SUNSET RELATED IMAGE CAPTIONING IN BENGALI WITH DEEP LEARNING

BY

Md. Jamiul Haque
Roll: 161-15-7538
AND
Md. Firoj Islam
Roll: 162-15-7825

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By
Ms. Israt Ferdous
Lecturer
Daffodil International University

Co-Supervised By

Mr. Sheikh Abujar
Sr. Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

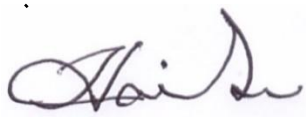
JULY 2020

APPROVAL

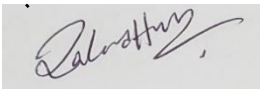
This Project titled "**Sunset related image captioning in Bengali**", submitted by Md. Jamiul Haque, ID: 161-15-7538 and Md. Firoj Islam, ID: 162-15-7825 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 09-07-2020.

BOARD OF EXAMINERS

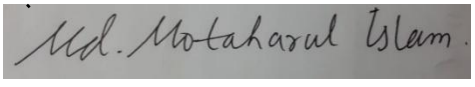
Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Associate professor &Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Md. Zahid Hasan — **Internal Examiner**
Professor
Department of Computer Science and Engineering
Daffodil International University



Dr. Md. Motaharul Islam — **External Examiner**
Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Ms. Israt Ferdous, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

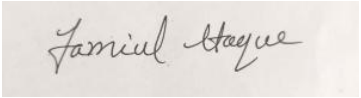
Supervised by:

Ms. Israt Ferdous
Lecturer
Department of CSE
Daffodil International University

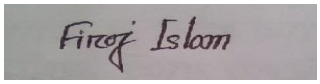
Co-Supervised by:

Mr. Sheikh Abujar
Sr. Lecturer
Department of CSE
Daffodil International University

Submitted by:



(Md. Jamiul Haque)
ID: 161-15-7538
Department of CSE
Daffodil International University



(Md. Firoj Islam)
ID: 162-15-7825
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Ms. Israt Ferdous**, **Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Deep Learning*" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. Syed Akther Hossain Professor and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Nowadays image captioning is a large area for research because of its social and commercial usage. To understand a picture, it is necessary to understand its features first. For this work, we have used a Convolutional Neural Network (CNN) which has trained images to generate words from its features. Then the words are arranged to form a sentence. The activation level of the CNN serves as an input for the Recurrent Neural Network (RNN) and generates a complete caption. These networks sequentially behave like an encoder and decoder. In existing work, the data used for this case study were not adequate and had a lack of different types of data. They don't have multiple captions. In this paper, we have introduced sunset related image-captioning methods in the Bengali language based on deep learning. To achieve better results, we have proposed a model, merged with the LSTM layer and the second last layer of the VGG16 model with a dense layer. We have achieved 78.26% accuracy with our proposed model for Sunset related image captioning in the Bengali language.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Artificial intelligence is a rapidly emerging sector in today's world. It deals with many common difficulties in our daily life. Of them, an artificial intelligence difficulty is generating captions where a textual statement ought to be generated for a given image or a photo. For describing a photo in internet or search it by image, we need to give it a good caption as an intermediate or final step. Basically, captioning an image involves generating a human-readable descriptive text for a given photo. In English, it is already an enriched section. But for Bengalis, it is still quite an ignored sector, which creates searching and other related difficulties for the people using Bangla languages.

Describing the feature observed in the given image is called image captioning. This way can be mentioned as a type of multi-class image classification and there lie a very huge number of classes. Through the use of machine learning techniques, image analysis can possible in the form of vectors. The training words are used by the machine learning methods to apply captions automatically for newly generated photographs. The initial steps subsisted between the features of images and training captions, which is covered the correlations. To achieve the sleek translation of the description-based vocabulary, these techniques were resolutely created and used by the machine-based translations. To caption an image is a simple matter for peoples, but it is very difficult for a machine because it involves the dataset will be in the (image as an input to the generated captions) appearance. Dataset consists of images that will use as input, the content of an image and how to translate into regular language. For captioning the photographs, we will need a massive number of datasets and their analogous output captions. The captions must be based on its topicality and put it into the system and just need to choose appropriate images. Fig. 1.1 show the scenario of our given data that is captioned by human.

(a) দুইসন্তানকেকোলেনিয়েছাতামাথায়একবাবাপাশেছাতামাথায়একট্র্যাফিকপুলিশ (A father has two children on his lap holding an umbrella and a policeman is standing beside him)

(b) পানিরবোতলহাতেদুইজনশিশুগ্রামেররাস্তাদিয়েহেটেযাচ্ছে (Two children walking down the village street holding water bottle)

Fig. 1.0.1: Example of Human generated caption

Bengali Image Captioning will be a great achievement for Bangladeshi peoples because it can be automatically used for indexing images. Also, it can be used in the field of military, biomedicine education, image searching, digital libraries, web searching, etc. In deep learning-based methods, features are trained automatically from training data. They can manage a big and different set of images.

1.1 Need of Deep Learning for Image Captioning

Image features are very important for understanding an image. The techniques used for this motive can be divided into two sections: **(a)** Deep machine learning, **(b)** Traditional machine learning [10]. Recently, deep learning techniques got state-of-the-art results on these types of problems. Research has been done for many years with image captioning. But the relevant techniques were limited for this and they weren't strong enough to manage the real world. The major problem is the limitations of heuristics or approximations for word-object relationships. To improve deep learning, in 2014 several high-profile AI labs began to release new approaches.

In the present world, deep learning is a very powerful area with so many applications. One of them is Image Captioning. It is a popular research field of Artificial Intelligence.

The advantage of deep learning is that it uses effective unsupervised or semi-supervised feature learning and layered feature extraction instead of man-powered feature extraction. The aim of feature learning is to seek better representations of data and to create a better model to learn these representations from large- scale unlabeled datasets.

Deep learning is a contemporary information science method capable of dealing with data effectively. The traditional machine learning technique needs domain expertise in many cases in order to reduce the complexity of the data. On the other hand, Deep Learning techniques try to learn high- level features from data in an incremental way.

1.2 Problem Statement

An image captioning consists of the process of generating textual statements from a given image which is based on the actions and objects in the image. The challenging problem in artificial intelligence demands both visual and linguistic understanding. In this paper, we have proposed deep learning-based image captioning.

In the field of Bengali Image Captioning, captions are generated by mapping to image. In this case, the caption that the trained model generates is limited to human-generated captions. But in our point of view, when the machine is able to generate a caption from the vocabulary list of words, it would be a better approach. In the previous approach, the memory requirement was high. But it is possible to use less memory in our approach.

1.3 Motivation

In our everyday life, we get acquainted with various types of images of different places from various sources. Among them, social media and articles are the most popular sources in this area. There are lots of images in those sources that have no explanation. The viewers have to understand the meaning of those particular images without any explanation by themselves. And in many cases, humans can understand without their detailed explanation. But it is the period of the modern era so the machine needs to illustrate the image captions if humans want an automatic image captioning.

In previous work, [1] the datasets they have been used are not enough for this case study. They also required more memory. The accuracy level of generated captions is also not so good. So, we want to work on this case study to generate more accurate captions for any image with more datasets. And this will require less memory. The majority of the images have subject looking at the camera so the model sometimes cannot generate proper caption. So, we decided to collect data from all angles.

1.4 Importance of Image Captioning

Through image captioning, people can easily understand the story. Image captioning is important in many ways. The method of mapping images to natural languages and contrarily is an effective way in this area. For example-

- **Medical Science:** Image captioning is useful in the field of medical science. By looking at image captions in the medical field, physicians can make the diagnosis easier. Due to automatic detection preventing disease can be more accurate and easier.

- **Indexing:** Useful in an automatic image indexing. It is important for Content-Based Image Retrieval (CBIR) [10]. It can be used in commerce, education, biomedicine, military, web searching, digital libraries, etc.

- **Self-driving cars:** Self-driving car is one of the ambitious challenges in the real world. By the blessings of image captioning, we can accurately caption the whole scenario around the vehicle. That gives collaboration to the autonomous car.

- **CCTV:** Nowadays CCTV cameras are in most of the places. And it usually contains several clips of our day to day working life so that if there is any suspicious activity held on, all previous history can be checked. In this sense, if we can generate contextual captions of those screenplays, then we could ring an alarm as soon as possible.

- **Helps the blind:** Image captioning is beneficial for the blind. By converting the scenario into the sequence of words and then to voice we can guide the blind people while traveling. In recent days both applications are famous in the field of deep learning.

- **Social Media Platforms:** Social media platforms can predict the captions of the images from the scene of what you wear, where you are at this moment.

- **Searching area:** Image captioning can help to make ‘Google Image Search’ as good as the popular ‘Google Search’. To do this each image needs to be converted into a meaningful caption and then the search can be accomplished based on the generated caption.

1.5 Background

The modern world is an area of research. It is a period of technology. Real-world has a remarkable proficiency in the area of deep learning-based image captioning. There are several works has been done on this case study. So, we can easily say that this field of research is quite popular in the present day. It is a good approach to provide text for any kind of image that arrives on the page so that an image can be heard or read as averse to just seen. [2] In previous work, dense captioning is very much popular in this sector. Fully Convolutional Localization Network is accomplished for addressing the text of a given image and also for localization. Creating a caption of an image by predicting objects is a very popular case study in the area of artificial intelligence. An image can be described by using detecting words, generating sentences and re-ranking those sentences through models and detectors. There are several approaches has been proposed in this area which gives us good result with better accuracy. But still, image captioning in the Bengali language did not get famous as the English language did. Only one case study has been done on image captioning in Bengali [1] which is very useful for future work like ours. To generate captions there are several networks have been used. Most of the cases Convolution Neural Network (CNN) acts as an encoder. On the other hand, Recurrent Neural Network (RNN) acts as a decoder.

1.6 Flow of the Research

Research activities have been carried out in several steps. At first, an extensive literature review was performed. Then the proposed idea and methodology for image captioning are described with the implementation of our model. At last, the proposed model was experimented. Figure 1.2 illustrates the flow of the research activities.

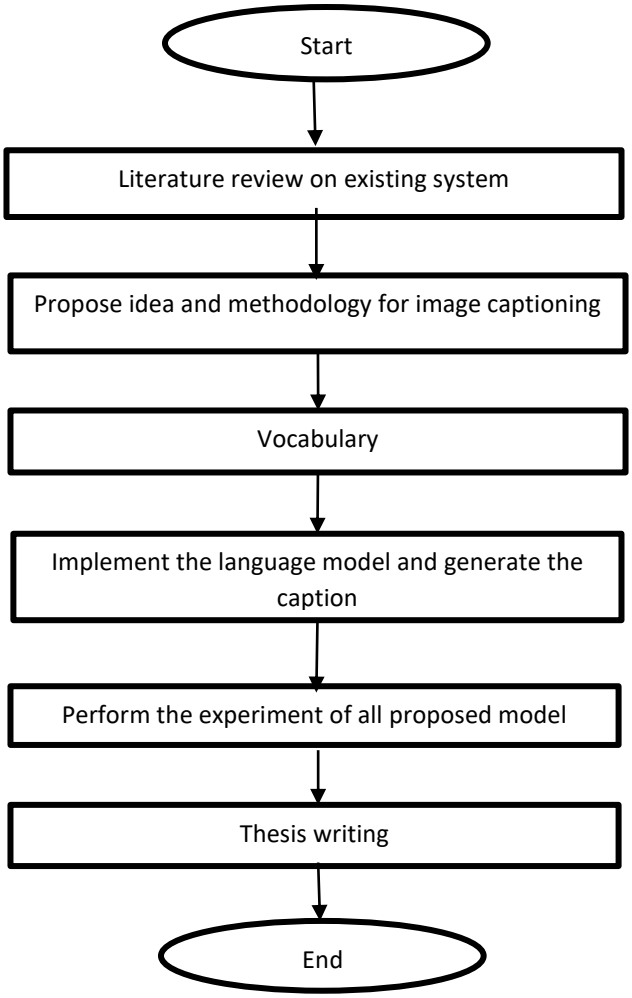


Figure 1.5.1: Flow of the research work

1.7 Thesis Organization

The rest of the paper is organized as follows.

Chapter 2 highlights the literature review of this case study.

Chapter 3 describes all about Artificial Neural Network for image captioning using deep learning approaches. Chapter 4 describes the working methodology.

Chapter 5 illustrates the models that we had used in our experiment and also describes the architecture and the solution we proposed. Then all the results and analysis part is shown of this work in Chapter 6.

CHAPTER 2

Literature Review

In [1], authors have used a pre-trained VGG16 image embedding model with stacked LSTM layers. 16, 000 Bangladeshi contextual images have been accumulated and manually annotated in Bengali. The output shows that VGG16 has successfully been able to learn a working language model and to generate captions of images quite accurately in many cases.

In [2], this paper introduced dense captioning. This paper proposed a Fully Convolutional Localization Network which is performed for addressing the description of an image and localization. On the Visual Genome dataset they evaluate their network.

The authors in [3], proposed the mechanism of bottom-up and top-down attention where the first one belongs to Faster R-CNN for image regions with an associated feature vector and the second one for feature weightings. They focused on large image regions.

In [4], another paper represented automatically generating image descriptions by using three steps: detect words, generate sentences and re-rank sentences through detectors and models. In their work, they showed that the generated captions are 34% equal to human captions.

This paper compared different approaches which are Detector Conditioned Models, Multimodal Recurrent Neural Network and k-Nearest Neighbor Model. They examined different types of issues such as overlapping issues, repetition of captions, etc. [5].

To produce tokens as an input for machine translation and image captioning need to train RNNs [6]. In this case, the unknown previous token is being replaced at an interface area. So, this paper introduced a method to change the training technique.

In [7], provides residual-networks. But still, this had a lower complexity. They are able to obtain 28% improvement on the COCO dataset.

In [8], it represents a Decoupled Novel Object Captioner (DNOC) framework that can disengage the language model from the object descriptions and also the words. It can generate by using the zero-shot novel object captioning. Another paper worked on a "cross-domain image captioner".

At interface, they proposed a "novel critic-based planning method". This can select high-quality sentences [9].

This paper provides a generic block diagram of the major groups of images, taxonomy of image captioning techniques. Evaluation matrices and datasets have been also a part of discussion [10]. This was done by deep-learning.

Another paper [11] provides us an automatic description generation from natural images using natural language processing and computer vision. During this work, CNN is used as the encoder and RNN as the decoder. Decoder is guided by a sequential guiding network during word generation.

The authors described the bi-directional mapping among images and their sentence-based statement. RNN is used to dynamically build a visual representation of the view as a caption is being generated [12]. This model is able to reform visual features given an image statement and generating novel captions given an image. The automatically produced captions are the same as (21.0 %) or peoples.

In [13], including 8,000 images that are each paired with five different captions this paper contains a frame sentence-based image annotation, a modern standard summation for sentence-based image statement and search. Their results displayed the importance of training on multiple captions per image, and of capturing syntactic and semantic characteristics of these captions. For this task separate human and automatic evaluation metrics and allowing us to augment our collection with additional relevance judgments of which captions describe which image and proposed strategies for collecting human judgments cheaply and on a very large scale.

In human caption evaluation semantic propositional content is a significant element. Automatically generating an image caption is an interesting task. Existing automatic evaluation metrics are primarily sensitive to n-gram overlap because evaluation is challenging. SPICE can answer questions for example which caption-generator better understands colors? And can caption-generators count? [14]

A paper contains automatic captioning of geo-tagged images by using dependency pattern models and n-gram language. They proposed that multiple web documents can summarize to automatically

generated image captions. They used n-gram language model for simpler representation of an object type model [15].

CHAPTER 3

Artificial Neural Network

3.1 Artificial Neural Network

An artificial neural network is a computational model network which is based on the composition and functions of the biological neural network. This network is known as a neural network. To receiving, processing and transmitting information, it behaves like an artificial human nervous method for in the field of computer science. Artificial Neural Network (ANN) plays a vital role in image captioning. The idea of learning and adapting to the flow of changing situations and produce information has always been one of the greatest research fields in the area of computer science. Not only that but also solving real-world problems the state - of - the- art achievement.

Artificial Neural Network is capable to caption an image based on its meaningful features. Feature extraction is a valuable part for captioning an image and also for image classification. If the features are not selected well then, the best classifier will not generate a good result. So, to achieve better performances, it is essential to minimize the dimension of the feature vector. Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN) are very important networks of artificial neural networks. These two networks are very much useful in the area of image captioning. We will discuss these networks later in brief.

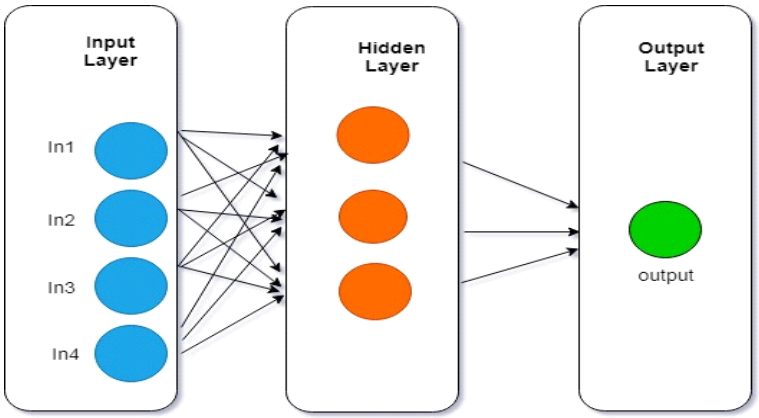


Fig. 3.1.1: Simple architecture of ANN

To understand the artificial neural network, we can take a look at the basic form of an ANN. It can be generated from 3 different layers: - Input Layer, Hidden Layer, and Output Layer

In Fig.3.1, the blue circle denotes the neurons on the input layer, the orange circle denotes the hidden layer and lastly, the green one is for the output layer. The arrow sign represents the connections between these three layer's neurons. The inputs are fed in the model through the input layer. Hidden layers are used for processing the inputs received from the input layer. It can be more than one. At the output layer, data has been made available after processing.

3.2 Deep Learning

Nowadays deep learning is a very important area for research and applications. It is a technique that teaches the computer to do those things that come naturally from humans. Deep learning is capable to establish correlations. It can be called as a static prediction. It is capable to establish correlations among present and future events. Deep learning can run regression among both for the past and future events. The future event is as like as the label in a sense. Deep learning does not compelled to care about time or the fact that something hasn't occurred yet. Deep learning might read a string of numbers and also predict the number relevant to occur next. Early methods disclosed by Hinton and collaborators concentrate on "greedy layer wise training" and unsupervised techniques such as autoencoders. By using the back-propagation algorithm deep learning is focused on training models like deep neural networks. In this case famous methods are:

- Convolutional Neural Networks
- Multilayer Perceptron Networks
- Long Short-Term Memory
- Recurrent Neural Networks

In our proposed model we used Convolutional Neural Networks, Recurrent Neural Networks and Long Short-Term Memory to accomplish our goal.

[10] Image captioning requires to identifying the important objects, their relationships and their qualities. It is necessary to produce not only syntactically but also semantically correct or appropriate sentences. Deep learning-based methods are a method that can able to handle the complexities and challenges of image captioning. Image captioning is one of the challenging applications of deep learning. It is the process of generating a description from a given image that is based on the actions and objects in the image.

Deep learning approaches have remarkable state-of-the-art results on generation captions. The most impressive thing about this method is an end-to-end model that is single can be defined to predict a caption, given an image. Deep learning is getting popular because of its accuracy level. It's getting high and high day by day. This technique is improving day by day to classifying the objects in images. Deep learning architectures such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been applied in computer vision.

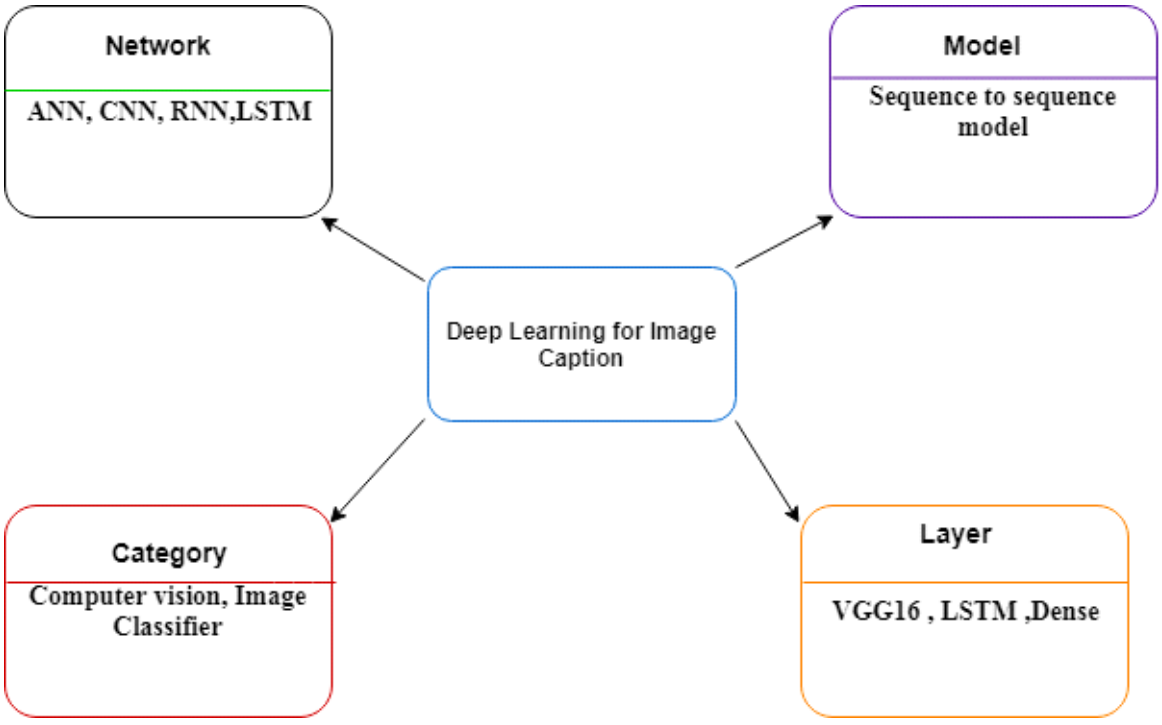


Fig. 3.2.0: Diagram of Research framework.

3.2.1 Convolutional Neural Network

A convolutional neural network (CNN) belongs to the deep learning neural network as a class. For different types of computer vision tasks, CNN is very popular. CNN refers to a remarkable breakthrough in image identification or recognition. In image classification, they're repeatedly working behind the scenes. A convolutional neural network is used to analyze visual imagery. The effect is so intense that it has been used from tagging photos of Facebook to self-driving cars. Its application to health care and protection systems is remarkable. Day by day this network is getting stronger in the field of deep learning. This network use very little preprocessing compared to other image classification algorithms. The pre-trained CNN used the ImageNet dataset. Mainly this network is trained on this dataset. Input images are transformed into a standard resolution. So that the input behaves like a constant for the model for any type of image that is given. In fig. 3.3, feature vector is created from a CNN. Technically, this vector is called embedding and the Convolutional neural network model is as like as an encoder.

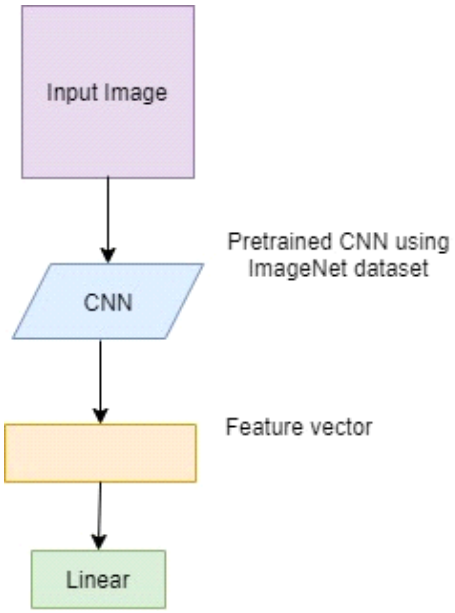


Fig. 3.2.1.1: Basic CNN

3.2.2 Recurrent Neural Network

In the Recurrent neural network, the output from the previous event is fed as input to the present event. Traditionally, inputs and outputs are independent. But when it need to predict the subsequent word of a sentence, it's required the previous word. So there is a need to remember the previous word in this technique. Long short term memory is a kind of recurrent neural network. We used this network in our proposed model.

CHAPTER 4

Methodology

4.1 Overview

Image captioning deals with a variable-length output series of words. To captioning the image, our proposed model learns from that particular image and geminated sentences which narrate the illustrated events in the field of natural language.

4.2 Research Methodology

There are several approaches for image captioning. But we choose Deep Learning based image captioning for its better accuracy. The algorithms based on deep learning can handle not only the complexities but also the challenges of image captioning in an impressive way. We can classify deep learning for image captioning methods such as Multimodal space-based, Language model-based, Novel object-based image captioning, Visual space-based, Supervised learning, Dense captioning, Attention-Based, Semantic concept-based, Stylized captions, Encoder-Decoder Architecture-based, Whole scene based, LSTM (Long Short-Term Memory).

For our research, we used dense captioning with Convolution Neural Network, the encoder and the Recurrent Neural Network with Long Short-Term Memory, the decoder. To implement this method, we used ‘Keras’(A python library). This process has been very beneficial for our work. Through CNN, we have generated words for captioning. Then, the words are arranged to form a sentence and this is done by a language model. For an image caption model, this embedding becomes a dense representation of the image. Every image has its own description. So, we need to detect the objects of those images which can be generalized by dense captioning. If the training dataset has the same types of sentence then generally the model returns the human caption. We propose a sequence to sequence model for image captioning, where the input is the images and the output is the sequence of words (y1, y2,..., yn). Our model draws attention to the conditional probability of an output sequence by detecting the maximum common object that is given in an input sequence. Basically, there is a mapping between the image and the caption. The sentence

and vocabulary list are gone through mapping with each image. Captions are generated recursively by training with that vocabulary list.

The impressive way of sequence to sequence captioning with an LSTM, Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). In existing work, datasets that have been used are not enough for this case study. And it also required more memory. The accuracy level of generated captions is also not so good. So, we wanted to work in this area to generate a more accurate caption for any image with more datasets. And this will require less memory. The majority of the images have a subject looking at the camera so the model sometimes cannot generate proper caption. So, we decided to collect images of all angles.

4.3 Environment

4.3.1 Python Environment

In this section, we will discuss our working environment that is used for generating captions. Basically python is a programming language. Our thesis paper is about the captioning image in Bengali. To do so we need to set up the environment with the necessary language and library.

To accomplish our task we used python 3 (programming language).

4.3.2 Keras

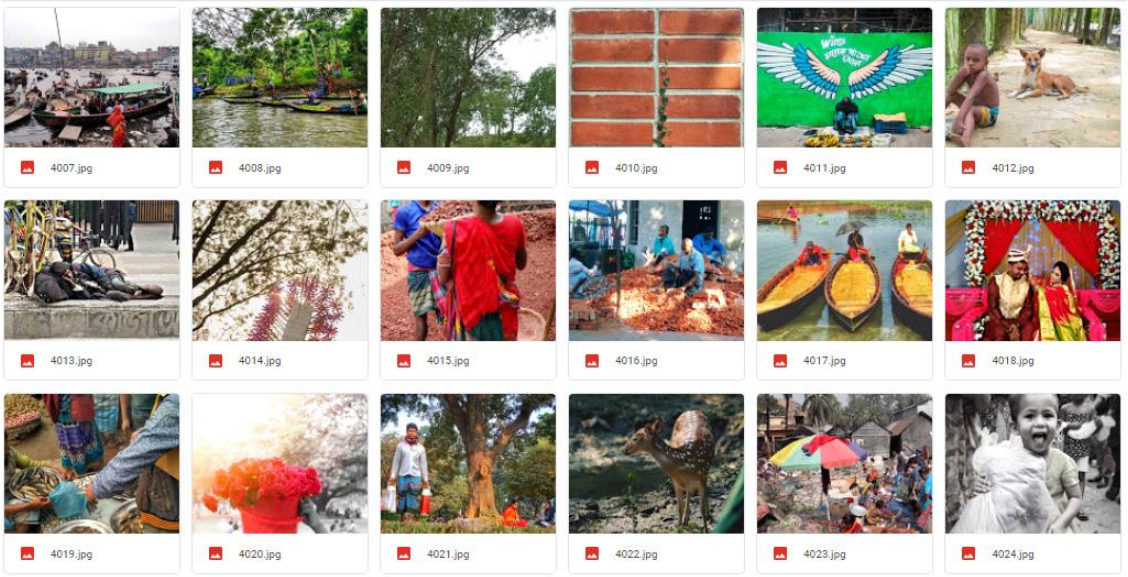
Keras is a high-level neural network API that is written in Python. It is able to run on top of "TensorFlow", "CNTK" or "Theano". Keras is exhibited as part of the research effort of the project named "Open-ended Neuro-Electronic Intelligent Robot Operating System".

Model is the core data structure of Keras. It is a way to organize the layers. The sequential model is a simple type of model. Keras is the python deep learning library which is used to run the models to generate captions. We have implanted our code with Keras.

CHAPTER 5

Data Collection

In the past, we have noted the huge effect of new datasets on research fields in Artificial intelligence. Flickr8K and MS COCO image datasets are one of the good datasets for image captioning. These datasets have images as well as consisting captions and it represents the foreign environment, culture, society, etc. These captions are written in the English language. But as our task is captioning the images in Bengali language so we could not use their captions for our case study. So we collect only images from Flickr8K and MS COCO as our dataset and captioned in the Bengali language. Moreover, we also need to add our native images. For this purpose, we used five diverse platforms to train and evaluate our model. These datasets containing real-world images and each image annotated with different types of captions. Our conducting dataset experiments on those platforms are news portal, www.flickr.com, www.pixabey.com, www.facebook.com, and www.unsplash.com, www.prothomalo.com. We used 00000 images as dataset, for training 0000, for validation used 0000 images and 0000 for testing our result. From the training set, we captioned 14000 images in the Bengali language as a set of input. In previous work, only one caption has been given for each image [1]. But in our data set, multiple captions have been used for an image like the MS COCO dataset. This helped us to get more accurate results. In our case study, we have been used images as data that are taken from various dimensions. So there are many variations in the Collective Data. Because of this variation of datasets human-generated captions also have a massive variety. This allows the machine to predict many types of objects and generate captions from the list of vocabulary. Fig. 5.1 and 5.2 show the sample dataset of our case study.



(a) Sample of Bangladeshi images

4007.jpg নদীর পাড়ে সারি সারি নৌকা, ট্রলার বাধা আছে যাত্রীর অপেক্ষায়

4008.jpg নৌকা বোঝাই করে পেয়ারা নিয়ে যাচ্ছে মাঝিরা

4009.jpg পাহাড়ী রাস্তায় হেঁটে চলেছে চা শ্রমিকেরা

4010.jpg ইটের লাল দেয়াল বেয়ে উপরে উঠছে পরগাছা

4011.jpg রঙিন দেয়ালের সামনে কলাবিক্রেতা কলা সামনে রেখে মাথা নিচু করে বসে আছে

4012.jpg ছেলেটি মাটির রাস্তায় বসে মুখে খাবার নিচ্ছে আর পিছনে একটি কুকুর বসে আছে

4013.jpg পাগল লোকটি রাস্তার আইল্যান্ডের মাঝে মাথার নিচে পোটলা রেখে ঘুমিয়ে আছে

4014.jpg একজন হকার খেলনা বিক্রি করতে বেরিয়েছেন

4015.jpg মহিলাটি ইটের ঝুড়ি ছেলেটির মাথায় তুলে দিচ্ছে

4016.jpg কয়েকজন লোক ইট ভাঙ্গছে

4017.jpg তিনটি নৌকায় তিনজন মাঝি বসে আছে

4018.jpg বউ বরের মাথার পাগড়ি ঠিক করে দিচ্ছে

4019.jpg মাছ বিক্রেতা মাছগুলো পলিথিন ব্যাগে ভরে দিচ্ছে

4020.jpg ছোট ছেলেটি মাথার উপর বালতিতে করে গোলাপ ফুল নিয়ে দাঁড়িয়ে আছে

4021.jpg পার্কে চা বিক্রেতা ফ্লাস্ক, চা নিয়ে ঘুরছে আর গাছের পাশে দুটি মেয়ে বসে আছে

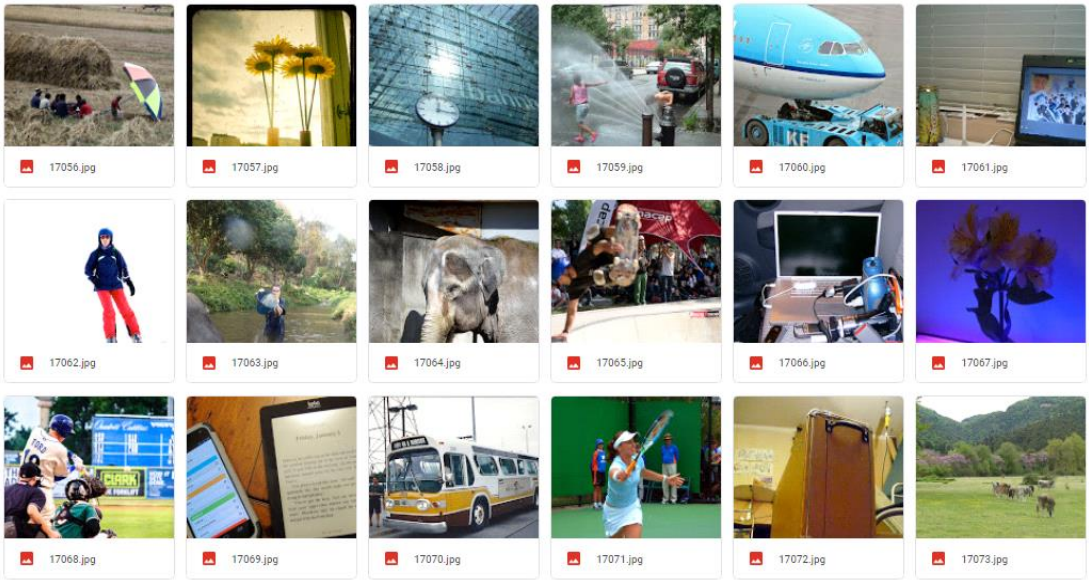
4022.jpg শিংবিহীন হরিণটি দাড়িয়ে আছে

4023.jpg বাজারে মহিলারা বসে মাছ কেটে দিচ্ছে

4024.jpg ছোট মেয়েটি ত্রান পেয়ে খুব খুশি

(b) Sample of human-generated captions

(c) Fig. 5.1: Dataset of Bangladeshi images



(a) Sample of foreign images

17056.jpg ক্ষেতের মাঝে কয়েকজন মানুষ বসে আছে এবং পাশেই একটি ছাতা খুলে রাখা হয়েছে

17057.jpg ফুলদানিতে সূর্যমূখী ফুল রাখা

17058.jpg কাচের দেয়ালের দালানের সামনে একটি ঘড়ি রাখা

17059.jpg পাইপ থেকে ছিটানো পানি দিয়ে বাচ্চারা ভিজছে

17060.jpg উড়োজাহাজের নিচে একটি গাড়ি রাখা

17061.jpg খোলা ল্যাপটপের পাশে একটি বাক্সে খাবার রাখা

17062.jpg বরফের উপর একজন মানুষ স্কেটিং বোর্ডের উপর দাঁড়িয়ে আছে

17063.jpg মহিলাটি খালের পানিতে নেমে বালতি দিয়ে পানি ছিটিয়ে দিচ্ছে

17064.jpg দেয়ালের কাছে হাতিটি দাঁড়িয়ে আছে

17065.jpg একটি ছেলে স্কেটিং করছে এবং তার পিছনে অনেক মানুষ তা দেখছে

17066.jpg গাড়ির সীটে ল্যাপটপসহ অনেক জিনিস রাখা

17067.jpg ফুলদানিতে ফুল রাখা

17069.jpg নোটপ্যাডের পাশে একটি মোবাইল রাখা

17070.jpg রাস্তায় বাসের পাশে একটি লোক দাঁড়ানো

17071.jpg একটি মেয়ে ব্যাডমিন্টন খেলছে

17073.jpg সবুজ মাঠে গরু ঘাস খাচ্ছে

(b) Sample of human-generated captions

Fig. 5.2: Dataset of foreign images

(a) Human generated captions

(c) Human generated captions

Fig. 5.3: Example of multiple captions

CHAPTER 6

Model

6.1 Overview

This section of the paper discusses the model that has been used in our system. There are many types of models in the world of deep learning that we can use in our research work. Our task is to give a picture as an input and get a perfect caption as an output. To do this we first trained the machine with various types of data. And to track this data, we need to have several networks. We will discuss about VGG16 pre-trained model and its dataset 'ImageNet', long short term memory model in brief.

6.2 Image Classification

In the present world, image classification has a great impact on artificial neural networks. It is one of the major concerns in recent days. Generally, it is the procedure of taking an input, for example, a picture and generating a class like "dog". In another sense, it is the probability of classifying images. VGG16 and ImageNet is the image classification model. Image classification was extended to the more challenging task of generating descriptions (captions) for images, often as a combination of CNNs and LSTMs.

6.2.1 VGG16 Pre-trained Model

In our work, we used the VGG16 (Visual Geometry Group) model which is previously trained on the ID (Image net dataset) and used as the previously trained image model in the proposed system with some slight adjustments. It is a convolutional network for classification and detection. This is the model of Keras with a 16-layer network. This is used by the VGG team members. There is also some other model (e.g. ResNet, ResNetV2, ResNeXt). These may give us a better result but we choose VGG16 for our work. Pre-trained VGG16 is quick and gives good performance. Fig. 6.1 illustrates the architectural structure of the VGG16 model. There are 16 layers, each layer used some filters. In layer 1, 64 filters are used in convolution layer, in layer 2, 64 filters are used in

convolution layer with Max pooling layer, in layer 3, 128 filters are used in convolution layer, in layer 4, 128 filters are used in convolution layer with Max pooling layer, in layer 5, 256 filters are used in convolution layer, in layer 6, 256 filters are used in convolution layer, in layer 7, 256 filters are used in convolution layer with Max pooling layer, in layer 8, 512 filters are used in convolution layer, in layer 9, 512 filters are used in convolution layer, in layer 10, 512 filters are used in convolution layer with Max pooling layer, in layer 11, 512 filters are used in convolution layer, in layer 12, 512 filters are used in convolution layer with Max pooling layer, in layer 13, 512 filters are used in convolution layer with Max pooling layer, in layer 14, 4096 nodes are connected fully, in layer 15, 4096 nodes are fully connected with Max pooling layer, in layer 16, Finally 1000 nodes are connected with the output layer which is activated with "softmax". This model has four different layers called convolution layer, max-pooling layer, fully connected layer at the end and one "softmax" layer and all of these have their own size. Overall this pre-trained model has a total of 16 layers. In our work, the VGG16 model is configured as D. This model achieves 7.5% in top-5 error on ILSVRC (2012) Val and 7.4% in top-5 error on ILSVRC (2012) test.

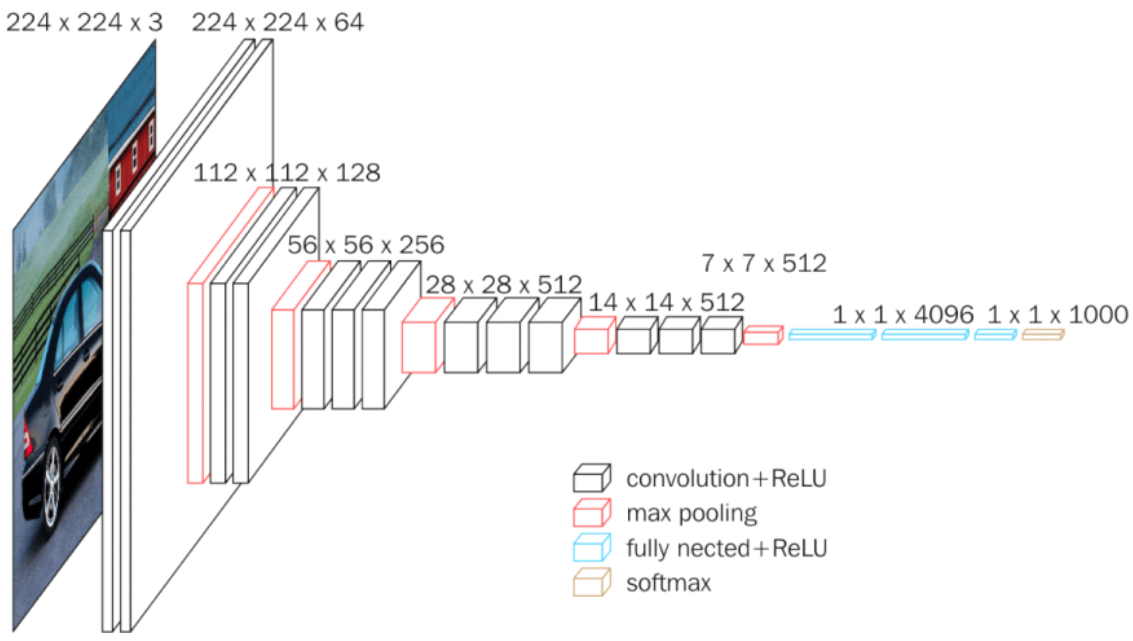


Fig. 6.2.1.1: VGG16 model's architecture at image level

6.2.2 Image Net

It is an image dataset which is used to pre-train the model like VGG16. Classifying image on ImageNet achieved a high-level accuracy in recent days. It helps to classify the pre-trained model. ImageNet has become a reference in the area of computer vision.

AI history has been transformed by the start of a deep learning revolution that is anchored by the October 2012 ImageNet victory.

6.3 LSTM

Long short-term memory (LSTM) is used in the area of deep learning. It is recurrent neural network (RNN) architecture. It can process individual data points as well as sequential data. Here individual data points can be referred to as images and sequential one can be referred to as speech or video.

These networks are suitable for classifying, processing and predicting based on data. In the area of a sentence language model, long short term memory predicted the next word in a sentence. Basically the LSTM is trained to predict the possible next word or value of any sequence. In a normal sense, it is like showing a person a sequence of images and asking to keep in mind the details. After a while show them a new picture that has a similar feature to the previous pictures and then ask for recalling the feature. This “recalling” and “remember” task is done by the LSTM network. CNN network is used with this LSTM model. The following figure illustrated the process of the LSTM model.

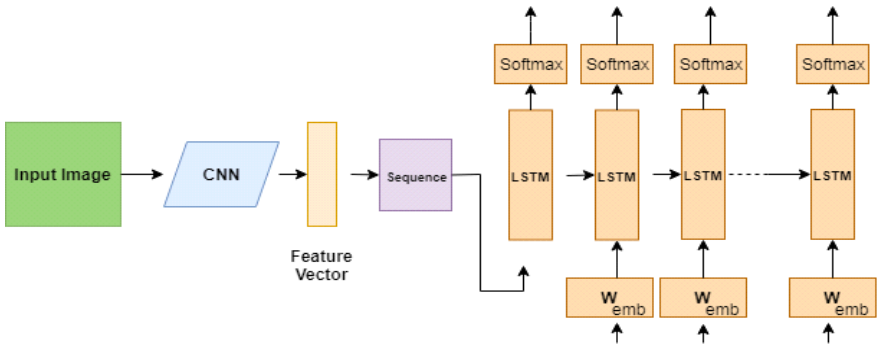


Fig. 6.3.1: LSTM model

6.4 Conditional Language Model

(LMs) Language Models estimate the correlative likelihood of various phrases. It is useful in different types of Natural Language Processing applications. It is the probability to a sequence of words.

Conditional language model is one kind of language model that assigns probabilities to a sequence of words given some conditioning context (a):

$$p(w_n \mid a, w_1 , \ldots, w_{n-1})$$

6.5 Proposed Model

In this paper, we have introduced image-captioning methods in Bengali language based on deep learning. To achieve better results, we have proposed a model, merged with the LSTM layer and the second last layer of the VGG16 model with a dense layer.

The main task is to encode the input image and decode from the mapping, one word at a time. In order to learn output sequences, first LSTM takes a special token, <startseq>, and then generates the next word to the nth sequence.

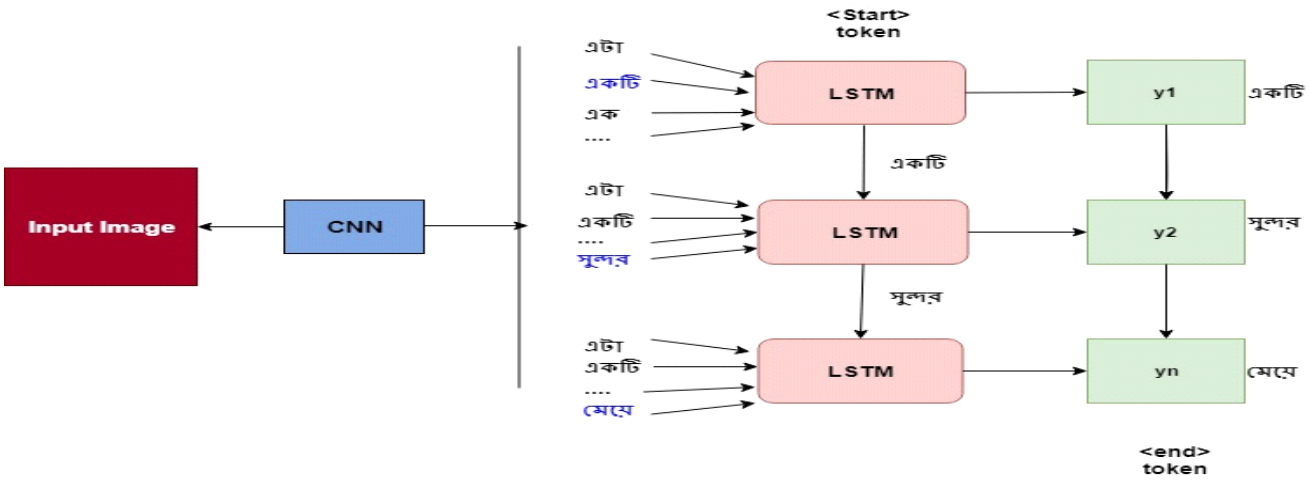


Fig. 6.5.1: Structural diagram

Sequence to sequence words will generate for an image. Those words will be concatenated to the end of each sequence. During this training session, <endseq> token is concatenated to the end of each sequences and the model returns a sentence for a particular image.

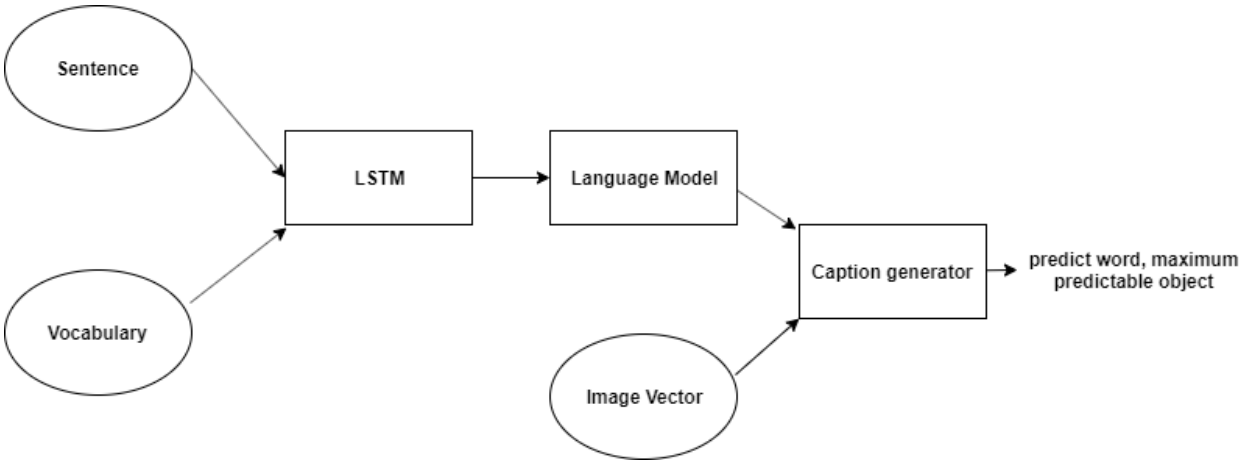


Fig. 6.5.2: Model Architecture

Fig. 6.5 illustrates our model architecture. Here we used a language model to predict words, next in the sequence of partial caption. At first, we created a language model from the captions we have. The work of this model is to train Bengali grammar and generate Bengali captions. For the training purpose, we gave some sentences and also a vocabulary list to the language model. So that it can learn how to generate a sentence from the given vocabulary. After that, we add image vector on that model. By using the VGG 16 model the machine takes a predictable object from the image vector. This model will generate captions based on the most predictive value. Our model is able to generate captions without mapping.

CHAPTER 7

Result Analysis

7.1 Overview

In this section, we evaluated our work in different scenarios and analyzed its acceptability and accuracy of our expected result. First, we described the Experimental setup of our work. Then we described the result analysis and showed how it holds up against existing solutions.

7.2 Experiment Setup

First of all, the whole experiment is implemented with the python deep learning library: Keras. After collecting a massive amount of data, it is used as a training dataset, validation and testing. We used merged with Long Short Term Memory (LSTM) layer and the second last layer of Visual Geometry Group (VGG16) model with a dense layer. After the machine has been trained with the training dataset, our proposed model has been compiled to generate the captions. This time a little amount of dataset is given to the machine to test whether the machine can make captions from its prediction level. No captions were given before for those datasets. The trained machine generates new captions by detecting objects and mapping words form Vocabulary. As a result in most cases, the machine is able to generate appropriate captions. But it also generates some inappropriate captions.

(a) নীলকাশেরনিচেনীলসমুদ্রেরতীরেকিছুসাম্পাননৌকাওমানুষআছে(Under the blue sky, there are some sampan boats and people on the shores of the Blue Sea)

(b) বিলেরকচুরিপানারমাঝেসারিসারিডিঙিনৌকাদিয়েরাস্তাবানানোহয়েছেএবংসেইরাস্তাদিয়েমানুষজননৌকায়চড়েবিলপারহচ্ছে(The road is made with cockleboat in the middle of the water hyacinth of bill and people are passing the bill by riding the boats)

Fig. 7.2.1: Human captions from the training set

Fig. 7.1 describes two images that are captioned by humans. Now if we need to generate a caption of an image from the machine then the machine will predict the word from the vocabulary list by mapping with the trained image and sequentially generate one by one word as a sequence of output for that image.

Let's have a look at the auto-generated caption of the image provided by the machine. As shown in Fig. 7.2, the machine-generated a caption by predicting some specific words. This automatic generated caption is relevant to the given picture.

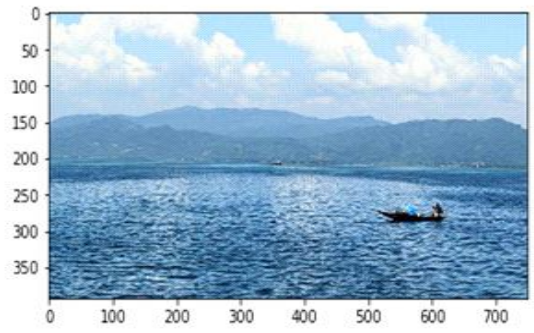
নদীতেঅনেকডিঙিনৌকাচলছেএবংনৌকায়মানুষচড়েছে

(There are many cockleboats on the river and people are riding on the boat)

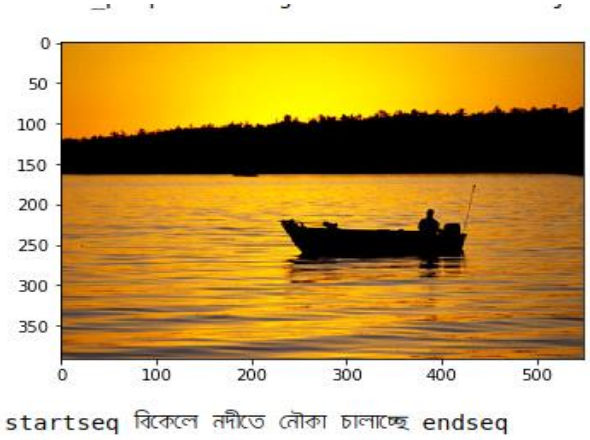
Fig. 7.2.2: Automatic generated captions

7.3 Result Analysis

Machine learning was developed widely, especially in the field of photography. After complete training, the proposed model generates an appropriate caption that makes sense. Those captions can explain the scenario of the given data. Our proposed model also generates two graphs that can capable to explain not only the model accuracy but also the model loss. In this section, we mentioned some appropriate examples and the generated graph by the proposed model.



(a) Blue sky and sailboat in the river and mountains far away

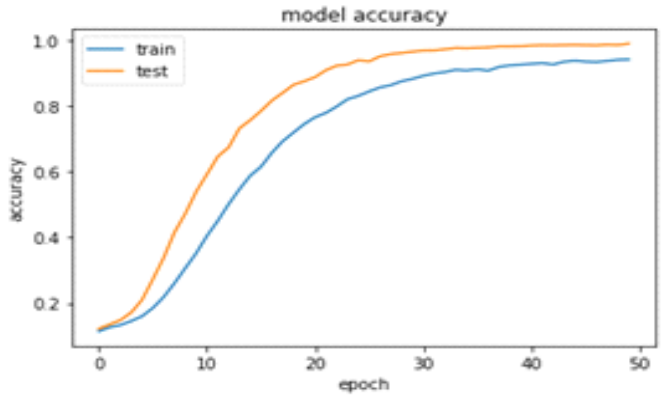


(b) In the afternoon, boats are running in the river

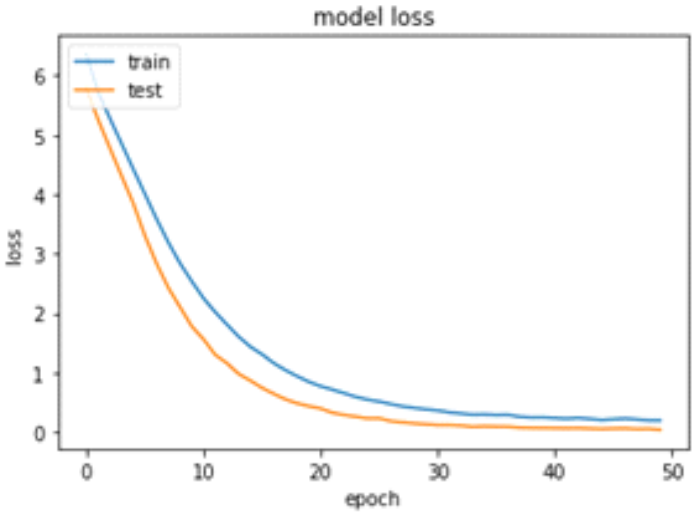
The man stood with an umbrella on his head

Fig. 7.3.1: Example of an appropriate generated caption by our proposed model

Nowadays, computer vision in some areas is so much fascinating. This can be trained to manage in a smoother way. And the captions are more appropriate and sharper than before. After training, our result showed 78.26% accuracy on this model. And it generates a graph. If we add more images as input, then the accuracy level will be growing up. Then we can achieve more accurate captions from a trained machine. Fig. 7.3 shows an appropriate result. Fig. 7.4 (a) shows the accuracy level which is increased on the other hand Fig. 7.4 (b) shows that the model loss is decreased. Generated graphs are given below-



Increasing accuracy level



- Decreasing model loss

Fig. 7.3.2: Accuracy and loss factor of our proposed model

Table-7.1 shows the outcome of our model. It shows the performance test of our model. 14000 images are used as input for testing the trained model. As we have used multiple captions, so the bleu score was calculated for 48000 sentences. The blue scores that are shown here are for four individual images. This value may change due to caption variations. Although the score is not perfect, the generated captions are almost accurate. And Table-7.2 shows the result of our sequence to sequence model.

Table-7.1 BLEU score

BLEU-1	1.007335
BLEU-2	1.285642
BLEU-3	1.328882
BLEU-4	1.492646

Table-7.2 Accuracy (%) of our model

Model Type	**Measure (%)**
Sequence to sequence Model	78.26

The result has been generated by the following equation.

Accuracy = 100- loss (a,b)

$$Loss(a,b) = \sum a \log(b)$$

Here, a= Human caption

b= Generated caption

CHAPTER 8

Conclusion

8.1 Discussion

Image captioning is performing a major role in different social platforms. People can easily understand the story if it is presented in their own language. And hence, we have proposed a model for captioning images in Bengali Language. Though there is an existing work [1] on this topic but our proposed model gives us a good result and better accuracy with the requirement of less memory. To do so, we have merged the LSTM layer and the 2nd last layer of the VGG16 model. Then we have used a dense layer. To evaluate our model, we have used various data sets from different sites and we have achieved a better result. Since we have taken the image of different angles as data, we have been able to collect varieties of data. In this sense, we will be able to measure the performance of different models by adopting this approach in our future work.

8.2 Future Work

We will increase our dataset for better performance. We will write 1, 00,000 captions for the future work. To get more accurate performance, we will use beam search instead of greedy search and also increase the level of unique words.

References

Conference/Journal Papers:

[1]. Motiur Rahman, Nabeel Mohammed, Nafees Mansoor and Sifat Momen: "*Chittron: An Automatic Bangla Image Captioning System*" (ScienceDirect, Volume 154, 2019, Pages 636-642)

[2]. Justin Johnson, Andrej Karpathy, Li Fei-Fei: "*DenseCap: Fully Convolutional Localization Networks for Dense Captioning".* Department of Computer Science, Stanford University. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3]. Justin Johnson, Andrej Karpathy, Li Fei-Fei: "*DenseCap: Fully Convolutional Localization Networks for Dense Captioning.*" Department of Computer Science, Stanford University. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4]. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang: "*Bottom-up and top-down attention for image captioning".* (Via. arXiv preprint arXiv: 1707.07998, 2017)

[5]. Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr DollÃąr, Jianfeng Gao, Xiaodong He, Margaret Mitchell, and John C Platt: "*From captions to visual concepts and back*." In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1473–1482, 2015.

[6]. Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell: "*Language models for image captioning: The quirks and what works.*" (arXiv preprint arXiv: 1505.01809, 2015)

[7]. SamyBengio, Oriol Vinyals, NaydeepJaitly and Noam Shazeer: "*Scheduled sampling for sequence prediction with recurrent neural networks*". In advances in Neural Information Processing System, pages1171-1179, 2015.

[8]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun: "*Deep residual learning for image recognition*". In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[9]. Yu Wu (CAI, University of Technology Sydney), Linchao Zhu (CAI, University of Technology Sydney), Lu Jiang (Google Inc.), Yi Yang (CAI, University of Technology Sydney, State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences): "*Decoupled Novel Object Captioner.*" (arxiv: v2 [cs.cv] 11 Aug 2018)

[10]. Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun: Show, Adapt and Tell: "*Adversarial Training of Cross-domain Image Captioner.*" In The IEEE International Conference on Computer Vision (ICCV), Vol. 2, 2017.

[11]. Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga: "*A Comprehensive Survey of Deep Learning for Image Captioning*" ACM Comput. Surv. 0, 0, Article 0. 36 pages, October 2018.

[12]. Daouda Sow, Zengchang Qin, Mouhamed Niasse, Tao Wan: "*A Sequential Guiding Network with Attention for Image Captioning*". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[13]. Xinlei Chenand C Lawrence Zitnick: "*Mind's eye: A recurrent visual representation for image caption generation*". In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2422-2431, 2015.

[14]. Xinlei Chenand C Lawrence Zitnick: "*Mind's eye: A recurrent visual representation for image caption generation*". In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2422-2431, 2015.

[15]. Micah Hodosh, Peter Young, and Julia Hockenmaier: "*Framing image description as a ranking task: Data, models and evaluation metrics*". Journal of Artificial Intelligence Research, pages 853–899, 2013.

[16]. Peter Anderson, Basura Fernando, Mark Johnson and Stephen Gould: Spice: "*Semantic propositional image caption evaluation*". In European Conference on Computer Vision. Springer, pages 382-398, 2016.

[17]. Ahmet Aker and Robert Gaizauskas: "*Generating image descriptions using dependency relational patterns*". In Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pages 1250–1258, 2010.

.

.

PLAGARISM REPORT

Image captioning in Bangla

ORIGINALITY REPORT

17%	7%	9%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

Submitted to Liverpool John Moores University 1 Student Paper	2%
MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz 2 Shiratuddin, Hamid Laga. "A Comprehensive Survey of Deep Learning for Image Captioning", ACM Computing Surveys, 2019 Publication	1%
www.cse.cuhk.edu.hk 3 Internet Source	1%

4	Yang Wang, Lan Wang, Feng Su, Jiahao Shi. "Video Text Detection with Fully Convolutional Network and Tracking", 2019 IEEE International Conference on Multimedia and Expo (ICME), 2019 Publication	1 %
5	allaboutmountainbikes.blogspot.com Internet Source	1 %
6	Submitted to National University Of Science and Technology Student Paper	1 %
7	**hal.archives-ouvertes.fr** Internet Source	1%
8	Submitted to Goldsmiths' College Student Paper	
9	Submitted to University of Newcastle Student Paper	

10	medium.com Internet Source	< 1 %
11	Submitted to iGroup Student Paper	< 1 %
12	hdl.handle.net Internet Source	< 1 %
13	Submitted to University of Lincoln Student Paper	< 1 %
14	github.com Internet Source	< 1 %

15	Submitted to Griffith College Dublin Student Paper	<1%
16	comtel.pe Internet Source	<1%
17	Submitted to IIT Delhi Student Paper	<1%
18	thesis.lib.ncu.edu.tw Internet Source	<1%
19	Salwa O. Slim, Ayman Atia, Marwa M.A., Mostafa-Sami M.Mostafa. "Survey on Human Activity Recognition based on Acceleration Data", International Journal of Advanced Computer Science and Applications, 2019 Publication	<1%

20 www.blog.arghh.net

Internet Source

21 E I OZCAN. "ARTIFICIAL NEURAL NETWORKS (A NEW STATISTICAL APPROACH) METHOD IN LENGTH-WEIGHT RELATIONSHIPS OF ALBURNUS MOSSULENSIS IN MURAT RIVER (PALU-ELAZIĞ) TURKEY", Applied Ecology and Environmental Research, 2019

Publication

22 Chuan-Jun Su, Yi Li. "Recurrent neural network based real-time failure detection of storage devices", Microsystem Technologies, 2019

Publication

23 Abhijit Ghatak. "Deep Learning with R", Springer Science and Business Media LLC, 2019

Publication

24	Submitted to University of Wolverhampton Student Paper	< 1 %
25	www.toptal.com Internet Source	<1%
26	www.omicsonline.org Internet Source	<1
27	Submitted to University of Strathclyde Student Paper	<1
28	en.wikipedia.org Internet Source	<1
29	Submitted to Bilkent University Student Paper	<1

30	Submitted to MunzurUniversitesi Student Paper	<1%
	Lecture Notes in Computer Science, 2015. 31 Publication	<1%
32	purehost.bath.ac.uk Internet Source	<1%
33	Submitted to The University of Manchester Student Paper	<1%
34	Submitted to Radboud Universiteit Nijmegen Student Paper	<1%
35	Submitted to University of Florida Student Paper	<1%
36	Submitted to California State University, Sacramento Student Paper	<1%

37	www.ukessays.com Internet Source	<1
	Rania Maalej, MonjiKherallah. "Improving the 38 DBLSTM for on-line Arabic handwriting recognition", Multimedia Tools and Applications, 2020 Publication	<1
39	Submitted to University of Bedfordshire Student Paper	<1
40	Submitted to University of Chichester Student Paper	<1
41	eprints.kfupm.edu.sa Internet Source	<1
	Submitted to CSU, San Jose State University 42 Student Paper	<1

43	Submitted to Afrihub Nigeria Ltd (NCC) Student Paper	<1
44	Submitted to University of Wales Institute, Cardiff Student Paper	<1
	Submitted to Concordian International School 45 Student Paper	<1

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		