.

MRC's ability for users to read professional literature on the e-commerce website through automated assistants such as customer-service chatbots or support systems is important. Here, most real-world documents are displayed in plain non-text formats (eg, HTML and PDF).

Search engines display ranking lists of relevant documents in response to keywords created by users based on various factors such as popularity measurement and keyword matching. , Frequency of access to documents, etc. However, users do not really accomplish the task of obtaining information as they need to check each document one by one to get the information they want, which becomes a time consuming process. Ideally, search engines should respond to relevant and concise sentences with relevant web links. Current QAS seeks to answer questions asked by users in natural languages after receiving and processing information from various data sources such as the Semantic Web. The current study focuses on MRC, specifically ignoring text-level comprehension, visual layout and content (text display, charts, tables). Studies have shown that the main class of questions asked by visually impaired users in their surroundings is reading the text of the film.

1.3 Research Gaps

1. While predicting ROIs of given document, Accuracy is quite low for unstructured documents i.e when image is surrounded by text, model is unable to segment the image properly.
2. If the gaps between the paragraph is different in different documents, it starts making multiple rois corresponding to one paragraph.
3. The frequency of some of the class labels is quite less in training dataset(unbalanced), model rarely predicts for that corresponding class.
4. Unable to classify some texts

1.4 Problem Definition and Scope

Building AI systems that answer natural language questions about a given image has been one of the most trending areas of research in recent years. Most of these questions can be answered by humans without major challenges or difficulties. "How many books are on the shelf?"

In fact counting books and answering such a question as "2" is not a difficult task for humans.The AI system is a very challenging task to understand the meaning of the question, the meaning of the image and the relationship between the question and the image and try to answer the question.

However, with increasing practice in the areas of in-depth learning, computer vision and natural language processing, it is now possible to build a system that can answer these questions accurately and with good results.

Overall, a common problem in answering a visual question is to build a system (algorithm) that will take (ideally) any image and question asked in natural language about that image. Takes, and the film gives a natural language answer to that question.

1.5 Assumption & Constraints

The assumptions and constraint are as follow:

Table 1.1 – Assumptions and Constraints

S NO.	ASSUMPTIONS AND CONSTRAINTS
1.	Support is only for typed and structured data and handwritten data
2.	Data cannot be processed by the system in the proposed state.
3.	The scanned document should not contain any artifacts or stray marks.
4.	Understanding of maps and documents depends on the complexity of the data as well as on each specific representation, so there must be an appropriate compromise between accuracy, model size and processing speed.

1.6 Standards

The standards used for building our project are as follows:

Table 1.2 –Standards

		Specifications
	IEEE 1233	Guide for Developing System Requirements Specification
Design	IEEE 1016	Standard for Information Technology- Systems Design- Software Design Description
Implementation and acquisition of tools	IEEE1462	Software acquisition
Website development	W3C Web Design and Applications standards	Standards for building and rendering Web pages, including HTML, CSS, SVG, device APIs, and other technologies for Web Applications
Testing	IEEE 829	Software test documentation
	IEEE 1008	Software unit testing
Machine Learning	IEEE 3653	Standard for Technical Framework and Requirements of Machine Learning

1.7 Approved Objectives

Our main contributions will be:

1. to analyze different existing techniques for question answering system.
2. to extract different features required for answering asked question to feed as input
3. embedding for the model building the encoder decoder model for question answering system to design framework for effective interaction with the user.

1.8 Methodology Used

New focus and language work to read and understand the text we have given Visual MRC as a document image. We can simply describe the end-to-end task as:

When the question q and the document image I are given, a model produces an answer a. The Visual MRC Task is a productive MRC task such as an article QA in which the answer is not limited to word duration in context context. Image capture can be divided into two sub-tasks:

Phase 1-Region-Of-Interest (ROI) detection
When an image I is given, the model will detect a set of ROIs. Each ROI consists of a bounding box bi and a semantic class label li.

Phase-2 OCR
Given a ROI ri , a model detects a set of word objects within the region. Each word object consists of a bounding box bi,j and a form wi,j . Optical Character (OCR) is a mature sub-field of computer vision. An important part of this work is to bake in inductive biases and specialized modules like OCR into models to endow them with the different skills (e.g. reading, reasoning) required by the all encompassing task of VQA.

Phase -3 Model building ,Training and implementing
We will build encoder-decoder model architecture with a saliency detector that could find important tokens relevant to the question and is trained at the same time the sequence-to-sequence task is being learned.

A sequence of embeddings would be passed to the encoder. Special tokens corresponding to the semantic classes of ROIs are used for the token and segment embeddings.

Phase-4 Evaluation of model
We will use BLEU , METEOR, ROUGE-L to assess the quality of the generated answers. Coco-caption toolkit was used to calculate these scores. We can also use the F1 score of BERT Score, which is highly correlated with human judgment.

1.9 Project Outcomes and Deliverables

In this work, we will describe the visual information directly by text which unifies the input information in advance in text domain. The system will answer a question alike to humans with the following aspects:

1. it will learn the visual and textual knowledge from the inputs (image and question respectively)
2. combine the two data streams
3. use this advanced knowledge to generate

1.10 Novelty of Work

This proposed solution is very unique as there are no current services with the ease that we aim to provide them with. The project is based on the idea of making things authentic and more robust in real world scenarios. The project has a very wide future scope of further development and expansion.

CHAPTER 2 – REQUIREMENTS ANALYSIS

2.1 Literature Survey

2.1.1 Theory Associated With Problem Area

Designing a system capable of natural language response has been a very challenging goal in the past. Almost every human being can do such a task comfortably without any problems, however, creating a program close to these capabilities has proven to be closer to science fiction than current artificial intelligence. In recent years, Deep Learning (DL) has made great strides and a lot of research has been done on the Visual Question Answering System (VQA), we can say that systems that have these capabilities are close to being developed. Generally we can define a VQA system as an algorithm that takes an image as input and uses that image to generate an answer to the natural language question based on the image. From the perspective of research, this issue is multifaceted and multidisciplinary. We need an understanding of the NLP to understand the question as well as to come forward with the necessary answer. An additional challenge in VQA is that both search and reasoning must take place on the content of the image provided to the program. For example, the software should be able to identify objects to answer questions if there are any pets in the picture. Or if it is cloudy, it requires visual classification and global knowledge to answer which teams are playing in the image of the game. In addition, the system requires some common sense reasoning and knowledge reasoning. Similar tasks occur in the field of computer vision (CV) such as object recognition, object detection, scene classification and there has been a lot of progress over the last few years. Therefore, a good VQA system should have the ability to find solutions for a variety of NLP and CV tasks, as well as reason about the content of the problem images. Obviously, this is a multidisciplinary and complex AI issue involving CV, NLP, Knowledge Representation and Reasoning.

VQA has a lot of potential uses and applications. Some of its very likely applications include helping the people on the blindness spectrum to gain information about their environment.

A VQA System would also facilitate interaction with things like webpages and social media, one aspect where it is almost impossible for visually challenged people to interact meaningfully. A VQA System would also help in image retrieval. This also has a lot of potential for making social media or e-commerce more accessible or even more interactive. VQA can also be used to bring about innovation and advancement in education or recreational purposes. So far progress and research in this field has hinged heavily on the available data sets and defined and existing metrics. And how to best evaluate a VQA System is still a question that we have yet to explore the answer to, and it is also a likely possibility that new datasets and defined metrics will further expand on and deepen the notion of quality. As with any other field there are still discussions and more deliberation to be had before we reach an advancement point. For example, it is fair to use Multiple Choice datasets considering that a VQA System could be right by chance? And how would the real world application be of such a system where such options aren't available. Also, how does the formulation of a question impact the answer that is obtained. Current results and performance has a lot of distance to cover before becoming quite human-like but the results are already at the level where they can be deployed in carious real life applications. Gradually, as VQA is integrated into various platforms, devices and tools, it will definitely have a very big impact on how people interact with technology and visual data.

2.1.2 Existing Systems and Solutions

There have been a number of proposed systems over the last few many years. All methods have common features such as-

1. Image Featurisation
2. Question Featurisation
3. Algorithmic computation of features to produce an answer.

Treating VQA as a classification problem is a general approach to generating answers. This framework involves treating both the image and question features as input to a classification system and treating each unique answer as a distinct category. Although, these systems can vary significantly as to how they integrate the question and image features.

There have also been a number of proposed models for VQA that have had varying amounts of success. Some of them are -

1. Baseline Models

Baseline methods help evaluate the difficulty of a dataset, and determine a baseline performance benchmark that more sophistic algorithms should outperform. For VQA, the simplest baselines are set using random guessing and guessing most repeated answers.

2. Bayesian Models

Bayesian algorithm are trained to model the spatial relationships of the objects, which is used to compute each answer's probability. This was the earliest known algorithm for VQA, but its efficacy was surpassed by simple baseline models. A partial reason for this was its dependence on semantic segmentation's results which were imperfect.

3. Attention Based Models

Using only global features may obscure task-relevant regions of the input space. Attentive models attempt to overcome this limitation. These models learn to 'attend' to regions of higher relevance in the input space. They have achieved great success in vision and NLP tasks, object recognition, captioning and machine translation. The basic idea behind these models is that certain visual regions in an image and certain words from the question are more informative for answering a given question than other words.

4. Bi-linear Pooling Models

There are two prominent VQA methods that have used bilinear pooling. Multimodal Compact Bilinear (MCB) pooling was a novel method for combining image and text features in VQA. Approximation of the outer-product between the image and text features is the basic idea which allows deeper interaction between the two modalities. Later, MCB was said to be too computationally expensive by the authors, even by using approximated outer-product.

A multi-modal low-rank bi-linear pooling (MLB) scheme which would use linear mapping and the Hadamard product to approximate bilinear pooling was proposed instead. MLB rivaled MCB at VQA when it was used with a spatial visual attention mechanism, although with a complexity reduction.

5. Compositional VQA Models

In VQA, questions may need multiple levels of reasoning steps to be answered properly. To solve VQA in a series of sub-steps, two compositional frameworks were proposed. External question parsers were used in a Neural Module Network (NMN) framework to uncover the sub-task from the question, whereas Recurrent Answering Units (RAU) training happens end-to-end and sub-tasks are implicitly learned.

Some solutions that have been explored upto different extents are-

- **Medical VQA**

AI has been applied to medical image understanding and related med-VQA and this has attracted increasing interest from researchers. This topic is now opening up unexplored scenarios to support medical staff in taking clinical decisions, as well as enhanced diagnosis and computer-generated second opinions.

- **VQA for Visually impaired people**

VQA has often been proposed as a tool to aid people with visual impairments or people on the blindness spectrum. The ability of VQA to help in answering a lot of daily questions that visually impaired people face and would greatly assist the visually impaired people in living without visual barriers.

- **VQA in Video surveillance scenarios**

VQA can also be used to assist operators monitoring visual surveillance feeds to enhance their understanding of scenes and help them take fair and quicker decisions.

- **VQA Education and cultural heritage**

VQA generates a lot of interest due to its high correlation with human perception. For example, via an educational robot using VQA to formulate questions and start an educational dialog taking inspiration from the surrounding environment, the user can directly ask questions he/she is interested in and this can provide a greater and more involved learning experience for the user.

- **VQA and Advertising**

An advertisement needs simplicity to be understandable and also be interesting and eye-catching. VQA is very helpful in this field as it can replicate human reactions and evaluate the advertising capability. Using VQA to understand the advertisement and, in particular, the underlying communicative strategy can be a very routine and useful application.

2.1.3 Research Findings for Existing Literature

After going through various research papers, following were the observations we come across

Table 2.1 – Research Findings for Existing Literature

Technique	Description	Result
Text Visual Question Answering (TextVQA)[1]	TextVQA requires reading and understanding the text in the pictures to answer a question. The model needs to read and reason the text in the picture to answer the questions in it. To do this job well, models must first identify and read the text in the pictures. The model should then argue about it to answer the question.	TextVQA is for applications involving aiding visually impaired users – answering questions about everyday images. that involve reading and reasoning about text in these images.

Multimodal Multi-Copy Mesh(M4C)[2]	The M4C is an approach to the textVQA task with answer estimation based on pointer-enhanced multimodal transformer architecture. When the question and image are given as input, we summarize the feature representations in three ways - the question, the visible objects in the image and the text in the image.	M4C model answer beyond a single classification step and predicts it through pointer augmented multi-step decoder.
T5[3]	T5 aims to re-frame all NLP tasks into a unified text-to-text-format, where input and output are always text strings. The text-to-text framework allows the use of a single model, loss function, and hypermeters in any NLP. Functions including machine translation, document summary, question answer and classification tasks. We can also apply T5 to regression tasks by training to estimate the string representation of a number instead of a number.	To generate realistic text, T5 relies on a fill-in-the-blanks type task with which it is familiar due to the pre-training. The model also adjusts its predictions based on the requested size of the missing text.
BART[4]	BART is an auto-encoder recommended for pre-training of the sequence-to-sequence model. Bart is trained by contaminating the text with an arbitrary noise function and by learning a pattern to reconstruct the original text. It uses the standard transformer-based neural machine translation architecture. It predicts many noisy processes, finds the best performance by randomly changing the order of the original sentences and using the novel in-filling scheme, where the scope of the text is replaced with a single mask token.	BART is particularly effective when fine tuned for text generation but also works well for comprehension tasks.
OCR[5]	OCR systems convert two-dimensional images of text, including machine-printed or handwritten text, into machine-readable text from its image representation. This remains a challenging issue when text is in an unrestricted environment such as landscapes due to geometric distortions, complex backgrounds and different fonts.	Identify and capture all the unique words using different languages from written text characters

Tesseract[6]	Tesseract is an open source OCR engine licensed under the Apache 2.0 License. It can use the API directly or (for programmers) to extract printed text from images. It supports a wide variety of languages. Tesract is compatible with many programming languages and frameworks through wrappers.	It is used with the existing layout analysis to recognize text within a large document.
Faster R-CNN[7]	In fast R-CNN, we feed the input image to CNN to create a convincing feature map. From the convolution feature map, we identify the field of proposals and square them and resize them to a constant size to feed them in a fully integrated layer using the Roy pooling layer. From the ROI feature vector, we use the Softmax layer to estimate the offset values of the square and boundary boxes of the proposed area.	This is faster than R CNN. It's because we don't have to feed 2000 region proposals to the CNN every time. Instead, the convolution operation is done only once per image and a feature map is generated from it.
Evaluation Metric 1) BLEU[8] 2) METEOR[9]	BLEU is an accurate-based evaluation metric that considers accurate n-gram matches. For a given value of n, the accuracy is calculated differently than n-grams in the product hypothesis, which corresponds to a few n-grams in the reference hypothesis. The final BLEU score is calculated as the geometric average of n-gram accuracy, which varies from n to 1 to N, where N is usually 3 or 4. METEOR uses both accuracy and recall, that is, it calculates the fraction of the hypothesis that matches the prediction (exact) as well as the fraction of the prediction in the hypothesis (recall). It also considers matching with keywords, synonyms and paraphrases. This function assigns different weights to match the words and the content words. The final score is	

3) ROUGE-L[10] 4) CIDEr[11]	the calculated accuracy and harmonic average of the recall based on these four matches. In addition, it also includes a fragmentation penalty for gaps and differences in word order. ROUGE-L is an F-measure based on the long common denominator (LCS) between the candidate and the target sentence. Looking at two scenes, a simple adaptation is a set of words that appear in the same order in two sequences, but unlike a n-gram the general adaptation does not have to be in a row. CIDer's goal is to automatically assess the image to see how well the candidate's sentence fits the consensus of the image description set.	

2.1.4. Problem Definition

There are many problem areas that still interfere with user machine interaction due to some limitations in the way machines receive input. A good VQA system can solve these problems. Current VQA systems are limited in that they take information from rich text, such as a museum or classroom or webpage that is commonly found in real life. This is a major obstacle to the integration of VQA across many application areas. It makes a lot of sense to consider solving this problem with the current state of VQA as the next step of user-machine interaction. The next advance in VQA is learning to draw doubts from the layout of text, images and other visual elements and their content.

There are a variety of problems which can help a lot to solve or correct. Search engines like Google do not refer to search results from images, which can lead to reliance on text with images to provide context, or the entire content as text. It does not have the ability to answer questions from text, pictures or image input. More recently, over the years, Google Assistant and Alexa have also become better question-and-answer systems to help users, but they are still far from VQA.

If this capability is added to search engines they can give more relevant results as well as answer questions directly instead of pattern, text matching and similar reverse image searches on the web. It also fundamentally changes the way blind or visually impaired people interact with technology and the way they access information. By questioning the VQA system they can get to know what is on or around their screen.

Currently visually impaired it is impossible for any visually impaired person to work and allow VQA to interact with such information once it has been implemented. This means a radical change in the lives of those with access to and access to the Internet. A good VQA system helps competent people, such as chatbot, to use context from the user's screen and reduce human dependence and increase efficiency in many tasks that are limited by the amount of data input that needs to be filtered. Does. Chatbots, simple VQA systems and other applications can have a huge impact on how we manage and process our data.

2.1.5 Survey of Tools and Technologies used

1. Pytorch

Pytorch is an open source (under revised BSD license) machine learning library based on the torch library that has applications in natural language processing and computer vision, primarily developed by Facebook's AI Research Lab. It is primarily an optimized tensor library used for deep learning applications using GPU and CPU.

2. Tensorflow

TensorFlow is a machine learning focused open source. It has applications in a wide variety of tasks, but its special focus is on the training and induction of deep neural networks. Tensorflow is a symbolic mathematical library based on data flow and differential programming. It is an open source Artificial Intelligence Library that uses data flow graphs to create models. It is used by developers to create multilayered large-scale neural networks. Tensorflow is mainly used for classification, perception, perception, discovery, evaluation and construction

3. Detectron

Detectron is a Facebook AI Research (FAIR) software system that implements sophisticated object detection algorithms, including Mask R-CNN. It is written in Python and is powered by Caffe 2 Deep Learning Framework. This library allows easy to use and build object detection, instance segmentation, key point detection and panoptic segmentation models. It has a simple, modular design that makes it easy to rewrite the script for another data-set.

4. Google Colab

Google Colab is an incredible online browser-based platform that allows you to run the Jupyter notebook environment entirely in the cloud. It is a powerful platform for learning and developing machine learning models in Python. Team members can also share and edit notebooks remotely simultaneously. Notebooks can also be published on GitHub and shared with the general public. Colab supports many popular ML libraries such as Tensor Flow, Pytorch, Keras and OpenCV.

5. Fast RCNN (part of the model)

The procedure is similar to the R-CNN algorithm. But, instead of feeding field proposals to CNN, we feed the input image to CNN to create a convincing feature map. From the convolution feature map, we identify the field of proposals and square them and resize them to a constant size to feed them in a fully integrated layer using the RoI pooling layer. From the ROI feature vector, we use the Softmax layer to estimate the offset values of the square and boundary boxes of the proposed area.

6. CNN

Convolutional neural network, or CNN, is an in-depth learning neural network designed to process structured sequences of data, such as images. Convolutional neural networks are widely used in computer vision and have become sophisticated for many visual applications such as image classification and have also had success in natural language processing for text classification.

7. OCR

OCR systems convert two-dimensional images of text, including machine-printed or handwritten text, into machine-readable text from its image representation. Optical character recognition can be challenging when text is in an unrestricted environment such as landscapes due to geometric distortions, complex backgrounds and different fonts. OCR as a process usually involves several sub-processes to perform as accurately as possible. Image preprocessing, text localization, character segmentation, character recognition, post processing.

2.1.6 Summary

In our vqa model we will try to train computer vision modality using faster rcnn by segmenting document into 9 ROIs and giving OCR along with positional segment and location embedding to T5 nlp question answering model to combine both modalities which is less explored or current topic of research.

2.2 SYSTEM REQUIREMENT SOFTWARE

2.2.1. Introduction

2.2.1.1. Purpose

This report provides technical documentation regarding a web application named “Smart DOC VQA Web Application”. This technical report goes through all features and details of the functionalities and working environment in which the application can be executed. This report also provides the overall description of the application with essential features and drawbacks in the application. It will explain the utility and a complete abstract for the development of system. It will also explain system constraints, interface and interactions with environment.

2.2.1.2. Intended Audience and Reading Suggestions

Audience who are familiar with HTML, CSS machine model implementation through Keras and TensorFlow such as Developers, testers will be able to comprehend this report comparatively better as this is a technical report. For Public Audiences or General Public in mind there is a Glossary at the end of the report which defines the technical terms used in the report for a better understanding of the report. The report is formatted in such a way, which starts with the documentation of SRS (software requirement specifications) and then with the application interfaces & features of the application.

Suggested reading of this technical report is to start with the overall description section, of this document which gives an overview of functionality of the product. It specifies the informal requirements and establishes context before the specification of technical requirements in the next chapter. This section is for the users, project managers, marketing staff and document writers.

The third chapter is the requirements specifications section of this document and is written primarily for the developers where it describes in the technical terms, the details of the functionality of the product.

The fourth chapter, other non- functional requirements section, of this document is written for the users and the developers as well and describes the safety and security policies.

2.2.1.3. Project Scope

Creating intelligent agents that can answer questions as well as people is a long-cherished goal of artificial intelligence. To achieve this goal, machine reading comprehension is a great challenge to enable a machine to read and understand natural language texts so that it can answer questions, asked by user related to the context.

Current studies in this domain, majorly focus on text-level understanding, while ignoring the visual layout and content like text appearance, charts, tables.

In studies it has been discovered that mostly question asked by visually impaired users involves images in the surrounding which involves reading text from image. But today's visual question answering system cannot read on its own.

The MRC capability can be valuable to users when employed by automated assistants like customer-service chat bots on e-commerce website or assistant systems for reading professional literature. In this many real-world documents are provided in non-plain text formats.

When we search on internet, we get relevant documents in order and ranked by user deduced keywords based on various aspects such as popularity measures, keyword matching, frequencies of accessing documents, etc. Howsoever, they don't really the fulfil the task of information retrieval as documents still have to be examined one by one to obtain the required information, it makes information retrieval a time consuming process. In an ideal situation, a search engine should return few relevant and concise sentences as answers together with their respective web links.

2.2.2. Overall Description

2.2.2.1. Product Perspective

Smart Doc VQA system is Deep learning Web application implemented using Keras and TensorFlow on Flask framework. The Web app is meant to answer the question asked by user related to document uploaded by the same user.

The main task of the algorithm can be divided into two sections: Firstly, it should accurately get the features of the document and question that is asked from user.

The second part of the algorithm is to get answer from features using Answer generating module. The features of the document are extracted firstly by dividing the Document into 9 ROI's and than extracting text from ROI 's along with image if there in document and is labeled by model. These text and image are processed and given in form in which it can be given to answer generating module.

Answer Generation Module is the main core of project which is trained on 10,000 images collected from different blogs, news article on internet. The input to the model will be features collected from first module.

This module is under domain of natural language processing unlike the upper module which was purely computer vision module.

2.2.2.2. Product Features

1) Register of the player

Those who are new to app are required to register and then login to get benefit of it.

2) Login in the web app:

For those who have been already registered, can enter their valid e-mail and can get benefit of it.

3) Upload the document:

The document from which user want to ask question need to be uploaded to the system.

4)ROI and OCR

The document uploaded by user will be processed at back-end where the document will be segregated into 9 ROI 's and using OCR corresponding text will be given to answer generation module.

5)Ask Question:

User can ask question that should be related to document.Answer Generating module will answer question from features collected from document.

6)Display answer

The answer to the asked question will get displayed in a DISPLAY BOX created in web app interface and also be mailed to the user 's email.

7)Reset

When u are over with one question and want to upload new document to ask question then reset your problem and start again by uploading new document and ask ques related to new document.

2.2.3. External Interface required

2.2.3.1. User Interfaces

The software provides good graphical interface for the user and the user can operate on the system, performing the required task like scanning, solving, getting help or getting the full solution.

2.2.3.2. Hardware Interfaces

Processor: Pentium(R) Dual-core CPU

Hard Disk: 40GB

RAM: 256 MB or more

Other requirements such as processor, memory which is already mentioned under Operating environment section.

2.3.3.3 Software Interfaces

The software package is developed using HTML and CSS, backend is done by Flask and the Deep Learning model is trained using Keras and implemented using TensorFlow.

Operating System: Window XP, Windows 7 and higher version

Language: HTML, Flask, CSS, HTML,

Database: SQ Lite

2.2.4 Other Non-Functional Requirements

2.2.4.1 Performance Requirements

The performance of the application is measured while it is played which means it evaluated on real time performance. From a user perspective, average page load must be less and slowest page load should take more than 1min. The app must be available most of the time.

2.2.4.2 Safety Requirements

Code safety ensures whether software is reliable and safe to use. Being an application, which contributes for the welfare of the society and also not demanding any other personal information except email-id for login, therefore any specific safety measure was not encountered.

2.2.4.3 Security Requirements

If security of the vendor is high priority he could use database partners carefully.

2.3 Cost Analysis

Following is the cost analysis of the project:

Table 2.2 – Cost Analysis

	Usage	Cost
GPU	For Running machine Learning Model	₹6000- ₹7000
AWS	Deploying web Application	₹500-₹1000

2.4 Risk Analysis

There are a many risk factors included in the building and completion of this project like: the device will use a large amount of processing power and will need some noticeable amount of time to respond. Another risk factor taken into consideration is that the model might produce some unnecessary text in some cases until it attains the desired accuracy. Even the faster R-CNN model has a limit to its accuracy. We need high computational power to train our model which can be counted as risk factor at time when we had to deploy and connect components with each other.

CHAPTER 3 - METHODOLOGY ADOPTED

3.1 Investigating Techniques

3.1.1 Document image layout analysis

There are many proposed methods for document image layout analysis.
Classification can be done into three different subgroups:

(i) region or block based classification methods- This will segment out document zones from a document image, and then categorise them into meaningful semantic classes
(ii) pixel based classification methods- This will take each individual pixel into account and use a classifier for labelled image generation with regions hypotheses.
(iii) connected component classification methods- Use local information to create object hypotheses that are further inspected, combined and refined, and finally classified.

When it comes to image classification, convolutional neural networks (CNNs)) have been widely adopted including document analysis. It inherent very intense computational burden usually limits the cost-benefit of using them in document storage and retrieval applications where low memory and fast processing are vital.

Block based classification method that consists of three steps:

i) pre-process a document input image and segment it into its blocks of content
ii) use their vertical and horizontal projections to train a CNN model for multi-class classification considering text, image and table classes
iii) using a pipeline including the trained CNN model to detect the layout of new documents.

Methodology

1) Segmenting blocks of content in the document image

i) Single pages are converted into gray-scale images.

ii) Then processed by the nonlinear, run-length smoothing algorithm to detect regions with high chance of containing information. The application of the algorithm is done in both horizontal and vertical directions and combination of the resultant binary images using the operator AND.

iii) Next, a 3 × 3 dilation operation is performed two times over the resulting binary image to create blocks of content.

iv) Finally, we iteratively detect the largest connected component in the binary image and denote it as a block of content. Until no more connected components are found in the image the detection process continues.

2) Classifying content blocks from the document image

Use the CNN model to classify into three different classes: text, table and image.

Two different CNN architectures: a two dimensional approach commonly used in different computer vision problems, which is used as the baseline and the fast one dimensional architecture proposed here allows very little data consumption and processing. One quantity uses estimates to give similar results over time.

3.1.2 Run Length Smoothing Algorithm (RLSA)

Run length smoothing algorithm (RLSA) is a method used for block segmentation and text differentiation. The method developed for the document analysis system consists of two steps. First, the partitioning process divides the area of a document into regions (blocks), each of which must contain the same type of data (text, graphic, halftone image, etc.). Next, some basic properties of these blocks are calculated.

The basic RLSA is applied to a binary sequence, represented by white pixels 0 and black pixels 1.

The algorithm converts the binary sequence x to the output sequence y according to the following rules:

- If the number of adjacent 0s is less than or equal to the pre-defined range C, then x is converted to 1 in y.
- Does not change in X to 1 y.

When applied to sample arrays, the RLSA combines the adjacent black areas with a minimum distance of C pixels. With proper selection of C, the joined fields become common data type fields.

The RLSA document is applied row-by-column and column-by-column, giving two different bit-maps. Because of the difference in the spacing of the document segments horizontally and vertically, we use different values of c for row and column processing. Bit-maps are combined into logic and operations. Additional horizontal sensitivity using RLSA gives the final splitting result.

3.1.3 Faster R-CNN

An RPN (Region Proposal Network) is a complete convolutional network that captures object bound and objectivity scores simultaneously at each location. RPN trains end-to-end to generate high-quality field resolutions, which are then used to identify fast R-CNN. We integrate RPN and Fast R-CNN into a single network, sharing their convincing features.

Selected search greedily merges super pixels based on low-level engineering features. Conventional feature maps used by area-based detectors such as Fast RCNN can also be used to generate field resolutions. On top of these convolutional features, we build the RPN by adding a few additional convoluted layers that reclaim the area and objectivity score at each location in the regular grid. RPN is therefore a kind of complete convincing network (FCN) and is trained end-to-end specifically for the task of formulating identification proposals. Object resolution methods are based on superpixels (eg selective search, CPMC, MCG) and sliding windows.

The R-CNN method trains CNN end-to-end to classify proposal fields into object categories or themes. R-CNN plays primarily as a classifier, and it does not override object boundaries (except for refinement by box regression boundary). Its accuracy depends on the performance of the field proposal module.

3.1.4 Detectron2

Detectron 2 is Facebook's new Vision Library, which allows you to easily build and build object detection, instance segmentation, key point detection and panoptic segmentation models. It has a simple, modular design that makes it easy to rewrite the script for another data-set.

It is a Pythorch based modular computer vision model library. This is the second iteration of the detector, in fact it was written in Cafe 2. The Detectron 2 system allows you to plug in custom positioning of custom computer vision technologies into your computer workflow. Detectron 2 includes all the models available in the original detector, including R-CNN, Mask R-CNN, Retinnet and Denspos. It also has several new models, including the Cascade R-CNN, the Panoptic FPN and the Tensor Mask.

3.2 Proposed Solution

For training object detection model for custom dataset, we need to start by building a model using a Feature Pyramid Network in combination with a Region Proposal Network if we choose region proposal based methods such as Faster R-CNN or one-shot detector algorithms like SSD and YOLO can be used. Implementation of either of them is complicated to work with if implementation is done from scratch. We need a framework where we can use state-of-the-art models such as Fast, Faster, and Mask R-CNNs with ease. Regardless, building a model from scratch is necessary to understand the math behind it.

Detectron2 is very useful if we need to train a model for object detection utilising a custom dataset. All the various models in the model zoo of the Detectron2 library are pre trained using the COCO dataset. The only requirement is to fine tune the custom dataset on the pre-trained model. Detectron2 is a total remake of the first Detectron that was released in 2018. Detectron was written on Caffe2, a Facebook-backed deep learning framework. Caffe2 and Detectron are now deprecated. Caffe2 now is integrated within PyTorch and Detectron2 is fully written on PyTorch. Detectron2 offers numerous types of Objection Detection. Object Detection can be done on any custom dataset using seven steps.

Step 1: Installing Detectron 2

This can be started with installing a couple dependencies like Torch Vision and COCO API and check if CUDA is available. CUDA is used to stay in track of the presently selected GPU and install Detectron2.

Step 2: Prepare and Register the Dataset

After importing a few necessary packages. All the datasets that can be used with detectron2 are listed in datasets. In case a custom dataset needs to be used alongside detectron2's loaders, then the dataset will have to be registered. There are a number of formats in which the data can be fed to a model including the YOLO format, COCO format etc. For detectron2 the COCO format is used to feed data. This format is in the form of a JSON file which contains all the details of the image such as size, annotations (i.e. coordinates of the bounding box), labels etc.

Step 3: Visualize the Training Set

3 random pictures can be selected from the train folder of the dataset and the bounding boxes can be visualized.

Step 4: Training the Model

In this step the configurations are provided and the model is now ready to get trained. The fine tuning of the model is done on the dataset as pre-training of the model is already done on the COCO model. For the purpose of object detection, there are a ton of models available in Detectron2's Model Zoo.

Step 5: Inference using the Trained Model

Now, we need to infer the results by testing of the model on the Validation Set. An output folder is saved in the local storage after successful completion of training for storing the final weights.

Step 6: Evaluation of the Trained Model

The evaluation of the model is usually done using the COCO standards of evaluation. To evaluate the performance of the model we use Mean Average Precision (mAP).

3.3 Work Breakdown Structure

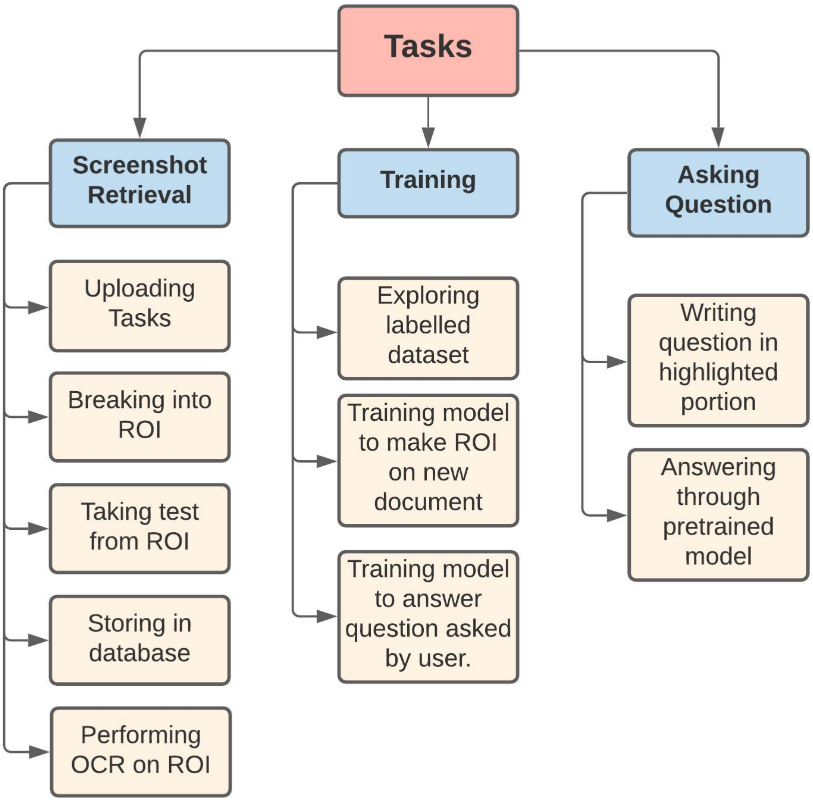

Tasks
Screenshot Retrieval
Training
Asking Question
Uploading Tasks
Breaking into ROI
Taking test from ROI
Storing in database
Performing OCR on ROI
Exploring labelled dataset
Training model to make ROI on new document
Training model to answer question asked by user.
Writing question in highlighted portion
Answering through pretrained model

Fig 3.1-Work Breakdown Structure

3.4 TOOLS AND TECHNOLOGY

R-CNN

This is used to overcome the problem of selecting a large number of fields. It uses optional search to extract only 2000 areas from an image designated as area resolutions. So, now, instead of categorizing a large number of regions, it only works with 2000 regions. These 2000 field proposals are made using the selective search algorithm. In addition to estimating the presence of an object in the area proposals, the algorithm estimates four values as offset values to increase the accuracy of the boundary box.

Faster-RCNN

The procedure is similar to the R-CNN algorithm. But, instead of feeding field proposals to CNN, we feed the input image to CNN to create a convincing feature map. From the convolution feature map, we identify the field of proposals and square them and resize them to a constant size to feed them in a fully integrated layer using the Roy pooling layer. From the ROI feature vector, we use the Softmax layer to estimate the offset values of the square and boundary boxes of the proposed area.

OCR

OCR systems convert two-dimensional images of text, including machine-printed or handwritten text, into machine-readable text from its image representation. Optical character recognition can be challenging when text is in an unrestricted environment such as landscapes due to geometric distortions, complex backgrounds and different fonts. OCR as a process usually involves several sub-processes to perform as accurately as possible. Image preprocessing, text localization, character segmentation, character recognition, post processing.

TextVQA

TextVQA requires reading and understanding the text in the pictures to answer a question. TextVQA is based on the custom pairing approach between a pair of two methods and is limited to a single predictive stage by transmitting textVQA as a classification task.

T5

Unlike the BERT-style model, which uses only the class label or input, the T5 is used to re-frame all NLP tasks into a unified text-to-text-format. You can output duration. The Text-to-Text Framework allows the use of a single model, loss function, and hyperparameters on any NLP task, including machine translation, document summary, query answer, and classification tasks (e.g., sentiment analysis). We can also apply T5 to regression tasks by training to estimate the string representation of a number instead of a number.

BART

BART is an auto-encoder recommended for pre-training of the sequence-to-sequence model. Bart is trained by contaminating the text with an arbitrary noise function and by learning a pattern to reconstruct the original text. It uses the standard transformer-based neural machine translation architecture. It predicts many noisy processes, finds the best performance by randomly changing the order of the original sentences and using the novel in-filling scheme, where the scope of the text is replaced with a single mask token.

M4C

The multimodal multi-copy mesh is an approach to text VQA work with a reproducible answer estimate based on a pointer-augmented multimodal transformer architecture. When the question and image are given as input, we summarize the feature representations in three ways - the question, the visible objects in the image and the text in the image. These three methods are referred to as the Query Word Attributes List, the Object Attributes list visible from the Off-the-Shelf Object Detector, and the OCR token attribute list based on the external OCR system. Model projects consist of representations of entities (query words, searched objects and identified OCR tokens) in three modes as vectors in the common embedding space learned. Then, a multi-layer transformer is applied to a list of features that override all features, enhancing their representation with intra- and intermodality references. The model learns to evaluate the answer by repeat decoding with a dynamic indicator network. During decoding, it feeds into the previous output to evaluate the next answer segment in an autoregressive manner. At each stage, it either copies the OCR token from the image or selects a word from its default answer vocabulary.

CHAPTER 4 - DESIGN SPECIFICATIONS

4.1. System Architecture

4.1.1. MVC Architecture

The controller is a agent between model and user which coordinates provides services asked by user. Model is main part of the project to generate answer and view provides interface of application.

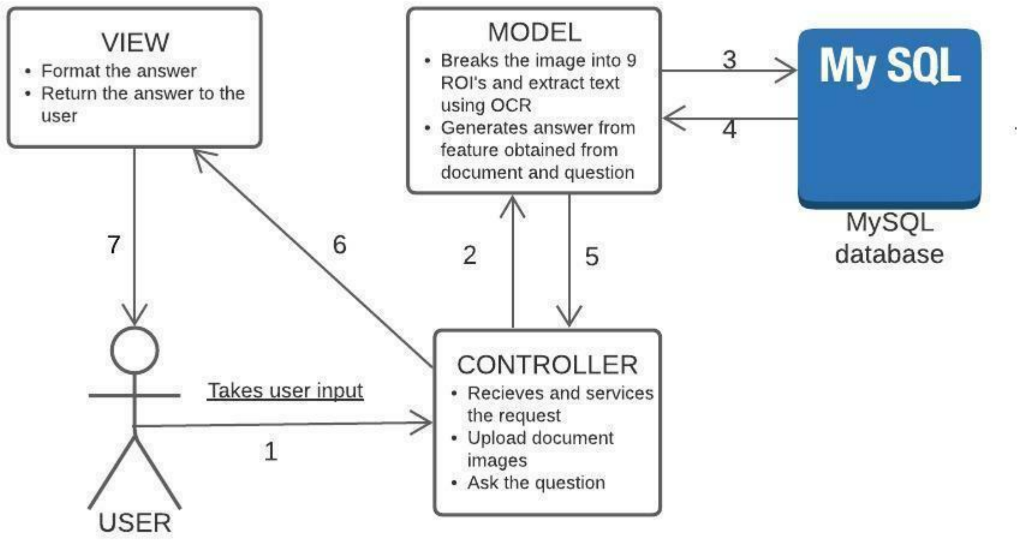

VIEW
• Format the answer
• Return the answer to the user
MODEL
• Breaks the image into 9 ROI's and extract text using OCR
• Generates answer from feature obtained from document and question
My SQL
MySQL database
3
4
7
6
2
5
CONTROLLER
• Recieves and services the request
• Upload document images
• Ask the question
Takes user input
1
USER

Fig 4.1- MVC Architecture

4.2. Design Level Diagrams

4.2.1. Data Flow Diagrams

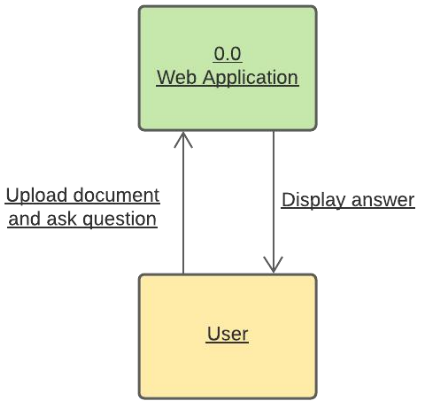


Fig 4.2 – Data Flow Diagram Level 0

Data-flow diagram is a way of representing a flow of data through a process or a system. It includes data given by user to data needed by the model.

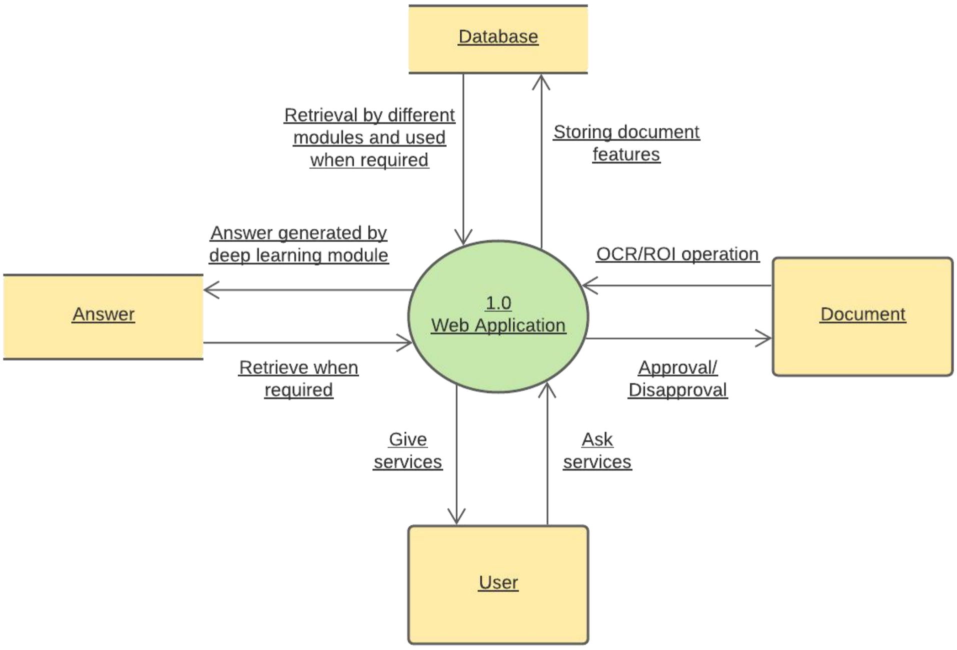


Fig 4.3 – Data Flow Diagram Level 1

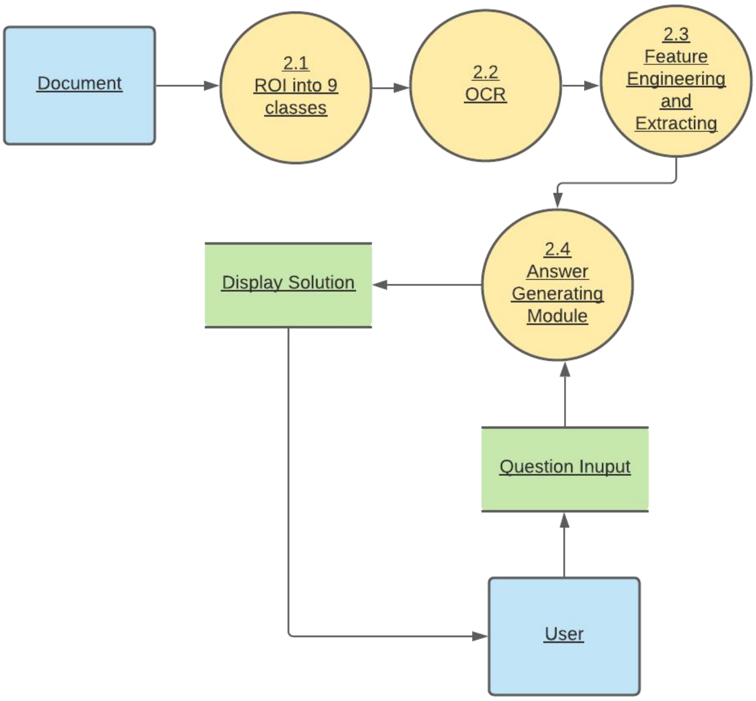


Fig 4.4 – Data Flow Diagram Level 2

4.2.2. Class Diagram

Class diagram shows the relationship between modules and type of relationship between them.

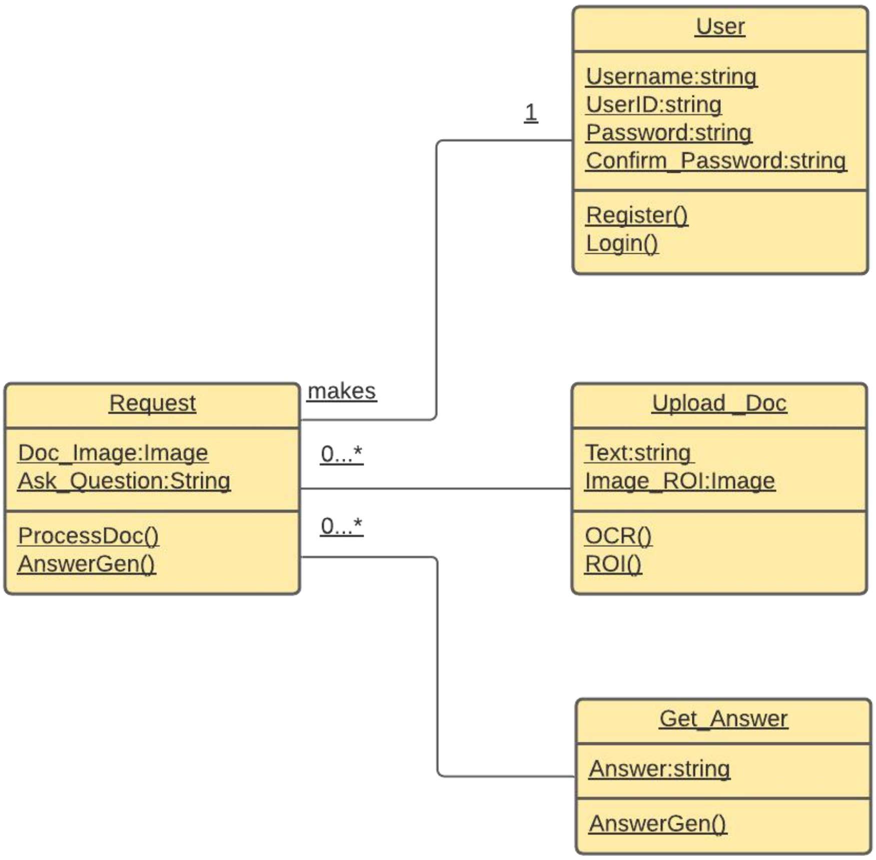


Fig 4.5 – Class Diagram

4.2.3. Sequence Diagram

Sequence diagram gives the sequence in which user will interact and calling of different modules at different time will be done. First the user login to application.If not registered than he first register to application. And than he upload document to portal with question that is related to document .

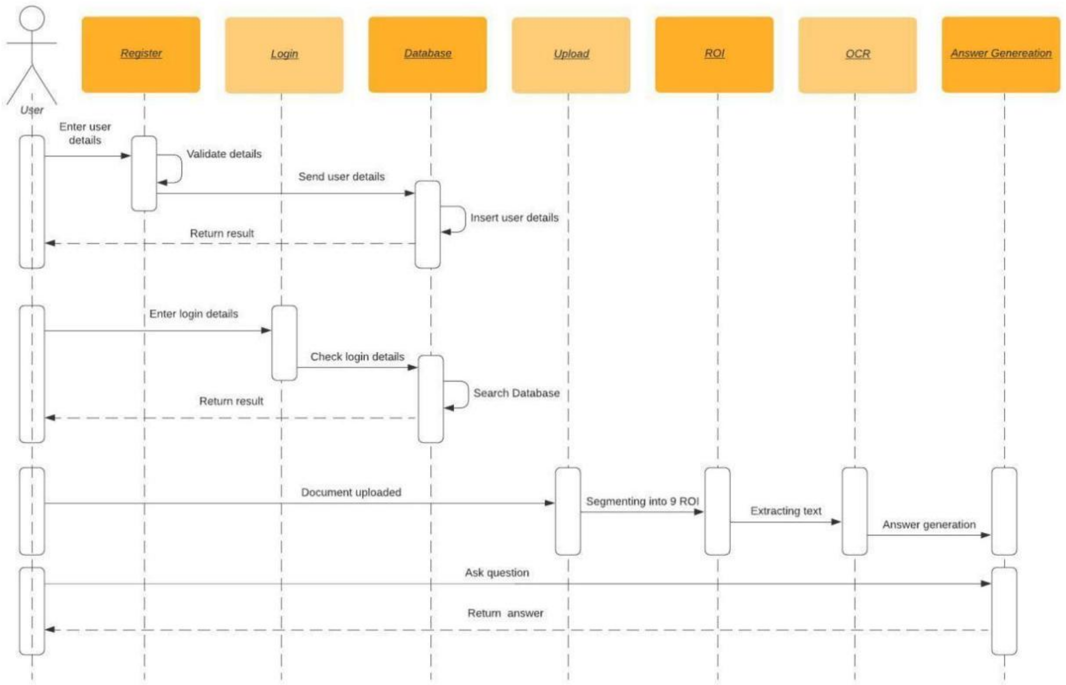


Fig 4.6 – Sequence Diagram

4.3. User Interface Diagram

4.3.1. Activity Diagram

User first login to profile or get register to application. Registered user than enter their credentials and get login to use application. Next, user uploads screenshot, the application validates it. Then user ask question related to document and application displays the generated answer.

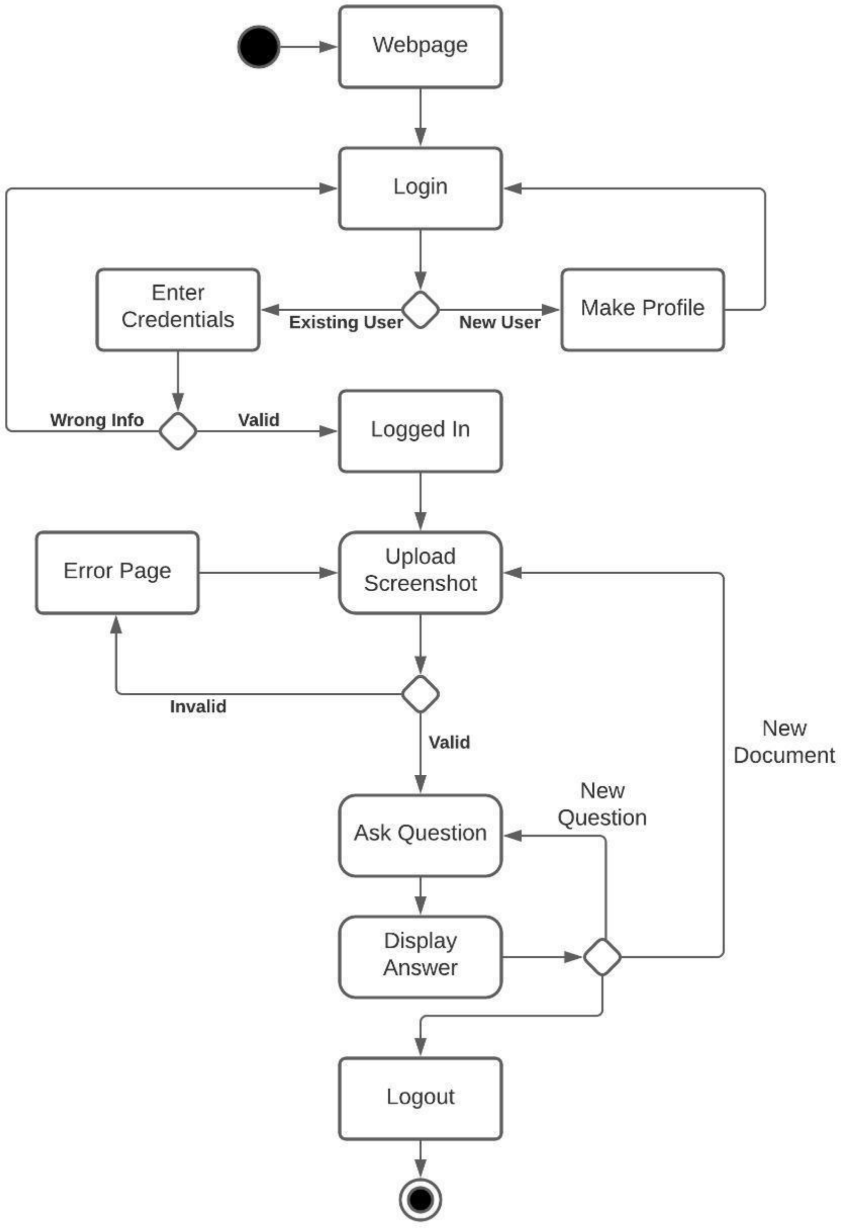
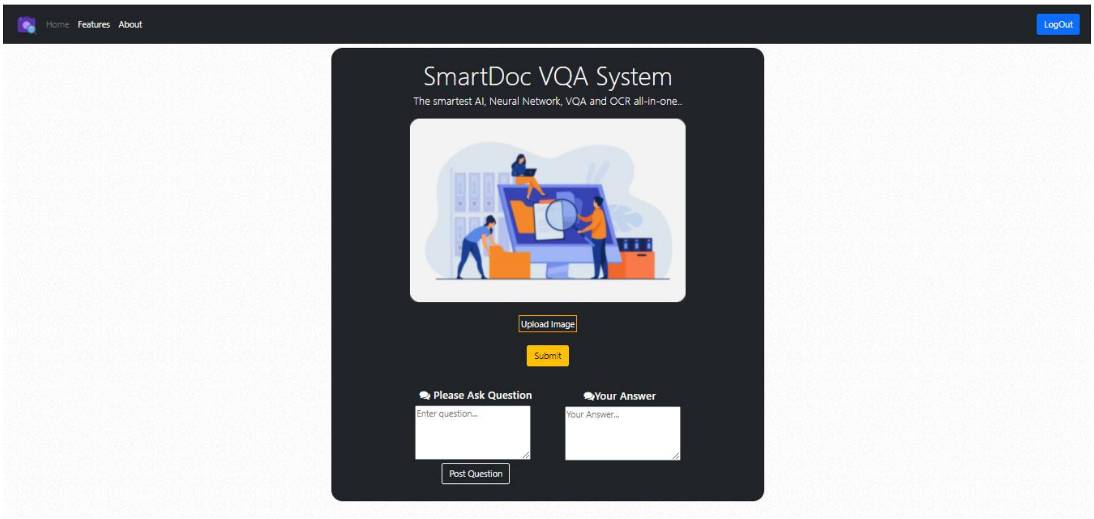


Fig 4.7- Activity Diagram

CHAPTER 5 -
IMPLEMENTATION AND EXPERIMENTAL RESULTS

5.1 Experimental Setup

The experimental setup for the proposed Visual MRC requires a device with (camera, mic and speakers) and good computational power to run the system properly. The device should also be able to prove the required support for running Python 3.0, pytorch and detectron2.


SmartDoc VQA System

The smartest AI, Neural Network, VQA and OCR all-in-one..

Fig 5.1 Experimental Setup of the Desktop App

5.2 Experimental Analysis

Experimental analysis is the process of evaluating the implemented system against several perspectives. For the proposed project, there are several desired outputs which the final system should produce. Thus, the experimental analysis basically checks the output of the system with respect to the expected output when it is provided with various inputs, i.e., the processed document image, text. During analysis, tests are performed to test the capacity of the system as well as its performance under various circumstances. The system is also checked in terms of the precision of the generated outputs.

5.2.1 Data

The document image is required for ROI and OCR. User is asked to upload document from which he wants to ask question. Faster RCNN module first divides the document into different ROI i.e., title, paragraph, table, caption etc. After dividing the image into 9 regions of interest, the OCR will get text with 9 labels and give it to as context to question answering module. Then answer is generated from the question asked by user related to document.

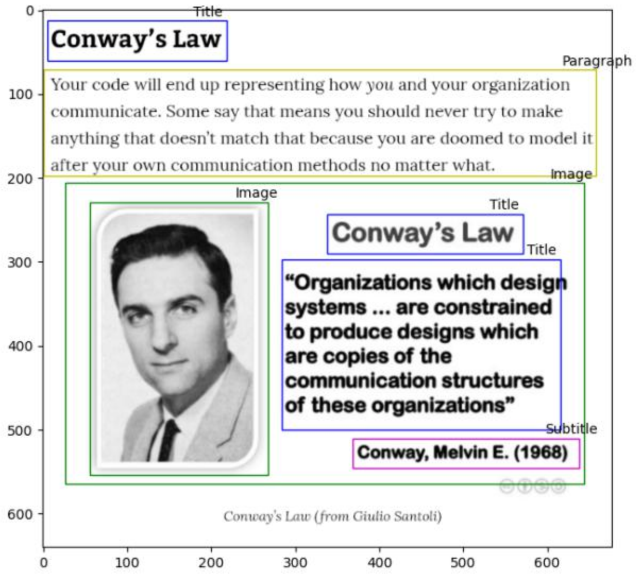
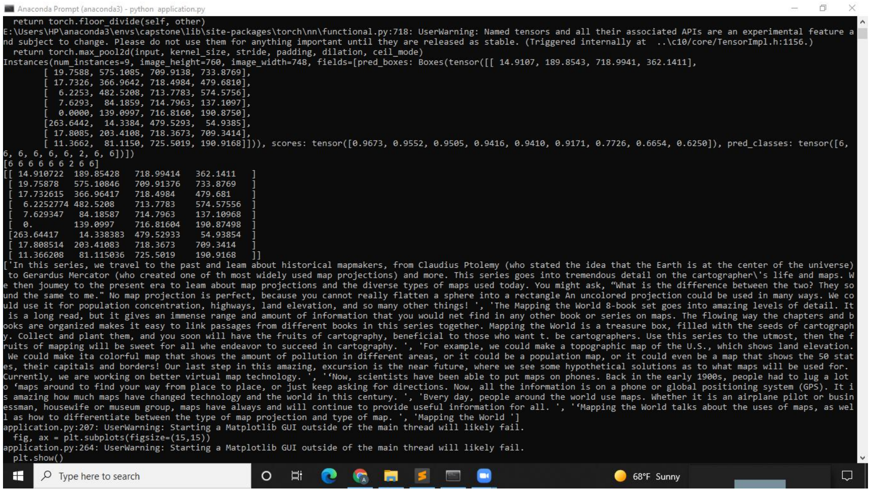

0
100
200
300
400
500
600
0
100
200
300
400
500
600
Title
Conway's Law
Paragraph
Your code will end up representing how you and your organization communicate. Some say that means you should never try to make anything that doesn't match that because you are doomed to model it after your own communication methods no matter what.
Image
Image
Title
Conway's Law
Title
"Organizations which design systems ... are constrained to produce designs which are copies of the communication structures of these organizations"
Subtitle
Conway, Melvin E. (1968)
Conway's Law (from Giulio Santoli)

Fig 5.2 Processed Image



```
  return torch.floor_divide(self, other)
E:\Users\HP\anaconda3\envs\capstone\lib\site-packages\torch\nn\functional.py:718: UserWarning: Named tensors and all their associated APIs are an experimental feature a
nd subject to change. Please do not use them for anything important until they are released as stable. (Triggered internally at  ..\c10/core/TensorImpl.h:1156.)
  return torch.max_pool2d(input, kernel_size, stride, padding, dilation, ceil_mode)
Instances(num_instances=9, image_height=760, image_width=748, fields=[pred_boxes: Boxes(tensor([[ 14.9107, 189.8543, 718.9941, 362.1411],
        [ 19.7588, 575.1085, 709.9138, 733.8769],
        [ 17.7326, 366.9642, 718.4984, 479.6810],
        [  6.2253, 482.5208, 713.7783, 574.5756],
        [  7.6293,  84.1859, 714.7963, 137.1097],
        [  0.0000, 139.0997, 716.8160, 190.8750],
        [263.6442,  14.3384, 479.5293,  54.9385],
        [ 17.8085, 203.4108, 718.3673, 709.3414],
        [ 11.3662,  81.1150, 725.5019, 190.9168]])), scores: tensor([0.9673, 0.9552, 0.9505, 0.9416, 0.9410, 0.9171, 0.7726, 0.6654, 0.6250]), pred_classes: tensor([6,
6, 6, 6, 6, 6, 2, 6, 6])])
[6 6 6 6 6 6 2 6 6]
[[ 14.910722  189.85428   718.99414   362.1411   ]
 [ 19.75878   575.10846   709.91376   733.8769   ]
 [ 17.732615  366.96417   718.4984    479.681    ]
 [  6.2252774 482.5208    713.7783    574.57556  ]
 [  7.629347   84.18587   714.7963    137.10968  ]
 [  0.         139.0997    716.81604   190.87498  ]
 [263.64417    14.338383  479.52933    54.93854  ]
 [ 17.808514  203.41083   718.3673    709.3414   ]
 [ 11.366208   81.115036  725.5019    190.9168   ]]
['In this series, we travel to the past and leam about historical mapmakers, from Claudius Ptolemy (who stated the idea that the Earth is at the center of the universe)
 to Gerardus Mercator (who created one of th most widely used map projections) and more. This series goes into tremendous detail on the cartographer\'s life and maps. W
e then joumey to the present era to leam about map projections and the diverse types of maps used today. You might ask, “What is the difference between the two? They so
und the same to me." No map projection is perfect, because you cannot really flatten a sphere into a rectangle An uncolored projection could be used in many ways. We co
uld use it for population concentration, highways, land elevation, and so many other things! ', 'The Mapping the World 8-book set goes into amazing levels of detail. It
 is a long read, but it gives an immense range and amount of information that you would net find in any other book or series on maps. The flowing way the chapters and b
ooks are organized makes it easy to link passages from different books in this series together. Mapping the World is a treasure box, filled with the seeds of cartograph
y. Collect and plant them, and you soon will have the fruits of cartography, beneficial to those who want t. be cartographers. Use this series to the utmost, then the f
ruits of mapping will be sweet for all whe endeavor to succeed in cartography. ', 'For example, we could make a topographic map of the U.S., which shows land elevation.
 We could make ita colorful map that shows the amount of pollution in different areas, or it could be a population map, or it could even be a map that shows the 50 stat
es, their capitals and borders! Our last step in this amazing, excursion is the near future, where we see some hypothetical solutions as to what maps will be used for.
Currently, we are working on better virtual map technology. ', '‘Now, scientists have been able to put maps on phones. Back in the early 1900s, people had to lug a lot
o ‘maps around to find your way from place to place, or just keep asking for directions. Now, all the information is on a phone or global positioning system (GPS). It i
s amazing how much maps have changed technology and the world in this century. ', 'Bvery day, people around the world use maps. Whether it is an airplane pilot or busin
essman, housewife or museum group, maps have always and will continue to provide useful information for all. ', '‘Mapping the World talks about the uses of maps, as wel
l as how to differentiate between the type of map projection and type of map. ', 'Mapping the World ']
application.py:207: UserWarning: Starting a Matplotlib GUI outside of the main thread will likely fail.
  fig, ax = plt.subplots(figsize=(15,15))
application.py:264: UserWarning: Starting a Matplotlib GUI outside of the main thread will likely fail.
  plt.show()
```

Fig 5.3 Context data from processed image

5.2.2 Performance Parameter

These are the parameters on which performance can be evaluated

●The correct ROIs : more the number of ROI generated by model are correct with correct coordinates , more optimized is the algorithm used and better is the performance.

● Text recognition (OCR): more the confidence factor gives to the result; more is the accuracy of the given result.

● Answer Generation: more the confidence factor gives to the result; more is the accuracy of the given result.

● Cost: the lesser the cost of this system compared to the cost of static parking system the better is the performance.

● Ease of use: the more user friendly the app is the better is the performance.

5.3 Working of the project

5.3.1 Procedural Workflow

Referring to figure 4.3.1 will explain the procedural workflow of our web application. First, we register our profile and then login with our profile to web page. Then, it will ask to upload document image. We will upload image relevant to question which we want to ask. After uploading document image, it will break it into 9 ROI s and OCR will convert it into text. Now, this will bring us to next page where we will enter question and model will generate answer and display on app.

5.3.2 Algorithmic Approaches Used

Following are the algorithmic steps for object detection on any custom dataset:

Step 1: Installing Detectron 2

Kickstart with installing a few dependencies such as Torch Vision and COCO API and check whether CUDA is available. CUDA helps in keeping track of the currently selected GPU and then install Detectron2.

Step 2: Prepare and Register the Dataset

Import a few necessary packages. Datasets that have support in detectron2 are listed in datasets. If you want to use a custom dataset while also reusing detectron2's data loaders, you will need to Register your dataset. There are certain formats in how data can be fed to a model such as a YOLO format, COCO format, etc. Detectron2 accepts the COCO Format of the dataset. COCO format of the dataset consists of a JSON file which includes all the details of an image such as size, annotations (i.e., bounding box coordinates), labels corresponding to it's bounding box, etc.

Step 3: Visualize the Training Set

We'll randomly pick 3 pictures from the train folder of our dataset and see how the bounding boxes look like.

Step 4: Training the Model

This is the step where we give configurations and set the model ready to get trained. Technically, we just fine-tune our model on the dataset as the model is already pre-trained on COCO Dataset. There are a ton of models available for object detection in the Detectron2's Model Zoo. Here, we use the faster_rcnn_R_101_FPN_3x model which looks in this way on a high level.

Step 5: Inference using the Trained Model

It's time to infer the results by testing the model on the Validation Set. An output folder gets saved in the local storage after successful completion of training in which the final weights are stored.

Step 6: Evaluation of the Trained Model

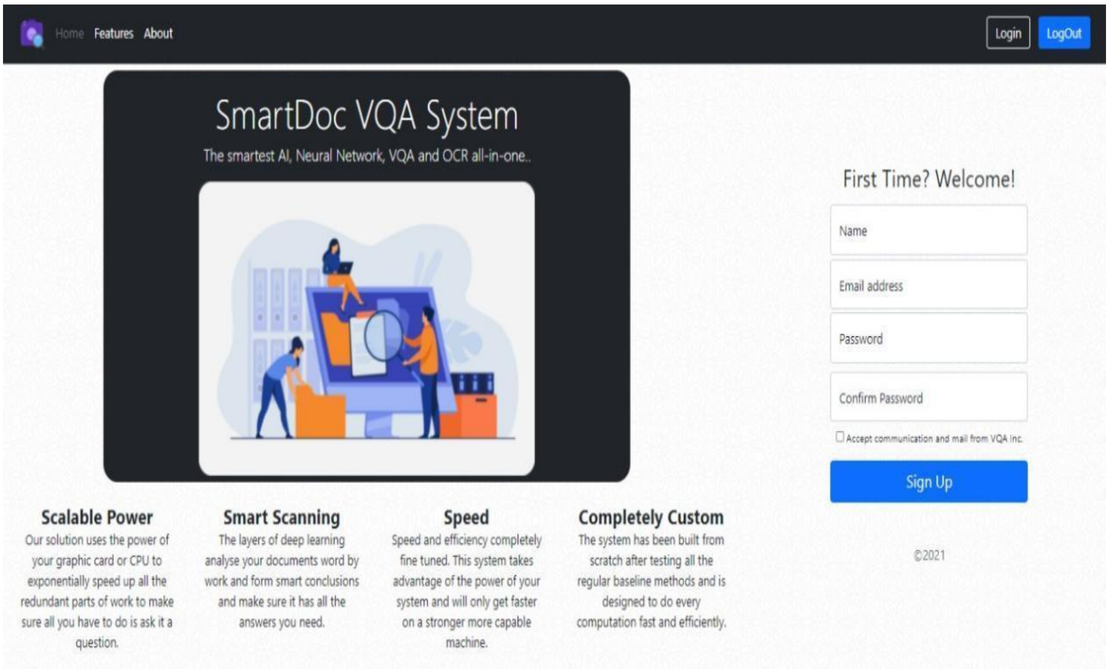
Usually, the model is evaluated following the COCO Standards of evaluation. Mean Average Precision (mAP) is used to evaluate the performance of the model.

5.3.3 Project Deployment

Various components used in our project are laptop, camera and clearing image module, age
estimation module, question answer database, speech to text and text to speech conversion
module and at last question generating module.

- Laptop: The main component of our learning helper which makes all other parts working
 and is the brain of our project. Inbuilt camera, mic, speaker and screen is used.
- Uploading module: It is used to upload and download the image for further processing.
- ROI module: It is used to estimate the age of the user from the scanned image
 which we get from the previous module.
- OCR module: used to extract text from the image.
- Answer Generating module: given context and question the module is able to generate answer
- Flask Web App: the integration of all modules and making it user friendly interface.

5.3.4 System Screenshots



SmartDoc VQA System

The smartest AI, Neural Network, VQA and OCR all-in-one..

First Time? Welcome!

☐ Accept communication and mail from VQA Inc.

©2021

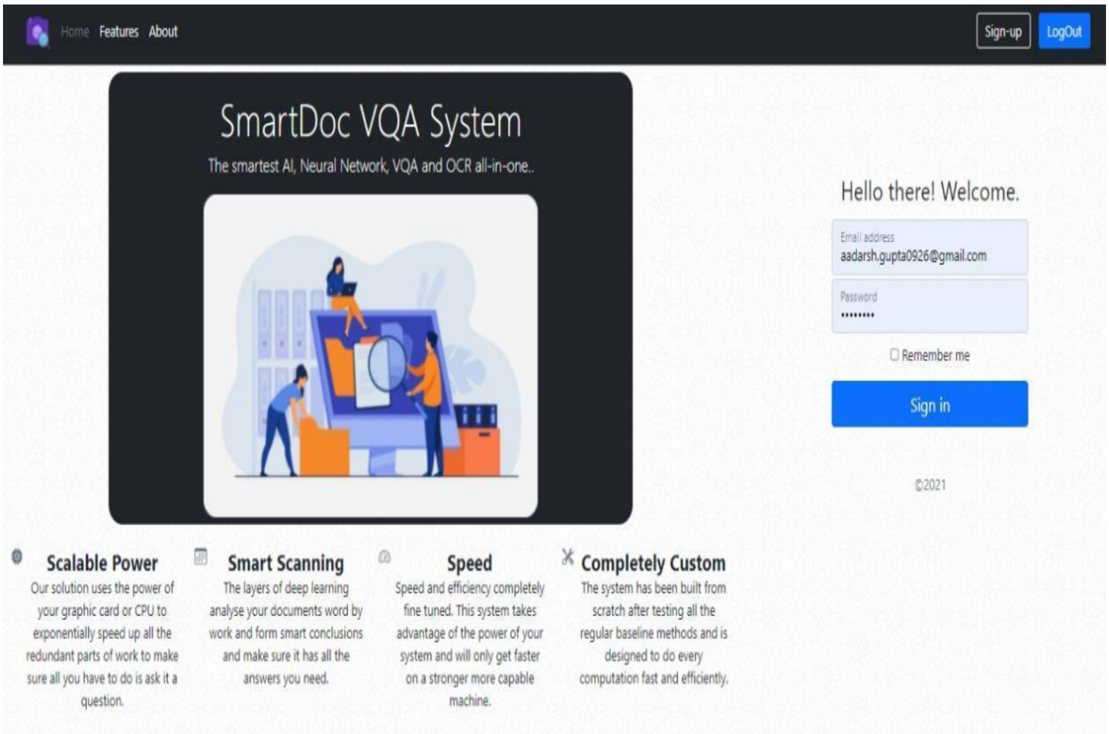
Scalable Power
Our solution uses the power of your graphic card or CPU to exponentially speed up all the redundant parts of work to make sure all you have to do is ask it a question.

Smart Scanning
The layers of deep learning analyse your documents word by work and form smart conclusions and make sure it has all the answers you need.

Speed
Speed and efficiency completely fine tuned. This system takes advantage of the power of your system and will only get faster on a stronger more capable machine.

Completely Custom
The system has been built from scratch after testing all the regular baseline methods and is designed to do every computation fast and efficiently.

Fig 5.4 Registration Page



SmartDoc VQA System

The smartest AI, Neural Network, VQA and OCR all-in-one..

Hello there! Welcome.

Email address
aadarsh.gupta0926@gmail.com

Password
••••••••

☐ Remember me

©2021

Scalable Power
Our solution uses the power of your graphic card or CPU to exponentially speed up all the redundant parts of work to make sure all you have to do is ask it a question.

Smart Scanning
The layers of deep learning analyse your documents word by work and form smart conclusions and make sure it has all the answers you need.

Speed
Speed and efficiency completely fine tuned. This system takes advantage of the power of your system and will only get faster on a stronger more capable machine.

Completely Custom
The system has been built from scratch after testing all the regular baseline methods and is designed to do every computation fast and efficiently.

Fig 5.5 Login Page

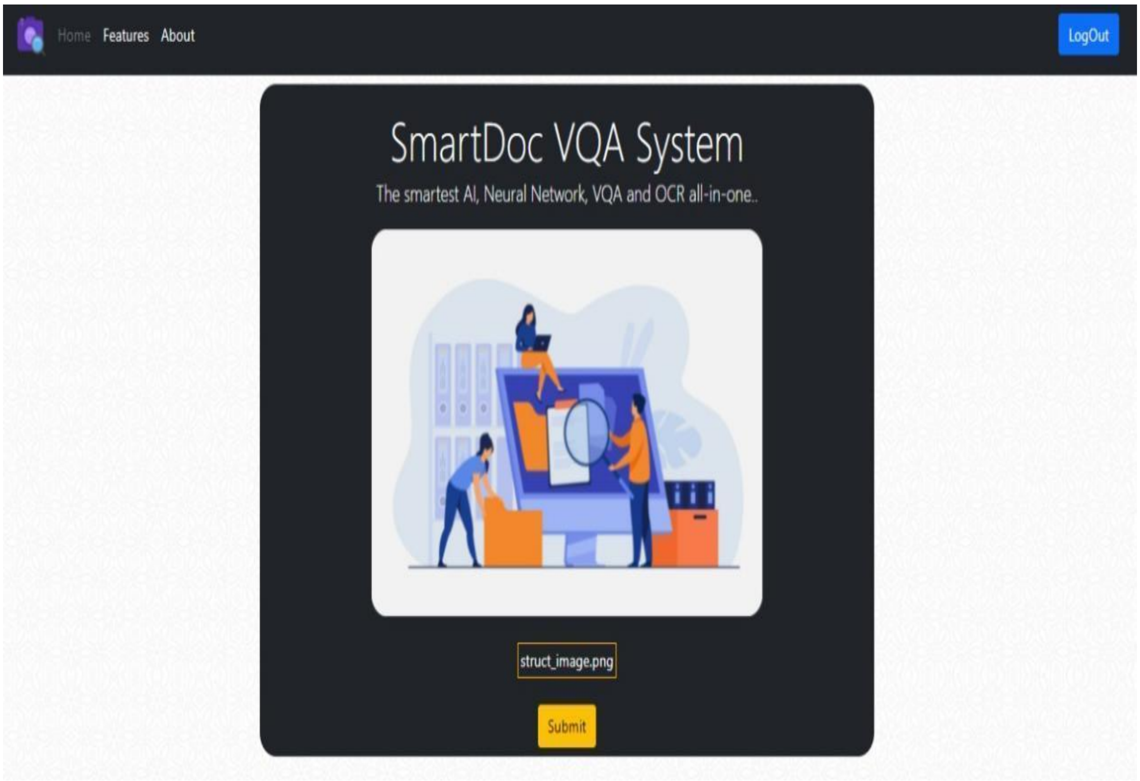
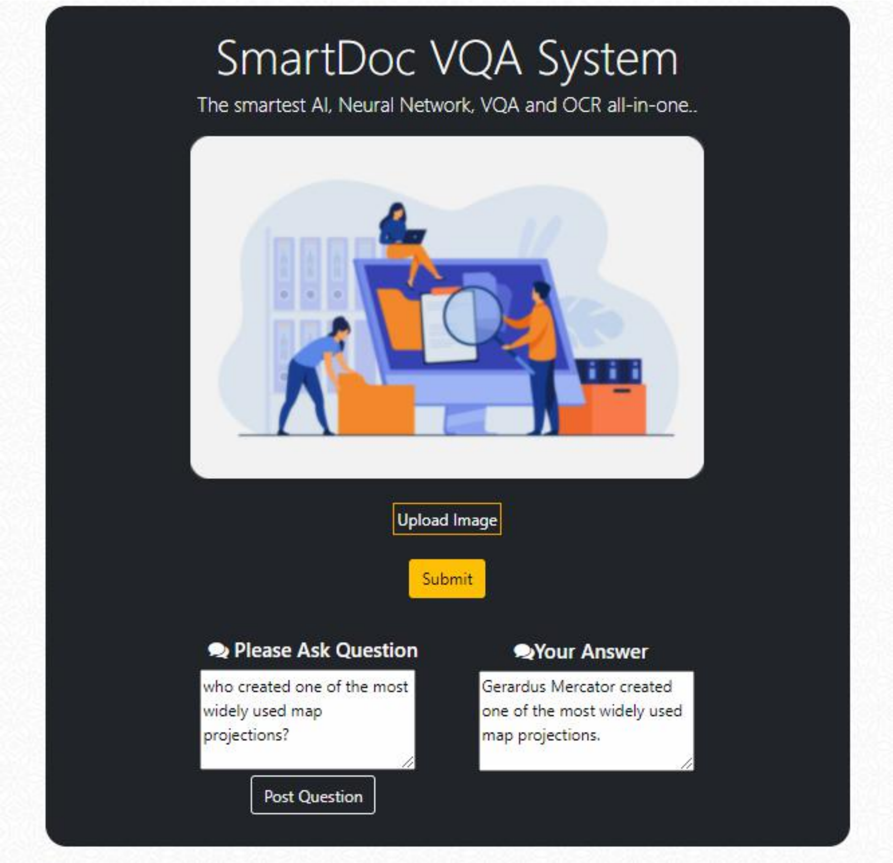


Fig 5.6 Select and upload image



who created one of the most widely used map projections?

Gerardus Mercator created one of the most widely used map projections.

Fig 5.7 Q&A system

5.4 Testing Process

5.4.1. Test Plan

We plan to perform the complete testing on our application to ensure that all the components are working individually and when integrated together. We will be carrying out unit tests to see if the code performs the desired tasks and meets the specifications. Our aim is to train our model on our dataset and see what label and ROI our model predict. Then we will see how accurate characters are recognized to form context vector. At the end the accuracy of answer generation module will be tested.

5.4.2. Features to be tested

We plan to test all the features of the web app as if they are working correctly i.e. the features like uploading document image, ROI, OCR and question answering module working correctly. Along with this we plan to test the algorithm of generating answers by taking question and context and faster RCNN module to divide into 9 ROIs.

5.4.3 Test Strategy

We will be testing each module separately and after that we will test the components all together. Test Strategy to validate the functionality of various system features are:

- Uploading Document: Check if the image of relevant document is uploaded without error and is given to further modules.
- Dividing into 9 ROIs: Check if document image is correctly segmented into ROI s that include paragraph, title, caption etc.
- OCR: from different ROI s to check if characters are recognized correctly.
- Answer Generating Module: Check if the answer generating module from given question and context could generate relevant answer.

5.4.4 Test Techniques

Black Box testing - The tester interacts with the user interface and provides inputs and checks the outputs returned without going unto the underlying code complexity.

White Box testing - The tester investigates the internal logic and structure of the code and examines if every single component is working fine or not.

Functional Testing - It involves testing the application against the business requirements. It incorporates all test types designed to guarantee each part of a piece of software behaves as expected by using use cases provided by the design team. The testing methods involve Unit Testing, Integration testing, System testing and Acceptance testing in order.

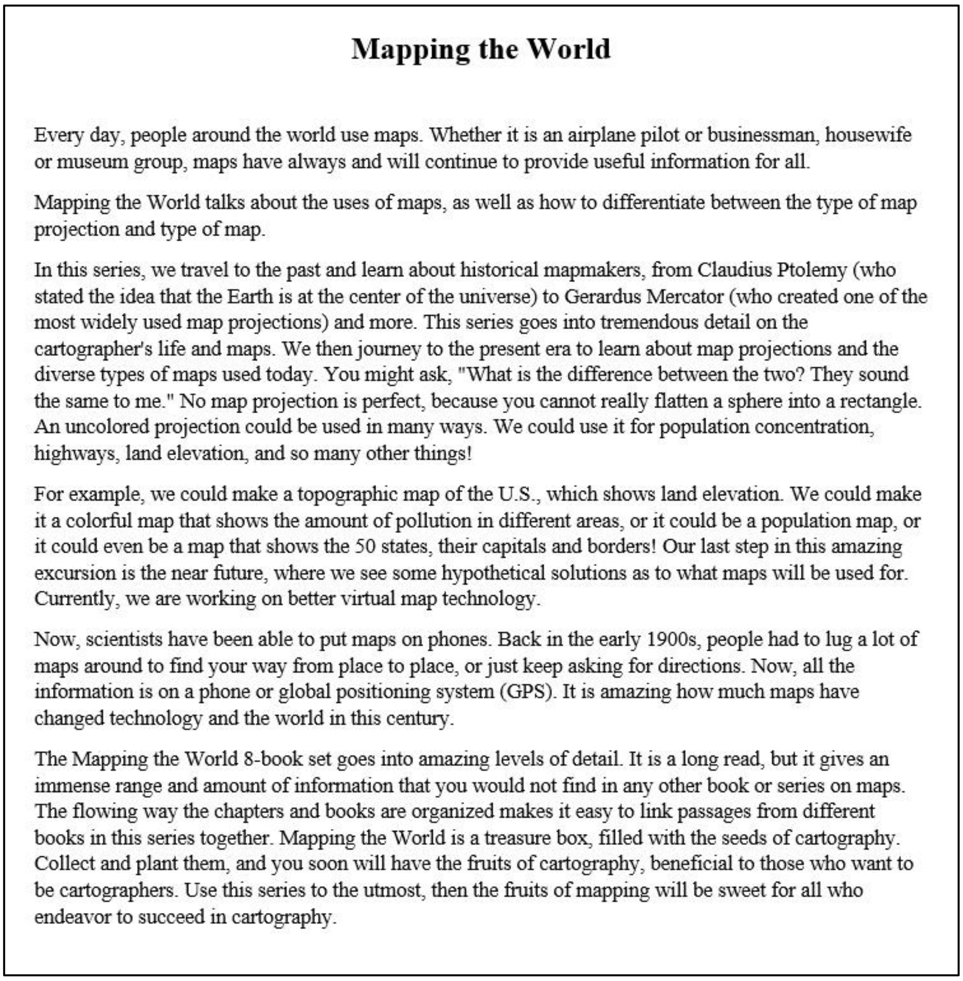
Non - Functional testing - Non-functional testing methods incorporate all test types focused on the operational aspects of a piece of software. Testing methods include performance testing, security testing, usability testing and compatibility testing.

- Uploading Document: simply by uploading by any document image
- Dividing into 9 ROIs: When uploaded the document image, print the processed image that we will show in next section
- OCR: From different ROI s to check if characters are recognized correctly, we will print text and confirm
- Answer Generating Module: Check if the answer generated is correct, we can use evaluation metrics.

5.4.5 Test Cases

Case 1: Testing ROI module

When user uploads the document image, at the back end the document should be segmented into 9 ROIs.



Mapping the World

Every day, people around the world use maps. Whether it is an airplane pilot or businessman, housewife or museum group, maps have always and will continue to provide useful information for all.

Mapping the World talks about the uses of maps, as well as how to differentiate between the type of map projection and type of map.

In this series, we travel to the past and learn about historical mapmakers, from Claudius Ptolemy (who stated the idea that the Earth is at the center of the universe) to Gerardus Mercator (who created one of the most widely used map projections) and more. This series goes into tremendous detail on the cartographer's life and maps. We then journey to the present era to learn about map projections and the diverse types of maps used today. You might ask, "What is the difference between the two? They sound the same to me." No map projection is perfect, because you cannot really flatten a sphere into a rectangle. An uncolored projection could be used in many ways. We could use it for population concentration, highways, land elevation, and so many other things!

For example, we could make a topographic map of the U.S., which shows land elevation. We could make it a colorful map that shows the amount of pollution in different areas, or it could be a population map, or it could even be a map that shows the 50 states, their capitals and borders! Our last step in this amazing excursion is the near future, where we see some hypothetical solutions as to what maps will be used for. Currently, we are working on better virtual map technology.

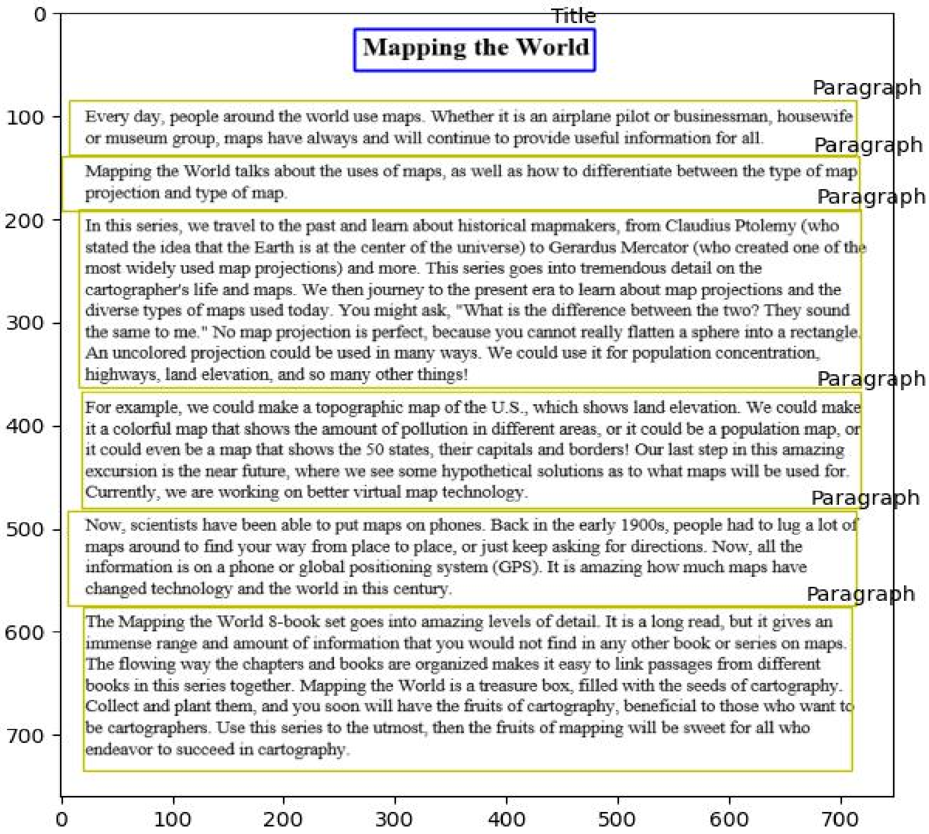
Now, scientists have been able to put maps on phones. Back in the early 1900s, people had to lug a lot of maps around to find your way from place to place, or just keep asking for directions. Now, all the information is on a phone or global positioning system (GPS). It is amazing how much maps have changed technology and the world in this century.

The Mapping the World 8-book set goes into amazing levels of detail. It is a long read, but it gives an immense range and amount of information that you would not find in any other book or series on maps. The flowing way the chapters and books are organized makes it easy to link passages from different books in this series together. Mapping the World is a treasure box, filled with the seeds of cartography. Collect and plant them, and you soon will have the fruits of cartography, beneficial to those who want to be cartographers. Use this series to the utmost, then the fruits of mapping will be sweet for all who endeavor to succeed in cartography.

Fig.5.8 Document image for testing ROI module

Case 2: Testing OCR module

Given the ROI bounding boxes coordinated along with the labels and confidence score, this module will generate the context vector out of the document image.



Title
Paragraph
Paragraph
Paragraph
Paragraph
Paragraph
Paragraph
0
100
200
300
400
500
600
700
0
100
200
300
400
500
600
700

Mapping the World

Every day, people around the world use maps. Whether it is an airplane pilot or businessman, housewife or museum group, maps have always and will continue to provide useful information for all.

Mapping the World talks about the uses of maps, as well as how to differentiate between the type of map projection and type of map.

In this series, we travel to the past and learn about historical mapmakers, from Claudius Ptolemy (who stated the idea that the Earth is at the center of the universe) to Gerardus Mercator (who created one of the most widely used map projections) and more. This series goes into tremendous detail on the cartographer's life and maps. We then journey to the present era to learn about map projections and the diverse types of maps used today. You might ask, "What is the difference between the two? They sound the same to me." No map projection is perfect, because you cannot really flatten a sphere into a rectangle. An uncolored projection could be used in many ways. We could use it for population concentration, highways, land elevation, and so many other things!

For example, we could make a topographic map of the U.S., which shows land elevation. We could make it a colorful map that shows the amount of pollution in different areas, or it could be a population map, or it could even be a map that shows the 50 states, their capitals and borders! Our last step in this amazing excursion is the near future, where we see some hypothetical solutions as to what maps will be used for. Currently, we are working on better virtual map technology.

Now, scientists have been able to put maps on phones. Back in the early 1900s, people had to lug a lot of maps around to find your way from place to place, or just keep asking for directions. Now, all the information is on a phone or global positioning system (GPS). It is amazing how much maps have changed technology and the world in this century.

The Mapping the World 8-book set goes into amazing levels of detail. It is a long read, but it gives an immense range and amount of information that you would not find in any other book or series on maps. The flowing way the chapters and books are organized makes it easy to link passages from different books in this series together. Mapping the World is a treasure box, filled with the seeds of cartography. Collect and plant them, and you soon will have the fruits of cartography, beneficial to those who want to be cartographers. Use this series to the utmost, then the fruits of mapping will be sweet for all who endeavor to succeed in cartography.

Fig.5.9 Document image for testing OCR module

Case 3: Testing answer generating module

Given a context and a question, module should generate the relevant answer to the provided question.

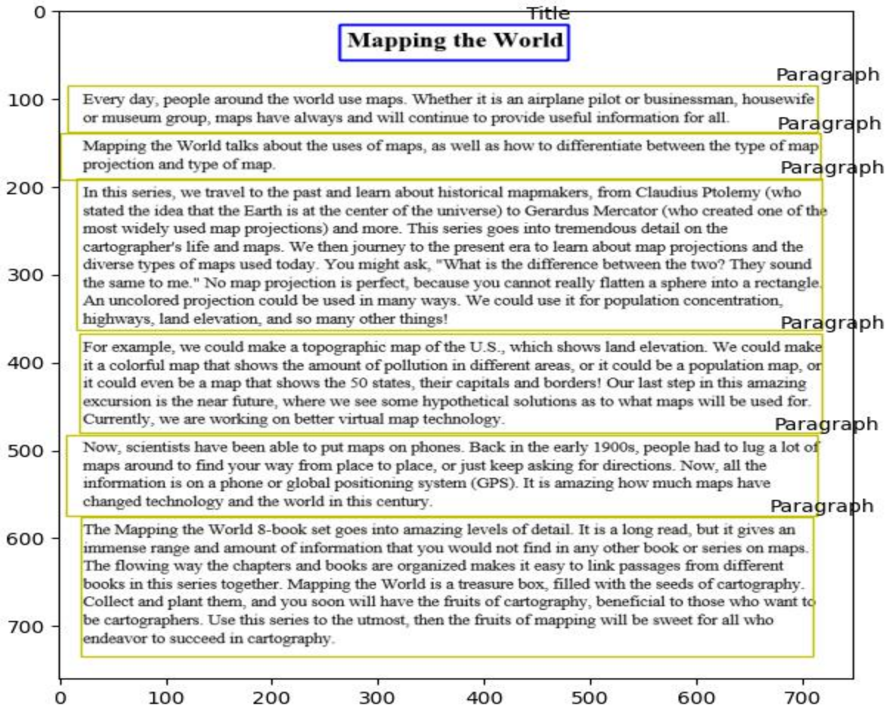
Table 5.4.5.1 – Context and questions

Context	Questions
In this series, we travel to the past and learn about historical mapmakers, from Claudius Ptolemy (who stated the idea that the Earth is at the center of the universe) to Gerardus Mercator (who created one of the most widely used map projections) and more. This series goes into tremendous detail on the cartographer's life and maps. We then journey to the present era to learn about map projections and the diverse types of maps used today. You might ask, “What is the difference between the two? They sound the same to me." No map projection is perfect, because you cannot really flatten a sphere into a rectangle An uncolored projection could be used in many ways. We could use it for population concentration, highways, land elevation, and so many other things! The Mapping the World 8-book set goes into amazing levels of detail. It is a long read, but it gives an immense range and amount of information that you would net find in any other book or series on maps. The flowing way the chapters and books are organized makes it easy to link passages from different books in this series together. Mapping the World is a treasure box, filled with the seeds of cartography. Collect and plant them, and you soon will have the fruits of cartography, beneficial to those who want t. be cartographers. Use this series to the utmost, then the fruits of mapping will be sweet for all the endeavor to succeed in cartography. For example, we could make a topographic map of the U.S., which shows land elevation. We could make it a colorful map that shows the amount of pollution in different areas, or it could be a population map, or it could even be a map that shows the 50 states, their capitals and borders! Our last step in this amazing, excursion is the near future, where we see some hypothetical solutions as to what maps will be used for. Currently, we are working on better virtual map technology. ‘Now, scientists have been able to put maps on phones. Back in the early 1900s, people had to lug a lot o ‘maps around to find your way from place to place, or just keep asking	Q1. Why map projections are not perfect? Q2. Who stated the idea that the Earth is at the center of the universe? Q3. Who created one of the most widely used map projects?

for directions. Now, all the information is on a phone or global positioning system (GPS). It is amazing how much maps have changed technology and the world in this century. Every day, people around the world use maps. Whether it is an airplane pilot or businessman, housewife or museum group, maps have always and will continue to provide useful information for all. ‘Mapping the World talks about the uses of maps, as well as how to differentiate between the type of map projection and type of map. Mapping the World.Q	

5.4.6 Test Results

Case 1: Results of ROI module



Mapping the World

Every day, people around the world use maps. Whether it is an airplane pilot or businessman, housewife or museum group, maps have always and will continue to provide useful information for all.

Mapping the World talks about the uses of maps, as well as how to differentiate between the type of map projection and type of map.

In this series, we travel to the past and learn about historical mapmakers, from Claudius Ptolemy (who stated the idea that the Earth is at the center of the universe) to Gerardus Mercator (who created one of the most widely used map projections) and more. This series goes into tremendous detail on the cartographer's life and maps. We then journey to the present era to learn about map projections and the diverse types of maps used today. You might ask, "What is the difference between the two? They sound the same to me." No map projection is perfect, because you cannot really flatten a sphere into a rectangle. An uncolored projection could be used in many ways. We could use it for population concentration, highways, land elevation, and so many other things!

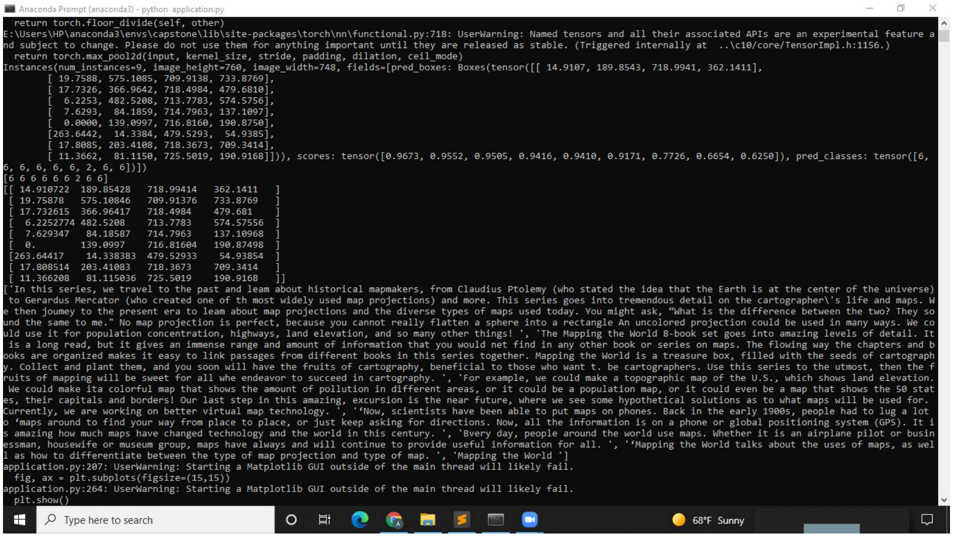
For example, we could make a topographic map of the U.S., which shows land elevation. We could make it a colorful map that shows the amount of pollution in different areas, or it could be a population map, or it could even be a map that shows the 50 states, their capitals and borders! Our last step in this amazing excursion is the near future, where we see some hypothetical solutions as to what maps will be used for. Currently, we are working on better virtual map technology.

Now, scientists have been able to put maps on phones. Back in the early 1900s, people had to lug a lot of maps around to find your way from place to place, or just keep asking for directions. Now, all the information is on a phone or global positioning system (GPS). It is amazing how much maps have changed technology and the world in this century.

The Mapping the World 8-book set goes into amazing levels of detail. It is a long read, but it gives an immense range and amount of information that you would not find in any other book or series on maps. The flowing way the chapters and books are organized makes it easy to link passages from different books in this series together. Mapping the World is a treasure box, filled with the seeds of cartography. Collect and plant them, and you soon will have the fruits of cartography, beneficial to those who want to be cartographers. Use this series to the utmost, then the fruits of mapping will be sweet for all who endeavor to succeed in cartography.

Fig 5.10 Results of ROI module

Case 2: Results of OCR module

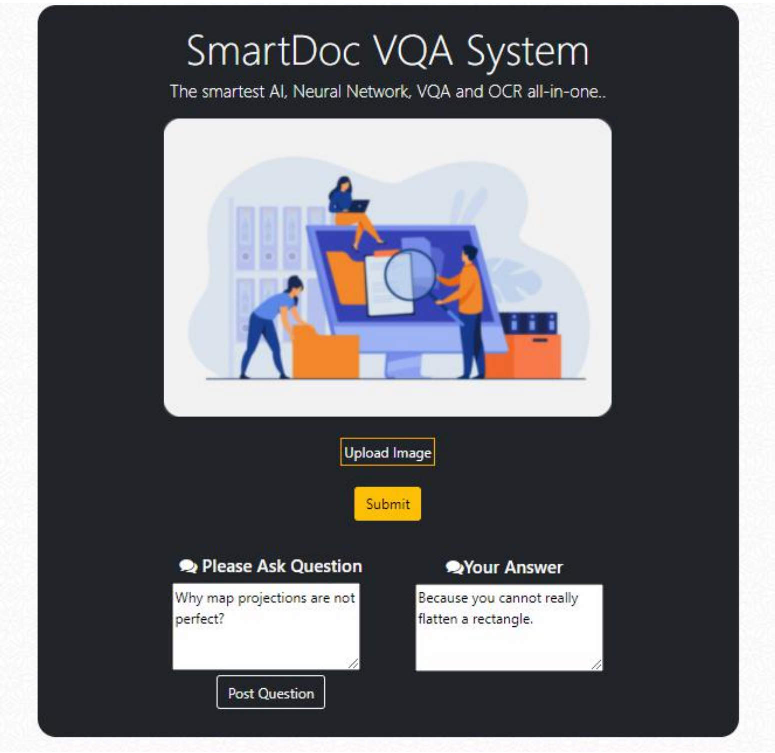


Anaconda Prompt (anaconda3) - python application.py

```
  return torch.floor_divide(self, other)
E:\Users\HP\anaconda3\envs\capstone\lib\site-packages\torch\nn\functional.py:718: UserWarning: Named tensors and all their associated APIs are an experimental feature a
nd subject to change. Please do not use them for anything important until they are released as stable. (Triggered internally at  ..\c10/core/TensorImpl.h:1156.)
  return torch.max_pool2d(input, kernel_size, stride, padding, dilation, ceil_mode)
Instances(num_instances=9, image_height=760, image_width=748, fields=[pred_boxes: Boxes(tensor([[ 14.9107, 189.8543, 718.9941, 362.1411],
        [ 19.7588, 575.1085, 709.9138, 733.8769],
        [ 17.7326, 366.9642, 718.4984, 479.6810],
        [  6.2253, 482.5208, 713.7783, 574.5756],
        [  7.6293,  84.1859, 714.7963, 137.1097],
        [  0.0000, 139.0997, 716.8160, 190.8750],
        [263.6442,  14.3384, 479.5293,  54.9385],
        [ 17.8085, 203.4108, 718.3673, 709.3414],
        [ 11.3662,  81.1150, 725.5019, 190.9168]])), scores: tensor([0.9673, 0.9552, 0.9505, 0.9416, 0.9410, 0.9171, 0.7726, 0.6654, 0.6250]), pred_classes: tensor([6,
6, 6, 6, 6, 6, 2, 6, 6])])
[6 6 6 6 6 6 2 6 6]
[[ 14.910722  189.85428   718.99414   362.1411   ]
 [ 19.75878   575.10846   709.91376   733.8769   ]
 [ 17.732615  366.96417   718.4984    479.681    ]
 [  6.2252774 482.5208    713.7783    574.57556  ]
 [  7.629347   84.18587   714.7963    137.10968  ]
 [  0.         139.0997    716.81604   190.87498  ]
 [263.64417    14.338383  479.52933    54.93854  ]
 [ 17.808514  203.41083   718.3673    709.3414   ]
 [ 11.366208   81.115036  725.5019    190.9168   ]]
['In this series, we travel to the past and leam about historical mapmakers, from Claudius Ptolemy (who stated the idea that the Earth is at the center of the universe)
 to Gerardus Mercator (who created one of th most widely used map projections) and more. This series goes into tremendous detail on the cartographer\'s life and maps. W
e then joumey to the present era to leam about map projections and the diverse types of maps used today. You might ask, “What is the difference between the two? They so
und the same to me." No map projection is perfect, because you cannot really flatten a sphere into a rectangle An uncolored projection could be used in many ways. We co
uld use it for population concentration, highways, land elevation, and so many other things! ', 'The Mapping the World 8-book set goes into amazing levels of detail. It
 is a long read, but it gives an immense range and amount of information that you would net find in any other book or series on maps. The flowing way the chapters and b
ooks are organized makes it easy to link passages from different books in this series together. Mapping the World is a treasure box, filled with the seeds of cartograph
y. Collect and plant them, and you soon will have the fruits of cartography, beneficial to those who want t. be cartographers. Use this series to the utmost, then the f
ruits of mapping will be sweet for all whe endeavor to succeed in cartography. ', 'For example, we could make a topographic map of the U.S., which shows land elevation.
 We could make ita colorful map that shows the amount of pollution in different areas, or it could be a population map, or it could even be a map that shows the 50 stat
es, their capitals and borders! Our last step in this amazing, excursion is the near future, where we see some hypothetical solutions as to what maps will be used for.
Currently, we are working on better virtual map technology. ', '‘Now, scientists have been able to put maps on phones. Back in the early 1900s, people had to lug a lot
o ‘maps around to find your way from place to place, or just keep asking for directions. Now, all the information is on a phone or global positioning system (GPS). It i
s amazing how much maps have changed technology and the world in this century. ', 'Bvery day, people around the world use maps. Whether it is an airplane pilot or busin
essman, housewife or museum group, maps have always and will continue to provide useful information for all. ', '‘Mapping the World talks about the uses of maps, as wel
l as how to differentiate between the type of map projection and type of map. ', 'Mapping the World ']
application.py:207: UserWarning: Starting a Matplotlib GUI outside of the main thread will likely fail.
  fig, ax = plt.subplots(figsize=(15,15))
application.py:264: UserWarning: Starting a Matplotlib GUI outside of the main thread will likely fail.
  plt.show()
```

Fig 5.11 Results of OCR module

Case 3: Result of answer generating module



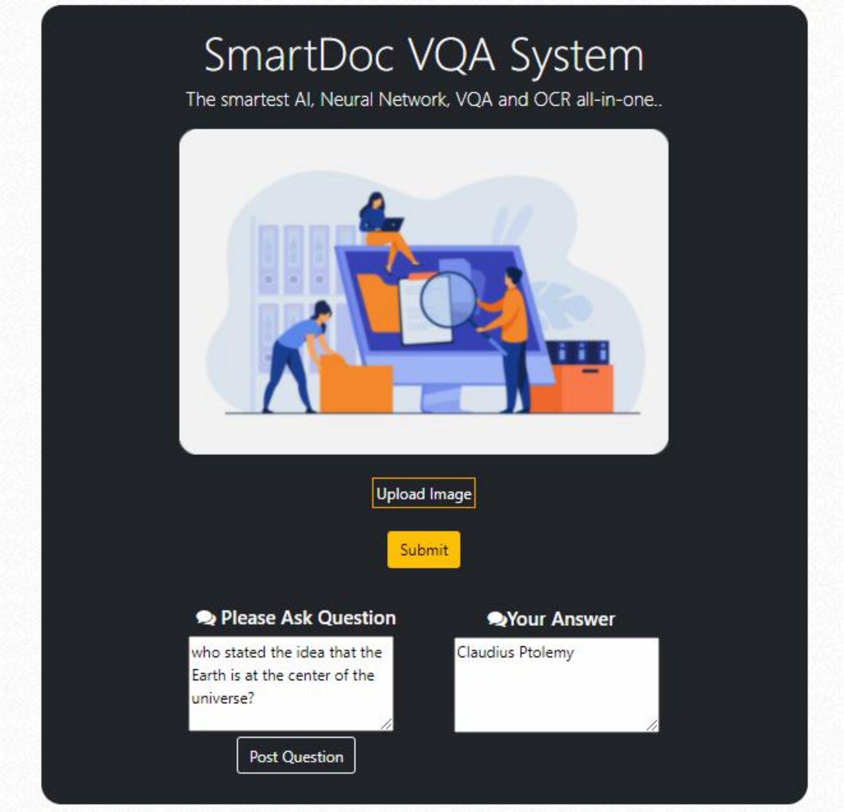
SmartDoc VQA System

The smartest AI, Neural Network, VQA and OCR all-in-one..

Please Ask Question: Why map projections are not perfect?

Your Answer: Because you cannot really flatten a rectangle.

Fig 5.12 Answer generation for question1



who stated the idea that the Earth is at the center of the universe?

Claudius Ptolemy

Fig 5.13 Answer generation for question2

5.5 Results and Discussions

After accomplishing testing phase of the product, the results of the testing are to be drawn out.

The proposed system was carried out against designed test cases were executed to check the performance measures of the system.

The system designed and test strategies implemented helped to check the real potential and performance of the system developed. The outcomes of the testing phase described that the product developed is working well against the desired features and delivering expected outcomes. The system is working give specific conditions such as power supply, web connectivity, NLP modules etc. The developed system is responding adequately and generating answers from document and question.

5.6 Inferences Drawn

- During the development of smart VQA system, it was observed that the accuracy of document segmentation the performance of OCR greatly depends and therefore effects the context vector.
- It is difficult to train such a large system. We had trained two models that are Faster RCNN and T5 but for few epochs because of less computation power.
- Customizing T5 model by adding location embeddings, segment embeddings, positional embeddings thus combining two modalities -computer vision and NLP

5.7 Validation of Objectives

Table 5.2 Validation of objectives

S.No.	**Objectives**	**Status**
1.	Analyzing different existing techniques for VQA.	Successful
2.	Extracting different features to feed as input embedding	Successful
3.	Building the encoder decoder model.	Successful
4.	Designing framework for effective interaction with the user.	Successful

CHAPTER 6 - Conclusions and Future Directions

6.1 Conclusions

1. Detectron2 is easy to use and create object detection, instance segmentation, keypoint detection and panoptic segmentation models. Its simplicity and modularity of design makes rewriting a script for another data-set quite easy.
2. Document Image layout analysis uses RLSA for finding the ROI's of the given document. On application of RLSA to pattern arrays, it has the effect of linking together neighboring black areas have a separation of less than given threshold pixels. This Threshold value is very important to link areas with common data type but different documents have different spacing between the lines making it difficult to decide the threshold value that can have 2 differnt ROIs for two paragraphs and have one ROI corresponding to one paragraph.
3. Faster R-CNN models can be used with differnt region proposal network(RPN) variants and different varients have differnt accuracy
4. By training different models, faster_RCNN_R_101_FPN_3x gives us best accuracy when when running for 650 iterations

6.2 Environmental, Economic and Societal Benefits

1. Promotes use of soft copy instead of hard copy.

2. Saves cutting of trees for paper.

6.3 Reflections

We achieved following from the project:

1. Picked relevant dataset for building up the HumanLike Visual Question answering system.

2. Fetched the relevant data from taken up screenshots.

3. Spliting the data into train,test and validation.

4. Fetched the relevant data from annotation file to draw rois on provided dataset.

5. Trained a faster r-cnn model using detectron2 facebook AI research's next generation library to get ROI's.

6. Trained model using different region proposal networks to get ROI's.
7. Made prototype for web application using flask.

We learned many new things from project that includes project documentation, implementation and coordinating between team members and mentor.

6.4 Future Work

We will first try to extract the text from the ROI generated above using the Tesract OCR. Once we have extracted the text, we are ready to train our question and answer system by applying various in-depth learning patterns such as VQA, TextVQA, M4C, T5 and BART. We check their accuracy using various evaluation metrics such as CIDer, BLEU, METEOR and ROUGE-L. We will first try to apply the question and answer system to text only and after achieving the required minimum accuracy, we will try to do the same for pictures and tables. Finally, we will try to combine the best of them with the fastest R-CNN model, creating a collective question-and-answer system on structured and non-structured images with 9 ROIs. Finally, we run the entire work into the web application.

CHAPTER 7 - Project Metrics

7.1 Challenges Faced

1. Finding or creating a concise dataset for the VQA System. Using the best available dataset was a big step in working on the most recent development. Using such a dataset also presented other challenges that shall be discussed further.

2. Splitting the dataset, pre-processing of the dataset and gathering the required screenshots. This was also a new challenge considering that the dataset is very vast and presents a major problem of computational complexity.

3. Working on this complex project with virtual collaboration and the differences in hardware available to the contributors also presents quite a big challenge as this project is completely software based.

4. The challenges of actually devising the model withstanding, a major other concern was how to evaluate the performance of the model on the entire dataset. This would require a very fast machine and hardware, which is a big challenge.

7.2 Relevant Subjects

Table 7.1 Relevant Subjects

UCS406	Data Structures and Algorithm	Creating all the different classes and storing information required data structures.
UCS503	Software Engineering	Helped in making various work flow diagrams and implementing techniques and correlating timings of the project.
UML501	Machine Learning	Training the model and testing the performances of various models tried.
UML602	Natural Language Processing	For generating the answers from the OCR text as well as understanding of the questions asked.

7.3 Interdisciplinary Knowledge Sharing

During the course of the project, members had regular group sessions to discuss the feasibility and state of the project. Through sharing knowledge members were able to learn many technologies by collaborating on this project. There were many things involved in making this project, documentation, web-app interface, developing the model, research of previous work. Collaboration by dividing the workload and discussing all the possible ways to accomplish our target, helped us in discovering the best way to do this work. Working on these technologies together as a team helped all of us to increase our knowledge across all these disciplines and refine our previous knowledge.

7.4 Peer Assessment Matrix

Table 7.2 Peer Assessment Matrix

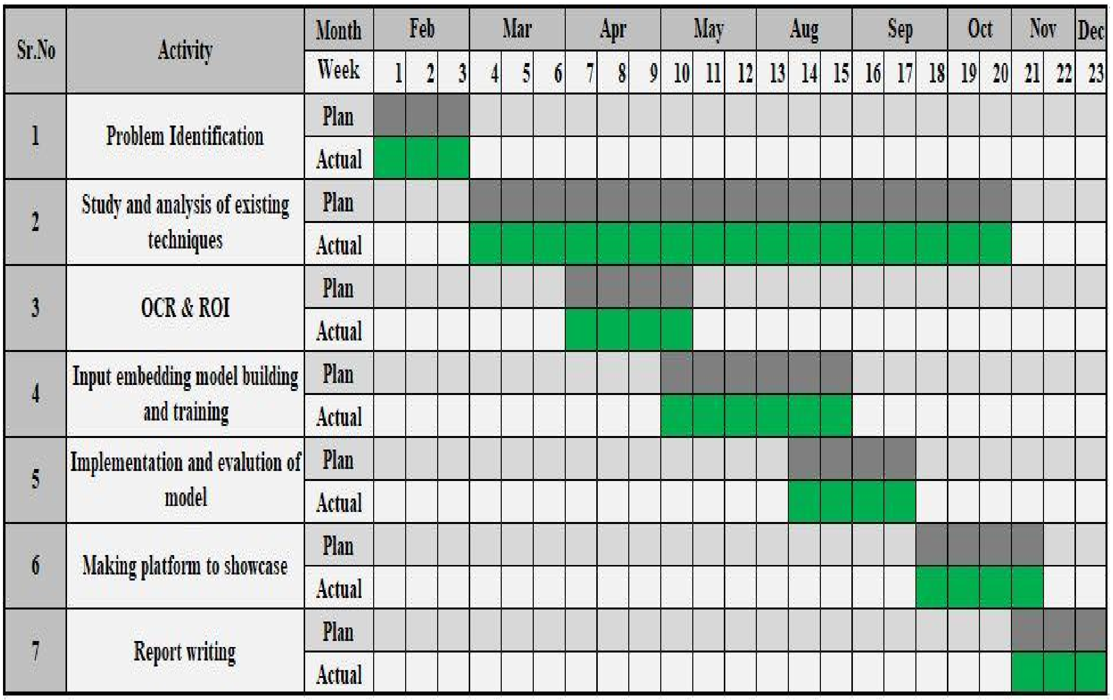
		Evaluation By			
		Ishan	Aadarsh	Harchetan	Sameep
Evaluation of	Ishan	5	4.5	5	4.5
	Aadarsh	4.5	5	4.5	5
	Harchetan	5	4.5	5	4.5
	Sameep	4.5	5	4.5	5

7.5 Role Playing and Work Schedule

Table 7.3 Role Playing

Project Members	Roles
Aadarsh Gupta	ROI, OCR, Faster R-CNN, Model Building, Website Interface, Question Answering, Report Writing
Ishan Manuja	Preprocessing, Model Building, Question Answering Module, Evaluation of Model, Research, Report Writing
Sameep Singh	Preprocessing, Model Building, Website Interface, Research, Question Answering Module, Report Writing
Harchetan Singh Chahal	ROI, OCR, Faster R-CNN, Model Building, Training and Implementation, Question Answering, Report Writing

Table 7.4 Work Schedule



Sr.No	Activity	Month	Feb			Mar			Apr			May			Aug			Sep			Oct		Nov		Dec
		Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	Problem Identification	Plan	■	■	■																				
		Actual	🟩	🟩	🟩																				
2	Study and analysis of existing techniques	Plan				■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			
		Actual				🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩			
3	OCR & ROI	Plan							■	■	■	■													
		Actual							🟩	🟩	🟩	🟩													
4	Input embedding model building and training	Plan										■	■	■	■	■	■								
		Actual										🟩	🟩	🟩	🟩	🟩	🟩								
5	Implementation and evalution of model	Plan														■	■	■	■						
		Actual														🟩	🟩	🟩	🟩						
6	Making platform to showcase	Plan																		■	■	■	■		
		Actual																		🟩	🟩	🟩	🟩		
7	Report writing	Plan																					■	■	■
		Actual																					🟩	🟩	🟩

7.6 Student Outcomes Description and Performance Indicators

Table 7.5 Student Outcomes Description and Performance Indicators

S.No	Description	Outcome
1.1	Ability to identify and formulate problems related to computational domain	The problem was identified as the lack of good VQA Systems that perform on par with a human,
1.2	Apply engineering, science, and mathematics body of knowledge to obtain analytical, numerical, and statistical solutions to solve engineering problems.	The concepts of science and mathematics were used to compute data points in training models.
2.1	Design computing system(s) to address needs in different problem domains and build prototypes, simulations, proof of concepts, wherever necessary, that meet design and implementation specifications.	The design of the model was built step by step with continuous testing and evaluation. A web-app prototype was also implemented alongside according to the design specifications.
3.3	Able to communicate effectively with peers in well organized and logical manner using adequate technical knowledge to solve computational domain problems and issues.	All the communication was very respectful and any critique was backed by logic and acknowledged and processed with support and proper discussions.
4.1	Aware of ethical and professional responsibilities while designing and implementing computing solutions and innovations.	The sources for information as well as the dataset were both ethically sourced with consent and were added to the references. All the documentation is prepared in a professional manner.
5.1	Participate in the development and selection of ideas to meet established objective and goals.	Ideas were brainstormed and regularly discussed. The whole team showed enthusiastic participation and the best ideas were selected according to the objectives

5.2	Able to plan, share and execute task responsibilities to function effectively by creating collaborative and inclusive environment in a team.	The whole project was built in collaboration and coordination of efforts alongside regular discussions with the mentor in punctual manner. The team is very respectful and solution-oriented.
6.1	Ability to perform experimentations and further analyze the obtained results.	The model was tested regularly with modification in the model across diverse testing and training data and results were used to improve it further
6.2	Ability to analyze and interpret data, make necessary judgement(s) and draw conclusion(s).	The outcomes of testing led to changing of various modules of the model during development and these conclusions were noted.
7.1	Able to explore and utilize resources to enhance self-learning.	The whole team participated in research and development of various parts of the model and stepped outside of their comfort zone to acquire new knowledge and skills.

7.7 Brief Analytical Assessment

Q1. What sources of information did your team explore to arrive at the list of possible Project Problems?

Ans: A lot of attention was given to problems in this field and the most recent developments were carefully analyzed. Different sources research papers, journal and data available on the internet etc. to be informed about project problems

Q2. What experimental methods did your project team use to obtain solutions to the problems in the project?

Ans: Various methods including changing platforms and tools and libraries were used alongside continuous testing. Different models were used as well as the dataset was converted into different ways of storing information. Other methods using different way of asking questions were also considered.

Q3. How did your team share responsibility and communicate the information of schedule with others in a team to coordinate design and manufacturing dependencies?

Ans: As most of the time was spent collaborating online, the team members mainly coordinated through whatsapp, google suite and emails. Meetings were carried out in zoom. Everyone was assigned work according to their roles and abilities and everybody was actively involved in planning as well as feedback for design.

Q4. What resources did you use to learn new materials not taught in class for the course of the project?

Ans: Most of the work done in this project was based on very recent technologies so most of the new information was acquired through research papers, documentation on online sites, repositories on libraries on github as well as learning material from youtube and online courses. Our mentor assisted us greatly in this process too.

Q5. Does the project make you appreciate the need to solve problems in real life using engineering and could the project development make you proficient with software development tools and environments?

Ans: This project helped us understand the wide variety of applications any development in AI or NLP has. It has a very wide scope of problems and solutions. This project makes us appreciate the need to solve problems in real life using engineering. AI, ML and NLP can make life simpler. Yes, the development of this project helped us become proficient with many software development tools and environments.

Q6. Did the project demand demonstration of knowledge of fundamentals, scientific and/or engineering principles? If yes, how did you apply?

Ans: Any problem requires a fundamental understanding of it before the solution can be found. In this project a lot of fundamentals were needed like Python to implement AI and ML as well as understanding of NLP was required. To develop the various modules the lessons learnt from Software Engineering were implemented and the web app was developed using the fundamentals of web development and the backend implementation was also done using Python. We built our knowledge upon these fundamentals to develop this project and achieve targets.

Appendix A: References

[1] Singh,A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards VQA Models That Can Read. In CVPR, 8317–8326.

[2] Hu, R.; Singh, A.; Darrell, T.; and Rohrbach, M. 2020. Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA. In CVPR, 9992–10002.

[3] Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Textto-Text Transformer. J. Mach. Learn. Res. 21(140): 1–67.

[4] Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In ACL, 7871–7880.

[5]Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In Proceedings of the International Conference on Document Analysis and Recognition, 2019.

[6]Smith, R. 2007. An Overview of the Tesseract OCR Engine. In ICDAR, 629–633.

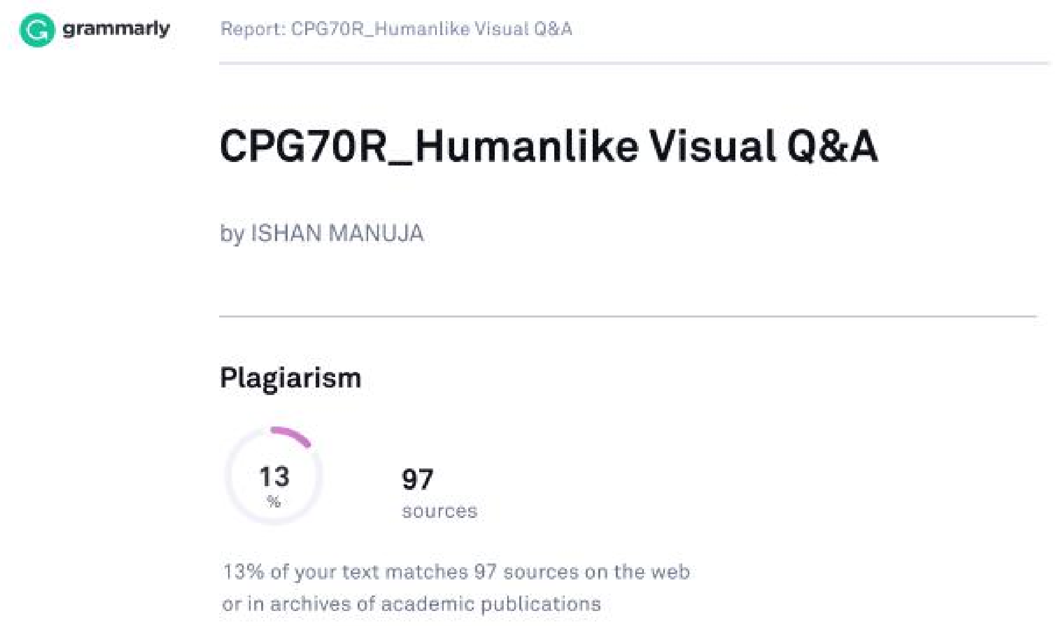
[7] Ren; Shaoqing; He; Kaiming; Girshicka; Ross; Sun; and Jian. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In NIPS, 91–99.

[8] Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In ACL, 311–318.

[9] Denkowski, M. J.; and Lavie, A. 2014. Meteor Universal:Language Specific Translation Evaluation for Any TargetLanguage. In WMT@ACL, 376–380.

[10] Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out@ACL, 74–81. [11] Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based imagedescription evaluation.

Appendix B: Plagiarism Report



Report: CPG70R_Humanlike Visual Q&A

CPG70R_Humanlike Visual Q&A

by ISHAN MANUJA

Plagiarism

13 %

97 sources

13% of your text matches 97 sources on the web or in archives of academic publications

Fig. Appendix B.1 Plagiarism report