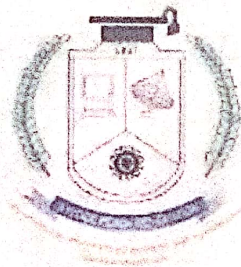


**AN ASSISTIVE VISION FOR VISUALLY CHALLENGED
PERSON USING ALEXA**

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and
Engineering with Specialization in Artificial Intelligence

By

**CHILAKAPATI DHEERAJ (Reg.no – 40731019)
KOTHAPALLI SUJITH REDDY (Reg.no – 40731048)**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Category-1 University by UGC

**Accredited with Grade "A++" by NAAC | 12B Status by UGC | Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,**

CHENNAI – 600119

April – 2024

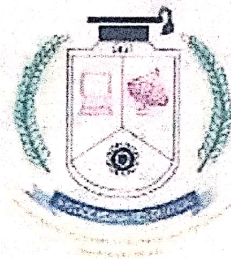


**AN ASSISTIVE VISION FOR VISUALLY CHALLENGED
PERSON USING ALEXA**

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and
Engineering with Specialization in Artificial Intelligence

By

**CHILAKAPATI DHEERAJ (Reg.no – 40731019)
KOTHAPALLI SUJITH REDDY (Reg.no – 40731048)**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Category-1 University by UGC

Accredited with Grade "A++" by NAAC | 12B Status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

CHENNAI – 600119

April – 2024





SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

CATEGORY-I UNIVERSITY BY UGC

Accredited "A++" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Chilakapati Dheeraj (40731019)** and **Kothapalli Sujith Reddy (40731048)** who carried out the Project entitled **"AN ASSISTIVE VISION FOR VISUALLY CHALLENGED PERSON USING ALEXA"** under my supervision from November 2023 to April 2024.

Internal Guide

Dr. R. RAJALAKSHMI M.E., Ph.D.

HEAD OF THE DEPARTMENT
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

Head of the Department

Jeppiaar Nagar, Rajiv Gandhi Salai,

Chennai - 600 119.

Dr. L. Lakshmanan M.E., Ph.D.

Submitted for Viva voce Examination held on 12-04-2024

Internal Examiner

External Examiner



DECLARATION

I, Chilakapati Dheeraj (Reg no - 40731019), hereby declare that the Project Report entitled 'AN ASSISTIVE VISION FOR VISUALLY CHALLENGED PERSON USING ALEXA' done by me under the guidance of Dr. R. Rajalakshmi M.E., Ph.D. is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering with Specialization in Artificial Intelligence.

DATE: 12-04-2024

PLACE: Chennai



SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to the **Board of Management of Sathyabama Institute of Science and Technology** for their encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph. D.,** Dean, School of Computing, **Dr. L. Lakshmanan M.E., Ph.D.,** and **Dr. S. Vigneshwari M.E., Ph.D.,** Heads of the **Department of Computer Science and Engineering** for providing me necessary support during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. R. Rajalakshmi M.E., Ph.D.,** for her valuable guidance, suggestions, and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

Globally, 1 billion people have a vision impairment that could have been prevented or has yet to be addressed. This 1 billion people include those with moderate or severe distance vision impairment or blindness due to unaddressed refractive error (123.7 million), cataract (65.2 million), glaucoma (6.9 million), corneal opacities (4.2 million), diabetic retinopathy (3 million), and trachoma (2 million), as well as near vision impairment caused by unaddressed presbyopia (826 million). In terms of regional differences, the prevalence of distance vision impairment in low- and middle-income regions is estimated to be four times higher than in high-income regions. With regards to near vision, rates of unaddressed near vision impairment are estimated to be greater than 80% in western, eastern and central sub-Saharan Africa, while comparative rates in high-income regions of North America, Australasia, Western Europe, and of Asia-Pacific are reported to be lower than 10%. Population growth and ageing are expected to increase the risk that more people acquire vision impairment. In this project in order to facilitate the blind we will be using deep learning algorithm such as efficientnet B3 to caption the image for the blind person in which the blind can know about the object detection, distance and position of object. This is been achieved by using advanced image captioning techniques implementing efficientnet B3 algorithms and tokenization methods where the scenes with different captions are learned by the machine. Whenever an image is captured via the camera are been recognized and predicted by the processor. The major objects are also predicted and the distances calculated from the camera. After the prediction, it is been sent to the Alexa microphone which will send an audio output to the user which can help them identify the distance and position of object. Thus, with the help of this project we provide an artificial vision to the blind, which can help them gain confidence while travelling alone. Ageing populations and population expansion are predicted to raise the likelihood that more people may have visual impairment. In this research, we will use deep learning algorithms, such efficientnet B3, to caption images so that the blind may read and understand information about object identification, object position, and distance. The distances between the main objects and the camera are also computed.

TABLE OF CONTENTS

Chapter No	Title	Page No
	ABSTRACT	i
	LIST OF FIGURES	iv
	LIST OF ABBREVIATIONS	v
1	INTRODUCTION	1
	1.1 General Introduction	1
	1.2 Technologies used	2
2	LITERATURE SURVEY	7
	2.1 A Text-Guided Generation and Refinement Model for Image Capturing	7
	2.2 SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning	7
	2.3 Topic-Oriented Image Captioning Based on Order-Embedding	7
	2.4 More is Better: Precise and Detailed Image Captioning using Online Positive Recall and Missing Concepts Mining	8
	2.5 Re-Caption: Saliency-Enhanced Image Captioning through Two-Phase Learning	8
	2.6 Cross-Domain Image Captioning via Cross Modal Retrieval and Model Adaptation	9
3	REQUIREMENT ANALYSIS	10
	3.1 Hardware requirements	10
	3.2 Software description	11
	3.3 Dataset Collection	13
4	DESCRIPTION OF PROPOSED SYSTEMS	15
	4.1 Existing Systems	15
	4.2 Disadvantages of Existing System	15

	4.3	Proposed System	15
	4.4	Advantages of Proposed System	16
	4.5	System Architecture	16
	4.6	Working	17
5		IMPLEMENTATION DETAILS	18
	5.1	Feature extraction using EfficientNet B3	18
	5.2	Tokenization	19
	5.3	Training with RNN Algorithm	20
	5.4	Prediction of scenes, distance and position of object	21
6		RESULTS AND DISCUSSIONS	22
7		CONCLUSION	26
	7.1	Conclusion	26
	7.2	Future work	26
		REFERENCES	27
		APPENDIX	30
		A. SOURCE CODE	30
		B. SCREENSHOTS	44
		C. RESEARCH PAPER	50
		D. CERTIFICATE	60

LIST OF FIGURES

FIGURE NO	FIGURE NAME	PAGE NO
1.1	Deep Learning Layers	2
1.2	Deep Learning Process	3
1.3	CNN	4
1.4	Types of CNN	5
3.1	Alexa Microphone	11
3.2	Google Colab Notebook	12
3.3	Dataset collect	14
4.1	Proposed System Architecture	16
5.1	Efficientnet B3	19
5.2	Tokenization	20
5.3	RNN architecture	21
5.4	Prediction of scenes, distance and position of object	21
6.1	Dataset Collection	22
6.2	Simple Pre-processing	22
6.3	Feature Extraction	23
6.4	Tokenization	23
6.5	Training using Recurrent Neural Network	24
6.6	Epochs vs Loss graph	24
6.7	Validating image caption	24
6.8	Object detection	25
6.9	Distance of object	25

LIST OF ABBREVIATIONS

CNN	Personally Identifiable Information
CNN	Convolutional Neural Networks
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IoT	Internet of Things
M2M	Machine-to-Machine
RNN	Recurrent Neural Networks
SCADA	Supervisory Control and Data Acquisition

CHAPTER 1

INTRODUCTION

1.1 GENERAL INTRODUCTION:

The human eye is like a camera that collects, focuses, and transmits light through a lens to create an image of its surroundings. In a camera, the image is created on film or an image sensor. In the eye, the image is created on the retina, a thin layer of light-sensitive tissue at the back of the eye.

Like a camera, the human eye controls the amount of light that enters the eye. The iris (the colored circular part of the eye) controls the amount of light passing through the pupil. It closes up the pupil in bright light and opens it wider in dim light. The cornea is the transparent, protective surface of the eye. It helps focus light, as does the lens, which sits just behind the iris. When light enters the eye, the retina changes the light into nerve signals. The retina then sends these signals along the optic nerve (a cable of more than 1,000,000 nerve fibres) to the brain. Without a retina or optic nerve, the eye can't communicate with the brain, making vision impossible.

Many people have some type of visual problem at some point in their lives. Some can no longer see objects far away. Others have problems reading small print. These types of conditions are often easily treated with eyeglasses or contact lenses. But when one or more parts of the eye or brain that are needed to process images become diseased or damaged, severe or total loss of vision can occur. In these cases, vision can't be fully restored with medical treatment, surgery, or corrective lenses like glasses or contacts.

The American Foundation for the Blind estimates that 10 million people in the United States are visually impaired. Visual impairment is a term experts use to describe any kind of vision loss, whether it's someone who cannot see at all or someone who has partial vision loss. Some people are completely blind, but many others have what's called legal blindness.

1.2 TECHNOLOGIES USED:

- Deep Learning

Deep learning is a computer software that mimics the network of neurons in a brain. It is a subset of machine learning and is called deep learning because it makes use of deep neural networks.

Deep learning algorithms are constructed with connected layers

- The first layer is called the Input Layer
- The last layer is called the Output Layer
- All layers in between are called Hidden Layers. The word deep means the network join neurons in more than two layers.

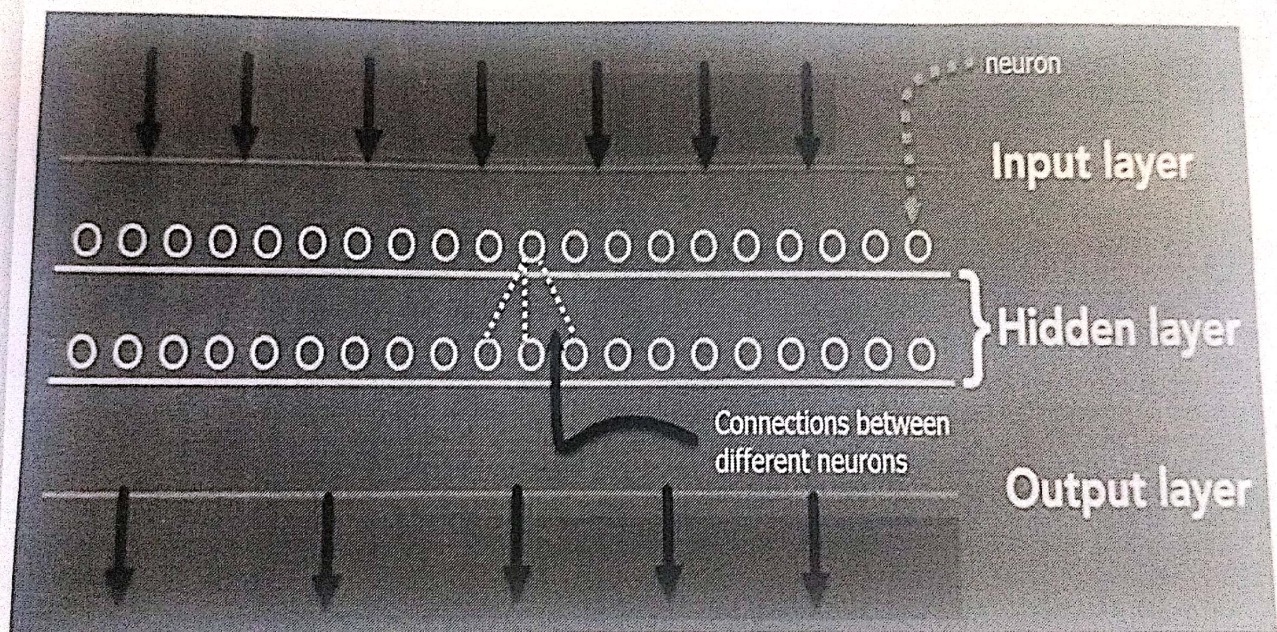


Figure 1.1 Deep Learning Layers

Each Hidden layer is composed of neurons. The neurons are connected to each other. The neuron will process and then propagate the input signal it receives the layer above it [1]. The strength of the signal given the neuron in the next layer depends on the weight, bias and activation function.

The network consumes large amounts of input data and operates them through multiple layers; the network can learn increasingly complex features of the data at each layer.

• IMPORTANCE OF DEEP LEARNING

Deep learning is a powerful tool to make prediction an actionable result. Deep learning excels in pattern discovery (unsupervised learning) and knowledge-based prediction. Big data is the fuel for deep learning. When both are combined, an organization can reap unprecedented results in term of productivity, sales, management, and innovation.

Deep learning can outperform traditional method. For instance, deep learning algorithms are 41% more accurate than machine learning algorithm in image classification, 27 % more accurate in facial recognition and 25% in voice recognition.

• DEEP LEARNING PROCESS

A deep neural network provides state-of-the-art accuracy in many tasks, from object detection to speech recognition. They can learn automatically, without predefined knowledge explicitly coded by the programmers.

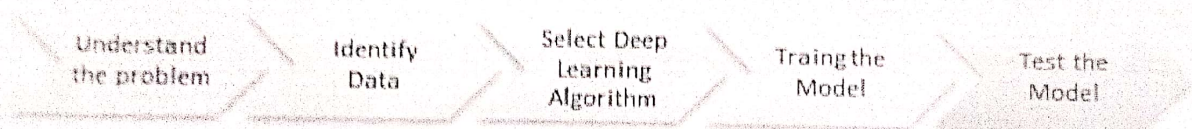


Figure 1.2 Deep Learning Process

To grasp the idea of deep learning, imagine a family, with an infant and parents. The toddler points objects with his little finger and always says the word 'cat.' As its parents are concerned about his education, they keep telling him 'Yes, that is a cat' or 'No, that is not a cat.' The infant persists in pointing objects but becomes more accurate with 'cats.' The little kid, deep down, does not know why he can say it is a cat or not. He has just learned how to hierarchies' complex features coming up with a cat by looking at the pet overall and continue to focus on details such as the tails or the nose before to make up his mind.

A neural network works quite the same. Each layer represents a deeper level of knowledge, i.e., the hierarchy of knowledge. A neural network with four layers will learn more complex feature than with that with two layers.

The learning occurs in two phases.

- The first phase consists of applying a nonlinear transformation of the input and create a statistical model as output.
- The second phase aims at improving the model with a mathematical method known as derivative.

The neural network repeats these two phases' hundreds to thousands of times until it has reached a tolerable level of accuracy. The repeat of this two-phase is called an iteration.

• RECURRENT NEURAL NETWORKS (RNN)

RNN is a multi-layered neural network that can store information in context nodes, allowing it to learn data sequences and output a number or another sequence. In simple words it an Artificial neural network whose connections between neurons include loops. RNNs are well suited for processing sequences of inputs.

- The RNN neurons will receive a signal that point to the start of the sentence.
- The network receives the word "Do" as an input and produces a vector of the number. This vector is fed back to the neuron to provide a memory to the network. This stage helps the network to remember it received "Do" and it received it in the first position.
- The network will similarly proceed to the next words. It takes the word "you" and "want." The state of the neurons is updated upon receiving each word.
- The final stage occurs after receiving the word "a." The neural network will provide a probability for each English word that can be used to complete the sentence. A well-trained RNN probably assigns a high probability to "café," "drink," "burger," etc.

• Common uses of RNN

1. Help securities traders to generate analytic reports
2. Detect abnormalities in the content of financial statement
3. Detect fraudulent credit-card transaction
4. Provide a caption for images
5. Power chatbots
6. The standard uses of RNN occur when the practitioners are working with time-series data or sequences (e.g., audio recordings or text).

• CONVOLUTIONAL NEURAL NETWORKS (CNN)

CNN is a multi-layered neural network with a unique architecture designed to extract increasingly complex features of the data at each layer to determine the output. CNN's are well suited for perceptual tasks.

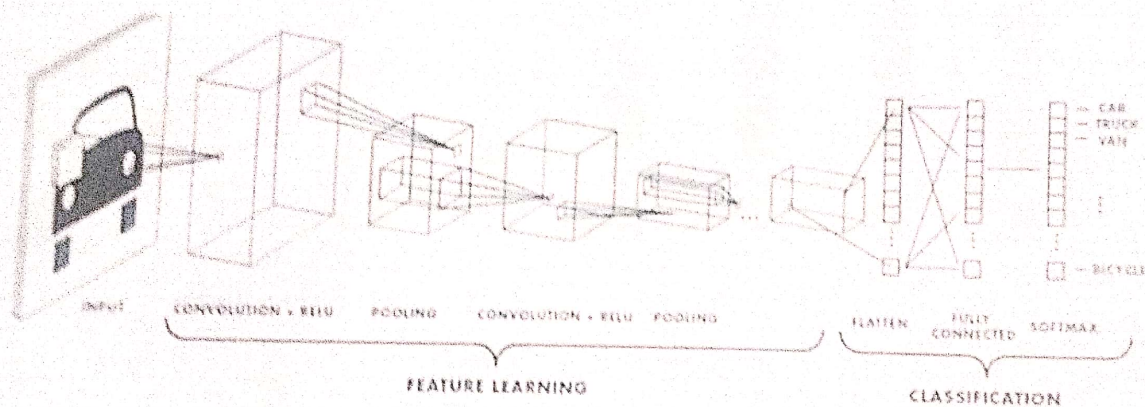


Figure 1.3 CNN

CNN is mostly used when there is an unstructured data set (e.g., images) and the practitioners need to extract information from it

For instance, if the task is to predict an image caption:

- The CNN receives an image of let's say a cat, this image, in computer term, is a collection of the pixel. Generally, one layer for the greyscale picture and three layers for a color picture.
- During the feature learning (i.e., hidden layers), the network will identify unique features, for instance, the tail of the cat, the ear, etc.

- When the network thoroughly learned how to recognize a picture, it can provide a probability for each image it knows. The label with the highest probability will become the prediction of the network

A Convolutional Neural Network (CNN) are a special kind of multi-layer neural networks, designed to recognize visual patterns directly from pixel images with minimal pre-processing. The ImageNet project is a large visual database designed for use in visual object recognition software research [23]. The ImageNet project runs an annual software contest, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), where software programs compete to correctly classify and detect objects and scenes.

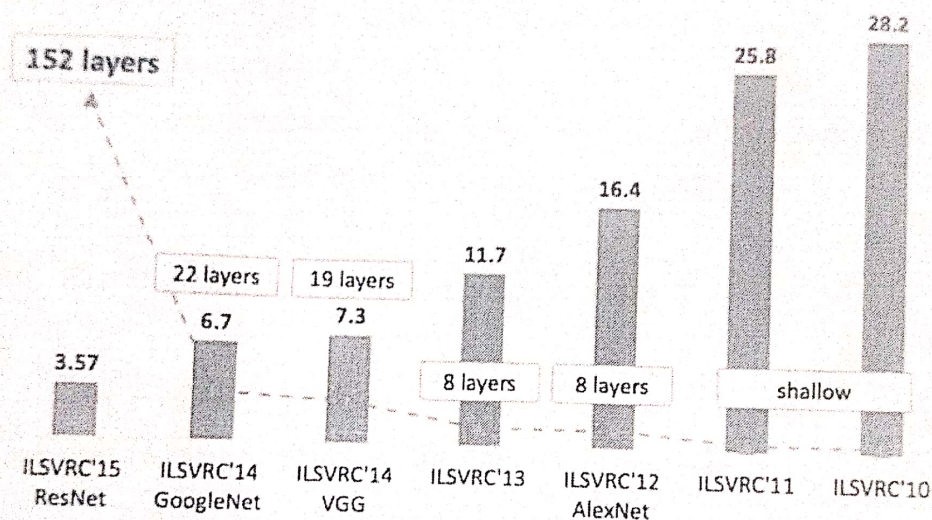


Figure 1.4 Types of CNN

- **Advantages of Convolutional Neural Networks (CNNs):**

1. Good at detecting patterns and features in images, videos, and audio signals.
2. Robust to translation, rotation, and scaling invariance.
3. End-to-end training, no need for manual feature extraction.
4. Can handle large amounts of data and achieve high accuracy.

- **Disadvantages of Convolutional Neural Networks (CNNs):**

1. Computationally expensive to train and require a lot of memory.
2. Can be prone to overfitting if not enough data or proper regularization is used.
3. Requires large amounts of labelled data.
4. Interpretability is limited, it's hard to understand what the network has learned.

CHAPTER 2

LITERATURE SURVEY

2.1 A Text-Guided Generation and Refinement Model for Image Captioning

A high-quality image description requires not only the logic and fluency of language but also the richness and accuracy of the content. However, due to the semantic gap between vision and language, most existing image captioning approaches that directly learn the cross-modal mapping from the vision to language are difficult to meet these two requirements simultaneously. Inspired by the progressive learning mechanism, we trace the "generating + refining" route and propose a novel Text-Guided Generation and Refinement (dubbed as TGGAR) model with assistance from the guide text to improve the quality of captions. The guide text is selected from the training set according to the visual content similarity, then utilized to explore the salient objects and extend the candidate words.

2.2 SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning

Deep neural networks (DNNs) have been recently found popular for image captioning problems in remote sensing (RS). Existing DNN-based approaches rely on the availability of a training set made up of a high number of RS images with their captions. However, captions of training images may contain redundant information (they can be repetitive or semantically similar to each other), resulting in information deficiency while learning a mapping from the image domain to the language domain. To overcome this limitation, in this article, we present a novel summarization-driven RS image captioning (SD-RSIC) approach. The first step obtains the standard image captions by jointly exploiting convolutional neural networks (CNNs) with long short-term memory (LSTM) networks [21].

2.3 Topic-Oriented Image Captioning Based on Order-Embedding

We present an image captioning framework that generates captions under a given topic. The topic candidates are extracted from the caption corpus. A given image's topics are then selected from these candidates by a CNN-based multi-label classifier [2]. The input to the caption generation model is an image topic pair, and the

output is a caption of the image. For this purpose, a cross-modal embedding method is learned for the images, topics, and captions. In the proposed framework, the topic, caption, and image are organized in a hierarchical structure, which is preserved in the embedding space by using the order embedding method. The lower bound pushes the images and captions about the same topic closer together in the embedding space.

2.4 More is Better: Precise and Detailed Image Captioning using Online Positive Recall and Missing Concepts Mining

Recently, a great progress in automatic image captioning has been achieved by using semantic concepts detected from the image. However, we argue that existing concepts-to caption framework, in which the concept detector is trained using the image-caption pairs to minimize the vocabulary discrepancy, suffers from the deficiency of insufficient concepts. The reasons are two-fold: 1) the extreme imbalance between the number of occurrence positive and negative samples of the concept, and 2) the incomplete labelling in training captions caused by the biased annotation and usage of synonyms [22]. In this paper, we propose a method, termed Online Positive Recall and Missing Concepts Mining (OPR-MCM), to overcome those problems. Our method adaptively re-weights the loss of different samples according to their predictions for online positive recall and uses a two-stage optimization strategy for missing concepts mining.

2.5 Re-Caption: Saliency-Enhanced Image Captioning through Two-Phase Learning

Visual and semantic saliency are important in image captioning. However, single-phase image captioning benefits little from limited saliency without a saliency predictor. In this paper, a novel saliency-enhanced re-captioning framework via two-phase learning is proposed to enhance the single-phase image captioning. In the framework, visual saliency and semantic saliency are distilled from the first-phase model and fused with the second-phase model for model self-boosting. The semantic saliency mechanism sheds some lights on the properties of words with part-of-speech Noun in a caption. Besides, another type of saliency, sample saliency is proposed to explicitly compute the saliency degree of each sample, which helps for more robust image captioning.

2.6 Cross-Domain Image Captioning via Cross-Modal Retrieval and Model Adaptation

In recent years, large scale datasets of paired images and sentences have enabled the remarkable success in automatically generating descriptions for images, namely image captioning. However, it is labor-intensive and time-consuming to collect a sufficient number of paired images and sentences in each domain. It may be beneficial to transfer the image captioning model trained in an existing domain with pairs of images and sentences (i.e., source domain) to a new domain with only unpaired data (i.e., target domain). In this paper, we propose a cross-modal retrieval aided approach to cross-domain image captioning that leverages a cross-modal retrieval model to generate pseudo pairs of images and sentences in the target domain to facilitate the adaptation of the captioning model. To learn the correlation between images and sentences in the target domain, we propose an iterative cross-modal retrieval process where a cross-modal retrieval model is first pre-trained using the source domain data and then applied to the target domain data to acquire an initial set of pseudo image-sentence pairs [20].

2.7 Chinese Image Caption Generation via Visual Attention and Topic Modeling

Automatic image captioning is to conduct the cross-modal conversion from image visual content to natural language text. Involving computer vision (CV) and natural language processing (NLP), it has become one of the most sophisticated research issues in the artificial-intelligence area [19]. Based on the deep neural network, the neural image caption (NIC) model has achieved remarkable performance in image captioning, yet there still remain some essential challenges, such as the deviation between descriptive sentences generated by the model and the intrinsic content expressed by the image, the low accuracy of the image scene description, and the monotony of generated sentences. In addition, most of the current datasets and methods for image captioning are in English. However, considering the distinction between Chinese and English in syntax and semantics, it is necessary to develop specialized Chinese image caption generation methods to accommodate the difference.

CHAPTER 3

REQUIREMENT ANALYSIS

3.1 HARDWARE REQUIREMENTS

• Alexa Microphone

Amazon Alexa, also known simply as Alexa, is a virtual assistant technology largely based on a Polish speech synthesiser named Ivona, bought by Amazon in 2013. It was first used in the Amazon Echo smart speaker and the Echo Dot, Echo Studio and Amazon Tap speakers developed by Amazon Lab126. It is capable of voice interaction, music playback, making to-do lists, setting alarms, streaming podcasts, playing audiobooks, and providing weather, traffic, sports, and other real-time information, such as news.

Alexa can also control several smart devices using itself as a home automation system. Users are able to extend the Alexa capabilities by installing "skills" (additional functionality developed by third-party vendors, in other settings more commonly called apps) such as weather programs and audio features.

It uses automatic speech recognition, natural language processing, and other forms of weak AI to perform these tasks [18]. Most devices with Alexa allow users to activate the device using a wake-word (such as Alexa or Amazon); other devices (such as the Amazon mobile app on iOS or Android and Amazon Dash Wand) require the user to click a button to activate Alexa's listening mode, although, some phones also allow a user to say a command, such as "Alexa" or "Alexa wake".

As of November 2018, Amazon had more than 10,000 employees working on Alexa and related products [17]. In January 2019, Amazon's devices team announced that they had sold over 100 million Alexa-enabled devices. In September 2019, Amazon launched many new devices achieving many records while competing with the world's smart home industry. The new Echo Studio became the first smart speaker with 360 sound and Dolby sound. Other new devices included an Echo dot with a clock behind the fabric, a new third-generation Amazon Echo, Echo Show 8, a plug-in Echo device, Echo Flex, Alexa built-in wireless earphones, Echo buds, Alexa built-in spectacles, Echo frames, an Alexa built-in Ring, and Echo Loop.



Figure 3.1 Alexa Microphone

3.2 SOFTWARE DESCRIPTION

The purpose of the Software Requirement Specification is to produce the specification of the analysis task and also to establish complete information about the requirement, behavior and also the other constraint like functional performance and so on. The main aim of the Software Requirement Specification is to completely specify the technical requirements for the software product in a concise and in unambiguous manner [16].

- **GOOGLE COLAB**

In this project, google colab is used as an open-source IDE.

Google Colaboratory is a free online cloud-based Jupyter notebook environment that allows us to train our machine learning and deep learning models on CPUs, GPUs, and TPUs.

It gives us a decent GPU for free, which we can continuously run for 12 hours. For most data science folks, this is sufficient to meet their computation needs.

Google Colab gives us three types of runtime for our notebooks:

- CPUs,
- GPUs, and
- TPUs

Colab gives us 12 hours of continuous execution time. After that, the whole virtual machine is cleared and we have to start again. We can run multiple CPU, GPU, and TPU instances simultaneously, but their resources are shared between these instances.

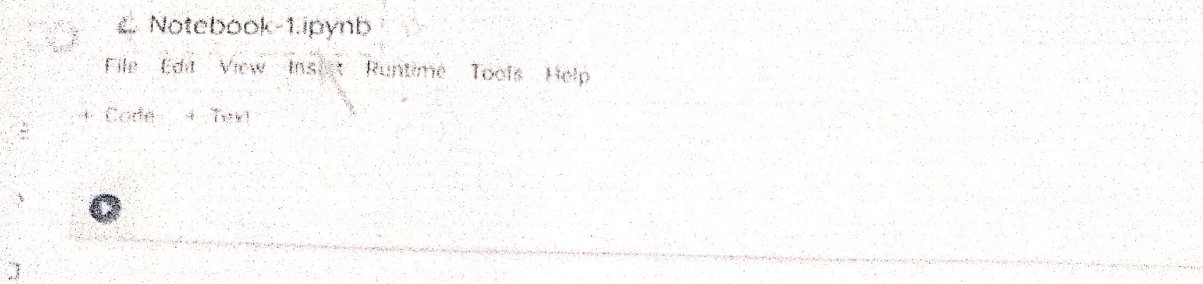


Figure 3.2 Google Colab Notebook

Colab notebooks allow us to combine executable code and rich text in a single document, along with images, HTML, LaTeX and more. When we create our own Colab notebooks, they are stored in our Google Drive account [15]. We can easily share our Colab notebooks with co-workers or friends, allowing them to comment on our notebooks or even edit them. To learn more, see Overview of Colab. To create a new Colab notebook we can use the File menu above, or use the following link: [create a new Colab notebook](#). Colab notebooks are Jupyter notebooks that are hosted by Colab.

As a developer, we can perform the following using Google Colab;

- Write and execute code in Python
- Create/Upload/Share notebooks
- Import/Save notebooks from/to Google Drive
- Import/Publish notebooks from GitHub
- Import external datasets
- Integrate PyTorch, TensorFlow, Keras, OpenCV
- Free Cloud service with free GPU

* PYTHON

In this project python is used as a programming language for development.

In technical terms, Python is an object-oriented, high-level programming language with integrated dynamic semantics primarily for web and app development. It is extremely attractive in the field of Rapid Application Development because it offers dynamic typing and dynamic binding options.

Python is relatively simple, so it's easy to learn since it requires a unique syntax that focuses on readability [14]. Developers can read and translate Python code much easier than other languages. In turn, this reduces the cost of program maintenance and development because it allows teams to work collaboratively without significant language and experience barriers [4].

Additionally, Python supports the use of modules and packages, which means that programs can be designed in a modular style and code can be reused across a variety of projects. Once a module or package is developed by an user, it can be scaled for use in other projects, and it's easy to import or export these modules.

One of the most promising benefits of Python is that both the standard library and the interpreter are available free of charge, in both binary and source form. There is no exclusivity either, as Python and all the necessary tools are available on all major platforms. Therefore, it is an enticing option for developers who don't want to worry about paying high development costs.

3.3 Dataset Collection

In this project, COCO dataset is used to train the model. A data set is a collection of data. Deep Learning has become the go-to method for solving many challenging real-world problems [12]. It's definitely by far the best performing method for computer vision tasks. The image above showcases the power of deep learning for computer vision. With enough training, a deep network can segment and identify the "key points" of every person in the image. These deep learning machines that have been working so well need fuel lots of fuel, that fuel is data. The more labelled data available, the better our model performs. The idea of more data leading to better performance has even been explored at a large-scale by Google with a dataset of 300 million images! When deploying a Deep Learning model in a real-world application, data must be constantly fed to continue improving its performance. And, in

the deep learning era, data is very well arguably the most valuable resource. There are three steps of collecting data

- **Scraping From the Web**

Manually finding and downloading images takes a long time simply due to the amount of human work involved. The task probably has some kinds of common objects are to be detected. And so that becomes the keyword for web-scraping. It also becomes the class name for that object [3]. Every *single pixel in the image is required*. To get those, it's best to use some really great image annotation tools that are already out there. Given a rough set of polygon points around an object, can generate the pixel labels for segmentation. Deep extreme cut is also quite similar except they use only the four extreme points around the object. This will then give some nice bounding box and segmentation labels. Another option is to use an existing image annotation GUIs.

- **Third-party:**

Since data has become such a valuable commodity in the deep learning era, many start-ups have started to offer their own image annotation services they'll gather and label the data. Given a description of what kind of data and annotations needed [13]. Mighty is one that has been doing self-driving car image annotation and has become pretty big in the space were at CVPR 2018 too. Payment AI are less specialized than Mighty AI, offering image annotation for any domain.

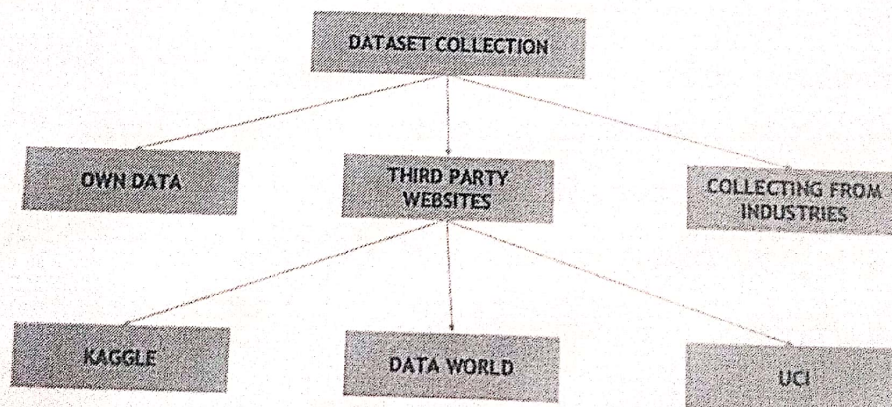


Figure 3.3 Dataset collect

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

4.1 EXISTING SYSTEM

A high-quality image description requires not only the logic and fluency of language but also the richness and accuracy of the content. However, due to the semantic gap between vision and language, most existing image captioning approaches that directly learn the cross-modal mapping from the vision to language are difficult to meet these two requirements simultaneously. Inspired by the progressive learning mechanism, we trace the "generating + refining" route and propose a novel Text-Guided Generation and Refinement (dubbed as TGGAR) model with assistance from the guide text to improve the quality of captions. The guide text is selected from the training set according to the visual content similarity, then utilized to explore the salient objects and extend the candidate words. Specifically, we follow the encoder-decoder architecture, and design a Text-Guided Relation Encoder (TGRE) to learn the visual representation that is more consistent with human visual cognition. Besides, we divide the decoder part into two submodules: a Generator for the primary sentence generation and a Refiner for the sentence refinement. Generator, consisting of a standard LSTM and a Gate on Attention (GOA) module, aims to generate the primary sentence logically and fluently. Refiner contains a caption encoder module, an attention-based LSTM and a GOA module, which iteratively modifies the details in the primary caption to make captions rich and accurate with help of the guide text [11]. Extensive experiments on the MS COCO captioning dataset demonstrate that our framework with fewer parameters remains comparable to transformer-based methods, and achieves state-of-the-art performance compared with other relevant approaches.

4.2 DISADVANTAGES OF EXISTING SYSTEM

- Not real time applicable as the accuracy obtained is less.
- Using an attention-based LSTM, it takes longer to train and require more memory to train.

4.3 PROPOSED SYSTEM

In this project in order to facilitate the blind we will be using deep learning algorithm such as efficientnet B3 to caption the image for the blind person in which the

blind can know about the object detection, distance and position of object. This is been achieved by using advanced image captioning techniques implementing efficientnet B3 algorithms and tokenization methods where the scenes with different captions are learned by the machine. Whenever an image is captured via the camera are been recognized and predicted by the processor [24]. The major objects are also predicted and the distances calculated from the camera. After the prediction, it is been sent to the Alexa microphone which will send an audio output to the user which can help them identify the distance and position of object. Thus, with the help of this project we provide an artificial vision to the blind, which can help them gain confidence while travelling alone.

4.4 ADVANTANGES OF PROPOSED SYSTEM

- Cheap and effective solution for the blind to become aware of the surroundings.
- Effective scene prediction model is developed.
- Provides an artificial vision to the visually challenged people.

4.5 SYSTEM ARCHITECTURE

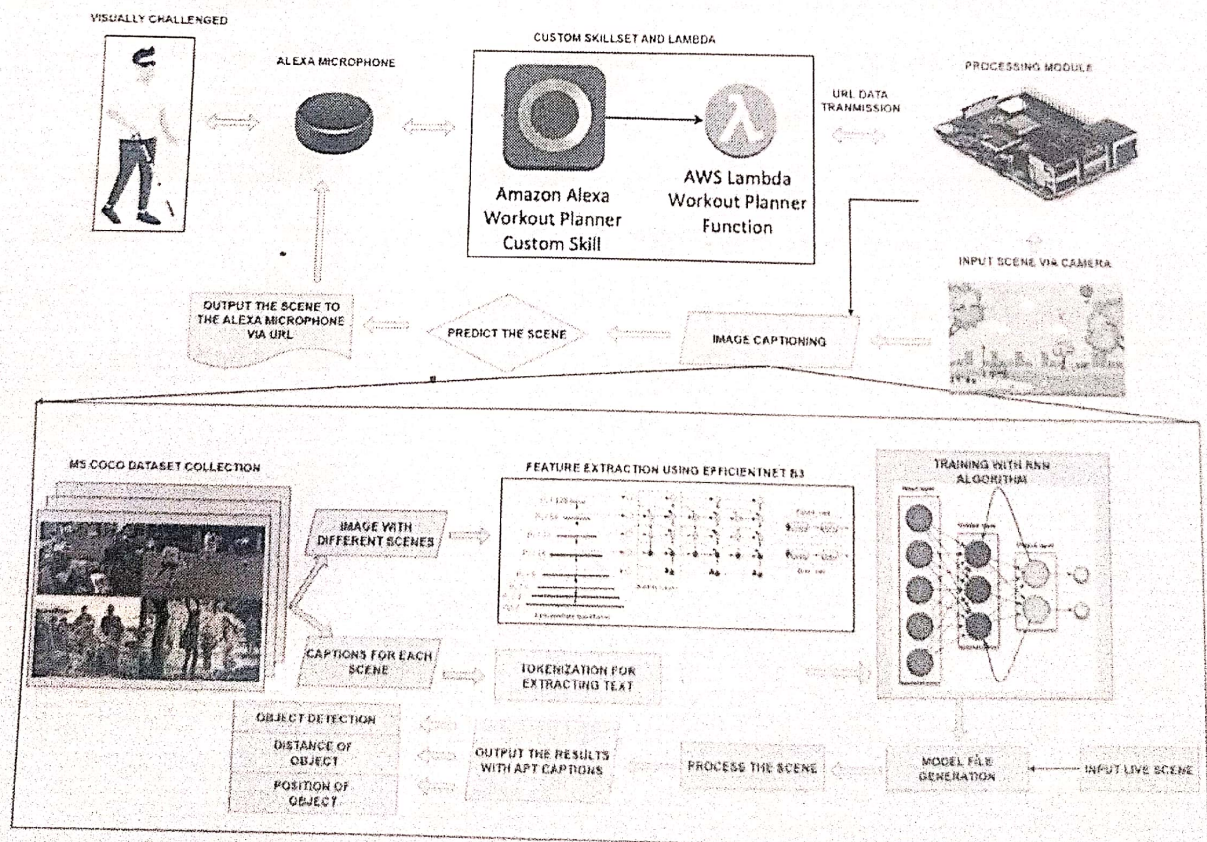


Figure 4.1 Proposed System Architecture

4.6 WORKING

In this project in order to facilitate the blind we have develop an artificial vision system with the help of deep learning techniques. Initially COCO dataset is collected to train the model, before training the model pre-processing of dataset is done so that the dataset can be directly applied to the deep Learning algorithms for further processes. After pre-processing feature extraction is done using efficientnet algorithm. Once the feature extraction is completed the model is trained using RNN algorithm. Recurrent neural networks (RNN) are the state-of-the-art algorithm for sequential data and are used by Apple's Siri and Google's voice search [10]. It is the first algorithm that remembers its input, due to an internal memory, which makes it perfectly suited for machine learning problems that involve sequential data.

It is one of the algorithms behind the scenes of the amazing achievements seen in deep learning over the past few years. After training the model it will be integrated with the Alexa device for developing its skills and recognizing the surrounding movements. Finally, whenever the user wants to know about the surrounding movements, they can ask Alexa to recognize what is happening around by saying any predefined keyword, so that Alexa can start recognizing the surrounding movements [9]. The major objects are also predicted and the distances calculated from the camera. This is achieved by using advanced image captioning techniques learned by the model. Whenever an image is captured via camera, the scenes are recognized and predicted by the machine. After the prediction, it is been sent to the Alexa microphone which will send an audio output to the user which can help them identify the distance and position of object. Thus, with the help of this project we provide an artificial vision to the blind, which can help them gain confidence while travelling alone.