

TRIBHUVAN UNIVERSITY

INSTITUTE OF SCIENCE AND TECHNOLOGY

PATAN MULTIPLE CAMPUS

A Major Project Report

On

Image Caption Generator

*For the partial fulfilment of requirement for the degree of*

*Bachelor of Science in Computer Science and Information Technology*

*(BSc.CSIT)*

Submitted By:

Aastha Oli (Exam Roll No.: 20001/075)

Prakriti Pudasaini (Exam Roll No.: 20063/075)

Sonika Acharya (Exam Roll No.: 20093/075)

Submitted To:

Department of Computer Science and Information Technology

Patan Multiple Campus

Patandhoka, Lalitpur

March, 2022



TRIBHUVAN UNIVERSITY

INSTITUTE OF SCIENCE AND TECHNOLOGY

PATAN MULTIPLE CAMPUS

A Major Project Report

On

Image Caption Generator

*For the partial fulfilment of requirement for the degree of*

*Bachelor of Science in Computer Science and Information Technology*

*(BSc.CSIT)*

Submitted By:

Aastha Oli (Exam Roll No.: 20001/075)

Prakriti Pudasaini (Exam Roll No.: 20063/075)

Sonika Acharya (Exam Roll No.: 20093/075)

Submitted To:

Department of Computer Science and Information Technology

Patan Multiple Campus

Patandhoka, Lalitpur

March, 2022

**SUPERVISOR’S RECOMMENDATION**

I hereby recommend that this project primed under my supervision by Aastha Oli (20001/075), Prakriti Pudasaini(20063/075) and Sonika Acharya (20093/075) entitled “Image Caption Generator” in partial fulfilment of this requirement for the degree of Bachelor in Science in Computer Science and Information Technology (BSc. CSIT) be processed for evaluation.

…………………………

Project Supervisor

Dadhi Ram Ghimire

Department of Computer Science and Information Technology, Patan Multiple Campus

**CERTIFICATE OF APPROVAL**

The undersigned certify that they have read and recommended to the Department of Computer Science and Information Technology, IOST, Patan Multiple Campus, a major project work entitled “Image Caption Generator” submitted by Aastha Oli, Prakriti Pudasaini, and Sonika Acharya in partial fulfilment for the award of Bachelor’s Degree in Computer Science and Information Technology (BScCSIT). The project was carried out under special supervision and within the time frame prescribed by the syllabus.

…………………………

Project Supervisor

Dadhi Ram Ghimire

Department of Computer Science and Information Technology, Patan Multiple Campus

…………………………

Program Coordinator

Dadhi Ram Ghimire

Department of Computer Science and Information Technology, Patan Multiple Campus

**ACKNOWLEDGEMENT**

We would like to express our sincere gratitude towards the Institute of Science and Technology, Tribhuvan University for the inclusion of the major project in the course of Bachelors in Computer Science and Information Technology (BSc.CSIT).

We are thankful to the Department of Computer Science and Information Technology, Patan Multiple Campus for constant guidance and supervision, as well as for providing all the necessary resources for the successful completion of the project.

We would like to extend our sincere regard to our supervisor Mr. Dadhi Ram Ghimire for his crucial supervision, constructive comments and suggestions, which gave the final shape of this project. We would also like to thank the faculty members of campus, all of our friends, seniors and colleagues for their valuable suggestions, comments and support.

Aastha Oli

Prakriti Pudasaini

Sonika Acharya

**ABSTRACT**

Image captioning is becoming one of the most frequently used technologies in the modern day. Deep neural network models are used to generate descriptions of the images. Image captioning is the process of creating a description for an image.

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. It requires both methods from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order .It requires recognizing the important objects, their attributes, and the relationships among the objects in an image. It generates syntactically and semantically correct sentences.

In this project, we present a model that generates a description of an image. The important aspect in implementing the image caption generator project includes data processing using Flickr 8k dataset, then extracting the features using CNN and generating captions using LSTM.

**Keywords:** *Convolutional Neural Network (CNN), Long short-term memory (LSTM), Image Caption Generator, Computer Vision, Natural Language Processing.*

**List of abbreviations**

BLEU Bilingual Evaluation Understudy

CNN Convolutional Neural Network

LSTM Long Short-Term Memory

RNN Recurrent Neural Network

**CHAPTER 1: INTRODUCTION**

# **1.1 Introduction**

Image caption generator is a process of recognizing the context of an image and annotating it with relevant captions. Indeed, a description must capture not only the objects contained in an image,but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in. The important aspects in caption generators for images involve CNN(Convolutional Neural Network) that extracts the features of the images and LSTM(Long short-term memory) which generates the caption from the extracted information of the image.

Initially, it was considered impossible that a computer could describe an image. With advancement of Deep Learning Techniques, and large volumes of data available, we can now build models that can generate captions describing an image.

The architecture of the model comprises two networks that are added together. Firstly, a CNN model, trained on an image classification task which is used to capture the details of the image. Moving forward, a Language generating model (RNN) is used to generate the appropriate caption given the image mappings. The task makes the CNN an image encoder and the fixed length vector mapping from the CNN is fed as the input to the RNN decoder. This decoder then generates the sentences.

In this project we are going to build one such annotation tool which is capable of generating very relevant captions for the image with the help of datasets.

# **1.2 Problem Statement**

Describing the content of an image using properly formed English language is a challenging task for a machine. The task requires an algorithm to not only understand the content of the image, but also to generate language that connects to its interpretation. Mimicking the human ability of providing descriptions for images by a machine is itself a remarkable step along the line of Artificial Intelligence. Traditionally, computer systems have been using predefined templates for generating text descriptions for images. However, these approaches do not provide sufficient variety required for generating lexically rich text descriptions. Neural network models have been put forward to suppress the given issue. The proposed concept could have great impact, for instance by helping visually impaired people better understand the content of images on the web.

# **1.3 Objectives**

* To generate a textual description of an image.
* To give an overview of images for captioning huge datasets.

**1.4 Scopes and Limitations**

**1.4.1 Scope**

**1.4.2 Limitation**

**1.5 Development Methodology**

**1.6 Project Roadmap**

**CHAPTER 2: BACKGROUND STUDY AND LITERATURE REVIEW**

**2.1 Background Study**

In the last few years, the problem of generating descriptive sentences automatically for images has gained a rising interest in Natural language processing (NLP). Image captioning is a task where each image must be understood properly and be able to generate suitable captions with proper grammatical structure.

There are a series of relevant research papers attempting to accomplish this task in the last decades, but they face various problems such as grammar problems, cognitive absurdity and content irrelevance. However, with the unparalleled advancement in Neural Networks, some groups started exploring Convolutional Neural Network (CNN) and Recurrent Neural Network (CNN) to accomplish this task and observed very promising results.

**2.2 Literature Review**

**CHAPTER 6: CONCLUSION AND FUTURE RECOMMENDATIONS**

**6.1 Conclusion**

**6.2 Future Recommendation**

We successfully completed what we mentioned in the project proposal, but used a smaller dataset (Flickr8k) due to limited computational power. There can be potential improvements if given more time. First of all, we directly used a pre-trained CNN network as part of our pipeline without fine-tuning, so the network does not adapt to this specific training dataset. Thus, by experimenting with different CNN pre-trained networks and enabling fine-tuning, we expect to achieve a slightly higher BLEU score. Another potential improvement is by training on a combination of Flickr8k, Flickr30k, and MSCOCO. In general, the more diverse training dataset the network has seen, the more accurate the output will be. We all agree this project ignites our interest in application of Machine Learning knowledge in Computer Vision and expects to explore more in the future.