Dissertation on

"VISUAL QUESTION ANSWERING ON STATISTICAL PLOTS"

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelor of Technology
in
Computer Science & Engineering

UE18CS390A – Capstone Project Phase - 1

Submitted by:

Sneha Jayaraman	**PES1201802825**
Sooryanath I T	**PES1201802827**
Himanshu Jain	**PES1201802828**

Under the guidance of

Prof. Mamatha H.R.
Professor
PES University

January - May 2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

‘VISUAL QUESTION ANSWERING ON STATISTICAL PLOTS’

is a bonafide work carried out by

Sneha Jayaraman	**PES1201802825**
Sooryanath I T	**PES1201802827**
Himanshu Jain	**PES1201802828**

in partial fulfilment for the completion of sixth semester Capstone Project Phase - 1 (UE18CS390A) in the Program of Study - Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Jan. 2021 – May. 2021. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 6th semester academic requirements in respect of project work.

Signature	Signature	Signature
Prof. Mamatha H.R.	Dr. Shylaja S S	Dr. B K Keshavan
Designation	Chairperson	Dean of Faculty

External Viva

Name of the Examiners **Signature with Date**

1. ____________________________ ____________________________

2. ____________________________ ____________________________

DECLARATION

We hereby declare that the Capstone Project Phase - 1 entitled **“VISUAL QUESTION ANSWERING ON STATISTICAL PLOTS”** has been carried out by us under the guidance of Prof. Mamatha H.R., Professor and submitted in partial fulfilment of the course requirements for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester January – May 2021. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

PES1201802825	**Sneha Jayaraman**	Sneha
PES1201802827	**Sooryanath I T**	
PES1201802828	**Himanshu Jain**	Himansh

ACKNOWLEDGEMENT

I would like to express my gratitude to Prof. Mamatha H.R., Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE18CS390A - Capstone Project Phase – 1.

I am grateful to the project coordinators, Prof. Sunitha R and Prof. Silviya Nancy J, for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Shylaja S S, Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support I have received from the department. I would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University for providing to me various opportunities and enlightenment every step of the way. Finally, this project could not have been completed without the continual support and encouragement I have received from my family and friends.

ABSTRACT

Question answering systems have been used in various domains and applications like dialog systems, and medical domains for interaction with patients. We have identified the domain of data analysis where question answering systems can be used. The statistical charts are used on a regular basis for data visualization to interpret the data and derive meaningful inference from them. This process can be automated using Question answering systems. Users can impose a question to the system, for a particular statistical chart, and the system must provide the answer to it in the most accurate manner. Building such a system requires the usage of right architecture, right frameworks and huge amounts of data. Once the model is built that satisfies the requirements, it can be deployed to a web application where a user can upload an image and, input a question, to generate the expected answer.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Statistical charts are an intuitive and simple way to represent data. They form a way of representing structured data in the form of graphical visualisations. Such graphical visualizations aid people in better interpreting features of data.

Deep Learning is a field that focuses on emulating human intelligence to develop new models that can reason figures and understand relationships that are intuitive to humans. Therefore, it is useful to build a model that can reason visual data in statistical plots. It is one step towards the improvement in machine reasoning capabilities.

Visual plots are commonly found in research papers, scientific journals, business records e.t.c. Therefore, automation of plot analysis through the means of question-answering aids an individual to draw statistical inferences quickly from them.

The most important benefit is that visual question answering models on charts will help data analysts question and reason plots on a large scale, and automate the decision-making capabilities in several sectors such as the financial sector.

Given this motivation, the aim of the project is to build a Visual Question Answering system which accepts statistical plots along with questions on the plot with respect to the elements of the plot (such as intersection of the curves, area under the curve, median value and few other varieties of such relational queries) and provides answers to the questions posed.

The system should discover relationships between elements of a plot and provide relational reasoning to answer questions on the plot.

Given an image of a statistical plot and a corresponding question, the model must be able to generate a representation of the image, parse and understand the query, and generate a suitable reply. Therefore, it involves an understanding of image and the query language to be able to provide for visual reasoning.

CHAPTER 2

PROBLEM STATEMENT

Statistical plots are used widely by academicians and business employees because they are a simple way to represent data. They can be easily analysed and interpreted.

What if we could build an automated system that can analyse, discover relationships between elements of a plot and provide for relational reasoning capabilities? Such a system would mark a step towards machine reasoning capabilities.

With this motivation, the project aims to build a Visual Question Answering system that accepts statistical plots along with user-specific questions concerning the elements of the underlying plot, such as the intersection of the plots, area under the curve and few other varieties of such relational queries, to provide answers.

The difficulty with statistical plots is that even though they are images, they contain both structured and unstructured data. In the case of natural images, there are just visual elements to handle. However, that is not the case with statistical plots, since they contain both visual elements in the form of bars/sectors and textual elements in the form of axis labels and ticks. To add to this, the size of objects that are there in natural images are constrained to small, medium and large in general. In the case of statistical plots, the aspect ratio is much more varied. For example, in the case of bar plots, there could be bars that are extremely small, and bars that are extremely large on the same plot.

The measure of the accuracy of prediction in the case of images is normally IOU (intersection over union). The success criteria for a correct prediction in the case of natural images is normally 50 per cent. The same rate is insufficient for statistical plots. This is because we want the prediction for a bar value (in the case of bar plots) to correspond to the actual value as seen on the graph, as close as possible. Therefore the adequacy criteria for a successful prediction is much higher.

The above-mentioned factors highlight the differences in natural images and statistical plots. Therefore, state-of-the-art object detection models do not suffice for this application. The aim here is to apply deep learning concepts to come to an acceptable solution to the problem of analysing statistical plots using machines.

CHAPTER 3

LITERATURE SURVEY

In the following subdivision , we present the current understanding and knowledge of the area along with reviewing substantial findings that help shape, inform and reform our study leading us to set the platform to further envisage the possible improvements that can be brought up.

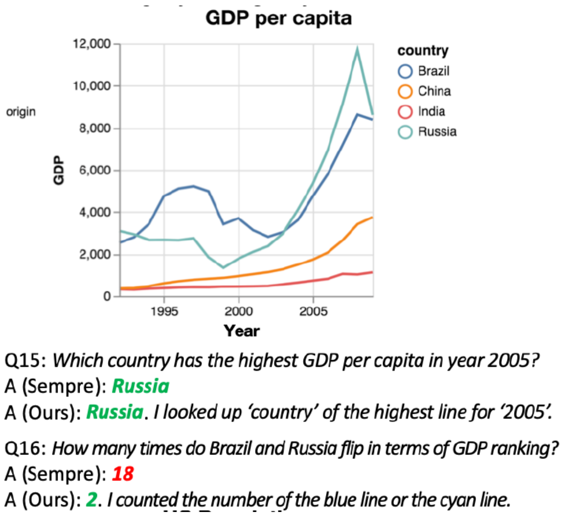
3.1 Background on Statistical Charts modelling for QA system

This section briefs the papers consulted and thoroughly reviewed to gain information on background, data used, the architecture style , the patterns and the proposed/existing methodologies being used in the domain of question answering system for statistical charts.

3.1.1 Answering Questions about Charts and Generating Visual Explanations

Summary

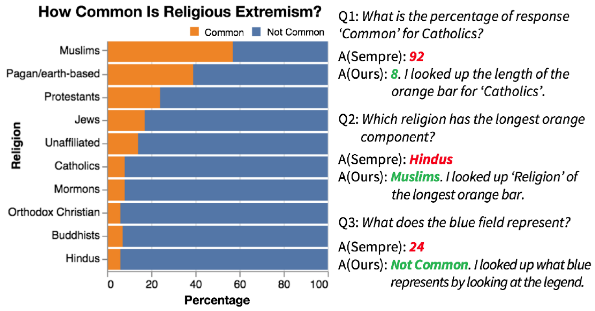
The paper under consideration proposes the chart question answering system that generates chart specific answers along with the explanation on how the answer was obtained. The visual attributes of the charts are transformed into references to the data. State-of-the-art ML algorithms are used to generate answers and its corresponding explanation.



Q15: *Which country has the highest GDP per capita in year 2005?*
A (Sempre): ***Russia***
A (Ours): ***Russia***. *I looked up 'country' of the highest line for '2005'.*

Q16: *How many times do Brazil and Russia flip in terms of GDP ranking?*
A (Sempre): ***18***
A (Ours): ***2***. *I counted the number of the blue line or the cyan line.*

Figure 1



Q1: *What is the percentage of response 'Common' for Catholics?*
A(Sempre): ***92***
A(Ours): ***8***. *I looked up the length of the orange bar for 'Catholics'.*

Q2: *Which religion has the longest orange component?*
A(Sempre): ***Hindus***
A(Ours): ***Muslims***. *I looked up 'Religion' of the longest orange bar.*

Q3: *What does the blue field represent?*
A(Sempre): ***24***
A(Ours): ***Not Common***. *I looked up what blue represents by looking at the legend.*

Figure 2

Figure 1 illustrates two question and answer pairs for a line plot. The results are compared between Sempre model and the model proposed.

Figure 2 illustrates three sample question and answer pairs for a stacked horizontal bar plot.

Dataset

The compilation of dataset consists of 52 charts, congregated from four contrasting sources:

- The Vega-Lite Example Gallery
- Graphical Charts in Pew Research Reports
- D3 charts that are accumulated from the internet
- Charts fabricated from tables present in the WikiTableQuestions data compilation.

The questions, answers and their explanations were manually generated.

Dataset Counts :

In union, the compiled data includes 5 line charts and 47 bar charts (32 simple, 8 grouped, 7 stacked).

In total, 52 charts are synthesized that includes 629 questions, 866 answers and 748 explanations for the answers generated.

Model

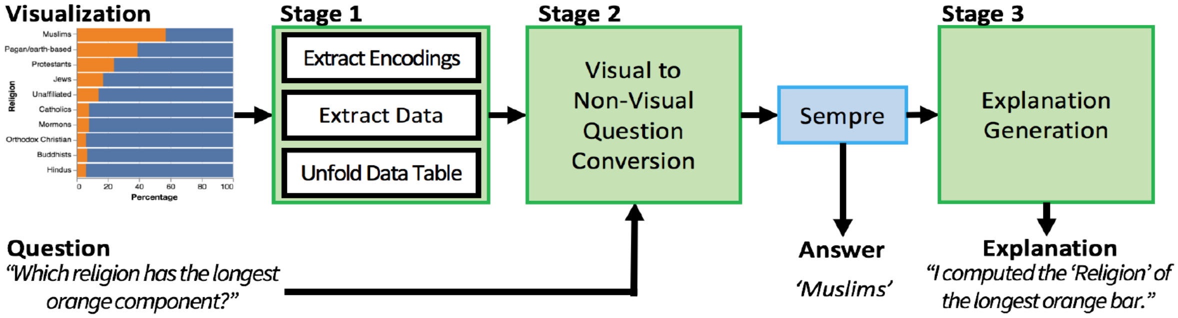
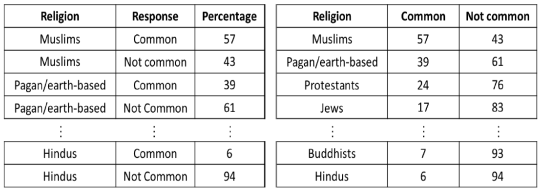


Figure 3



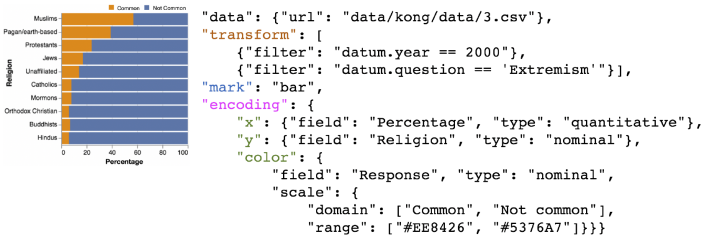
Religion	Response	Percentage
Muslims	Common	57
Muslims	Not common	43
Pagan/earth-based	Common	39
Pagan/earth-based	Not Common	61
⋮	⋮	⋮
Hindus	Common	6
Hindus	Not Common	94

(a) Flat data table

Religion	Common	Not common
Muslims	57	43
Pagan/earth-based	39	61
Protestants	24	76
Jews	17	83
⋮	⋮	⋮
Buddhists	7	93
Hindus	6	94

(b) Unfolded data table

Figure 4



```
"data": {"url": "data/kong/data/3.csv"},
"transform": [
    {"filter": "datum.year == 2000"},
    {"filter": "datum.question == 'Extremism'"}],
"mark": "bar",
"encoding": {
    "x": {"field": "Percentage", "type": "quantitative"},
    "y": {"field": "Religion", "type": "nominal"},
    "color": {
        "field": "Response", "type": "nominal",
        "scale": {
            "domain": ["Common", "Not common"],
            "range": ["#EE8426", "#5376A7"]}}}
```

Figure 5

Figure 3 illustrates the pipeline of this model for question answering system

Figure 4 shows the dataset format

Figure 5 shows the unfolded data table

Methodology proposed

Firstly, visual encodings like the pinnacle aka height of the bar, color/shade grading of the line, etc. are extracted from the charts. The input question is transformed, replacing any visual references made by chart elements to the non-visual references to the data fields and data values. The Unfolded table is passed through Sempre (QA algorithm that functions with relational data tables instead of any statistical charts) to generate the answer. Sempre model converts the input question into a lambda expression. It then performs query execution on the data table generated to produce the answer.

The lambda expression obtained is transformed to visual explanation for the answers by a method known as template-based translation.

Formal Steps Stated in the paper

- **Stage 1**: Extract Data Table and Encodings
- **Stage 2**: Visual to Non-Visual Question Conversion
  - Step 1: Mark detection
  - Step 2: Dependency parsing
  - Step 3: Visual attribute detection
  - Step 4: Visual operation detection
  - Step 5: Apply encodings
  - Step 6: Natural language conversion
- **Stage 3**: Explanation Generation
  - Step 1: Natural language conversion
  - Step 2: Implicit field recovery
  - Step 3: Redundancy Cleanup
  - Step 4: Sentence Completion
  - Step 5: Encoding application

Merits

The paper not only provides accurate answers to the questions but also provides an explanation on how the answer was obtained. The model produces valid answers and their corresponding explanations as opposed to the Sempre model that could not answer the questions correctly.

Demerits

The system cannot handle certain types of questions that involve synonyms of the features present in the chart. There is scope for improving the explanation provided for the answers.

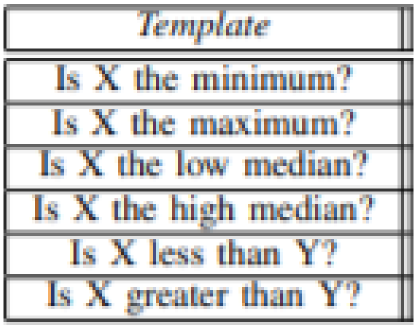
3.1.2 FigureNet: A Deep Learning model for Question-Answering on Scientific Plots

Summary

This model uses a CNN with depth-wise convolutions, LSTM and feed-forward NN to handle the task of answering questions on plot such as pie and bar on a dataset that's named FigureQA.

Dataset

The dataset used is the FigureQA dataset that contains more than a million questions with answers on various types of scientific plots. This dataset has plots with elements that are color-coded. There are a total of 100 colors that are used across both training and test datasets. Therefore, it is possible to distinguish between elements without the need of character recognition for text. Additionally, it provides for pre-annotated data with bounding boxes.



Template
Is X the minimum?
Is X the maximum?
Is X the low median?
Is X the high median?
Is X less than Y?
Is X greater than Y?

Figure 6

Figure 6 shows the template of questions in the FigureQA dataset.

Model

The FigureNet architecture, as proposed, can handle the task of answering relational questions on pie charts and bar plots. It uses the FigureQA dataset, that consists of statistical plots with plot elements that are color coded. Additionally, it is guaranteed that the plot consists of no more than 11 plot elements, and that there are 100 different colors that are used to represent plot elements.

The end goal of the FigureNet model is to be able to answer questions in a binary yes/no manner. To be able to do this, the authors have divided the task into subtasks as follows.

- Spectral Segregator Module - Identify plot elements and color of the plot elements
- Order Extraction Module - Identify and quantify the values associated with each plot element, and then sort it into increasing order.
- Question Encoding - Provide an encoding for the question.
- Question Color encoding - Identify mentions of color in the question.

Methodology proposed

Spectral Segregator Module:

This module is used to identify individual elements and color of these elements of the plot. 128 x 128 x 3 image is passed as input to a CNN that uses depth-wise convolutions to identify colors and separate channel information. This way we don't just get an aggregate map of the image. The output here is a 512 dimensional image representation. This image representation is passed to a 2-layer LSTM to get the most probable color for each element. There can be at most 11 elements in a plot.

Order Extraction Module:

This module is used to identify and quantify the statistical values of each plot element and their relative order. It is similar to that of the previous module except that now the output of the LSTM will be the ordering for each of the plot elements starting from 1.

Question Encoding and Question Color encoding:

This module uses 2 layers of LSTM cells (many to one model) to produce a question encoding.

Final feed-forward NN:

All the four modules are concatenated and passed onto a feed forward NN to produce a binary (Yes/No) answer using the Sigmoid Activation function for the output layer.

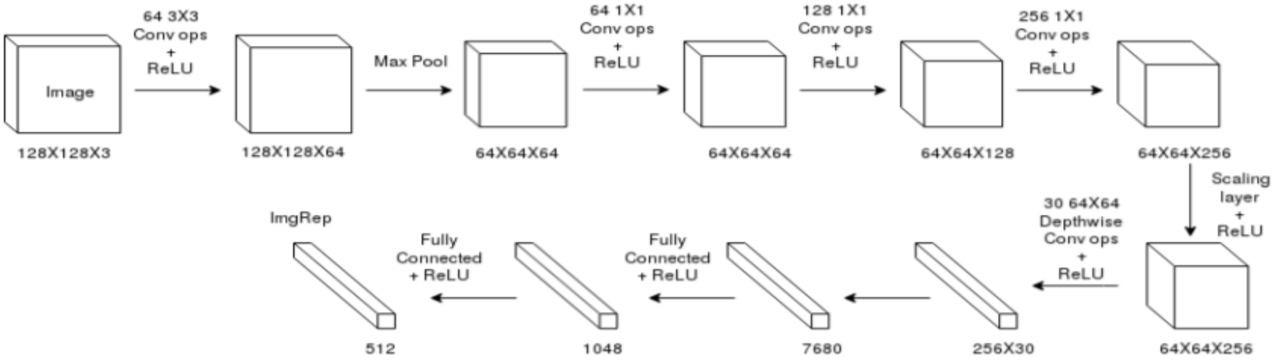
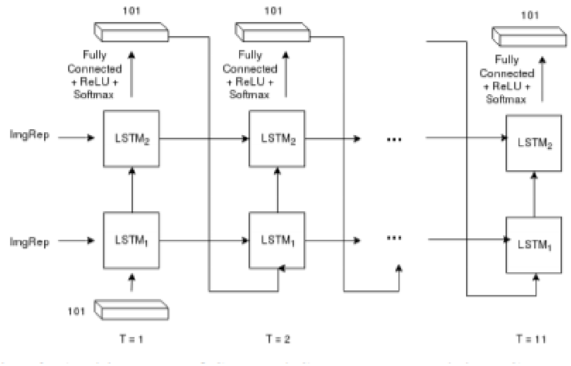
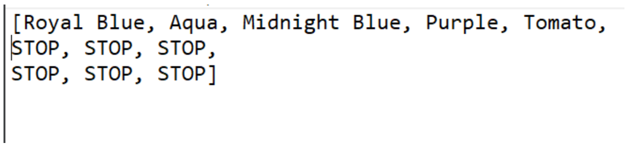


Figure 7

Figure 7 shows the architecture of the Spectral Segregator Module that uses layers of convolution and max pool, followed by depthwise convolutions and feed forward layers.


101
101
101
Fully Connected + ReLU + Softmax
Fully Connected + ReLU + Softmax
Fully Connected + ReLU + Softmax
ImgRep
ImgRep
LSTM$_2$
LSTM$_2$
LSTM$_2$
LSTM$_1$
LSTM$_1$
LSTM$_1$
...
...
101
T = 1
T = 2
T = 11



```
[Royal Blue, Aqua, Midnight Blue, Purple, Tomato,
STOP, STOP, STOP,
STOP, STOP, STOP]
```

Figure 8

Figure 9

Figure 8 shows the rest of the spectral segrator module that uses a custom LSTM architecture. The input here is the image representation that is obtained as the output from the architecture in Figure 7.

Figure 9 shows a sample output as obtained from the architecture in Figure 8.

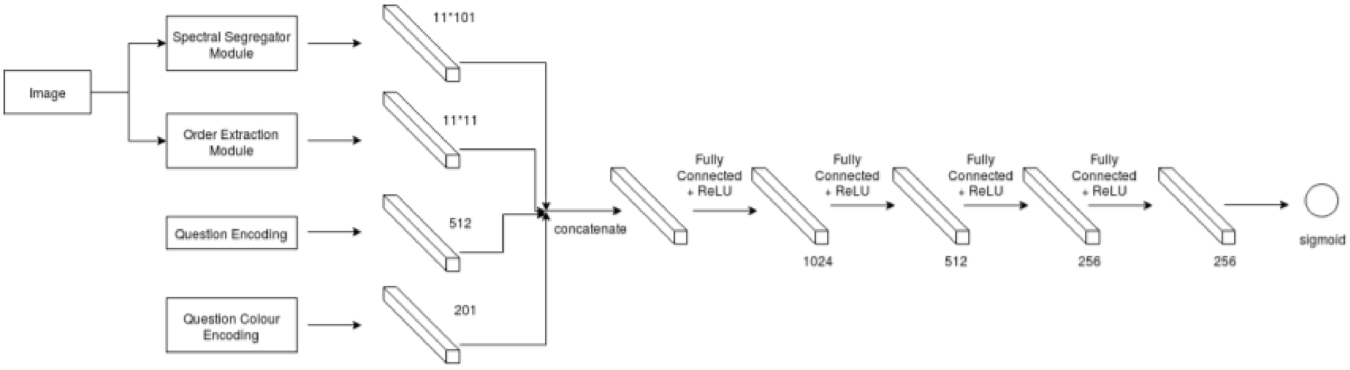


Figure 10

Figure 10 shows the final feed forward architecture.

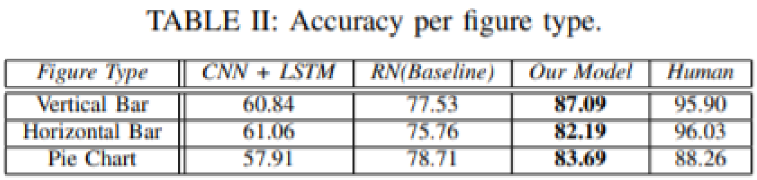


TABLE II: Accuracy per figure type.

Figure Type	*CNN + LSTM*	*RN(Baseline)*	*Our Model*	*Human*
Vertical Bar	60.84	77.53	**87.09**	95.90
Horizontal Bar	61.06	75.76	**82.19**	96.03
Pie Chart	57.91	78.71	**83.69**	88.26

Figure 11

Figure 11 depicts a table comparing accuracy values for each type of plot.

Merits

The model performs significantly better than the baseline models. This is because the architecture doesn’t use the traditional CNN, instead uses depthwise convolutions. Additionally, the model used lesser training time as articulated in the paper.

Demerits

The model works on only bar plots and pie charts. It is capable of only binary reasoning, and not capable of answering open-ended questions. It makes use of the FigureQA dataset, thereby making use of the property of the charts being color coded.

3.1.3 ChartNet: Visual Reasoning over Statistical Charts using MAC-Networks [3]

Summary

Proposed paper solves the problem of reasoning over charts (only bar and pie charts) using MAC-Network (Memory, Attention, and Composition). The model is capable of answering open-ended questions and gives chart-specific answers. The classification layer of MAC is substituted by the regression layer and constructs a boundary for the text that corresponds to the answer. OCR is used to read the text and display the answer.

Dataset

The data synthesized by the model consists of bar-charts and pie-charts was made. The bar charts dataset consists of vertical bars and was created by varying the height and number of colors of bars. Data for pie-charts is created by varying the colors of sectors and also the angles between them. The annotations of the bounding box that are present over the chart images are saved to give chart specific answers. In total, 20k, 5k and 5k image question pairs for training, validation and testing are created, for both bar charts and pie charts.

ChartNet: Visual Reasoning over Statistical Charts using MAC-Networks

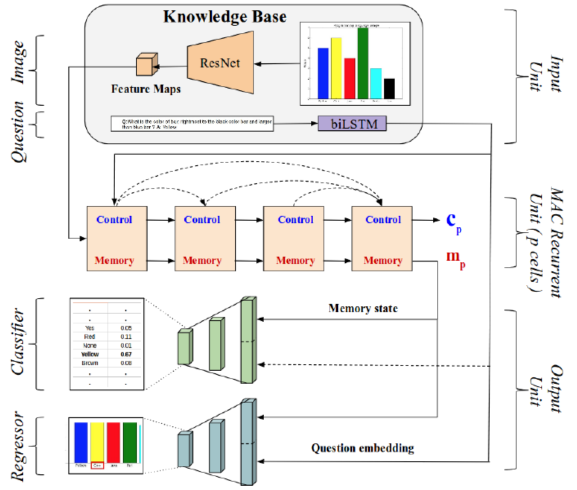


Fig. 2. Flowchart showing proposed architecture of *ChartNet* for visual reasoning over bar and pie charts. The Knowledge base consists of visual feature maps extracted using a ResNet-101 [5] pre-trained model. The question is encoded using a Bidirectional LSTM. A recurrent MAC layer is used to generate the reasoning output at each step, based on the question and two fully connected branches perform classification over a generic set of answers and regress the coordinates of image specific answers.

Figure 12

Figure 12 shows the architecture of the model proposed.

Methodology proposed

ChartNet network consists of three layers: Input unit, MAC Cell, Output unit.

Input Unit

Bar or pie chart is given as an input and corresponding question. Features from the images are extracted using ResNet101 deep CNN architecture. Knowledge base is defined to depict the height and the width image. Questions are converted into word embeddings and they are further processed using the biLSTM model.

MAC Cell

It represents a recurrent unit which consists of three components : Control, Read and Write. It is defined to reason the questions posed and also to implement them.

Output Unit

This unit consists of two networks : Classifier and Regressor. Classifier network predicts the probability distribution over all of the predefined answers by using softmax normalization. The regressor network is used to provide chart-specific answers.

Merits

Automated method for question answering over open-ended questions. MAC-Network included with the regression layer helps the model make prediction over unseen answers.

Demerits

The model is not generic and works only for vertical bar charts and pie charts. Model cannot answer questions that require numerical operations.

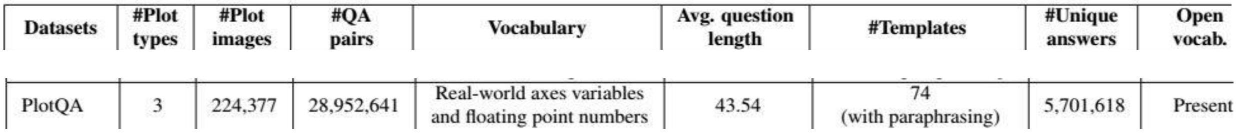
3.1.4 PlotQA: Reasoning over Scientific Plots [4]

Summary

A step towards developing a holistic plot based visual question answering model , which can handle both in vocabulary and open ended queries using a hybrid approach.

Dataset

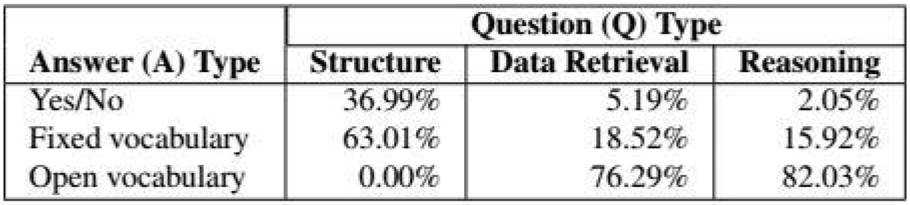
The graphical summaries are produced from data provenanced from the organizations like the World Bank, government maintained sites to name a few , thereby having a large vocabulary of graph parameters like ticks , and a wide variety of range in data instances. Out of vocabulary questions are generated and they are not straight forward as they are generated on the basis of 70 plus patterns extracted from 7,000 public flock questions asked by data collectors on a sampled set of 1000+ plots.



Datasets	#Plot types	#Plot images	#QA pairs	Vocabulary	Avg. question length	#Templates	#Unique answers	Open vocab.
PlotQA	3	224,377	28,952,641	Real-world axes variables and floating point numbers	43.54	74 (with paraphrasing)	5,701,618	Present

Figure 13

Figure 13 summarises the dataset used.



Answer (A) Type	Question (Q) Type		
	Structure	Data Retrieval	Reasoning
Yes/No	36.99%	5.19%	2.05%
Fixed vocabulary	63.01%	18.52%	15.92%
Open vocabulary	0.00%	76.29%	82.03%

Figure 14

Figure 14 shows the long range distribution of Query and Response types from the data compilation in the PlotQA data compilation.

Model

Existing Works :

Existing solutions for VQA fall under two categories: (i) extricate the response from the graphical data input (like in LoRRA) or (ii) respond with an answer to the query posed based on the existing vocabulary (like in SAN and BAN). Such approaches seem to work well for data compilations as portrayed in DVQA , but under fit for PlotQA with a considerable majority of out of vocabulary queries.

Plot QA's Model :

This is a composite model encompassing the below features and entities:

(i) a binary categorizer for decision making as to whether the input query be responded with from an existing vocabulary or demands an advanced treatment.

(ii) a simpler question categorizer to respond to queries of the simpler or complex treatment type..

(iii) The process pipeline with the CNN and LSTM combination setup.

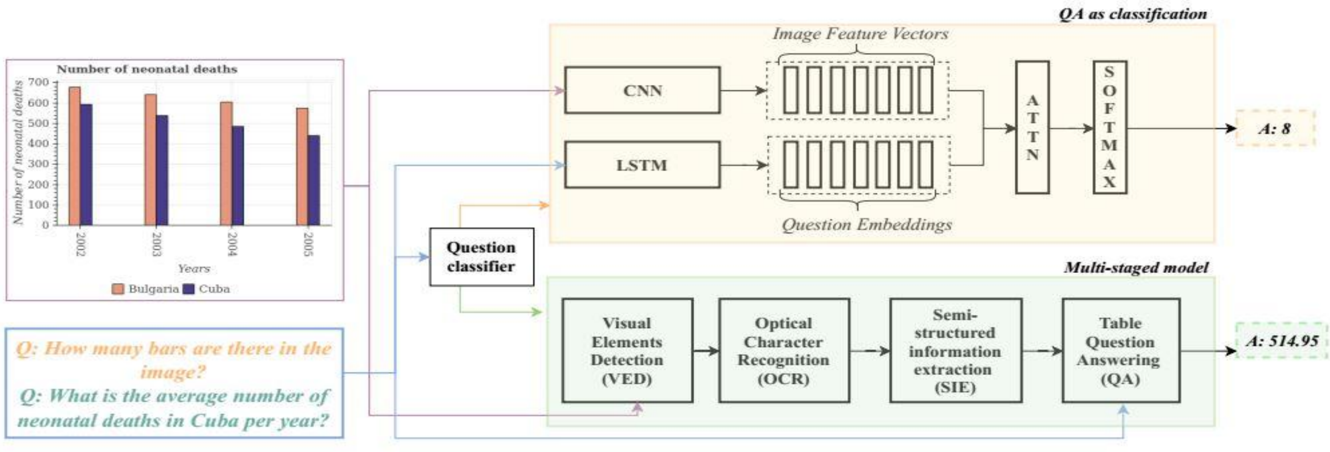

QA as classification
Image Feature Vectors
Number of neonatal deaths
700
600
500
400
300
200
100
0
Number of neonatal deaths
2002
2003
2004
2005
Years
Bulgaria
Cuba
CNN
LSTM
A
T
T
N
S
O
F
T
M
A
X
A: 8
Question Embeddings
Question classifier
Multi-staged model
Q: How many bars are there in the image?
Q: What is the average number of neonatal deaths in Cuba per year?
Visual Elements Detection (VED)
Optical Character Recognition (OCR)
Semi-structured information extraction (SIE)
Table Question Answering (QA)
A: 514.95

Figure 15

Figure 15 shows the architecture of the model proposed. Observe the two pipelines.

Methodology proposed

This is a composite mix match model encompassing the the detailed entities as follows: (i) a binary classifier for categorizing the complexity of the question and to provide decision about if it can be handled by in vocabulary or does it need assistance of out of vocabulary processing pipeline (ii) a simpler question categorizer model to respond to queries of types as mentioned in (i),finally (iii) presence of a multi-staged model encompassing four software modules as briefed in the following section to deal with the lower half of the pipeline which is the out of vocabulary questions.

1. Visual Elements Detection Module

Primary task to perform is to extricate the visual entities by demarcating / annotating bounding boxes around those entities and categorize them into the appropriate groups.

Upon comparing all methods, it is observed clearly that the Faster R-CNN model in union /combination with Feature Pyramid Network (FPN) outperforms the existing architecture combinations and hence that becomes an apt fit for the visual element detection module.

2. Object Character Recognition Module:

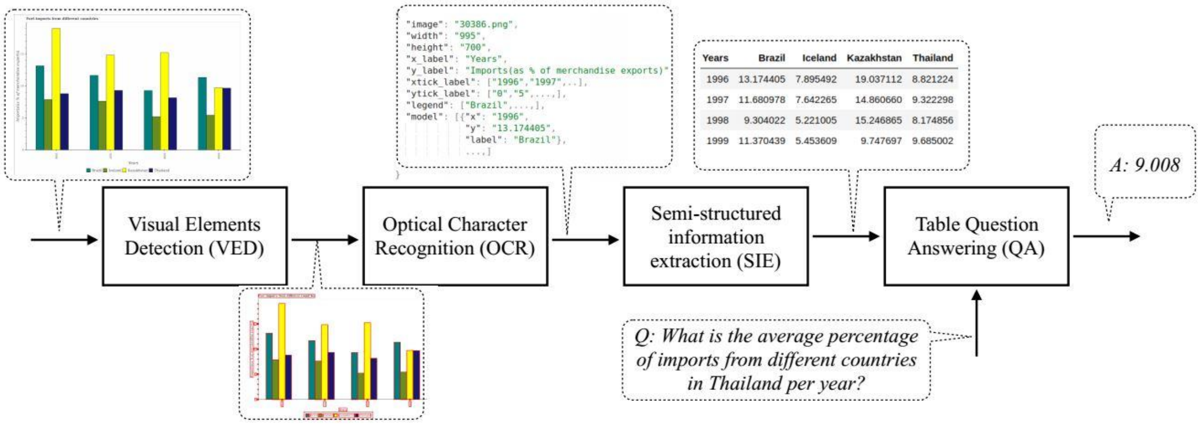
The common visual entities of the graphical summaries like legends, tick labels to name a few . accommodate numerical and textual data. For the purpose of extricating this data from bounding boxes annotations , the avant-garde OCR model is used.

3. Semi-Structured Information Extraction Module:

The penultimate phase of the pipeline. The data captured as results in the form of json/dictionary from the previous phase is formatted into a table structure using this module.

4. Table Question Answering Phase:

The ultimate final phase of the processing pipeline is to answer queries by superimposing them on the semi-structured pivoted table version of the image which has now been dissected. This is akin to answering questions from the WikiTableQuestions dataset , so the similar process is recreated to obtain results.


Visual Elements Detection (VED)
Optical Character Recognition (OCR)
Semi-structured information extraction (SIE)
Table Question Answering (QA)
A: 9.008
Q: What is the average percentage of imports from different countries in Thailand per year?

```
"image": "30386.png",
"width": "995",
"height": "700",
"x_label": "Years",
"y_label": "Imports(as % of merchandise exports)"
"xtick_label": ["1996","1997",..],
"ytick_label": ["0","5",....],
"legend": ["Brazil",....],
"model": [{"x": "1996",
           "y": "13.174405",
           "label": "Brazil"},
           ....]
}
```

Years	Brazil	Iceland	Kazakhstan	Thailand
1996	13.174405	7.895492	19.037112	8.821224
1997	11.680978	7.642265	14.860660	9.322298
1998	9.304022	5.221005	15.246865	8.174856
1999	11.370439	5.453609	9.747697	9.685002

Figure 3: Our proposed multi-staged modular pipeline for QA on scientific plots.

Figure 16

Figure 16 shows the proposed multi-staged modular pipeline.

Merits

Can handle out of the vocabulary questions (OOV) along with in-vocabulary question types. The data collected to prepare graphs in the dataset are from various financial and business resources. Blurs the line of difference between computerized-data plot datasets and real life data summarized in graphs and query patterns.

Demerits

The model is not generic and works only for bar charts , line charts and dot plots. There exists a want/development in regard to more precise visual element detecting (VED) modules to enhance responses over the queries posed on the plots.

CHAPTER 4

DATA

The following chapter describes the data used to build the question answering system for statistical plots. Building the right questionnaire with right visualization is critical to build a high accuracy model.

4.1 Overview

Statistical plots are used to represent the data and help in deriving insights from them. Some of the basic charts are bar charts, pie charts and line charts. Visual aid is required to analyze these charts and answer some of the questions related to them.

4.2 Data Format

Statistical plots are used to learn about data and their important features. Building a visual question answering system requires statistical plots and well-defined questions and answers. The detailed description of the charts and the question answer pair is provided in further sections.

4.3 Statistical Charts

Statistical charts form one of the inputs to the visual question answering system. The types of charts that we have constrained to are bar charts, pie charts and optionally line charts for further research.

A typical bar chart will include:

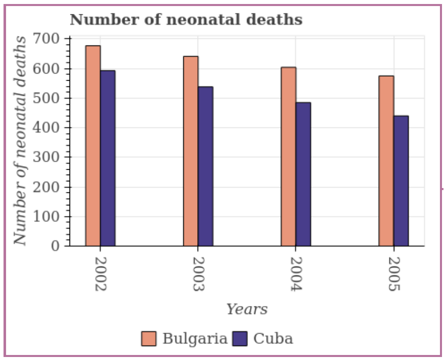
- Title
- Vertical axis - Numerical labels

- Horizontal axis - Categorical labels
- Legends

There are different types of bar charts like vertical charts, horizontal charts, stacked charts, and group charts. The aim of this research is to address some of these types and use them to correctly answer the questions.

A pie chart includes different sectors of varying proportion and color coded according to the legends mentioned in the charts. Identifying different portions of these charts according to the questions asked is one of the challenges of this research work.

Shown below are the examples for types of plots.



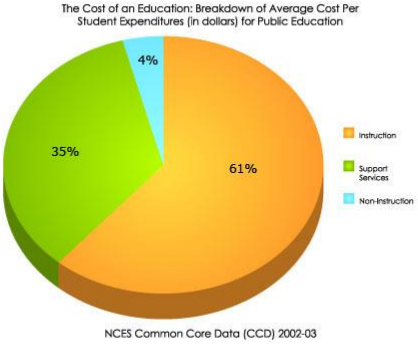


Figure 17 Bar Chart

Figure 18 Pie Chart

(Image Courtesy:PlotQA: Reasoning over Scientific Plots)

(Image Courtesy:https://nces.ed.gov/nceskids/help/user_guide/graph/pie.asp)

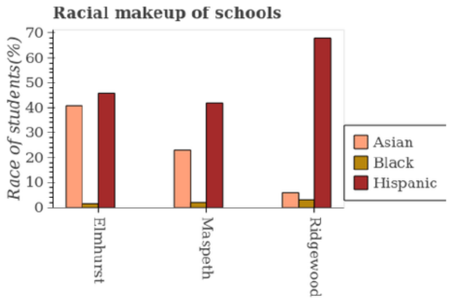
The dataset will include these charts in a well-ordered, properly named format. Each image will be annotated using image processing to identify the axis labels, legends, the bar heights (in case of a vertical bar chart), area of the sector (in case of a pie chart) and many more fields required to answer the questions.

4.4 Question Answer

Every image has a list of questions and their answers stored in the repository. Descriptive questions or Open-ended questions are generated for each image. The answers for them are also stored. Some of the examples of the question answers are shown in the figure.

During the training of the model, the chart, list of questions and their corresponding answers are passed as the input. The model learns to predict the answer. During testing, only the chart and a question is taken as the input, and the model predicts the answer.

Charts are annotated and mapped with the correct references made in the question. For example, if one of the categories in the bar chart is Hispanic and the question involves some kind of reference to this category, then the model must be able to identify this and draw a bounding box around it. This will help the model to correctly analyze the chart and extract required numerical information from it and answer the open-ended question correctly.



Q: What is the average number of Hispanic students in schools? A: 51.67

Figure 19 Bar Chart with one example of question and answer

(Image Courtesy:PlotQA: Reasoning over Scientific Plots)

CHAPTER 5

SYSTEM REQUIREMENTS SPECIFICATION

1. Project Scope

Goal

The system should be able to answer questions on Vertical and Horizontal Bar Graphs, Pie Charts, and Line Plots.

Limitations

The system will not be able to answer questions on the smoothness or the roughness of the plots. It can only answer relational queries – queries with respect to the other elements of the plot.

2. Product Perspective

Visual question answering specific to statistical plots, has less prevalent work done. It is one step towards the improvement of machine reasoning and pattern identifying capabilities.

2.1 Product Features

1. Input: The input to our model is an image of a statistical plot and a corresponding question on the plot.
2. Model: Our model will take in the input, process the visual image and parse the query, concatenate the inferences from both and produce an output. Thus, it combines the technicalities of image processing and query language understanding.
3. Output: The output of the model is the answer to the question.

The product consists of the model, and an interface to the model.

2.2 Operating Environment

The model will be made available as a web application; hence it does not depend on the underlying Operating system and its versions. It only requires a browser support of HTML5 and above. Use of the model can be done via the internet.

On the server side, the model can be saved and loaded when necessary. Therefore, the platform must be reliable and available.

2.3 General Constraints, Assumptions and Dependencies

The project focuses on model building and providing for accurate and reliable answers to the questions posed on the statistical charts. Hence, there is less focus on the security considerations. However, our solution will consist of an interface to the model that facilitates image upload and questions on the image.

2.4 Risks

Operational risk in terms of management and support for the product is a possible risk case.

Functional Requirements

Question Answering System for Charts take in statistical charts and questions related to them as input. Visual features from the charts have to be extracted and preprocessed. This can be done using techniques like image processing and Optical Character Recognition (OCR). The questions provided as an input need to be pre-processed using NLP techniques. Important details from the image and the questions need to be mapped and the corresponding answer should be predicted. Deep learning methods and architectures will be employed to predict the output.

Inputs are validated by the system to recognize only statistical charts. Any other images will be rejected by the system.

The system will be trained on huge amounts of data and the results will be validated against the true answers. The parameters for the model will be tuned to provide the most optimal answer as the output.

3. External Interface Requirements

3.1 User Facing Interfaces (UI)

The user interface is a web application which will take a chart and user-specific question. Users will be allowed to upload an image from their local drive. A text box will be provided that will accept the questions. The model will take the images and questions as the input and run in the backend. It will return the most probable answer and display it on the screen. Additional data like the prediction accuracy can also be displayed. A trained model will be deployed on the web server, and hence, the output should be produced within a few seconds.

3.2 Hardware Requirements

Deep learning tasks are compute intensive. The hardware requirements to build and train the model will require a minimum of 8GB RAM. Optionally, the model can be trained on the cloud to avoid physical hardware limitations. Once the model is trained, the testing phase can be done on any commodity hardwares. Web application development places no hard constraints on the hardware requirements.

3.3 Software Requirements

Question Answering Systems are built using the deep learning models and NLP techniques. Python (Version: 3.6 or more), NLP libraries, image processing tools and deep learning frameworks are required to build the model. Web frameworks like React and Flask are required to build and deploy the web application. The system can be built on any operating system like Windows or Ubuntu. Versioning of the model will be done using GitHub. Updated versions will be merged to the main branch and any testing and feature engineering will be done separately.

3.4 Communication Interfaces

Communication interfaces are not required to build the question answering system as the model is tested and run on the local machine.

4. Non-Functional Requirements

4.1 Performance Requirement

- **Usability** : The trained model must be available at ease to use it ,just by choosing an image input from the local drive or file-system . Similar to drag and drop or attaching files. The user must be able to navigate through the interface even with minimal exposure towards computing technologies.

- **Reliability** : As the product is developed by following practices in deep-learning , machine learning and imaging , there is no certain fixed level of reliability that can be set for the product. Reliability is not completely independent of the inputs to the model , hence reliability varies with respect to the context and type of inputs passed in.

- **Maintainability** : As the product is just a model , maintaining the model isn't a difficult task , it only requires a machine to reside on and a browser / interface to access.

- **Performance** : The model is expected to draw statistical inferences based on the question and the input passed with a good and acceptable accuracy level.

- **Robustness** : The model must be robust enough to classify any of the graph images and questions related to its scope of operation. The model must yield good performance for any input from the wide range of possible inputs pertaining to the scope of the model.

4.2 Safety Requirements

Safety requirements pertaining to this product are minimal as it isn't deployed onto a live physical environment . The decision process for users , such as passing image inputs or attaching them to gain access to model /product and data must follow a need-to-know principle, which states that access to internal data must be available only to the designers of the model and shouldn't be exposed to the end users.

4.3 Security Requirements

Data Sharing : As there are lots of graphical images to be collected , the data and statistics storage will be done to maintain the correct functioning of the model and to reconstruct what went wrong in case of any system - failures by constructing checkpoints. The datasets if artificially generated need to be secured locally and not be made available for commercial usages.

Model security : The model trained needs to be protected on a local machine or online if deployed onto the cloud . Modification of firewall rules or enabling regular scans may help in securing the build model.

CHAPTER 6

SYSTEM DESIGN

6.1 High Level Design:

Question Answering System for Charts take in statistical charts and questions related to them as input. Visual features from the charts have to be extracted and preprocessed. This can be done using techniques like image processing and Optical Character Recognition (OCR). The questions provided as an input need to be pre-processed using NLP techniques. Important details from the image and the questions need to be mapped and the corresponding answer should be predicted. Deep learning methods and architectures will be employed to predict the output.

Inputs are validated by the system to recognize only statistical charts. Any other images will be rejected by the system. The system will be trained on huge amounts of data and the results will be validated against the true answers. The parameters for the model will be tuned to provide the most optimal answer as the output.

6.2 Current System

There have been attempts in the recent past to improve machine reasoning capabilities through visual question answering systems on graphical plots. The RN architecture and the CNN-LSTM architecture form baseline comparison models with an accuracy of 75% and 60% respectively. This in comparison to human accuracy falls short by a large margin.

A recent paper publication introduced the FigureNet architecture that was able to achieve an accuracy of approximately 85% on an open-source dataset. This however, only gives a yes/no binary output to a question posed, and is limited to only bar and pie charts.

Another adaptation to this model, showed significant enhancements in terms of being able to answer open-ended questions on a different synthesised dataset.

This domain of visual question answering on statistical plots, however has a lot of scope of improvement in terms of future enhancements of these models. There is a possibility of expanding the types of charts to those beyond bar and pie charts or even improving on accuracy through model adaptation.

6.3 High Level System Design

Component Diagram:

Input: There are two inputs to our model that the user needs to provide. The first input is an image – that depicts a statistical plot and the second input is a relational question on the image.

Our Model: The design consists of 3 primary components.

1. **Image Encoding Module**

   This is a module that takes in the image as its input and produces image feature vectors as its output. Feature vectors are a vector representation of the image that encodes all of the relevant information in the image.

2. **Question Encoding Module**

   The input to this module is the question (English Language) and the output is a question embedding. The question embedding captures all of the relevant information in the question in a format that is suitable for further modelling.

3. **Result Estimation Module**

   This module takes in as input the output produced by both of the previous modules. It produces as an output the final answer to the question. Thereby, it finds the correlation between the image and the question based on previous training on a variety of image-question pairs.

Output: The output is an answer to the question that was posed on the image.

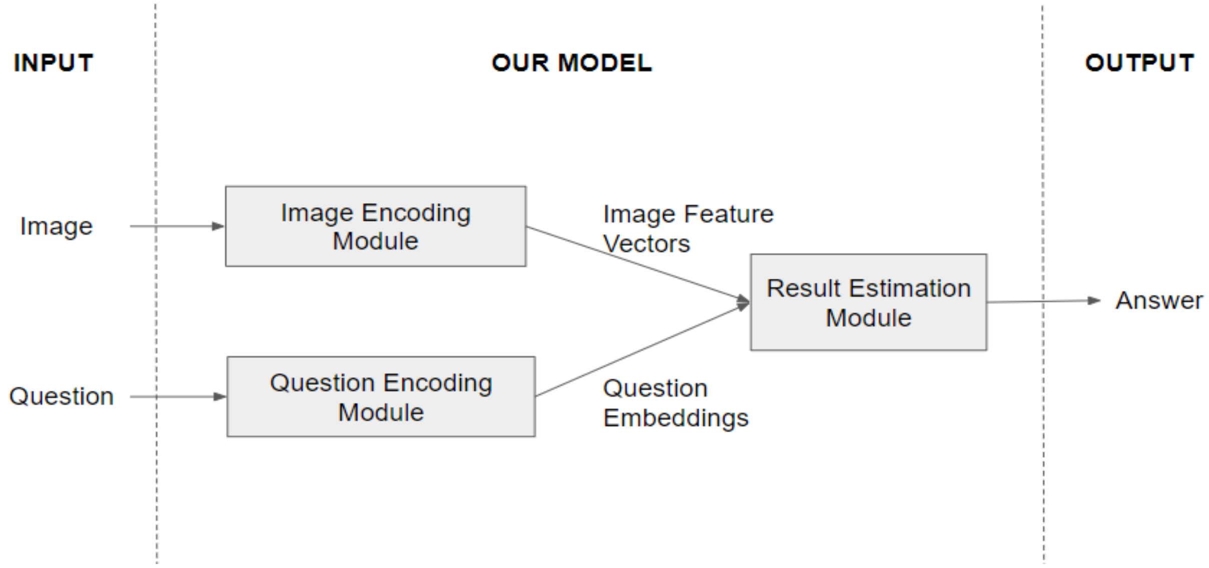


Figure 20

Figure 20 shows the high level design of the proposed model.

Image Encoding Module:

This consists of a sequence of convolution operations and pooling operations, followed by fully connected layers that would output a flattened image feature vector.

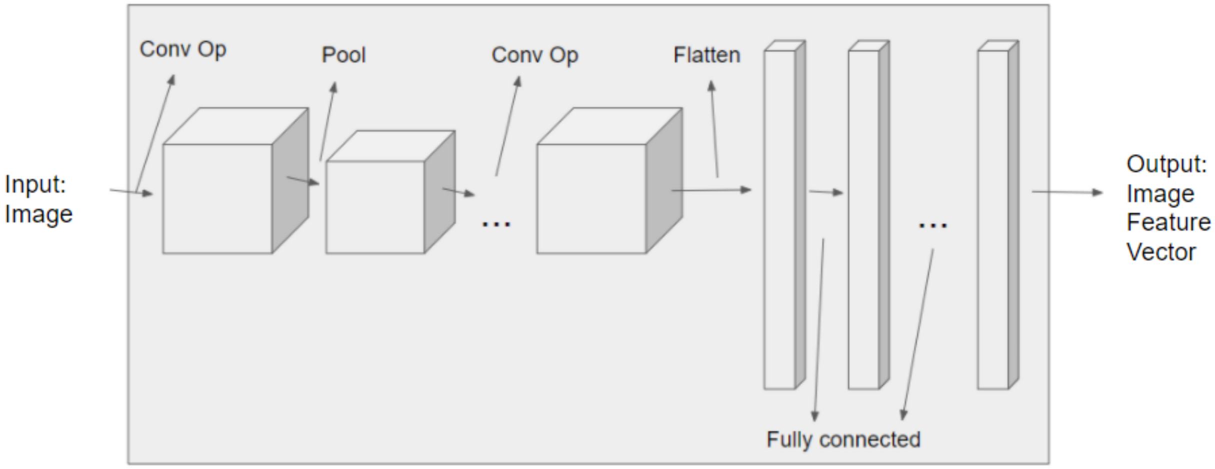


Figure 21

Figure 21 shows the high-level architecture of the image encoding module.

Question Encoding Module:

This consists of a series of recurrent cells, that would take each word of the input question as its input, to finally output a question embedding.

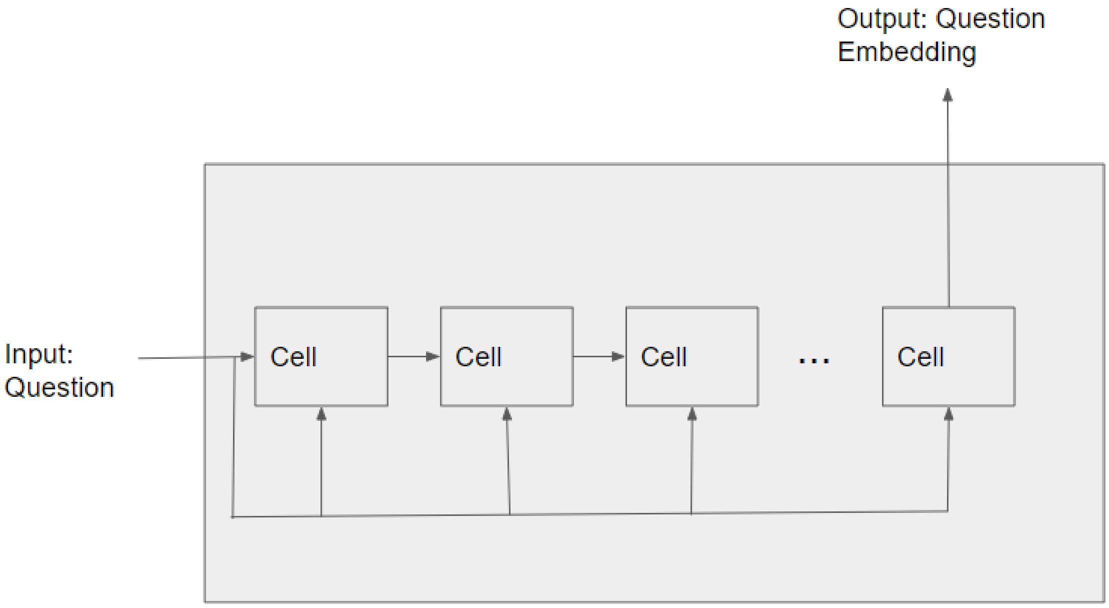


Figure 22

Figure 22 shows the high-level architecture of the question encoding module.

Result Estimation Module:

This module concatenates the image feature vector and the question embedding, and passes that on to fully connected layers. The output is the answer to the question.

CHAPTER 7

CONCLUSION OF CAPSTONE PROJECT PHASE-1

Capstone project phase 1 was conducted during Jan 2021 - May 2021. The problem statement was defined and the domains are identified. A thorough literature survey of the problem statement was done and the novelty and uniqueness of the problem statement was established. The data required for the project was explored and required knowledge was obtained the type and format of data required for the model. Various state of the art models that have already been implemented have been explored. The data and the architecture used by them have been thoroughly analyzed and documented.

CHAPTER 8

PLAN OF WORK FOR CAPSTONE PROJECT PHASE-2

The second phase of the capstone project will be conducted during Summer Term (May-2021 to July-2021) or during the seventh semester (August-2021 to December-2021) depending on the situation. Implementation of the problem statement identified will be done during this phase. Results and outcomes of the project will be explained and thoroughly analyzed.

REFERENCE / BIBLIOGRAPHY

[1] Kim, Dae Hyun, Enamul Hoque, and Maneesh Agrawala. "Answering questions about charts and generating visual explanations." Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020.

[2] Reddy, Revanth, et al. "Figurenet: A deep learning model for question-answering on scientific plots." 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019.

[3] Sharma, Monika, et al. "ChartNet: Visual reasoning over statistical charts using MAC-Networks." 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019.

[4] Methani, Nitesh, et al. "Plotqa: Reasoning over scientific plots." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.

APPENDIX A DEFINITIONS, ACRONYMS AND ABBREVIATIONS

Acronyms	**description**
VQA	Visual Question Answering
ML	Machine Learning
NN	Neural network
CNN	Convolutional Neural network
LSTM	Long Short Term Memory
MAC	Memory Attention Composition
OCR	Optical Character Recognition
NLP	Natural Language Processing