Image Caption Generator for Visually Impaired Individual

A report submitted in fulfillment for the award of the degree of

Bachelor of Technology

in

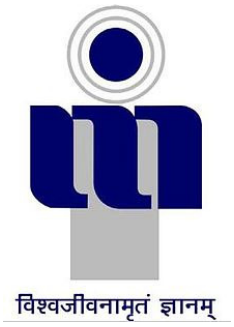
Computer Science and Engineering

By

Muskan Debnath : 2021BCS-046

Under the Supervision of

Prof. Shashikala Tapaswi



ABV-INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT GWALIOR

GWALIOR, INDIA

DECLARATION

I hereby certify that the work, which is being presented in the report/thesis, entitled IMAGE CAPTION GENERATOR FOR VISUALLY IMPAIRED INDIVIDUAL, in fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering and submitted to the institution is an authentic record of my/our own work carried out under the supervision of Prof. Shashikala Tapaswi. I also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

Dated: **Signature of the candidate**

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dated: **Signature of supervisor**

Acknowledgements

I would like to express my sincere thanks to all those people who made this seminar report possible. First and foremost, I would like to express my profound respect and gratitude to my supervisor, Prof. Shashikala Tapaswi, who has been guiding force behind this work. I am greatly indebted for his invaluable guidance, constant encouragement, and his valuable comments on my work. I am fortunate enough to have such an advisor who gave me the freedom to think independently and explore new ideas. More importantly, I would like to thank for the patience he has shown in carefully reading and commenting on the manuscripts, and countless revisions of this dissertation. His commitments and dedication to research have been and will continue to be a constant source of inspiration for me. I have no doubts that finishing my degree in proper and timely manner was impossible without his help. I am highly privilege to have got an opportunity to work with such a wonderful person.

Finally, I would like to thank the Almighty God for bestowing me this opportunity and showering his blessings on me to come out successful against all odds.

Muskan Debnath

Abstract

Image description generation is a problem of predicting an information rich description of input image. It is a very complex task that involves the field of Computer Vision along with Natual Language Processing. Previous researches have proposed the work on image captioning using three major methods namely, 1) template based method, 2) retrieval based method, and 3) encoder-decoder based method. Currently, encoder-decoder based architecture have gained popularity in this task as they not only have achieved good performance, but has the capability of generating descriptions like humans.

In our study, we will propose a neural network based architecture which will have good performance and will be easier to train too. We will conduct experiments on various encoder and decoder architectures. A complete validation and experimental analysis of all the suggested models for encoder and decoder will be given in this study. The proposed architecture will be trained and tested on the flickr8k dataset. Furthermore, the model will be deployed on the cloud and logic of generating the description for the blind person will be implemented on a device through a native mobile application.

Keywords: Image, Encoder-Decoder, Neural Network, CNN, RNN, LSTM, Caption, Image Description

Contents

List of Figures

List of Tables

List of Acronyms

CV	Computer Vision
NLP	Natural Language Processing
3D	Three Dimensional
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
WHO	World Health Organisation
LSTM	Long Short Term Memory
VGG	Visual Geometry Group
GRU	Gated Recurrent Unit
BLEU	BiLingual Evaluation Understudy
MSCOCO	Microsoft Common Objects in Context
DAA	Dual LSTM Adaptive Attention
METEOR	Metric for Evaluation of Translation with Explicit ORdering
AICRL	Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention
ResNet	Residual Network
ML	Machine Learning
API	Application Program Interface
TTS	Text-To-Speech
MOS	Mean Opinion Score
Seq2Seq	Sequence to Sequence

1

Introduction

1.1 Introduction

Image description generation has recently gained the interest of many researchers in the Computer Vision (CV) and Natural Language Processing (NLP) field to assist visually impaired individuals. Generation of the image's description can help the blind person view the world like a normal person. Advancements in computer vision can help blind people accomplish otherwise complex tasks like education, everyday chores, shopping, accessing the internet or gadgets, and many more.

Human beings have gained the ability to give a detailed description of the image after years of learning and experience. The machines can also learn to do the same by training on massive datasets. But, the problem of generating an image description is much more complex than detecting an object in the image or classification of the image. Image description generation involves detecting the essential objects in the given image and understanding the relationship between those detected objects. The sequence of words generated as a description should correctly capture the action and attributes. The captions generated for the image should be syntactically and semantically as correct as possible.

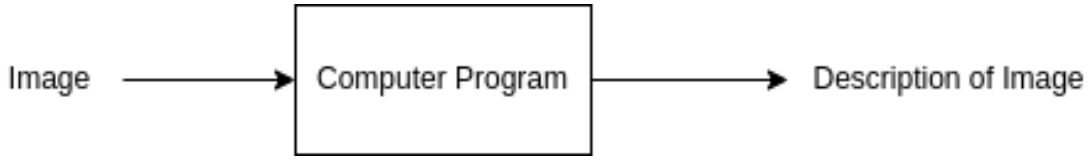


Figure 1.1: Block Diagram for basic idea behind image description generation

Many researches have been conducted previously to give the description of the image. The task of generating the description of the image can be classified into three broad categories:

(i) Retrieval based method

(ii) Template based method

(iii) Encoder-Decoder based method (Neural Network based method)

1.1.1 Retrieval based method

Figure 1.2: Block diagram of retrieval based method for image captioning [1]

As the name suggest, retrieval based method generate the description of the image by retrieving the images similar to the input image. It fetches the sentences from the pool of sentences used for describing the similar images. After we have the similar images and the description sentences of those images, it uses the similarity method to find the description of the input image. The relevant description can be one of those similar description or the combination of those similar description.

This method work on the belief that there always exist an similar image and certain set of description that can be used to generate the required description. It is not helpful in generating totally new caption or description.

1.1.2 Encoder-Decoder based method

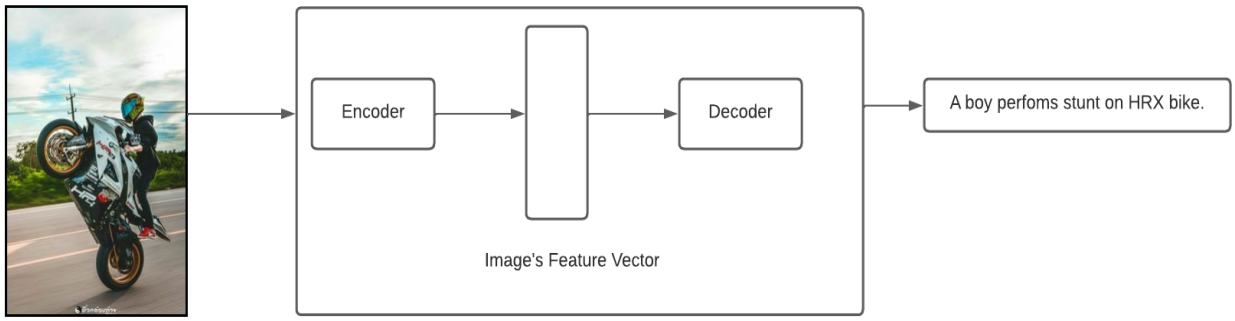


Figure 1.3: Block diagram of encoder-decoder based for image captioning [1]

This method is also known as Neural Network based method for generating the captions. As the name suggests, the method is made up of two parts encoder and decoder. Convolution Neural Network is used as an encoder where features are extracted from the input image, basically converting 3D input image into a feature vector. Recurrent Neural Network serves the purpose of the decoder by generating the sequence of words for the description of the image. The main challenge here is to figure out which CNN and RNN to use.

This neural network architecture has been adopted now to conduct majority of the image description generation task. It is successfully able to give a information rich description of the image because of its neural network like architecture.

1.2 Motivation

According to the report from WHO, more than 2.2 billion people have vision-related difficulties [6]. Conducting regular day-to-day chores like shopping, walking on the road and many more is extremely difficult for them, making them dependent on others. Now the world is becoming more accepting, as many researches are being conducted on developing vision-based technologies for assisting blind people in their daily activities. Image description generation can help individuals with visual difficulties to view the world like normal humans.

The task of generating image descriptions like humans is not easy. There are complexities involved in this task. The sentences should be grammatically correct, and the generated description should be rich in information. The proposed trained model should refrain from reporting the bias of overfitting the data. It should be general enough to give unique descriptions for the different images.

1.3 Report Organisation

The report consists of the following chapters with a brief knowledge about them:

(i) **Chapter 2** shows the background research related to the project problem, along with critical research for the project.

(ii) **Chapter 3** identifies the disadvantages and flaws in the existing methods and also identifies the problem statement and lists the objectives.

(iii) **Chapter 4** explains the methodology which is proposed to implement and overcome the shortcomings of the research gaps.

(iv) **Chapter 5** discusses the work done till now and the preliminary results obtained.

(v) **Chapter 6** discusses the conclusion and the future tasks left regarding the project.

2

Literature Review

This chapter presents the related work in the area of image description generation. The methodologies employed in diverse research are evaluated critically to determine the advantages and disadvantages of using a specific approach.

2.1 Literature Review

At present, the image description task is mainly focused on encoder-decoder-based or neural network-based architecture. Convolutional Neural Network (CNN) is used as an image encoder, representing the image as a feature vector. Recurrent Neural Network (RNN) is used as a decoder where the feature vector is taken as input, and the network generates a sequence of words as image descriptions. The model's neural network architecture helps obtain a more generalizable and variable-length description of the image. This method does not force any boundation on the number of predefined templates. It is also rich in information and hence, closer to human perception, which is not the case in template-based and retrieval-based image caption generation.

Vinyals *et al.*, the authors presented an end-to-end neural network architecture to generate the image description. This was one of the earliest encoder-decoder, single model, state of the art method to achieve the image description. The authors identified the similarity between the image description and machine translation task. The machine translation method needs to generate the T, the translated sentence when we are given with S, the source sentence or the input sentence. The objective of the machine language translation is to maximise $P(T|S)$, probability of the translated sentence given the source sentence. RNN is used as an encoder, which converts the source sentence S to a vector of a given length. Then that vector representation is fed as input into the RNN (decoder) to get the translated sentence T. They successfully replaced the encoder section of the above machine translation model with the Convolutional Neural Network, while decoder part of the network was implemented using the Long Short Term Memory (LSTM). Convolution Neural Network (CNN) has always been successful in representing the image as a feature

vector. The problems associated with the RNN was also addressed, as the context should be taken into consideration only for small time frame. Hence, they used LSTM which has a forget gate and learns to forget the previous words. They proposed their model as Neural Image Caption (NIC) [7]. The literature gap in this proposed method was that CNN only converted the input image into the single feature vector and does not emphasise on the spatial representation of the features of image.

In the paper [7], the entire feature vector produced as output of CNN is used to generate the word at the given time in LSTM decoder. Though, special attention should be laid on certain region of the image to generate the words at a time. Kelvin *et al.*, the authors used this important observation while proposing a new end-to-end neural network architecture with the visual attention mechanism on LSTM. This mechanism of adding attention forces the word to belong to a certain region. They achieved BLEU-1 score 67 on Flickr8k dataset and BLEU-1 score of 71.8 on MSCOCO dataset [8].

In attention based mechanism for image description generation, LSTM serves two purpose. First, LSTM is used redefine the visual representation of the feature matrix and also used as a sequence of words generation model. Xiao *et al.* proposed to use two LSTM along with adaptive attention separately to handle each of these task. Similar to the previous models discussed, CNN is used to obtain the feature representation of the image. One of the LSTM ($LSTM_A$) is used to get the visual representation matrix, a fine-grained hidden state information. This visual representation matrix and the hidden state of $LSTM_A$ is used to obtain the attention weights of the image (α_t), which will be the input of the second LSTM ($LSTM_B$). The hidden state of $LSTM_A$ and the updated attention weights from the adaptive attention module is fed as input to the $LSTM_B$ and then sequence of words are obtained as output at time t. They proposed their model as Dual LSTM with Adaptive Attention (DAA). This architecture was tested too on the most popular image captioning dataset: Flickr and MSCOCO. DAA mechanism obtained

BLEU-1 score of 75.8 on MSCOCO dataset and 68.6 on Flickr30k dataset. The same mechanism obtained METEOR score of 27.1 on MSCOCO dataset and score 21.5 on Flickr30k dataset [9].

In the paper [10], authors also proposed a neural network architecture for generating captions from the images. They used Flickr8k dataset with five description for each image to test the model. Similar to other models, they used Convolutional Neural Network as encoder and Long Short Term Memory model as decoder to generate the descriptions. Visual Group Geometry (VGG) which is available in 16 and 19 layers is used as a pre-trained model for the fetching the image feature representation vector. The probability of maximum likelihood is used to generate the sequence of words as captions further. BLEU score of 53.35 was achieved on the Flickr8k dataset.

Yan *et al.*, the authors proposed single AICRL model for automatic image captioning. The model uses ResNet50 architecture which is a Convolutional Neural Network based architecture as encoder. The encoder maps the three-dimensional representation of image to one-dimensional feature vector representation. The decoder part of the model utilizes LSTM plus soft attention which with each iteration pays attention to a certain region of the image and predicts the next word in the image description. The depth of the neural network plays significant role in the accuracy obtained by the model. ResNet architecture has much required depth and also it does not have too much parameters, making it a good Convolutional Neural Network to be selected as an encoder. They obtained the fine-tuned hyperparameters by extensively training the model. The model displayed significant good result on image caption generation. They achieved BLEU score of 61.1 on Flickr dataset and 73.1 on MSCOCO dataset [11].

In the paper [12], a comparative analysis of different encoder-decoder based architecture for image captioning was performed. In the paper [7] they used VGG architecture for

the encoder while in the paper [11] they used ResNet architecture for the encoder. The different architectures used in the encoder certainly results in different BLEU score. The complexities of the architecture which is number of parameters along with the accuracy of the model should be taken into consideration while selecting the pre-trained model. They performed experiment on various level of depth of ResNet, DenseNet and VGG encoder architecture. All the above architecture were also tested after applying the attention mechanism onto it. ResNet and DenseNet are low level complexity neural network and also gave good BLEU score on the Flickr and MSCOCO dataset while performing image captioning.

In the paper [1], also proposed an encoder-decoder based architecture for image captioning. They focused on increasing the accuracy of the model using the VGG16 Hybrid Places 1365 as encoder and LSTM as decoder. VGG16 Hybrid Places 1365 was used as encoder because this pre-trained model was trained on ImageNet and Places dataset both, giving 365 more classes than VGG16 CNN model. The proposed model outperformed the state-of-the-art architectures for image captioning. They achieved a BLEU score of 0.7350 on the MSCOCO dataset.

Table 2.1: Summary of Literature Review

S. No.	Title and Authors	Year	Key Takeaways
1	Show and Tell: A Neural Image Caption Generator [7] O. Viyals, A. Toshev, S. Bengio, and D. Erhan	2015	Simple encoder-decoder based architecture with CNN as encoder and LSTM as decoder.
2	Show Attend and Tell: Neural Image Caption with Visual Attention [8] Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Ben- gio	2015	Neural network architecture with attention mechanism to get the image description.
3	DAA [9] F. Xiao, X. Gong, Y. Zhang, Y. Shen, J. Li, and X. Gao	2019	Instead of using a single LSTM in the decoder component of the model, two LSTM along with adaptive attention was used to enhance the performance of the model.
4	Image Captioning Using Deep Learning [10] C. Amritkar and V. Jabade	2018	Instead of using a single LSTM in the decoder component of the model, two LSTM along with adaptive attention was used to enhance the performance of the model.
5	Automatic Image Captioning based on ResNet50 and LSTM with soft attention [11] Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang	2020	Model uses ResNet50 architecture as CN and decoder part of the model utilizes LSTM plus soft attention to predict next part of the image description.
6	Automatic Image Caption Generation Using Deep Learning A. Verma, A. K. Yadav, M. Kumar, and D. Yadav [1]	2022	They focused on increasing the accuracy of the model using the VGG16 Hybrid Places 1365 as encoder and LSTM as decoder.
7	Comparative Analysis of Encoder-Decoder based model S. Katiyar and S. Kumar [12]	2021	Compared VGG16, ResNet, DenseNet encoder architecture for image captioning technique

3

Statement of Problem based on Identified Research Gaps

This chapter covers the research gaps and disadvantages of the papers that were mentioned in the earlier chapter. This chapter also includes the problem statement that is formed to remove research gaps.

3.1 Research gaps

Some of the research gaps encountered after an extensive study of the previous literature are as follows:

(i) Not much improvement is done in the Convolutional Neural Network part of the model. Encoder part of the architecture can be improved by introducing the channel attention mechanism in CNN.

(ii) The language generation model uses sequence based neural network. Most of the proposed neural network based architecture uses LSTM [5]. LSTM overcomes the shortcoming of handling the long-term dependency efficiently, but it is much complex and requires more time to train to get the required weights.

(iii) In the above researches, there is uncertainty regarding the selection of the encoder for the neural network based architecture for image caption generation.

3.2 Problem Statement

It has been established that blind people find it difficult to conduct normal day-to-day activities. Many studies has already been conducted for developing an encoder-decoder based architecture for generating the description of the images. Majority of the previous architecture are either are complex or do not have good performance. There is a need of developing an neural network based architecture which is simple yet efficient in generating the description of the image. This study will try to experiment by combining various pre-trained encoders like VGG16, Inception V3 and ResNet50 with decoders for sequence generation such as GRU and LSTM with and without attention mechanism. LSTM model has been previously adopted by various studies but GRU needs experimentation as

it simpler and easier to train than LSTM. The study will focus on reducing the training time and automating the process of selecting the encoder based on the labels fetched from the object detection algorithm.

The problem is to generate the description of various images for assisting the blind people. There is a need of deploying the neural network on the cloud and implementing the logic on any device. The study will try to build a native mobile application for predicting the description of the images.

3.3 Thesis Objective

(i) To study and develop a solution for predicting the description of the images in real-time with good accuracy.

(ii) To further modify the existing models on image description generation with new modifications to reduce the complexity and training time of the proposed model.

(iii) To further automate the process of selecting the encoder for the neural network architecture to generate the description for the image.

(iv) To develop a native mobile application for fetching the description given the image.

4

System Overview and Methodology

This chapter covers the details about the framework of the project. This chapters also includes details about the dataset used.

4.1 System Design

4.1.1 Neural Network Architecture for Image Description Generation

The current study involves designing the architecture of the image description generation module using the encoder-decoder based method. Encoder will output the feature vector after taking image as the input. The feature vector will be given to the attention module and then a simple and efficient decoder will be selected to output the sequence of words as the image description. An overview of the proposed architecture is given in the figure 4.1.

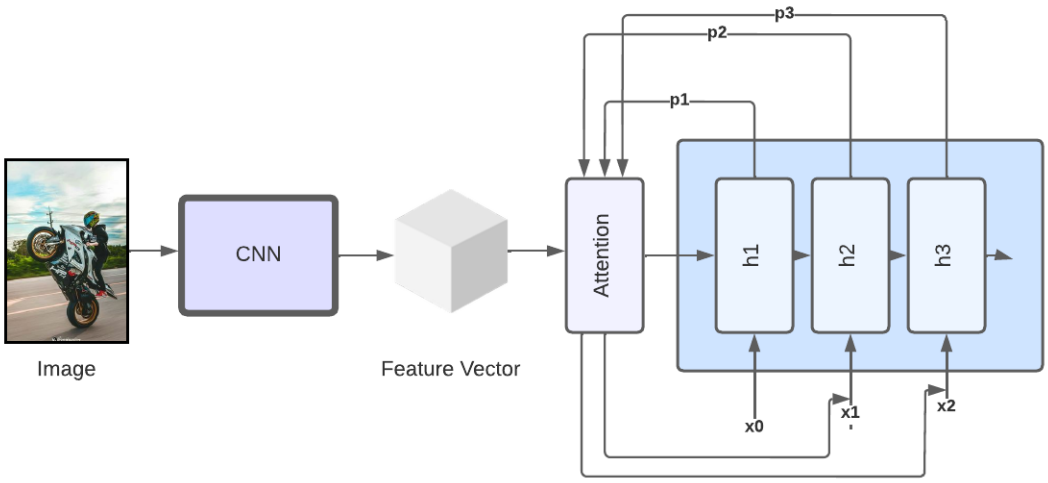


Figure 4.1: An overview of the proposed architecture of the image description generation module.

The encoder (CNN) in Figure 4.1 will return a feature vector of image which will be fed as input to the attention module. The attention module will compute attention weights for each iteration using the context and the feature vector. The attention weights will be multiplied with the feature vector and then will fed to sequence model (LSTM/GRU).

4.1.2 Automating Encoder Selection

There are various encoders that can be used to fetch the image vector representation. To deal with the uncertainty of the selection of encoder which gives the best result for image description generation. We will deploy all the models with various encoders and automate the selection of the encoder based on the matching labels obtained from object/label detection algorithm.

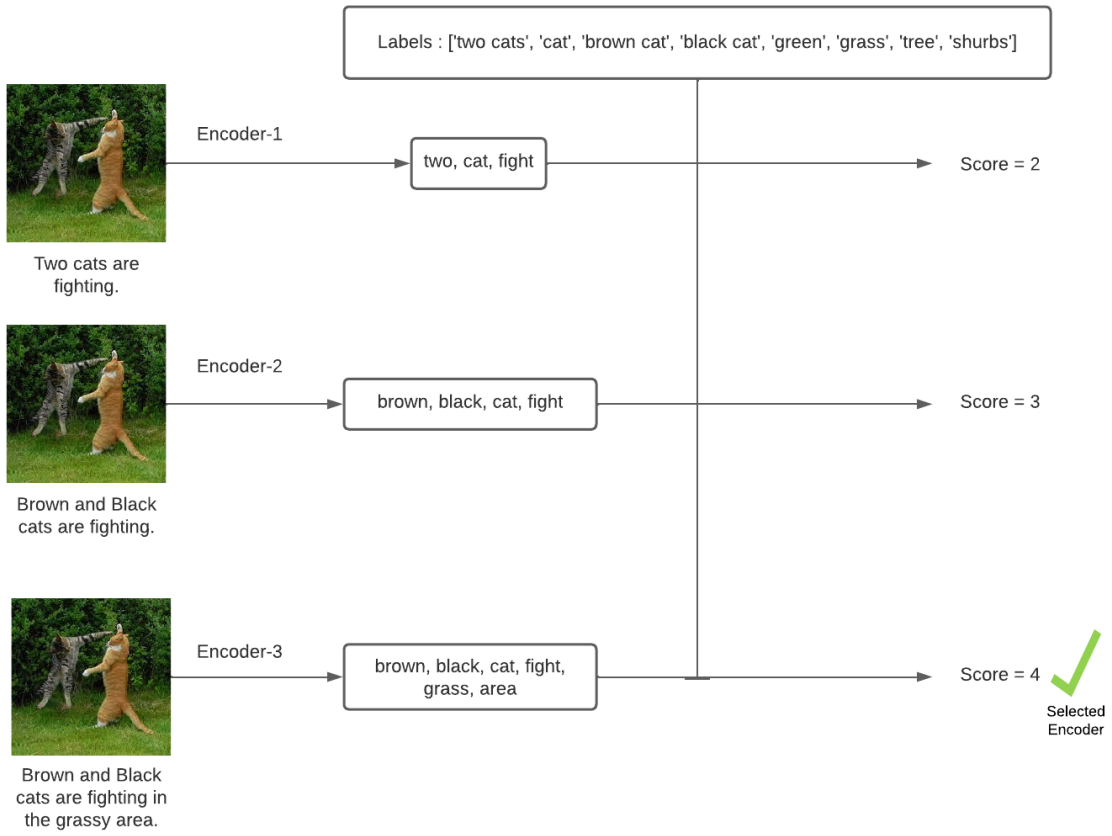


Figure 4.2: Automating the selection of different encoders based on the labels obtained from the object detection algorithm.

Figure 4.2 describes, the method for automating the object detection. The labels will be generated by the object detection and caption will be generated by the various encoder-decoder model. Root word will be extracted from the captions and the score will be calculated based on number of root word matching with the labels. The model with the best score will be selected for sending the image description.

4.1.3 Design for Web Application

After, we have our efficient and simple model for generating image description. We will deploy our model on the cloud platform and will build an mobile application that will help the blind individual to click a photo and get the audio of the image description generated.

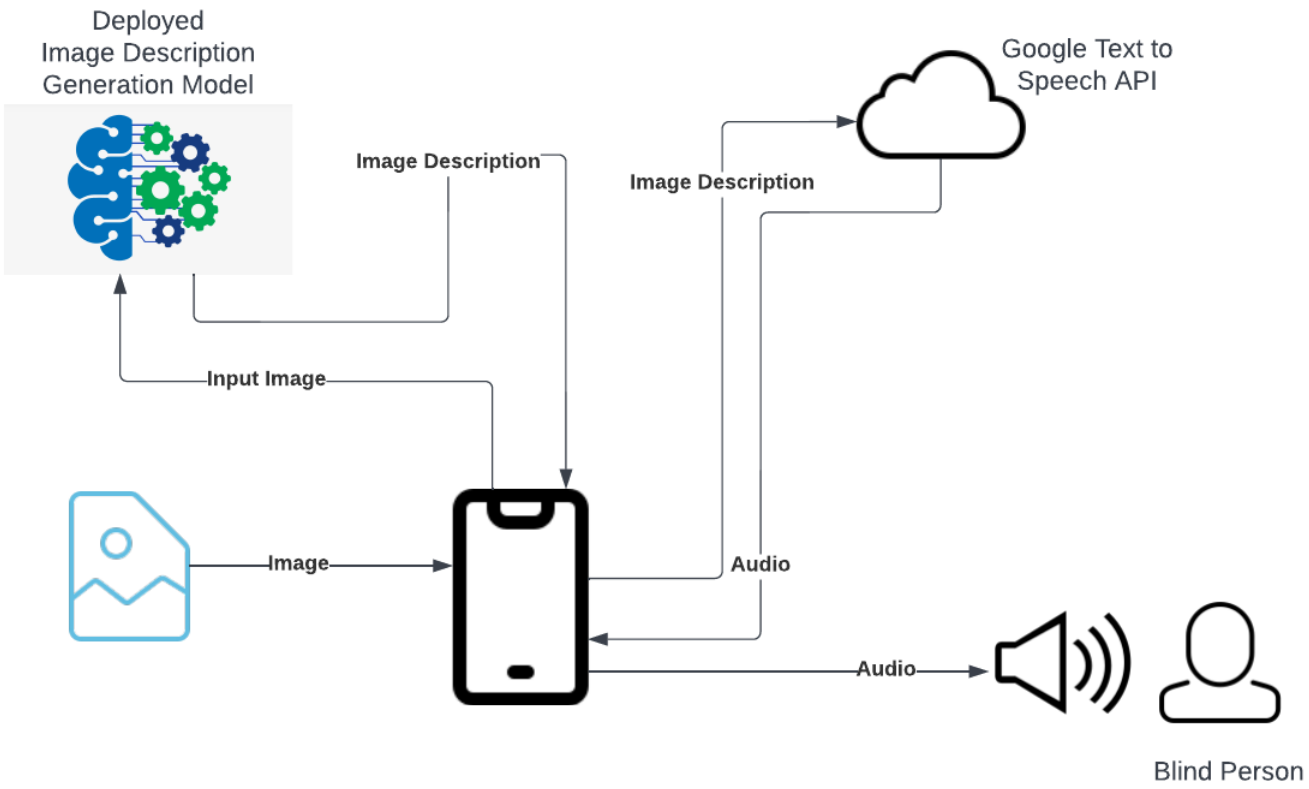


Figure 4.3: An overview of the proposed architecture of the image description generation module.

The flow of the logic of the system described in the Figure 4.3 is as follows:

(i) The input will be taken from the native mobile application.

(ii) The user will request for the description of the image from the deployed ML model on the cloud.

(iii) The model will return the description sentence to the client side of the application.

(iv) Google's Text to Speech API will be used to convert the image's description to the audio.

(v) The audio will be played for the blind person to tell them about the description of the image.

4.2 Proposed methodology

The objective of the present study is to generate image description with good performance. The task of generating the image caption can be divided into two parts:

(i) Feature Extraction

(ii) Description Generation

4.2.1 Feature Extraction

Most of the studies conducted on image captioning have proved that encoder-decoder based architecture gives a rich description of the image with high performance. The three dimensional image is given as input to the encoder part of the architecture and image's feature representation is extracted as output of the encoder. Convolutional Neural Network (CNN) serves the purpose of the encoder. Generally, CNN performs manipulations on the image using the Convolution layer and Pooling Layer. It effectively reduces the dimensionality of the image without losing the necessary information or features captured in the image. Many pre-trained Convolutional Neural Network based architecture (ResNet, VGG, Inception) are present, which can reduce the time of training the image description generation model.

The first task is to choose an efficient CNN model, considering the complexity and the performance of various CNN models.

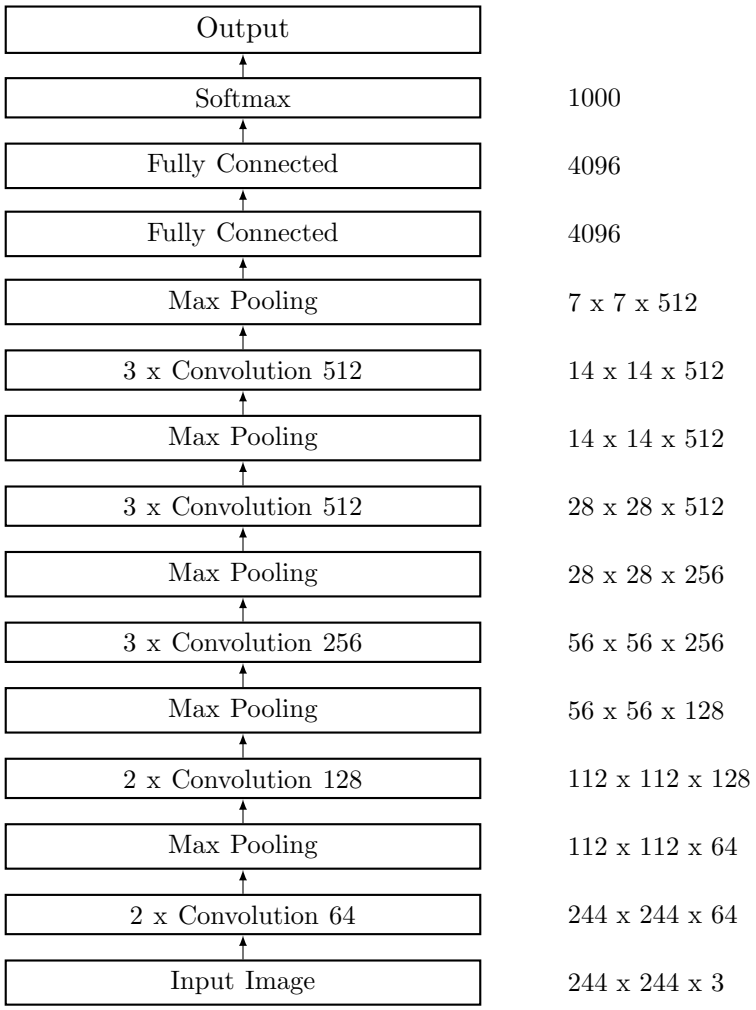


Figure 4.4: The architecture of the pre-trained CNN model, VGG16 [2]

4.2.1.1 VGG16

Visual Geometry Group (VGG) is a Convolutional Neural Network (CNN) architecture proposed by Karen *et al.* [2]. VGG16 has 16 number of layers where weights are learned while training the model. VGG19 is also a Convolutional Neural Network model available which has 19 layers with weights.

The VGG16 model takes a standard input image with size 244x244x3. It uses kernel filter of 3x3 size with stride 1. They use maxpool of size 2x2 with stride 2. The size of the kernel filter, maxpool filter and padding is constant throughout the model. The detailed architecture of VGG16 is given in the figure 4.4.

4.2.1.2 Inception

Inception V3 is also a CNN pre-trained model published in 2015 [3]. It performed better than its subsequent parts like V1 and V2. There were many major advancements that led to increase in the performance of Inception V3. It factorized the convolutions into many smaller convolutions, a convolution was spillted to many asymmetric convolution to reduce the computation. In total, the Inception V3 consisted of 42 layers. A constrained view of the Inception model is given in the figure 4.5.

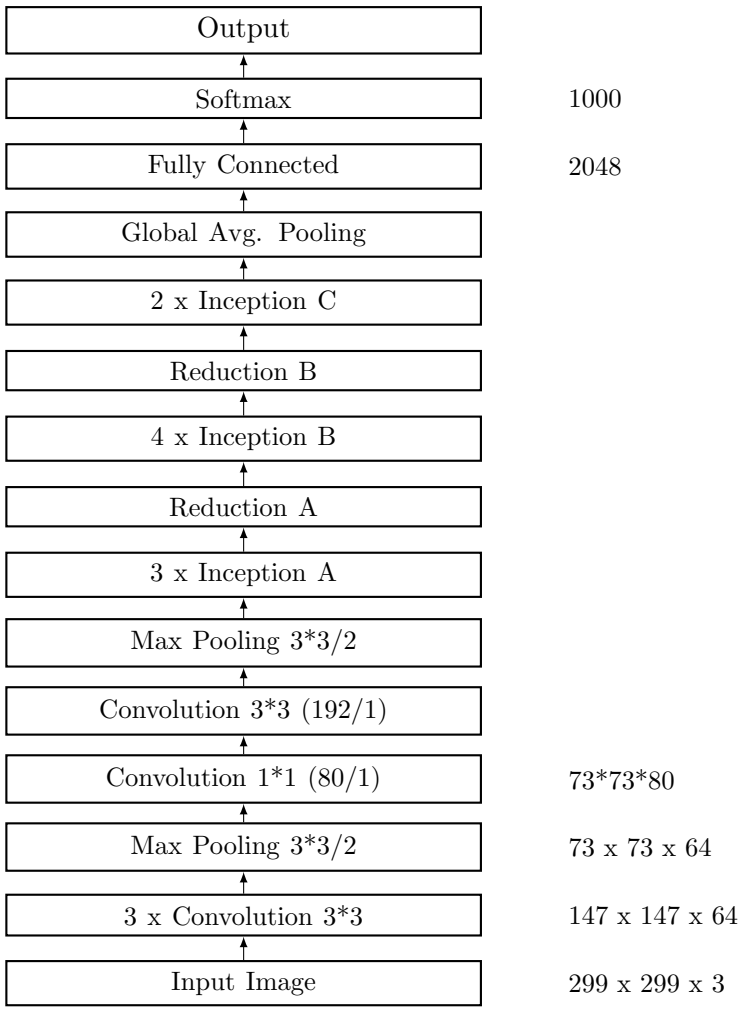


Figure 4.5: The architecture of the pre-trained CNN model, Inception V3 [3]

4.2.1.3 ResNet50

The ResNet architecture was proposed to solve the problem of the vanishing gradient commonly seen in the CNN models. The problem was resolved using the skip connections between the different layers [4]. ResNet50 architecture was 50 layers deep and had fewer parameters than the VGG16 and Inception V3 architecture. Though, the model will be difficult to train because of the increase in the depth but will certainly have better performance. The architecture of ResNet50 is given in the figure 4.6.

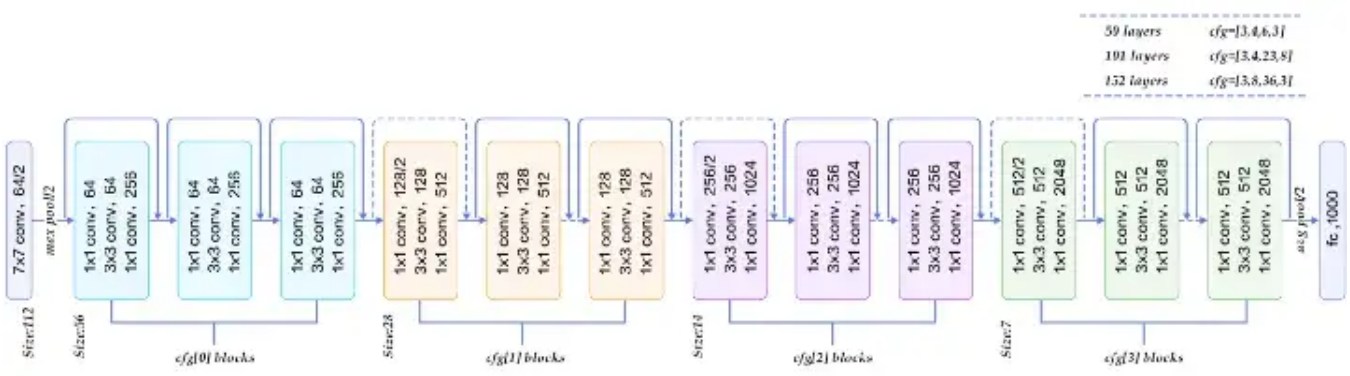


Figure 4.6: The architecture of the pre-trained CNN model, ResNet50 [4]

4.2.1.4 Feature Vector Representation of Image

Consider, image I is fed into any selected convolutional neural network model. The output of the encoder will be the visual representation of the features of the image. Let's say F is the feature vector received as output after fully connected layer of CNN and V is the spatial vector received as output from the last convolution layer of CNN.

$$F = C_{fc}(I) \tag{4.1}$$

$$V = C_{conv}(I) \tag{4.2}$$

Here, C_{fc} is the output of the fully connected later of CNN and C_{conv} is the output of the last convolution later of CNN.

$$V = \{v_1, v_2, ..., v_{k^2}\} \tag{4.3}$$

The output of the last convolution layer V is the visual k x k grid which can map to the previous convolution layer outputs and can perfectly define the relative feature positions. The fully connected layer output F will be given as input to the decoder or language model in model where no attention mechanism is adopted, while the F will be the input to the attention module in the attention mechanism model of image description generation.

4.2.2 Description Generation

The task of image description generation deals with the generation of sequence of words. The idea of generation of sequence of words make us use the Recurrent Neural Network (RNN).

Consider I is the input image and θ is the model parameter to be trained. The aim of the model is to generate sentence S by maximising the likelihood of the expression.

$$\theta^* = \arg\max_{\theta} \sum (I, S) \log p(S|I; \theta) \tag{4.4}$$

$$h_{t+1} = f(h_t, x_t) \tag{4.5}$$

The generation of the next hidden state (h_{t+1}) is a non-linear function (f) of the previous hidden state and the context word (h_t). The naive RNN network fails to deal with the long sequence of words because of the exploding gradients words get vanished with time. Many researches conducted on image caption generation adopted LSTM as the language generation model as it solves the problem discussed aforementioned.

In our study we will be experimenting with both LSTM and GRU model. GRU is simpler RNN model with not only solves the above mentioned problem but also requires lesser parameter to train. It is an effort to reduce the complexity of the decoder section of the architecture.

4.2.2.1 LSTM - Long Short-Term Memory

LSTM is a variation of the recurrent neural network. It consist of three gates: input gate, output gate and the forget gate. The value of previous hidden state, h_{t-1} and the current

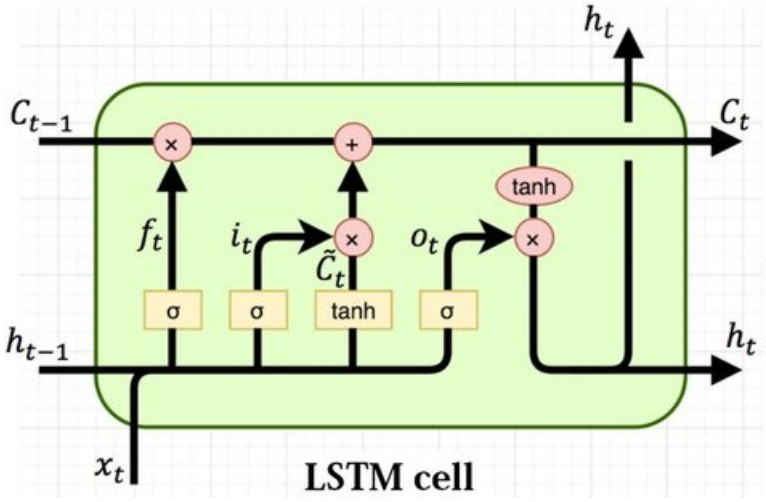


Figure 4.7: The basic architecture of the repeating module of LSTM [5]

element of the sentence x_t is passed through the sigmoid function.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{4.6}$$

Now, we need to decide through the input layer, i_t which values needs to be updated and set of all the possible candidates that could be the next potential C_t.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{4.7}$$

$$\tilde{C_t} = \tanh(W_C.[h_{t-1}, x_t] + b_c) \tag{4.8}$$

C_{t-1} will be updated to C_t using the formula given below.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C_t}$$

The output of the hidden state of this iteration will be decided using the sigmoid and the hyperbolic function.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{4.9}$$

$$h_t = o_t * \tanh(C_t) \tag{4.10}$$

The Ct and h_t will be fed as input to the LSTM at $t + 1$ iteration. At every iteration, the h_t given to the softmax function to get the distribution of probability of every word

from the dataset.

4.2.3 Visual Attention

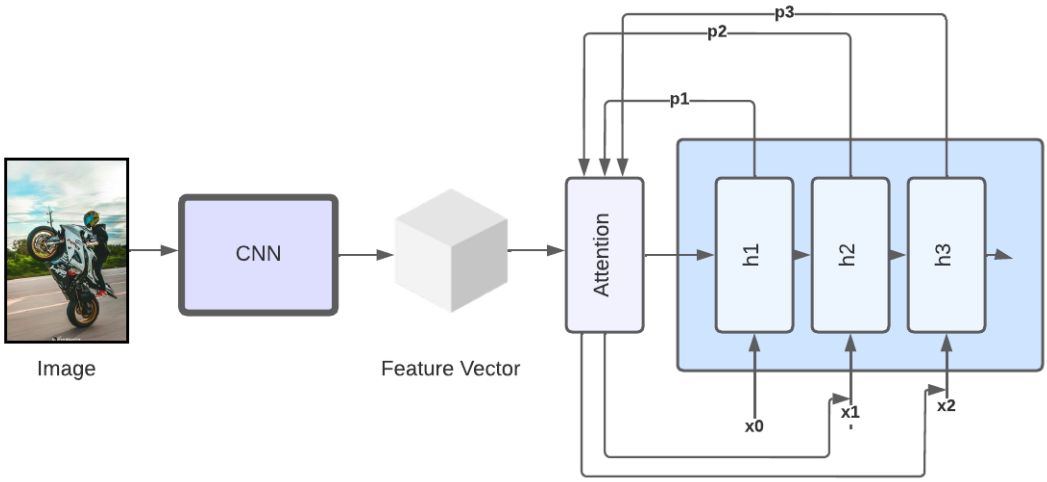


Figure 4.8: Encoder-Decoder based architecture for image description generation using attention module

The idea of the attention mechanism comes from the attention used for the image classification, as all of the pixels are of no use to classify the image. In attention-based image description generation, the output of the convolutional fully connected layer fed into the attention module. With attention gate, the decoder concentrate on only those pixel of the image that are relevant. In each iteration, the feature vector representation of the image and previous context of the LSTM is given to the attention module. The attention module then calculates the weight of the attention vector. The attention mechanism is differentiable. The training of the attention mechanism can be done using the backpropagation.

$$a_t = \sum_{1}^{n} S_j Y_j \tag{4.11}$$

Here, a_t is the attention vector obtained after training. S is the function that gives the output after performing the softmax operation. Y is the feature representation of the image.

4.3 Evaluation metrics

We evaluate our model using the BLEU score, which measures how well the predicted description aligns with the reference description. BLEU calculates n-gram overlaps between the predicted and true descriptions without considering structural accuracy [13]. Originally designed for machine translation evaluation, a valid model should achieve a BLEU score above 0.5.

4.4 Dataset

Flickr8k dataset will be used to evaluate the performance of the image description generation and also for training the model. The dataset consist of 8000 images which can be splitted in the ratio of 6:1:1 for training, validation and testing purpose respectively. In the dataset, for each and every image there are five descriptions. **MSCOCO** dataset is also similar with respect to the structure, it also contains five descriptions for each image. Total number of images in the dataset is around 164K. The training set can have 84K images, validation set and testing set can contain 40K images each [14]. Due, to hardware constraint we will restrict our training with Flickr dataset.

4.5 Text-To-Speech (TTS) Conversion

In our study, the description obtained from the neural network model will be converted to the waveforms, audio for assisting the blind person about the image. We will be using Google's Text-To-Speech API for performing the mentioned task.

The API works on the Tacotron 2 [15], proposed by the engineers of Google. It is a neural network architecture for generating the audio version of the text (speech synthesis). The neural network architecture works on Seq2Seq model, where inputs are the sequence of word embeddings and output is mel-scale spectrograms, which is sequence of waveforms for audio. The architecture comprises of two major components, (i) RNN Seq2Seq Model with attention module which takes input as the sequence of words and produces mel spec-

trogram sequence, and (ii) WaveNet [16] which takes predicted mel spectrogram frames as input and convert it to waveforms for raw audio which is much more closer to the human speech. This model achieved mean opinion score(MOS) of 4.53 much closer to professional human speech.

5

Results and Discussions

This chapter details about the work done till now and the results obtained.

5.1 Activities Completed

(i) Filtered recent research papers on image description generation using various architecture. From the literature review, found out the literature gaps in recent researches and generated the problem statement.

(ii) Completed the image description generation task using Flicker8k dataset with various pre-trained encoders: VGG16, Inception V3 and ResNet50 combined with RNN decoders: LSTM with and without attention.

(iii) Selected LSTM as the decoder based on the accuracy of the model.

(iv) Built the native web application that takes the input as image and predict the description. Converted the generated text description to audio using the Google Text-To-Speech API.

5.2 Results

5.2.1 Encoder-Decoder Image Description Model

In our study we adopted encoder-decoder based architecture for generating the image description. We performed an experimental analysis by combining various pre-trained CNN models (encoders) : VGG16, Inception V3, ResNet50, with various sequence generating models (decoders): LSTM with and without attention mechanism.

The Table 5.1 gives detail of the BLEU-1, BLEU-2 score and Table **??** gives detail of the METEOR and GLEU score of above mentioned architectures without attention while training on the Flickr8k dataset. The Table 5.2 gives detail about the time required for training with various model for 20 epochs and 32 batches without attention.

Table 5.1: BLEU Score for image description model with various encoder and decoder without attention on Flickr8k dataset.

S.No.	Encoder	Decoder	Attention	BLEU-1 Score
1	VGG16	LSTM	No	0.536422
2	Inception V3	LSTM	No	**0.565980**
3	ResNet50	LSTM	No	0.558754

Table 5.2: Training time in min(s) for image description model with various encoder and decoder without attention on Flickr8k dataset.

S.No.	Encoder	Decoder	Training Time (min(s))
1	VGG16	LSTM	94
2	Inception V3	LSTM	92
3	ResNet50	LSTM	92

To deal with the uncertainty between different encoders and improving the accuracy of the image description, we automated the process of selection of encoders. For each image, the encoder with the maximum matching root words with the labels obtained from label detection algorithm will be considered best.

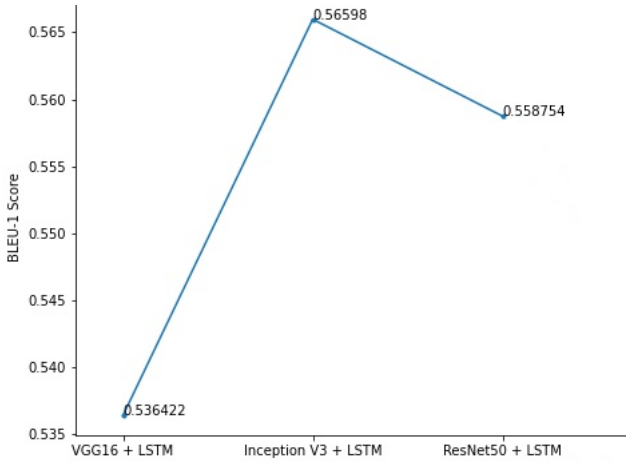


Figure 5.1: Graphical comparision of BLEU score of different encoder-decoder models without attention on Flickr8k dataset.

After experimenting with the automated neural network architecture, we achieved a good BLEU-1 score of 0.698747.

A web interface was developed to visualize the output of our model. Users can upload images and receive real-time captions. The interface integrates encoder selection and the

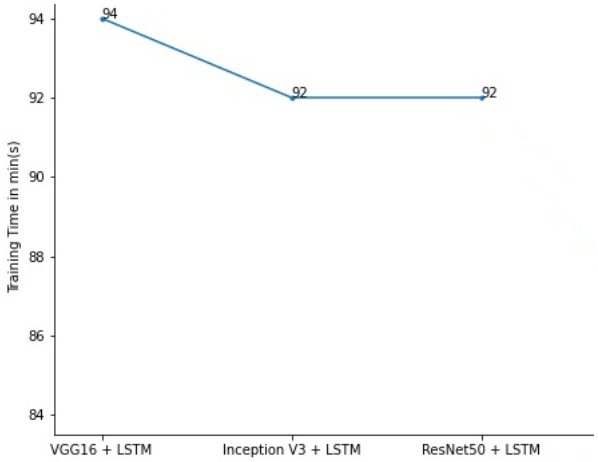


Figure 5.2: Graphical comparision for training time of different encoder-decoder models without attention on Flickr8k dataset.

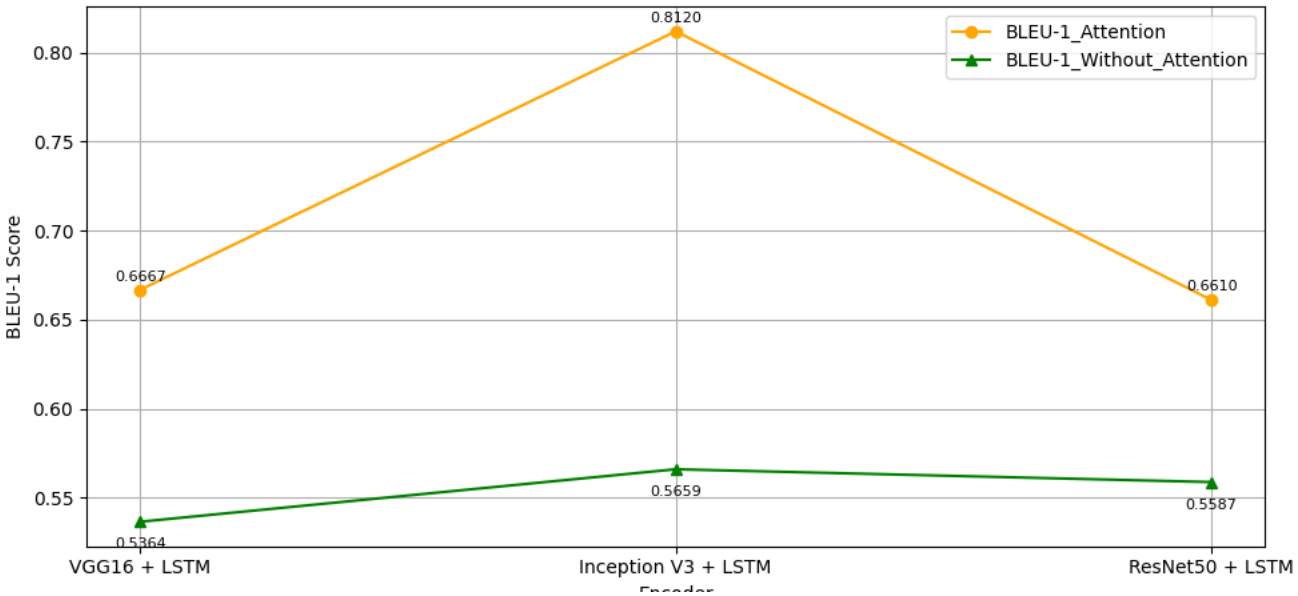


Figure 5.3: Graphical comparision of BLEU-1 score of different encoder-decoder models with and without attention on Flickr8k dataset.

captioning pipeline. A sample output is shown below.

```python
# Initialize lists to store actual and predicted captions
actual_captions_list = []
predicted_captions_list = []

# Loop through the test data
for key in tqdm(test):
    # Get actual captions for the current image
    actual_captions = image_to_captions_mapping[key]
    # Predict the caption for the image using the model
    predicted_caption = predict_caption(model, loaded_features[key], tokenizer, max_caption_length)

    # Split actual captions into words
    actual_captions_words = [caption.split() for caption in actual_captions]
    # Split predicted caption into words
    predicted_caption_words = predicted_caption.split()

    # Append to the lists
    actual_captions_list.append(actual_captions_words)
    predicted_captions_list.append(predicted_caption_words)
```

```
[ ] print("BLEU-1: %f" % corpus_bleu(actual_captions_list, predicted_captions_list, weights=(1.0, 0, 0, 0)))

100% [=======================================>] 810/810 [25:39<00:00, 1.53s/it]
BLEU-1: 0.698747
```

Figure 5.4: BLUE score 0f the final model

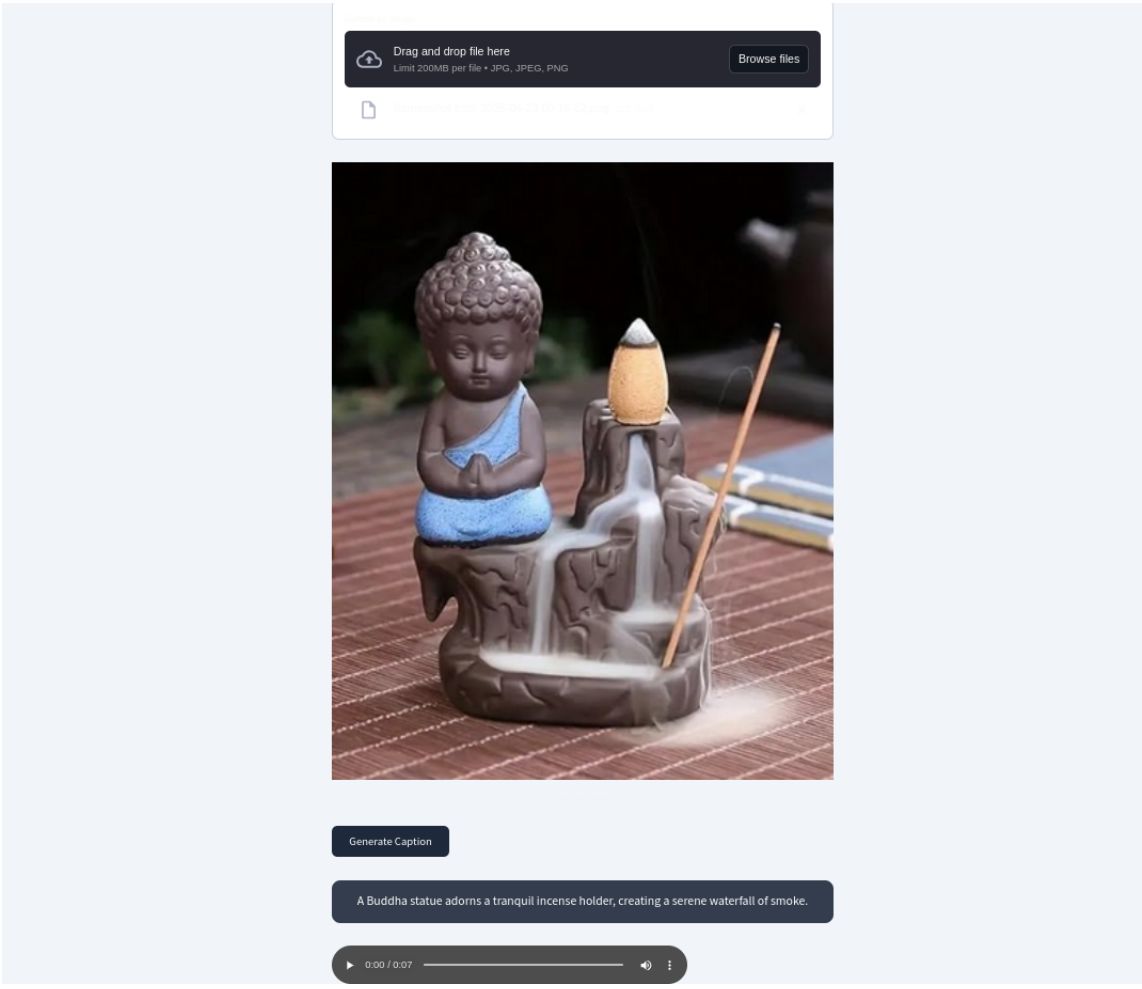


Figure 5.5: Screenshot of the image captioning web application

6

Future Work

This chapter details the conclusion of the project and the future scope of the thesis.

6.1 Conclusion

The current study focused on proposing an encoder-decoder based architecture with attention for image description generation task. It also helped in automating the selection of the best model based on the labels obtained from object-detection algorithm. Considering the training complexity we selected LSTM as decoder, and then automated the selection of encoder based on the matching labels and root word of the caption generated, to increase the accuracy of the generated caption. Attention module will be embedded between the encoding and decoding layer to enhance the performance. A web application will be built which will help in assisting the blind person by generating the description of the image. A complete experimental analysis has been given for the architecture proposed for the image description generation.

6.2 Future Work

A complete analysis of previous methods and approaches used for image description generation has been done in the literature review process.

(i) Training the final model on Flickr30k, MS-COCO dataset to increase the performance of the model further.

(ii) Selecting a better way to score the performance of the encoders for automating the selection of encoders for image description generation.

(iii) Optimizing the result fetch time for generating the image description on the image.

Bibliography

[1] A. Verma, A. K. Yadav, M. Kumar, and D. Yadav, "Automatic Image Caption Generation Using Deep Learning," *Springer Nature*, pp. 228–231, 07 2022.

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks For Large-Scale Image Recognition," *Visual Geometry Group, Department of Engineering Science, University of Oxford*, 2015.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Computer Vision and Pattern Recognition*, 2015.

[5] Hochreiter, Sepp, and J. Schmidhuber, "Long Short-term Memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[6] "World health organisation: Blindness and vision impairment report," https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

[7] O. Viyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[8] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *International Conference on Machine Learning*, 2015.

[9] F. Xiao, X. Gong, Y. Zhang, Y. Shen, J. Li, and X. Gao, "DAA: Dual LSTMs with Adaptive Attention for Image Captioning," *Neurocomputing*, 2019.

[10] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning," *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation*, 2018.

[11] Y. Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang, "Automatic Image Captioning based on ResNet50 and LSTM with soft attention," *Wireless Communications and Mobile Computing*, 2020.

[12] S. Katiyar and S. Kumar, "Comparitive Evaluation of CNN Architectures for Image Caption Generation," *International Journal of Advance Computer Science and Applications*, 2021.

[13] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," 10 2002.

[14] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," https://cocodataset.org/.

[15] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis By Conditioning WaveNet On Mel Spectogram Predictions," *Google, Inc., University of California, Berkeley*, pp. 228–231, 07 2018.

[16] A. van den, O. Sander, D. H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *Google DeepMind, London, UK*, pp. 228–231, 07 2016.