A Step-One Visual Learning App for children leveraging Knowledge-Aware Visual Question Answering Framework

By

Ahmed Jamshed

FALL-2018-MS-CS 00000277947 SEECS

Supervisor

Dr. Muhammad Moazam Fraz

Department of Computing

School of Electrical Engineering and Computer Science (SEECS)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

(February, 2022)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "A Step-One Visual Learning App for children leveraging Knowledge Aware Visual Question Answering framework" written by AHMED JAMSHED, (Registration No 00000277947), of SEECS has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Advisor: Dr. Muhammad Moazam Fraz

Date: 21-Dec-2021

Signature (HOD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Approval

It is certified that the contents and form of the thesis entitled "A Step-One Visual Learning App for children leveraging Knowledge Aware Visual Question Answering framework" submitted by AHMED JAMSHED have been found satisfactory for the requirement of the degree

Advisor : Dr. Muhammad Moazam Fraz

Signature: _____

Date: 21-Dec-2021

Committee Member 1:Dr. Omar Arif

Signature: _____

22-Dec-2021

Committee Member 2:Dr. Muhammad Shahzad

Signature: _____

Date: 21-Dec-2021

Signature: _____

Date: _____

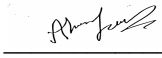
Dedication

I dedicate this effort to my family who always supported me and pushed me in any way possible to become what I am today. Their sacrifices seeded my success especially my MOM DAD who always supported me and pushed me throughout this journey.

Certificate of Originality

I hereby declare that this submission titled "A Step-One Visual Learning App for children leveraging Knowledge Aware Visual Question Answering framework" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name: AHMED JAMSHED

Student Signature: 

Acknowledgments

Glory be to Allah (S.W.A), the Creator, the Sustainer of the Universe. Who only has the power to honour whom He please, and to abase whom He please. Verily no one can do anything without His will. From the day, I came to NUST till the day of my departure, He was the only one Who blessed me and opened ways for me, and showed me the path of success. Their is nothing which can payback for His bounties throughout my research period to complete it successfully.

Firstly, I would like to express my sincere gratitude to my research supervisor, Dr. Muhammad Moazam Fraz, for giving me the opportunity to do research and providing invaluable guidance throughout this research. It was a great privilege and honor to work and study under his guidance. I am extremely grateful for what he has offered me.

Beside, my supervisor, I would like to thank Dr. Faisal Shafait who also has helped me with the methodology to carry out the research and to present the research works as clearly as possible. Without his support and help this work wouldn't have been completed. My sincere thanks also goes to GEC members, Dr.Omar Arif and Dr. Muhammad Shahzad for their support and encouragement.

At last but not least, I am extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future.

Ahmed Jamshed

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

VQA	Visual Question Answering
CNN	Convolutional Neural Network
KRR	Knowledge Reasoning and Representation
NLP	Natural Language Processing
CV	Computer Vision, Electroencephalographic
MCQ	Multiple Choice Questions
DL	Deep Learning
ML	Machine Learning
RNN	Recurrent Neural Network
LSTM	Long short-term memory
GRU	Gated Recurrent Unit
MCQ	Multiple Choice Questions
SQuAD	Stanford Question Answering Dataset
IC	Image Captioning
OOSC	Out Of School Children

Abstract

Visual Question Answering (VQA) is a complex cognitive multimodal inference problem in which the latest techniques from the fields of NLP (Natural Language Processing) and CV (Computer Visions) are merged with the goal of answering open-ended and free-formed natural language questions by understanding the visual analogies. Previously, the performance of VQA models used to suffer when asked knowledge-based and commonsense aware questions but it all changed with the introduction of transformers, as transformer-based language models now possess some degree of knowledge and commonsense implicitly. Additionally, we can also provide external knowledge explicitly using Knowledge Reasoning and Representation (KRR) techniques to further enhance the performance benchmarks. In order to train and benchmark these knowledge-aware VQA models, several datasets like OK-VQA, GQA, KVQA etc. are introduced in which questions require some sort of cognitive inference from available external knowledge. These datasets are quite capable as they are carefully crafted for the intended purpose but they are static as they have limited questions and images only.

This paper presents the framework which is capable of producing multiple relevant knowledge-aware MCQs associated with each unique image, using the knowledge-rich corpus from Wikipedia. These MCQs can be used for preparing dynamic knowledge-aware VQA datasets. We can also use this framework by developing a visual learning app to educate children in an interactive manner, especially in the remote areas of developing countries where they seldom get a chance to learn new concepts in a proper school environment.

Keywords: *Visual Question Answering, Image Semantic Understanding, Natural Language Processing, Transformers, Automated Datasets*

Chapter 1

Introduction and Motivation

1.1 Introduction

Artificial Intelligence (AI) is transforming the world around us very quickly as it is helping humanity in many different ways whether be it the smart automation of hectic routines or be it the prediction of stocks and whatnot. It is all done by using Machine learning (ML) algorithms and Deep learning (DL). Generally, ML is aimed to carry out more specific and direct tasks while DL is aimed to tackle more general and complex tasks as it utilizes Artificial Neural Network (ANN) which mimics the functionality of the human brain. The task of knowledge-aware VQA requires true cognitive capabilities as it combines the challenges of Computer Vision (CV), Natural Language Processing (NLP) and Knowledge Representation and Reasoning (KRR) [11].

Transformers are a perfect match for such types of challenges as they can easily operate on any type of sequential data using similar processing blocks, allowing us to mix inputs from multiple modalities, whether it be an image, video, text or any other data in sequential format. The sense of textual data always comes from the sequence of words in a sentence and similarly, the sense of an image depends upon the relationship of objects present in the image. So, the language data is inherently a neat sequence of related words giving a context to the sentence but a visual image can be made sequential by dividing it into a series of patches from left to right and top to bottom [12] which is then transformed into vectors and treated as normal words in the same old language transformer.

A Transformer uses attention mechanism to enhance the most important and relevant

parts of the input data while fading away the rest. It gives them the ability to encode the most important relationships among different parts of the multi-modal data. These Transformers have replaced the slower Recurrent Neural Networks (i.e. LSTM and GRU) using the technique of positional encoding which has enabled them to process the sequential data in parallel thus enhancing the training speeds by many folds. Transformers are very data hungry as they require a huge amount of data for getting trained on. Infact, the corpus of GPT3 (a transformer based language model) contained almost the whole Internet. Although language models like GPT3 understands the languages very well but they are blind for the tasks like VQA. Language models can provide rich textual embeddings to multimodal transformers but the task of understanding visual analogies and extracting the relevant knowledge from knowledge bases makes it hard to answer knowledge aware visual questions which is why the performance benchmarks on datasets like OK-VQA [5] are very low. OK-VQA [5] is a static dataset and it has only limited set of categories and questions .

The idea is to prepare a framework that can generate multiple relevant knowledge-aware MCQs dynamically from a single image using the information from the relevant text corpus or Wikipedia. It can help us to train and benchmark knowledge-aware VQA models in multiple domains but the potential of such a framework could be far-reaching since we have an opportunity to utilize the beneficial aspects of knowledge-sharing. This framework can also be used to educate children in the remote areas of developing countries where they don't have the facility of schools and teachers and even elders don't have much education.

This is where our proposed Step-One app comes in as a medium to educate and empower children in those remote areas where they don't have anyone to guide them about their surroundings and environment. Often young children in such far-flung areas can't afford to attend school and they don't even have access to some kind of informal education. Children found in such social settings don't have a serious attitude towards learning and education as it appears burdensome. So the problem is compounded as there is also lack of will and aptitude to learn. Rather they want to spend their time in playful activities, hence we developed our app in such a fun-filled manner which would motivate their young mind to actually engage in learning activities without considering them academically boring. All this was previously unthinkable since static data sets were compiled with the aid of crowd workers which obviously had its limitation with regard

to its ability to benchmark VQA models. With latest transformer models we can easily generate dynamic data-sets which gives us knowledge-aware questions that require the learner to consult external knowledge sources in order to answer them.

1.2 Motivation

Pakistan demographically consists of majority younger population and it is very unfortunate to see so many children out of school. This out of school children number is not settling down in Pakistan by any means[1] [13]. There could be many reasons accounting for out of school children, sometimes the students do not want to continue with their studies, other times they may have an elder or parent who did not allow. Most of the times financial condition of the family could not support education rather they want all hands to earn and bring in money. Other reasons may include the fact that school is too far away and it is impossible to make that long a journey. Obviously many parents don't send their children to schools because its too expensive and its unthinkable to pay those hefty tuition fees while at the same time also make both ends meet. The OOSC numbers for each province can be observed in fig 1.1

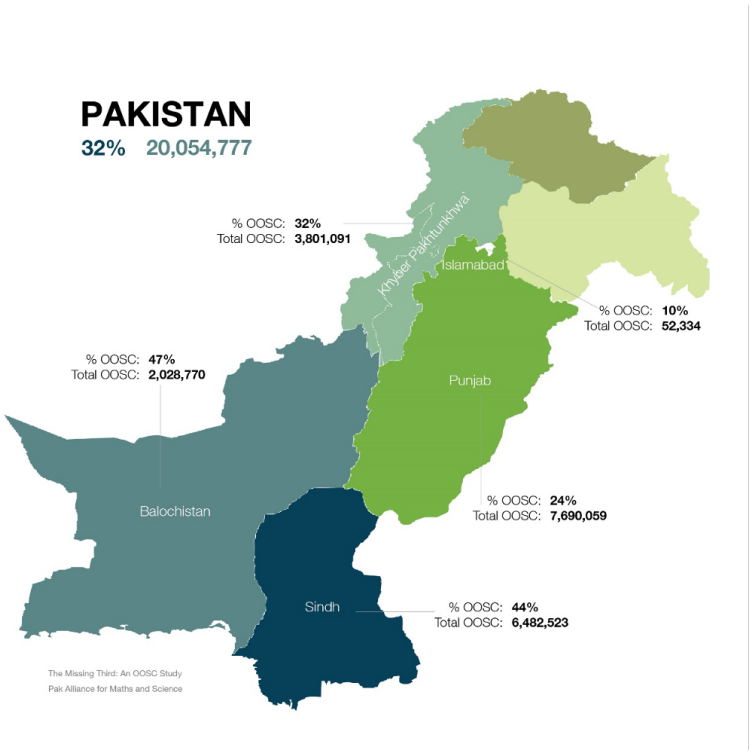


Figure 1.1: Out Of School Children (OOSC) in Pakistan. [1].

Besides out of school children, we often see that classrooms, academics and homework is something that evokes feelings of dread and students just drag themselves, especially through subjects they are not particularly fond of. There is just no motivation whatsoever and consequently learning suffers. Even if a child has access to learning

opportunities they may not have cooperative adults who would be patient enough to quell their curiosity. After all, a curious mind needs information about anything and everything that comes under its observation and if no one is willing to invest their time and energy to answer the queries the child would develop inaccurate concepts or worse, abandon their quest for knowledge.

If this youth is not provided with the means to learn and grow themselves intellectually we could be facing a lot of population that is not a productive member of the society which in turn would lead to further disparity between people and social unrest. There is a need to curb this increasing illiteracy and absence of any medium to bridge this gap between children of remote areas to education.

No.	Benefits of quizzes
1.	The testing effect: retrieval aids later retention
2.	Quizzing identifies gaps in knowledge
3.	Quizzing causes students to learn more from the next learning episode
4.	Quizzing produces better organization of knowledge
5.	Quizzing improves the transfer of knowledge to new contexts
6.	Quizzing can facilitate retrieval of information that was not tested
7.	Quizzing improves meta-cognitive monitoring
8.	Quizzing prevents interference from prior material when learning new material
9.	Quizzing provides feedback to student itself
10.	Frequent quizzing encourages students to study

Table 1.1: Ten benefits of quizzes [10]

What if we could revolutionise the way knowledge is communicated, convey it in a manner that is not only gripping but also accurate. This is where we come in, utilising rich media to make an appealing presentation that is intellectually fulfilling and exciting enough to keep the learner motivated for exploring new horizons. Of course accessibility is not an issue since most of the people have smartphones and internet and that is all which is needed to bring the treasure trove of knowledge in the palm of your hand. An

individual will be presented with an array of pictures or they could even upload pictures of anything that pique their interest. Our app will analyse the images and provide all the basic information about it by utilising computer vision to provide best content online. Not only that but there will be a quiz taken at the end to further solidify whatever has been learned.

There are many reasons to opt for quizzes [14], primarily because they are fun and improves the quality of learning [15]. The mind views it as a puzzle to be solved and stimulates our thought process to recall and apply whatever has been learned. Besides, if a person successfully solves a quiz also leads them to believe in their ability to have a reasonable command over the subject.They have confidence in the fact that they have retained the knowledge gained, and could recall it any time they want. Another reason quizzes have been incorporated is that it helps in breaking down the subject into little learning exercises which provides a logical progression through the subject. The learner is fully aware of their ability and knows what is the next step as they progress through a topic. Not only is quiz a good way to gauge personal performance but it is also a great tool for revision. More quizzing benefits is shown in the table 1.1

Moreover, the deep-learning based knowledge-aware VQA models are somewhat comparable to a child's brain as they both require a lot of intensive training, from various perspectives, in order to learn about new concepts in the world [16]. Even the most advanced VQA models find OK-VQA [5] and kVQA [6] datasets quite challenging and difficult to tackle. The questions picked from such knowledge-aware datasets cannot be easily answered by just looking at the contents of the image. Rather they require some sort of meaningful specific relevant external information to answer such type of open-ended and free-formed questions. The proposed VQA generator framework can also be used to generate the Knowledge-aware VQA datasets, like KVQA and OK-VQA, for any domain in a dynamic or adversarial manner, provided the query-able textual corpus containing all the relevant information about that specific domain.

Previously such type of data sets were collected by employing crowdsourcing strategies, for example using Amazon Mechanical Turk to manually compile the data which consequently was static in nature. With the advent of Transformer models we can generate dynamic data sets which have none of the deficiencies of a static data set, and is capable of generating knowledge-aware questions in run-time.

Chapter 2

Related Work

2.1 Related Datasets

In this section, we will discuss some of the related work and datasets in the field of Visual Question Answering VQA. Most of these datasets are static and do not require any kind of external knowledge to answer the asked questions. Although the questions in these datasets require common sense reasoning and spatial understanding but they can simply be answered by merely looking at the visual data. Previously, such a large amount of knowledge-aware visual questioning answer data is collected through crowd-sourcing strategies mainly by employing the crowd workers from Amazon Mechanical Turk.

2.1.1 Image Captioning by Asking Questions

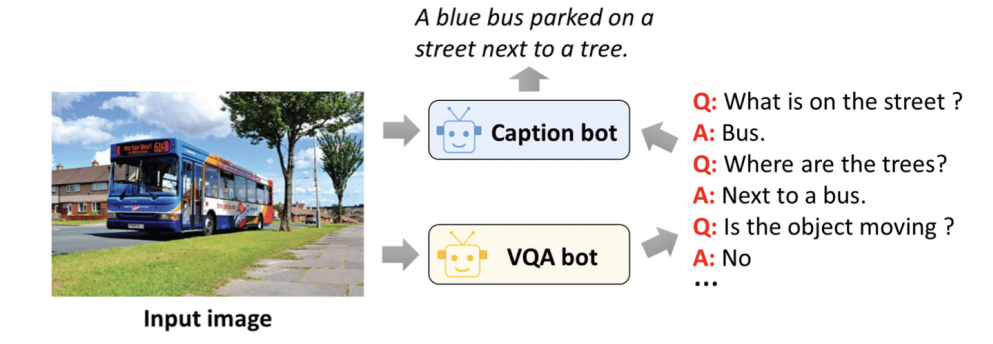


Figure 2.1: Image Captioning by Asking Questions [2].

The task of Image Captioning (IC) is quite similar to the task of VQA in terms of

methodology. Both of them need to understand the image first, using CV techniques and then they need to process the sentence using NLP. However, VQA needs to understand the visual analogies more deeply than IC as there can be multiple perspectives involved when interpreting the same image due to multiple truths. This deeper understanding from VQA task can be exploited in Image Captioning to extract more accurate, detailed and meaningful captions from the images.

This paper [2] attempts to improve the performance of image captioning model by fusing VQA-grounded feature and the attended visual feature extracted from the image (see Fig. 2.1). It shows that simple image-captioning models improved through the grounded features of Visual Question Answering produce more detailed and better captions than conventional image-captioning methods on large-scale public datasets.

2.1.2 Generating Natural Questions from Images for Multimodal Assistants

Generating obvious and common-sense questions from visual images is easier but they are not suitable for multimodal assistants as humans would typically ask them the more natural and diverse, knowledge-aware questions.

Figure 2.2: List of potential questions a user might ask the digital assistant [3].

It is very crucial for multimodal assistants to understand and generate such type of questions by properly understanding the visual content of the image so that they can easily handle more diverse and natural type of questions from humans in real-world scenario. It studies the image to such an extent that the questions are not just surface

level, rather they are intricate and complex just like we humans would ask. The line of inquiry will almost be in line with curious queries which are formed in a human mind.

This paper [3] attempts to generate more meaningful and diverse natural language questions from the images using the example questions provided by the annotators. (see Fig. 2.2).

2.1.3 GQA

GQA [4] dataset consists of 22M automatically generated compositional questions based on real-world images from COCO [17] and Flickr dataset using scene graphs from Visual Genome [18] dataset. It is good for benchmarking spatial understanding and commonsense reasoning carried out by a VQA model.

Figure 2.3: Image from GQA dataset [4].

The questions asked in this paper [4] are related to commonsense and spatial understanding which are the good parameters to benchmark the VQA models but very little outside knowledge is involved in this dataset and most of the questions can easily be answered purely based on the content of the image as shown in Table 2.1).

No.	Questions	Answers
Q1.	What is the woman to the right of the boat holding?	Umbrella
Q2.	Is the jacket blue?	No
Q3.	What color is the object the woman is holding?	Purple
Q4.	On which side of the photo is the woman, the right or the left?	Right
Q5.	Are there any men to the left of the person that is holding the umbrella?	No
Q6.	Are the shoes red?	Yes

Table 2.1: GQA[4] Questions generated for Fig.2.3

2.1.4 OK-VQA

OK-VQA [5] proved to be a very challenging dataset because it required outside knowledge to answer its questions. Integrating knowledge into VQA models is very different from asking simple image-related questions as the VQA model is first required to extract meaningful related information from the knowledge bases [19] [20] [21] and only then it can answer the questions rightly. These knowledge bases are mainly consisted of triplets like <*Car, CapableOf, Crash*> OR <*Wood, UsedFor, Boats*> which can easily be turned into knowledge graphs.

Figure 2.4: Examples from OK-VQA dataset [5].

A VQA model like KRISP [22] can extract related information from these knowledge

graphs in order to predict answers of knowledge-aware questions which are only partially dependent on the visual information. The questions in this dataset are divided into 10 knowledge domains as shown in Fig 2.4).

2.1.5 KVQA

Generally, in VQA, the questions asked are about common nouns like (dogs, cars, guitars, etc.) but K-VQA [6] shifted the focus of VQA questions to named entities like (Barack Obama, Amir Khan, Serena Williams etc.) in which some background knowledge about these entities is required. This dataset contained 183K question-answer pairs about 18K celebrities present in 24K images where questions require multi-entity, multi-relation and multi-hop reasoning over Knowledge Graph to predict the answers.

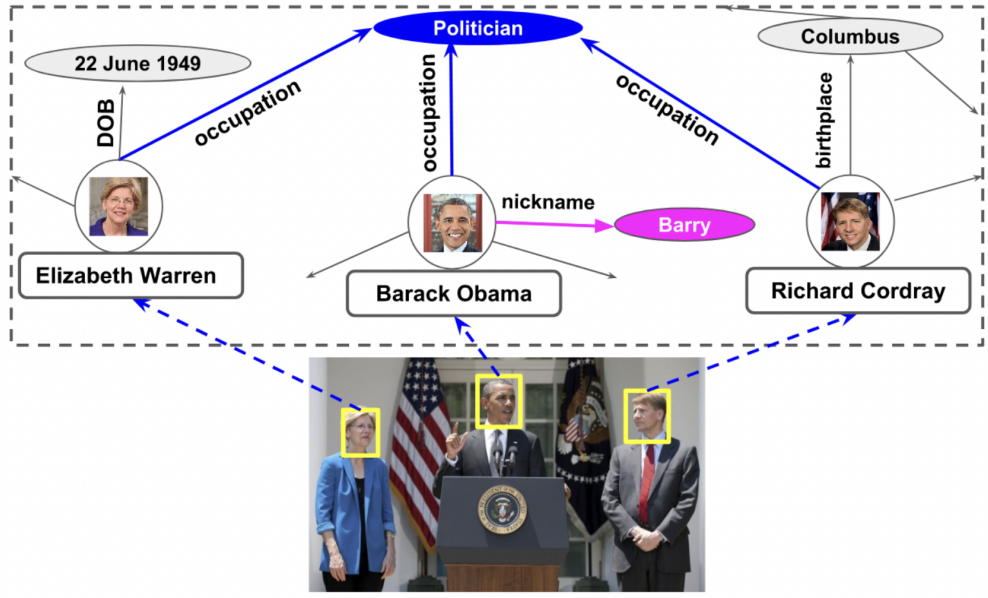


Figure 2.5: Example image from KVQA dataset [6]

No.	Questions	Answers
Q1.	Who is to the left of Barack Obama?	Richard Cordray
Q2.	Do all the people in the image have a common occupation?	Yes
Q3.	Who among the people in the image is called by the nickname Barry?	Person in the center

Table 2.2: Questions generated for Fig.2.5

In KVQA, an exhaustive list of celebrities including athletes, politicians and actors is first compiled manually and then their images along with the captions are extracted from their Wikipedia pages. After acquiring images and corresponding captions from relevant Wikipedia pages they utilized human annotators to refine the images by counting, identifying and assigning the order of appearance (spatial understanding) of the people in the images while also removing all the useless and duplicate images to further refine this dataset. Once the named entities and the order of their appearance are identified, the annotators extracted the relevant ground-truth answers from Wikidata using templated-questions. The questions are also paraphrased to make this dataset challenging and closer to the real-world.

KVQA analysis reflected that it is challenging to use for both vision community (recognising faces and poses) as well as language understanding community (answering questions requiring multi-hop and quantitative reasoning over multiple entities in Knowledge Graph). The questions in this dataset about real world entities are shown in Table 2.2).

2.2 Related Apps

The proposed Step-One app enables the children to learn about their favourite subjects using the latest techniques from the domain of Computer Vision. In the past, some attempts have been made to solve similar types of problems using different approaches making it necessary to briefly discuss them too.

2.2.1 CognitiveCam

CognitiveCam [23] is a VQA based application designed to help the visually impaired. It uses the IBM Watson Bluemix APIs to recognize everyday objects and age/gender of the person. It takes the question as voice input and then classifies it from three classes i.e. Object Identification, Text Reading and age-gender description. Once the question is classified, it analyzes the images and produce the answer as voice output.

2.2.2 Google Lens

Google Lens [24] is a powerful multipurpose AI-based app that uses reverse image searching to dig out the most relevant stuff from the internet. It is capable of scanning and translating the text shown to it. It can identify the items of daily use like outfits, furniture, home-decor, vehicles etc to get the available relevant information about them and even helps you to buy them from online stores. It can also identify the places around like restaurants, landmarks, storefronts etc to provide the information like ratings, hours of operation, historical facts and much more. This app operates on the true potential of artificial intelligence.

2.2.3 Other available trivia apps

Many existing trivia apps on mobiles like Trivia Star Quiz App [7], General Knowledge Quiz App[9], Knowledge Trainer [8] randomly select questions from the pool of predetermined hard-coded questions which usually are very simple and general. These apps do not help children in learning rather these are only made to test the general IQ of a person. A child does not have any learning opportunity using these apps.

We established this position due to the simple fact that such apps do not encourage

reading or gaining any sort of prior knowledge or information. They just make use of a child's prior experiences and ask basic common sense questions which do not involve critical thinking. It may be good for stimulating the brain but ultimately it is still using past information and experiences of the learner. It is not furnishing the mind with new information, it is not promoting learning or reading. The mind is not being actively engaged with any kind of new information which would later be tested with trick questions.

Step-One puts the user in the driving seat and let them decide for themselves what topic they want to pursue. This serves two important things, first it lets the interest of the user dictate their learning path and this liberty is absent in trivia apps. Secondly, this autonomy over topic selection will ensure that anyone pursuing their favorite subjects will always look forward to learning about it more, and motivated to actually acquire new knowledge in a fun-filled manner.

Features	Step-One	Trivia apps
Initiate Image based learning	✓	X
Dynamic Topics	✓	X
Control over topic selection	✓	X
Availability of learning material	✓	X
New knowledge and concepts	✓	X
Detailed learning	✓	X
Dynamically generated quizzes	✓	X
Cheat Context	✓	X
Fun Factor	✓	✓
AI-based automatically generated questions	✓	X
Learning assistance to children	✓	X
Instant feedback	✓	✓
Useless frills	X	✓

Table 2.3: Step-One comparison to other trivia apps out there

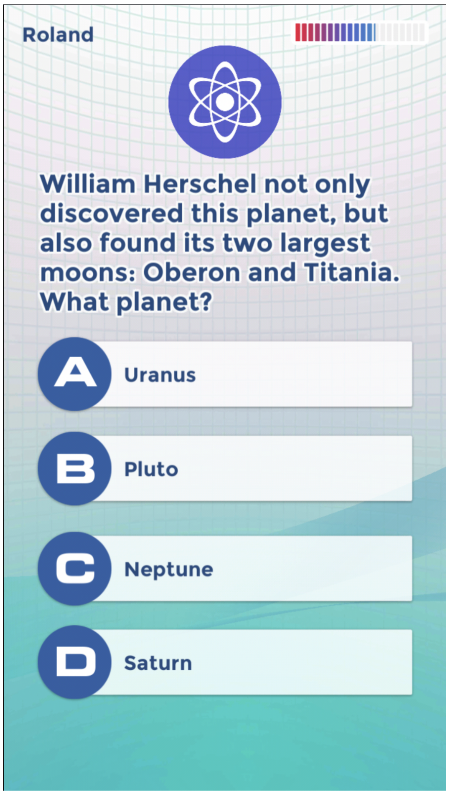
Step-One, keeping true to its aim of furnishing young minds with new knowledge and concepts, also has a comprehensive 'cheat context' option for questions that may prove too difficult. Basically it gives the user essential hints at what could be the right answer

by providing a complete knowledge context to the question. This reinforces learning and reading in the user and once they are done with the question they are completely aware of the reason behind the right answer. As compared to trivia apps which at best tell user about the answer either being right or wrong without providing any reasoning or rationale behind it, providing just a surface level knowledge without any background or context.

In short Step-One retains the interest of its users by making them decide for themselves with primary focus on learning and educating while avoiding useless frills like coins based reward system, lifelines or leaderboard of trivia apps which has little to do with actual learning. A brief comparison between Step-one and other available trivia apps is exhibited in table 2.3

(a) Quiz Screen **(b)** Static Questions

Figure 2.6: TRIVIA STAR Quiz Games developed by Super Lucky Games LLC [7]



(a) Quiz Screen

(b) Score Screen

(c) Static Categories Screen

Figure 2.7: Knowledge Trainer: Trivia developed by The Binary Family [8]

(a) Quiz Screen **(b)** Score Screen

Figure 2.8: General Knowledge Quiz developed by TIMLEG [9]

Chapter 3

Architecture

The whole architecture can be divided into four modules (see Fig. 3.1). If these modules are used jointly, we get the relevant knowledge-aware dataset according to the uploaded image but we can also use them separately in order to obtain the result after each step which can then be used in the Step-one learning app.

To summarize this interaction between these four modules we first start with the images. As soon as the image is provided, it would extract all the relevant labels from that image and then on the basis of those labels the app would gather relevant knowledge corpus from the internet/Wikipedia and then it converts the raw HTML to understandable blocks of information, including all the content below headings and their sub-headings. Answer-aware questions are then extracted using Google's T5 Transformer model [25] which is currently the state-of-the-art for such NLP tasks. Once we get the questions and answers and their relevant context, we can then generate the relevant options using sense2vec [26] along with other techniques to handle the cases of failure.

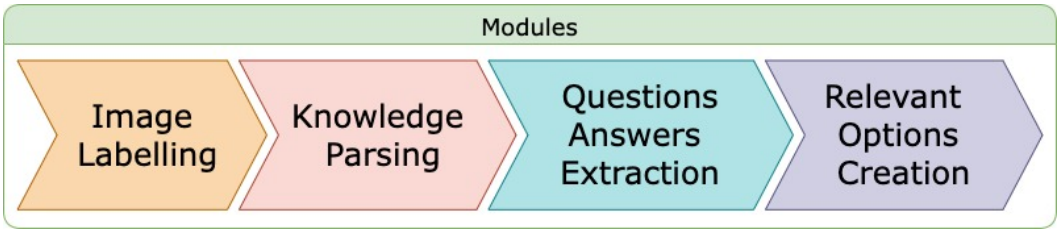


Figure 3.1: Modules.

3.1 Image Labelling Module

This module is responsible for extracting multiple labels from the given image. We used the Image Labelling API from Google Cloud Vision to extract the labels from the image as it is backed by Google's enormous data and it can identify thousands of real-word objects with a great accuracy. The biggest advantage of using the Image Labelling API instead of Object Detection API from Google Cloud Vision is that it not only returns the descriptions of the detected objects rather it also returns descriptions of some of the related concepts too. Normally, those related concepts have their own Wikipedia's page.

For example, when the image of some animal is selected, it not only returns the name of that animal but it also returns some more related attributes and concepts about that animal like its classification name i.e mammal or the names of its body parts i.e trunk, ivory etc.

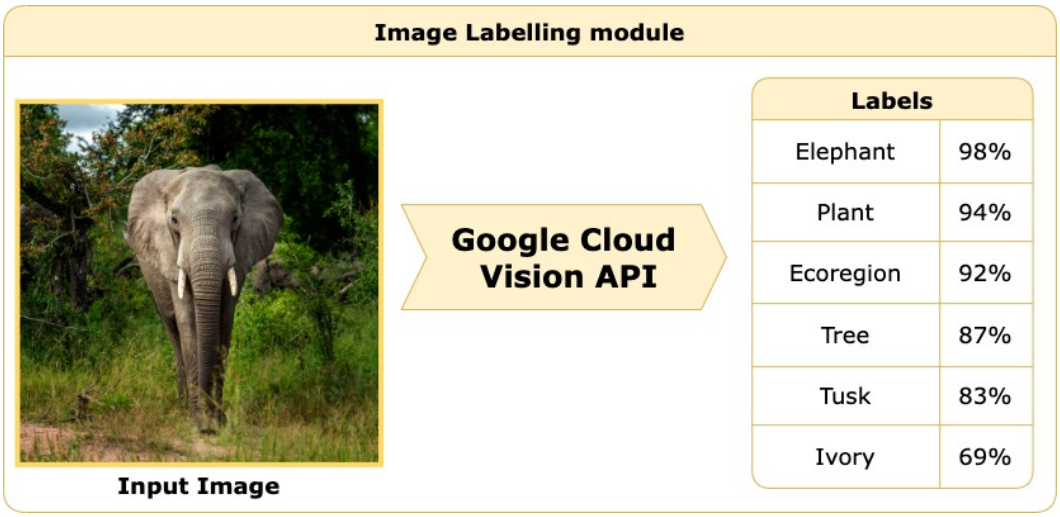


Figure 3.2: Image Labelling Module.

Image Labelling API from Google Cloud Vision returns the machine-generated identifier (mid) along with description and score prediction against each label assigned to the image. Information against each mid can be retrieved using Google Knowledge Graph Search API. The most important piece of the information, returned through Google Knowledge Graph Search API, is the link of related Wikipedia's page from where we can extract the further information about each label (see Fig. 3.2).

3.2 Knowledge Extraction Module

This module is responsible for extracting relevant knowledge against each label provided by the Google's image labelling API. We get the relevant Wikipedia's page URL against each label from the previous module. For the Step-one app, we also require an image and a description against each label so that the presented topics can be made easily digestible and interesting for a child. So, in the first request to Wikipedia's API, we extract the main image and relevant description against each page and then we extract the rest of the page's HTML in the second request. Dividing these requests into two, makes the execution of requests faster for the Step-one learning app.

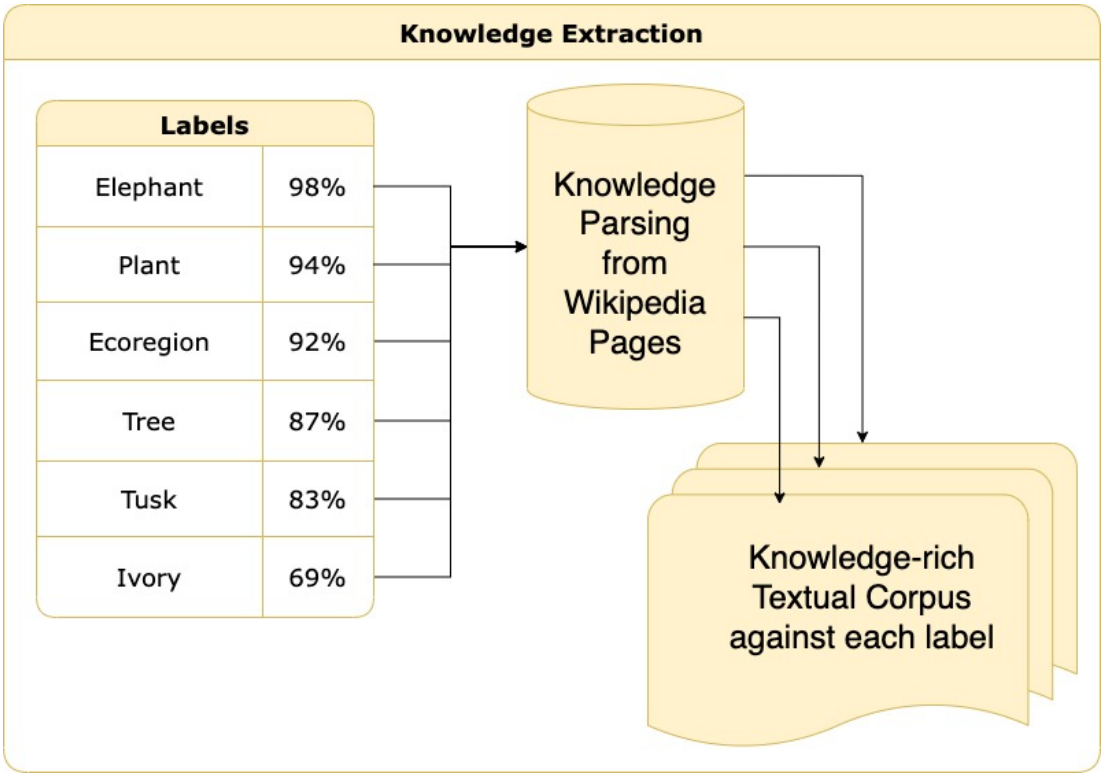


Figure 3.3: Knowledge Extraction Module.

Before feeding this output to next module, we need to convert the page's raw HTML into some meaningful text. Therefore, this HTML is parsed to extract headings, subheadings, paragraphs and images (see Fig. 3.3).

3.3 Question Answer Extraction Module

This module is responsible for generating multiple Questions and answers from the given text. For this purpose, we used Google's T5 [25] model which is fine-tuned on SQuAD v1.1 [27] dataset for Answer-aware Questions Generation [28] by appending the answer string at the start of the context string. For Example: *Elephants <SEP> Elephants are the largest existing land animals.*

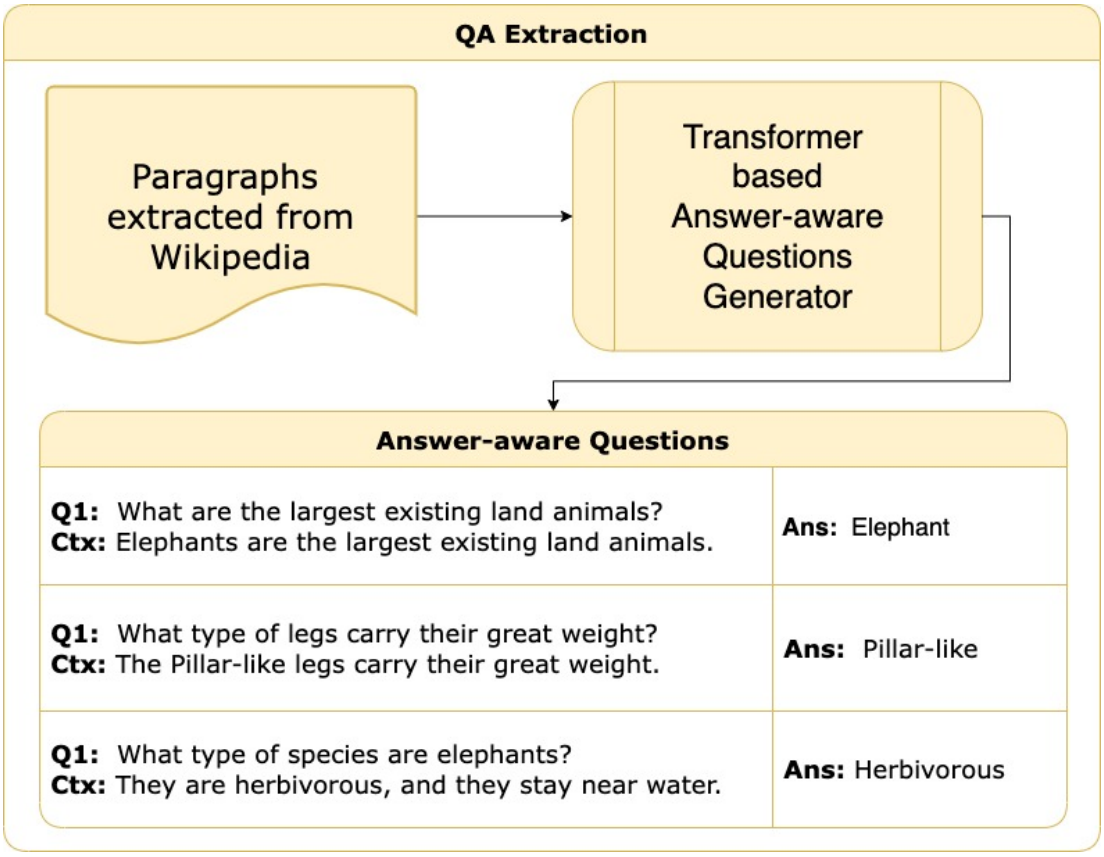


Figure 3.4: Answer aware Questions Extraction Module.

Each paragraph from the previous module is used to extract multiple answer-aware questions [1] along with its context which will be used to extract relevant options in the next module (see Fig. 3.4). Extracted Answer can be a single word or it can also be the combination of words.

[1] I would like to express my utmost gratitude towards Patil Suraj for his exceptional work in the field of answer-aware question generation which provided me with an impetus to take it further to develop this framework.

3.4 Relevant Options Creation Module

This module is responsible for creating multiple relevant options against each generated Question-answer pair which is obtained from the previous module along with the context. We used the Spacy pipeline to generate meaningful tokens from the context. If there is some named entity present in the context, it will easily recognize that. e.g. If Apple is present in the context of company the it will extract the Apple as named entity and if the apple's context is fruit then it won't extract it as named entity. It also assigns the Part-of-Speech tag to each token (see Fig. 3.5).

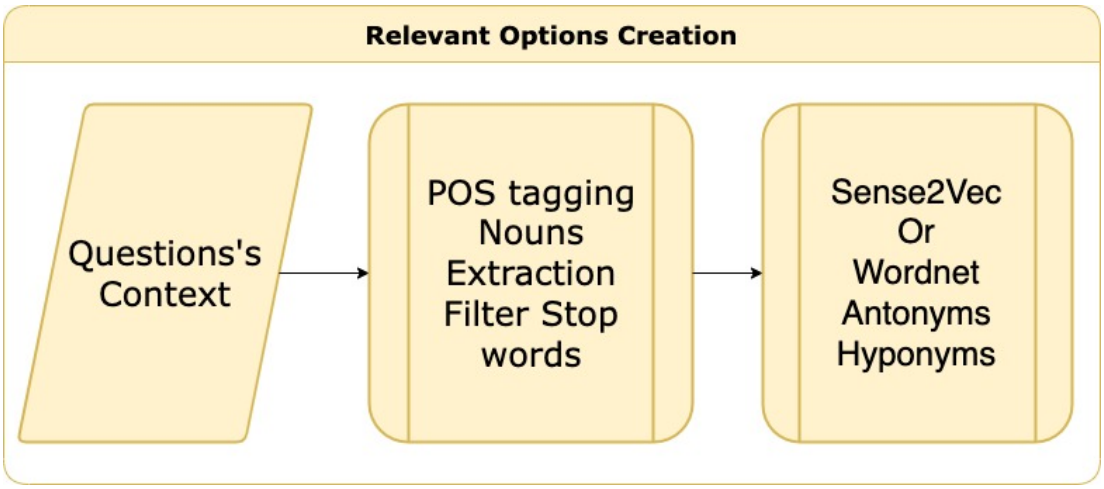


Figure 3.5: Relevant Options Creation Module.

If some named entity is found in the context sentence and that part is included in the answer then it is directly passed to Sense2Vec [26]. If there is no named entity present then stopwords are first removed from the context and then priority is given to numbers and adjectives as these are good candidates for generating options. If some number is found in the context and the answer then first it's shape is extracted which can be 13,600 represented as dd,ddd or 10-20 represented as dd-dd and then random number from 1-9 is substituted in the place of 'd' which generates similary shaped random numbers for options.

If there is no number or adjective is present in the context or answer then random word is picked from the context which should also be the part in the answer and then it is passed to sense2vec. If sense2vec fails to find any contextually similar words then some other token is passed. If sense2vec still fails to find any suitable option then the antonyms or co-hyponyms are extracted from Wordnet. But there are still some cases,

where all the above strategies fails like in case of scientific name of species etc. and we don't have any meaningful related options for that word. In that case of failure we just produce misspelled words by changing some random character in the answer.

Chapter 4

Analysis

The proposed knowledge-aware question generator framework is capable of generating a huge amount of high-quality knowledge-aware Multiple Choice Questions which can be used to train and benchmark the knowledge-aware VQA models and can also be used as a mode of informal education in the form of Step-one learning app developed for less fortunate children in remote areas. Let's examine the output of this framework in a bit more detail.

4.1 Quantitative Analysis

This framework can generate a huge amount of questions corresponding to each detected label in the image. The relevant knowledge is gathered from the Wikipedia's page using all the detected labels and then almost all of the extracted knowledge is converted into answer-aware knowledge-rich multiple choice questions, making the quantity of generated questions per image quite large.

We used Google Vision's Image labelling API which can assign the labels to images by quickly classifying them into millions of predefined categories where each category has it's corresponding Wikipedia article and as of 17 January 2022 [29], there are 6,439,978 articles in the English Wikipedia containing over 4 billion words (giving an average of about 626 words per article), and 55,008,151 pages and these articles continue to grow at the rate of 17,000 new articles per month. These articles cover the areas of business, science, biology & health/medicines, geography, history, biography, society, culture & arts etc. So theoretically, this framework has the potential to convert most

of the available information into knowldege-aware MCQs or we can also generate such knowledge-aware MCQs on demand according to our requirement or according to the visual data present in the dataset of images. It can also be used in an adversarial manner, where advanced VQA models will try to answer open-ended and knowledge-aware questions posed by this framework.

4.2 Qualitative Analysis

As established in Quantitative Analysis this framework is not limited by the quantity of the questions produced but of course quality of the generated questions/answers should be a major concern here.

Table 4.1 and 4.2 reflect the true potential of such knowledge-aware question generator framework. This framework is also capable of producing ground-truth answers and also relevant options along with the knowledge-rich questions. Moreover, the context is also provided along with each question to further clarify the quality of the produced questions using Google's small T5 Transformer [25] model with 60 million parameters, trained on SQuAD dataset[27]. We used such a small model to increase the speed of question generation when the quiz is requested from the Step-one learning. But we can also use larger T5 models to further improve the quality of generated questions where timing is not a constraint like while producing questions for knowledge-aware VQA dataset.

Considering this image of lion, following sample questions were generated from sample topics.

Sample Topics	Sample Questions	Answer with a blank	Sample Options	Context
	What is the fifth largest order of mammals?	_____	Acanthopterygii, Acarina, Actiniaria, ***Carnivora***, Actinomyxidia etc.	The order Carnivora is the fifth largest order of mammals, comprising at least 279 species.

Carnivora	Where is the hyper-arid region of Carnivora?	the Sahara _____	forest, woodland, timberland, grassland, battlefield, ***desert***, battleground etc.	Carnivorans live on every major landmass and in a variety of habitats, ranging from the cold polar regions to the hyper-arid region of the Sahara Desert to the open seas.
	What are the feliforms?	cats, the_____, the mongooses and the viverrids	hippos, elephants, ***hyenas,*** weasels, otters, gazelles, bears etc.	The feliforms include families such as the cats, the hyenas, the mongooses and the viverrids.
	What is the name of the lion that is native to Africa and India?	_____	Abrocoma, Acinonyx, ***Panthera,*** Acrobates, Adenota, Aepyceros, Ailuropoda etc.	The lion (Panthera leo) is a large cat of the genus Panthera native to Africa and India.
	What is the end of the lion's tail?	_____	knot, swat, ***tuft,*** comb, tussock etc.	It has a muscular, deep-chested body, short, rounded head, round ears, and a hairy tuft at the end of its tail.

Lion	What do female lions usually prey on?	_____	domesticated cats, great apes, wild cats, carnivores, mammals, ***Ungulates*** etc.	Groups of female lions usually hunt together, preying mostly on large ungulates.
	Where were cultural depictions of lions prominent?	_____	ancient world, ***Ancient Egypt,*** middle ages, Roman times, Mesopotamia, Medieval Europe, ancient Rome etc.	Cultural depictions of lions were prominent in Ancient Egypt, and depictions have occurred in virtually all ancient and medieval cultures in the lion's historic and current range.
	What are the biggest causes of the decline of the lion?	habitat loss and conflicts with _____	other species, living things, ***humans,*** intelligent beings, sentient species etc.	Although the cause of the decline is not fully understood, habitat loss and conflicts with humans are the greatest causes for concern.
	What is the protruding portion of an animal's face?	A _____	beak, tail, paws, whiskers, ***snout,*** lower jaw etc.	A snout is the protruding portion of an animal's face, consisting of its nose, mouth, and jaw.

Snout	What is the wet furless surface around the nostrils of the nose of many mammals called?	_____	rhicarium, rhanarium, rhinafium, ***rhinarium,*** rhiqarium etc.	The wet furless surface around the nostrils of the nose of many mammals is called the rhinarium (colloquially this is the "cold wet snout" of some mammals).
	What is the rhinarium associated with?	_____ sense of olfaction	same strength, stronger, tougher, ***stronger,*** weaker, more endurance etc.	The rhinarium is often associated with a stronger sense of olfaction.

Table 4.1: Sample MCQs from some selected topics generated using the the image of the lion

Considering this image of canal, following sample questions were generated from sample topics.

Sample Topics	Sample Questions	Answer with a blank	Sample Options	Context
	When were the earliest boats found?	_____ years ago	2,000–10,000, 8,000–30,000, ***7,000–10,000,*** 3,000–80,000, 8,000–60,000	The earliest boats are thought to have been dugouts, and the oldest boats found by archaeological excavation date from around 7,000–10,000 years ago.
	What is the oldest recovered boat in the world?	Pesse _____	kayak, pontoon, boat, small boat, jet ski, ***canoe,*** dinghy, raft etc.	The oldest recovered boat in the world, the Pesse canoe, found in the Netherlands, is a dugout made from the hollowed tree trunk of a Pinus sylvestris that was constructed somewhere between 8200 and 7600 BC.
	Where is the Pesse canoe exhibited?	Drents _____	Archive, Bank, ***Museum,*** Drop, Library, Storyhouse, Repertory etc.	This canoe is exhibited in the Drents Museum in Assen, Netherlands.

Boat (History)	Where has a 7,000-year-old seagoing reed boat been found?	_____	Libya, Syria, Lebanon, Iran, Iraq, Afghanistan, Yemen, Sudan ***Kuwait*** etc.	A 7,000-year-old seagoing reed boat has been found at site H3 in Kuwait.
	What type of wooden ship was constructed solely of?	_____	mahogany, wood, cedar, birch, spruce, hardwoods, ***teak,*** pine etc.	This type of mammoth wooden ship was constructed solely of teak, with a transport capacity of 400 tonnes.
	When were boats used in Sumer?	between 4000 and _____ BC.	1500, 4500, ***3000,*** 6000, 500, 5000.	Boats were used between 4000 and 3000 BC in Sumer, ancient Egypt and in the Indian Ocean.
	What does water do?	Allowing _____ compounds to react in ways that ultimately allow replication	Inorganic, artificial, natural ingredients, ***organic,*** synthetic, processed, free-range etc.	It carries out this role by allowing organic compounds to react in ways that ultimately allow replication.

	What is the sum total of anabolism and catabolism?	_____	catabolism, digestion, ***metabolism,*** hormone production, ketone production, insulin production, leptin etc.	Metabolism is the sum total of anabolism and catabolism.
	In catabolism, water is used to break bonds in order to do what?	to generate _____ molecules	Larger, bigger, same size, tinier, wider, tiny, narrower, similar size ***smaller*** etc.	In catabolism, water is used to break bonds in order to generate smaller molecules (e.g. glucose, fatty acids, and amino acids to be used for fuels for energy use or other purposes).
	What is water fundamental to?	photosynthesis and _____	Oxygenation, skeletal muscle, oxidative phosphorylation, perfusion, ***respiration,*** excretion, oxygen consumption etc.	Water is fundamental to photosynthesis and respiration.

Water (Effects on Life)	What do living cells use to capture the sun's energy?	oxidize the hydrogen and _____	argon, ***carbon,*** nitrogen, silicon, boron, methane etc.	All living cells use such fuels and oxidize the hydrogen and carbon to capture the sun's energy and reform water and CO2 in the process (cellular respiration).
	What distinguishes towers from masts?	lack of _____-wires	Dude, chick, douche bag, ***guy,*** he etc.	Towers are distinguished from masts by their lack of guy-wires and are therefore, along with tall buildings, self-supporting structures.
	What improves the visibility of the clock?	the height of a clock _____	Teleport, turret range, TPing, creep wave, ***tower,*** enemy base etc.	For example, the height of a clock tower improves the visibility of the clock, and the height of a tower in a fortified building suchas a castle increases the visibility of the surroundings for defensive purposes.

Tower	Towers may also be built for observation, leisure, or what?	_____	physical infrastructure, communications, information services, ***telecommunication*** radio spectrum, public utility etc.	Towers may also be built for observation, leisure, or telecommunication purposes.

Table 4.2: Sample MCQs from some selected topics generated using the image of the boat in the water canal.

Chapter 5

Step-One App Flow

Step-one, learning app for children, is a practical demonstration of the answer-aware question generation framework discussed above. We believe visual stimulus in the form of images is best to grasp the attention of young impressionable minds. Nothing arouses curiosity more than an image of a subject you want to learn about. The app would not put off a curious mind by forcing it to read long paragraphs about the subject, rather only essential information regarding the subject would be presented to the students in each chapter. Such interesting information will be displayed in a manner that would hold the attention of the reader to the last word since we realize the attention span of a child is very short.

We would add the challenge of a quiz at the end to make the competitive learner look forward to how they performed and gauge their ability by putting the newly acquired information to test. It would not only help them recall essential information in a test setting but also give them that confidence of having strong general knowledge about every day things that surround them in their daily lives.

This quiz based approach helps to teach children about new things while keeping away the feeling of dread from their little minds. It makes learning long lasting, easy and fun for them. The app pushes the learners to test their newly assimilated knowledge by offering them automatically generated multiple-choice questions extracted from the knowledge-rich textual corpora of the selected topic. Answering such multiple-choice questions and getting scores out of them, provides instant feedback to the learners on their responses in the quiz.

The app is designed in such a way that even young children could use the app in a very

easy and engaging manner without getting overwhelmed by the information presented. The content is laid out with a focus on making learning approachable and easy to understand without losing interest in the subject at hand. This is achieved by following the policy of less is more, as information is presented with simplicity and depends on the learner if they want to read about a subject in greater detail. This allows the reader to digest new information before delving deeper into the subject.

This whole quiz-based learning app is mainly divided into these four screens:

5.1 Image Input Screen

(a) Image Picker Screen **(b)** Image Selection Screen

Figure 5.1: A child can easily initiate the learning process by selecting any image using these screens

This screen will help children to easily initiate the learning process. They can choose any image of their own liking from the phone's image gallery. Once the image is chosen from the phone's gallery, it will replace the add image icon for the confirmation of selection and then the child can press the next search button to actually upload the image to server for analysis and topics extraction (see Fig. 5.1a).

If the child is being lazy and don't want to access the phone's gallery to upload the image or if nothing interesting is going in his/her mind then the image can also be selected from the predefined list of images. We have planned to update this list of predefined images on daily basis (see Fig. 5.1b), We have hosted all these images on Google Cloud Storage to comply with the requirement of Google Cloud Vision API.

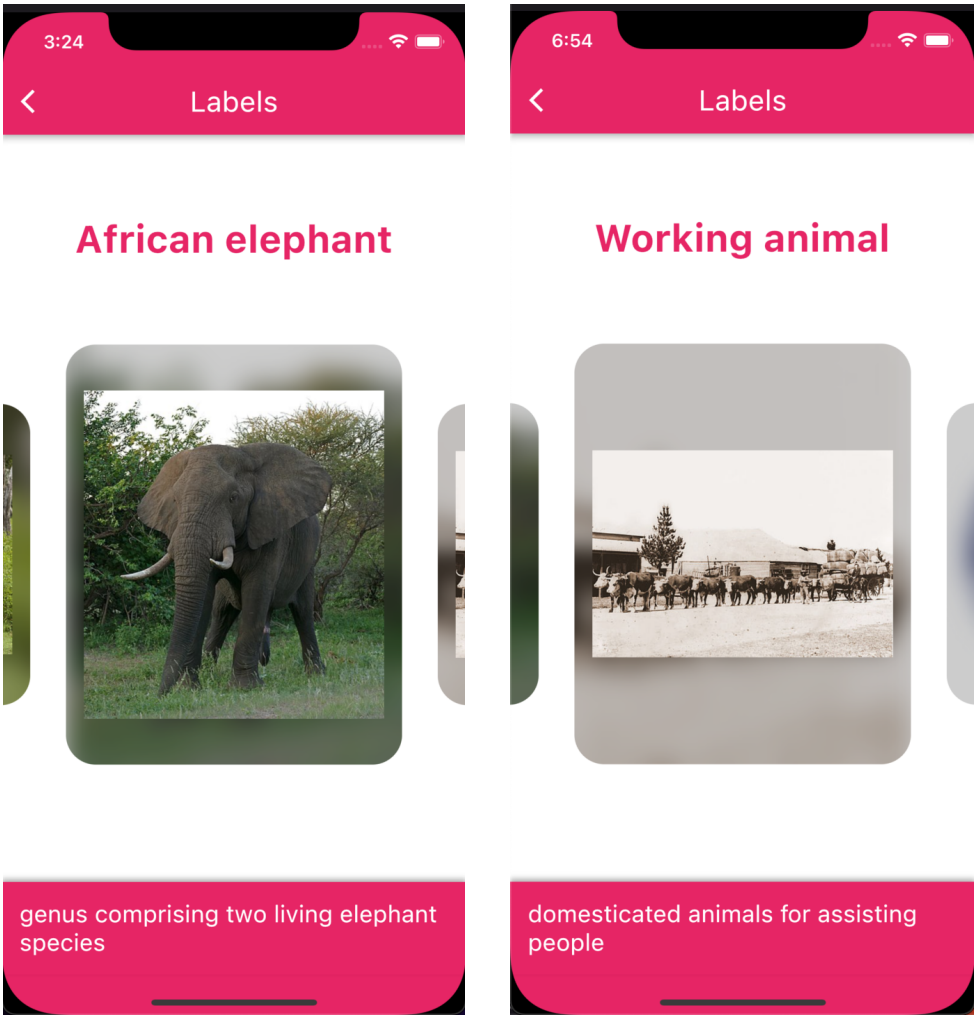
Since children are curious by nature and they are always trying to make sense of their environment, primarily through observation. It is all about capitalizing on their natural urge to learn about anything new in their observation of nature and providing them information close to natural settings like animals that share our planet with us would definitely pique their curiosity, hence our major focus would be to add more images of animals.

5.2 Topic Selection Screen

This screen will display all the extracted topics against the selected image using the Image Labelling Module (see Sec. 3.1). Google Cloud Vision API returns only label strings extracted from the selected image but these label strings could be very boring for the child to look at. We used these labels to extract the extra content, including main Image and relevant description, from the corresponding Wikipedia's page which are then displayed to the child as shown in Fig. 5.2a and Fig. 5.2b.

The best thing about using the Image Labelling API instead of Object Detection API from Google Cloud Vision is that it not only returns the machine identifier and description of the detected object rather it also returns machine identifiers and descriptions of all the related concepts too.

For example, when the image of elephant is selected, it not only returned the label "Elephant" but it also returned the labels like "Working Animal" since elephants are often domesticated for pulling the heavy objects in various parts of the world. These



(a) Main Topic **(b)** Supporting Topic

Figure 5.2: A child can select the main topic or he can also learn about supporting topics.

sort of labels will also help to further enrich the conceptual knowledge about the objects present in the selected image.

5.3 Reading Screen

This screen will display all the extracted Knowledge corresponding to the selected topic using the Knowledge Extraction Module (see Sec. 3.2). First, raw HTML is requested against the selected topic using Wikipedia's API which is then parsed to obtain headings and subheadings along with their content and images. A reading book is prepared using the parsed content in which the cover page displays the title, image, description and other important stuff against the selected topic. The headings are displayed as chapters

(a) Main Descriptive details **(b)** Extra Details

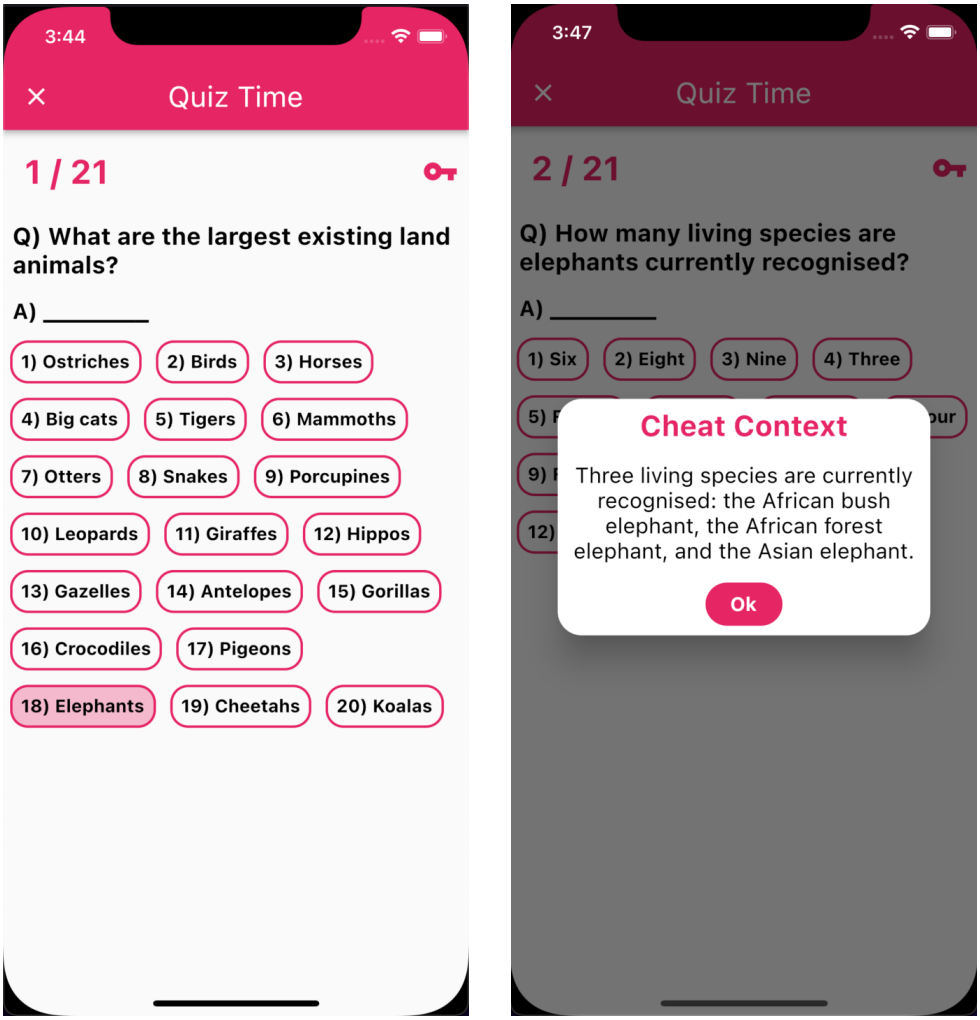
Figure 5.3: A child can read the main details along with all the supporting details.

of the book and each chapter contains its subheadings. A child can jump to any chapter he/she wants using the sidebar which is vertically scroll-able. A special page-turning animation is used to create the feeling of reading a book for the child. Each chapter can have the compact mode in which only the context of the extracted questions will be displayed. This will make the information easily digestible for the child. A child can take a quiz not only from each chapter of this book but also from sub-headings under that chapter. (see Fig. 5.3b).

This would enable a child to understand a subject more thoroughly by breaking it down into smaller, easier to understand parts. This would not only prevent overwhelming the child with information overload but also empower them to undertake learning activity at their own pace. It helps them to prepare for a topic in a manageable, efficient

manner and provide them a momentum to scale a subject where they decide how deeply they want to learn about a topic. Each chapter with its sub-headings have their own individual quiz questions at the end to make a child scale each topic and earn that feeling of reward by cementing their progress by scoring well.

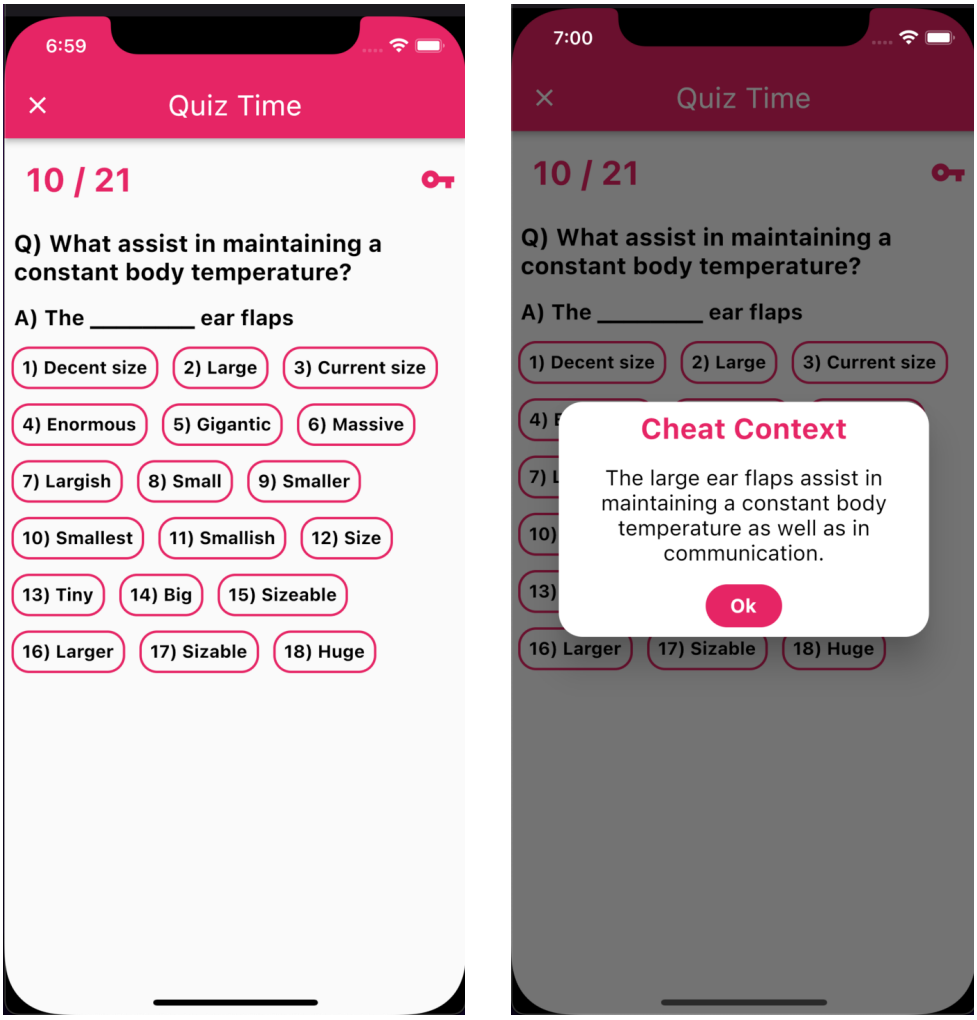
5.4 Quiz Screen



(a) Quiz question from selected Topic **(b)** Context to help children

Figure 5.4: A child can take a quiz while looking at the cheating context for help.

This screen is responsible for displaying the quiz against the selected Chapter using the Question Answer Extraction Module (see Sec. 3.3). Once a child feels well-versed in any of the chapter, they can request to generate a quiz based on that chapter. Once a quiz is generated, they can then try to attempt it by gathering all the knowledge they have



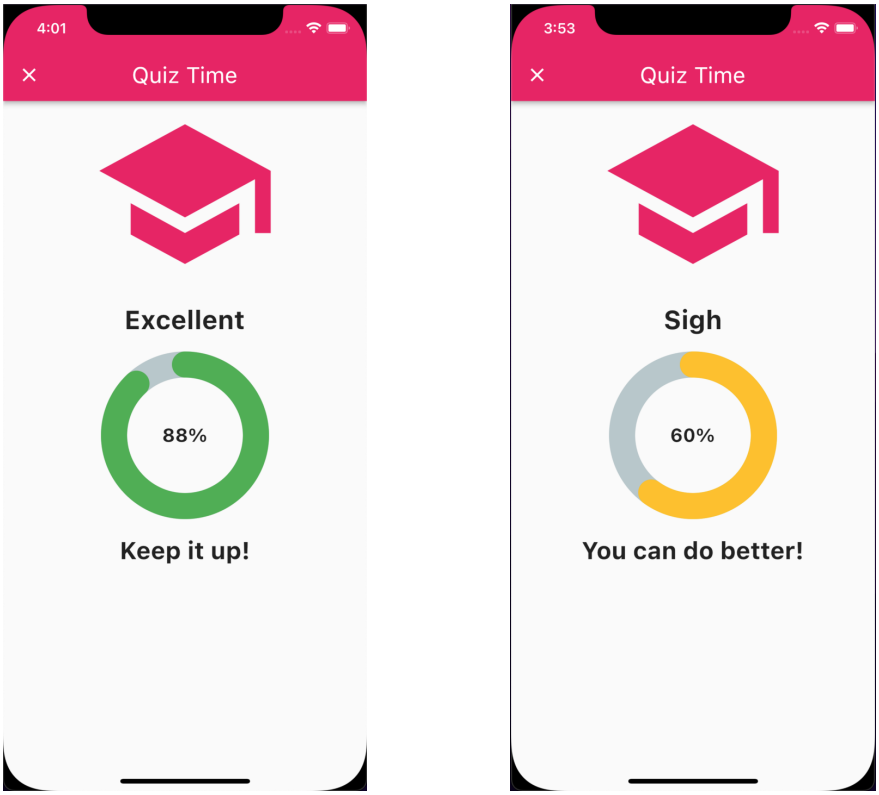
(a) Quiz question from selected Topic **(b)** Context to help children

Figure 5.5: A child can take a quiz while looking at the cheating context for help.

just learned. If a child is feeling confused in any of the question, they can cheat a bit by pressing the special cheat button which is placed on the screen to show the relevant context of the question (see Fig. 5.4). This is done to encourage learning rather than rote learning as Step-One's main aim is to educate. If there are multiple words present in the answer, mostly some random word is selected from the answer sentence and then the Part of speech (POS) tag is extracted for that word using the context of the question and then the relevant options are generated considering that word only. The number of relevant options is also high here so that a student couldn't pick an answer by guess-working. This would challenge the learning of the student as each option would seem relevant and only a mind with complete clarification of concepts will be able to choose the right answer.

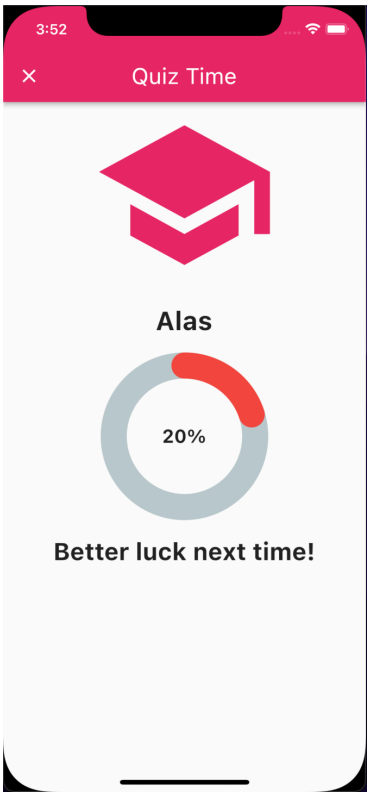
5.5 Score Screen

This screen is responsible for displaying the score of the attempted quiz. Step-One app aims at encouraging learning and making it fun and in no way would provide feedback to demotivate a user. Therefore we tried to distinguish between each score in a way that would still push a user to not lose hope and keep attempting the quiz until they improve. If the score is less than 40% then progress bar will be shown in red color by displaying "Better luck next time!" message and if the score is between 40% to 70% then progress bar will be shown in yellow color by displaying the "You can do better!" message. If the score is greater than 70% then the progress bar will be shown green by displaying "Well done, Keep it up!" message. The aim is to provide learner immediate feedback and let them be the judge of their own performance. They could easily gauge how well they did against the questions and whether it needs improvement or not. These little interactions will develop the interest of the child in learning and scoring big and of course push them to give their best. (see Fig. 5.6). Accurate and immediate feedback would give the user that much needed confidence of moving forward in a topic and having the peace of mind that they have done well or in other case that resolve to learn with better focus and attempt it again.



(a) Score above 70%

(b) Score b/w 40% - 70%



(c) Score below 40%

Figure 5.6: Score screens where a child can check the result of the attempted quiz

Chapter 6

Impact

We believe tapping the potential of artificial intelligence is the need of the hour to make education easily accessible to all, and we are fully confident that this will contribute to the literacy levels of the masses. Especially the children from marginalised sections of society will have a chance at learning and becoming knowledgeable in topics they choose out of their own interest. This will empower them to become productive members of the society, brimming with confidence of knowledge and general know-how of things around them.

It goes without saying that such an initiative would be beneficial for the children of developing countries, where poverty curbs an important intellectual and social development of these young impressionable minds. There are a number of other problems that may deny the indispensable education to the children of third world countries besides poverty. The non-serious attitude of most developing countries' governments towards education is for all to see with the meagre amount allocated to the education budget. Add to it the malnutrition and children often forced to work to support their families and we could truly appreciate the impact of reaching out to these children. It would not be less than a miracle to be able to educate a child after all these unfortunate ground realities surrounding them. If we talk about Pakistan specifically, there are around 63 million children between the ages of 5 and 16 and only 27 million are catered to by the public schooling system [1].

We believe we can make our positive contribution to the digital literacy of these children and at least give them an opportunity to excel and perform. It may not be enough but it will be something significant enough to nurture the critical thinking skills of these

children and they may as well exceed expectations and do wonders with the knowledge gained.

Techniques like Image Captioning by Asking Questions [2] see Sec 2.1.1 show us that various cognitive tasks can be improved using grounded features from VQA task and benchmarks can further be enhanced by using knowledge-aware VQA. If we carefully think about it, we would realize that it's the most natural thing to apply in AI domain as children of human beings learn about new things the same way. The proposed framework can produce knowledge-aware questions dataset from the given images on demand. This dataset can then be used to train and benchmark the VQA models. This dynamically produced dataset can then assess the performance and true capabilities of the VQA models.

We would be putting the latest technology of Transformer models to generate knowledge-aware questions data, something which was previously only done via crowd workers, who were manually required to compile and verify the data that was deficient since it was not so knowledge-aware as well as static in nature. With the proposed model, we will not only generate dynamic data sets but they would be fully knowledge-aware and hence the opportunity to utilize it for training and benchmarking the VQA models and also for educating the masses. It will not only generate knowledge-aware questions about the main object present in the image but it will also try to cater some of the related concepts as well which could be very beneficial for the VQA model answering free-form and open-ended knowledge-aware questions.

Chapter 7

Future Directions and Conclusion

7.1 Future Directions

Step-One is a very crucial learning app for children as it enables them to learn without any instructors around but we all know that most of the children will only be comfortable in using this app if it is made available in their mother tongue. Some of the school dropouts may have a little understanding of English but children who never attended school would have no chance against English language hence they would not be able to take advantage of this knowledge. This app can be made available in multiple languages using Google's translation and we are planning to launch such a version too.

We can also use NLTK summarizers to extract the most meaningful information only which can then be used to produce questions from. This would not only help the user to zero in on the most essential information but also help them focus in preparing for the questions. We can also introduce a compact mode in the app where only the quiz related contextual information will be shown to the student. We can also make use of more related images from the internet in order to provide interesting visual stimuli to the students.

We can also incorporate text to speech and speech to text functionality in the app to provide a more engaging experience to the children since a child will always be more comfortable in processing the information by just listening to it. It is obvious that reading requires certain command over language without which a child could not proceed to understand what is being communicated but on the other hand listening does not involve any such pre-requisites and a child could swiftly assimilate new information.

This functionality of text to speech could be further upgraded to highlight the text being spoken out loud so that the student is completely aware of not just the pronunciation of the word but also the spellings. This would aid the learner to not just have good command over spoken language but also how to spell it out properly.

Since an individual could switch between many languages as Google's translation will be used, it would be possible for them to learn phonetics and spellings of another language if they are willing. Obviously this is ambitious but the opportunity is still there if anyone wants to go that one step further and learn.

We can also use this framework in an adversarial fashion where our question generator will keep posting questions from the images dataset and in turn the VQA model will try to answer those knowledge-aware questions. This was something previously lacking since static datasets used to be compiled using different crowd sourcing strategies, for example utilizing Amazon Mechanical Turk (AMT) and as a result the data was inherently static and straight-forward to be used for benchmarking and training of VQA models.

7.2 Conclusion

There is no doubt that the world is moving towards digitization and the accessibility of smartphones is more widespread than we think. Tapping such a medium's potential to alleviate literacy of thousands of children is not so far fetched in this time and age. The impact would only be felt after saving this generation from the scourge of illiteracy as well as packaging knowledge in a way that motivates individuals to learn on their own.

Step-one child app can change the lives of millions of children who are dropped out of schools and not having enough funds. Utilizing the potential of latest technologies, for example the advent of transformers and the rise of knowledge-aware VQA models we saw an opportunity of leveraging it for the betterment of society by making informal education and general knowledge accessible to millions of students and children outside of school. The situation is completely hopeless for such children since they have no schools as well as complete absence of any mentor who could guide them, and Step-One attempts to fulfill that gap.

We saw a number of related datasets, some of them were targeting common sense or spatial questions while others were even incorporating external knowledge but those all datasets were either generated by human annotators or they were static, based on surface level knowledge.

We discussed how different modules of this app come together to give us a powerful VQA generator framework which derives labels from the selected image and then uses to extract relevant knowledge from Wikipedia. Later on that knowledge corpus is utilized to extract questions via transformer and then options are generated to convert those questions into MCQs (Multiple choice Questions).

In our analytical findings we established that relevant external knowledge is very crucial to answer the generated questions via the proposed framework and we were able to produce a quiz-based learning app for children using our knowledge-aware question generator. We reflected in our app flow that how user-friendly and engaging this platform is. It is full of fun-filled activities while at the same time educating children in an exciting manner.

We discussed how we could progress in the future in upgrading our app and evolving it to cater more and more people and making it such that its easier to use.

This was an attempt to bridge that gap of education and knowledge between those students who have access to formal learning as compared to those who can't access it in any shape or form. You can say its an initiative to make the less fortunate young minds of our society come at the same level of understanding of the world around them as those who could afford formal education.

The proposed framework can help the researchers to train and benchmark the Knowledge-aware VQA models as this framework can also be used in an adversarial fashion.

References

[1] Pak Alliance for Maths and Science. The missing third, 2021. URL https://mathsandscience.pk/publications/the-missing-third/.

[2] Xiaoshan Yang and Changsheng Xu. Image captioning by asking questions. *ACM Trans. Multimedia Comput. Commun. Appl.*, 15(2s), jul 2019. ISSN 1551-6857. doi: 10.1145/3313873. URL https://doi.org/10.1145/3313873.

[3] Alkesh Patel, Akanksha Bindal, Hadas Kotek, Christopher Klein, and Jason Williams. Generating natural questions from images for multimodal assistants. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2270–2274, 2021. doi: 10.1109/ICASSP39728.2021.9413599.

[4] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[5] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8876–8884, 2019.

[7] Super Lucky Games LLC. Trivia star quiz games, 2021. URL https://play.google.com/store/apps/details?id=com.trivia.star.android&hl=en&gl=US.

[8] the binary family. Knowledge trainer: Trivia, 2021. URL https://play.google.com/store/apps/details?id=com.thebinaryfamily.knowledgetrainerfree&hl=en&gl=US.

[9] TIMLEG. General knowledge quiz, 2021. URL https://play.google.com/store/apps/details?id=com.timleg.quiz&hl=en&gl=US.

[10] Henry L Roediger III, Adam L Putnam, and Megan A Smith. Ten benefits of testing and their applications to educational practice. *Psychology of learning and motivation*, 55:1–36, 2011.

[11] Ahmed Jamshed and Muhammad Moazam Fraz. Nlp meets vision for visual interpretation - a retrospective insight and future directions. In *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, pages 1–8, 2021. doi: 10.1109/ICoDT252288.2021.9441517.

[12] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[13] Unicef Pakistan. Education - giving every child the right to education, 2017. URL https://www.unicef.org/pakistan/education.

[14] Promote Your School. 10 reasons why you should use quizzes with children, 2021. URL https://www.promoteyourschool.co.uk/blog/10-reasons-why-you-should-use-quizzes-with-children.

[15] Cassandra Willyard Science. The benefits of pop quizzes - test-taking helps students develop better clues for remembering, 2010. URL https://www.science.org/content/article/benefits-pop-quizzes.

[16] Dr. William Corvey. Grounded artificial intelligence language acquisition (gaila), 2021. URL https://www.darpa.mil/program/grounded-artificial-intelligence-language-acquisition.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects

in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016.

[19] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[20] Sumithra Bhakthavatsalam, Kyle Richardson, Niket Tandon, and Peter Clark. Do dogs have whiskers? a new knowledge base of haspart relations, 2020.

[21] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[22] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa, 2020.

[23] Keshav Kolluru, Shreyans Shrimal, and Sudharsan Krishnaswamy. Cognitivecam: A visual question answering application, 11 2017.

[24] Google. Google lens, 2021. URL https://play.google.com/store/apps/details?id=com.google.ar.lens&hl=en&gl=US.

[25] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[26] Andrew Trask, Phil Michalak, and John Liu. sense2vec - a fast and accurate method for word sense disambiguation in neural word embeddings, 2015.

[27] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

[28] Patil Suraj. Question generation using transformers, 2021. URL https://github.com/patil-suraj/question_generation.

[29] Wikipedia. Wikipedia: Size of wikipedia, 2021. URL https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.