INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

Dharwad, Karnataka- 580009

A Project Report On

"Image caption generator in Hindi language"

Submitted by

Ch Samyuktha (19BCS030)

M Joshasree (19BCS069)

S Pranay Sai Teja (19BCS102)

M Anupama (19BCS123)

Under the Guidance of

Dr. Deepak K T

Asst. Professor

Electronics and Communication Engineering

Indian Institute of Information Technology Dharwad

CERTIFICATE

This is to certify that the Project Work entitled **— "Image caption generator in Hindi language"** is a bonafide work carried out by Ch. Samyuktha (19BCS030), M. Joshasree(19BCS069), S. Pranay Sai Teja(19BCS102), M. Anupama(19BCS123) in fulfillment for the Deep learning Project of Bachelor of Technology in Computer Science & Engineering of the Indian Institute of Information Technology Dharwad during the year 2022-2023. The Project Report has been approved as it satisfies the academics prescribed for the Bachelor of Technology degree.

Signature of Supervisor(s)

Name(s)

Department(s)

(Month, Year)

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Ch. Samyuktha
19BCS030

M. Joshasree
19BCS069

S. Pranay Sai Teja
19BCS102

M. Anupama
19BCS123

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project.

It is our privilege to express our sincerest regards to our project coordinator, **Dr. DEEPAK K T**, Asst. Professor, Department of Electronics and Communication Engineering, Indian Institute of Information Technology Dharwad, for his valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of our project.

We take this opportunity to thank all our lecturers who have directly or indirectly helped our project. We pay our respects and love to our parents and all other family members and friends for their love and encouragement throughout our career. Last but not the least we express our thanks to our friends for their cooperation and support.

ABSTRACT

One of the core issues in artificial intelligence is the autonomous generation of image captions from an image's content. The work of "Image caption generation" requires expertise in both computer vision and natural language processing. Many studies have been conducted in this area, but they have mostly centered on producing English-language image descriptions because English-language image caption databases represent the majority in this sector. Language should not, however, constitute a restriction on the model for the visual description generator. A challenge with the absence of image captioning datasets in languages other than English is that they are often morphologically rich, like Hindi. With the help of Google Cloud Translator, we have written a python script for whole existing english dataset and created a Hindi picture description dataset based on images from the Flickr8k dataset, and the translated captions are saved in the captions_withcomma.txt file. This dataset is used in the study's encoder-decoder neural network model training.We present an image caption generator for our regional language that is Hindi using inceptionV3 and LSTM with attention module. Model achieved a BLUE score of 0.594435.

CONTENTS

CHAPTER 1

INTRODUCTION

Identifying the numerous objects in an image and creating a meaningful connection between them are both challenging tasks when captioning pictures.

Effectively completing this work will have numerous advantages in many different fields. Be it the healthcare, retail, IT, or a number of other yet-undiscovered areas.

People create descriptions of images by adding arbitrary elements like feelings or artistic impacts. Additionally, when there are several things in an image, the conventional structure is unable to convey the relationships between the various objects.

Rather than condensing a complete image into a static representation, several image captioning models have recently begun employing attention mechanisms, which highlight prominent aspects dynamically as needed and also increase the accuracy of description.

1.1 Problem Statement

Our main objective is to design an optimized deep learning model (with attention) to generate a Hindi caption for an input image.

CHAPTER 2

LITERATURE SURVEY

2.1 Introduction

In this section, we are interested to elaborate the main articles that we have followed to implement the project. This section contains two subsections having explanations and observations from the articles that used required algorithms.

2.2 Review 1

From the article **Deep learning approach for Image captioning in Hindi language, MSc Research Project, Data Analytics,Ankit Rathi** we learned about:

- Base architecture of image captioning models in languages other than english.
- Multimodal learning.
- Encoder-Decoder framework.

- Attention mechanism.

For reference : Deep learning approach for Image captioning in Hindi language

2.3 Review 2

From the article,**"Image Caption Generator in Hindi Using Attention,Abhishek SETHI, Aditya JAIN, Chhavi DHIMAN,Electronics and Communication Engineering (ECE),Delhi Technological University (formerly DCE),Delhi, India"**,we learnt about the actual structure of encoder-decoder using attention model using in image captioning in Hindi.

For reference: (PDF) Image Caption Generator in Hindi Using Attention

CHAPTER 3

ENCODER-DECODER ARCHITECTURE

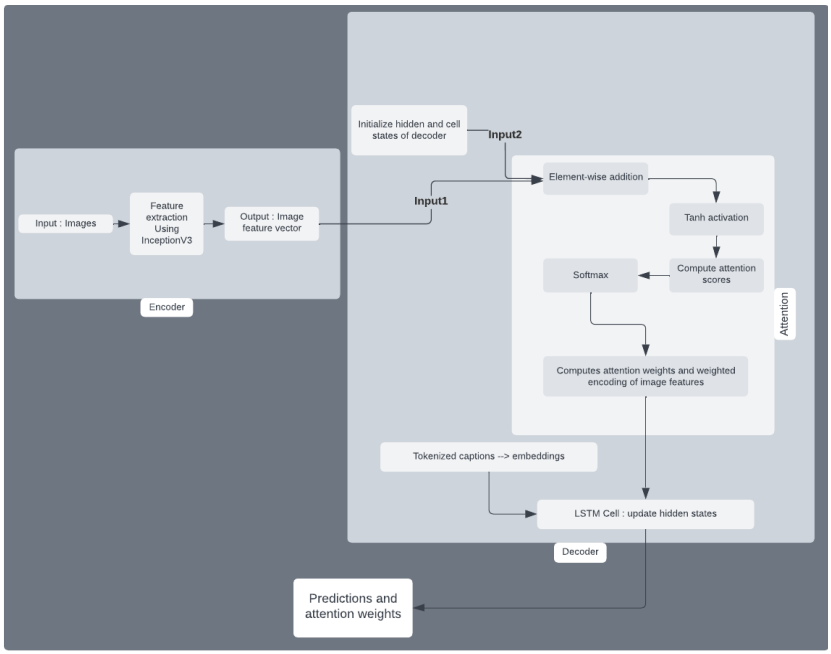

Figure 3 - Architecture for encoder-decoder model during training

CHAPTER 4

PROCEDURE

4.1 Data Description

Data description involves both data collection and data preprocessing.

4.1.1 Data Collection

The first step in the deep learning workflow is to collect data for training the DL model. The accuracy of DL systems' predictions is only as good as the data used to train them. Some of the data sources on which we can rely on for data collection are :

- Dataset search by Google
- Open ML
- UCI: Machine Learning Repository
- Public datasets on GitHub
- Visual Data Discovery
- Kaggle
- Amazon Datasets

We have collected our data from Kaggle.

For reference :

- Flickr 8k Dataset | Kaggle

4.1.2 Data Translation

We have converted english captions into Hindi captions by using python script that integrates with a google translator and saved them with their image id and separated them by a comma.

4.1.3 Data Preprocessing

We have created a pandas dataframe to analyze the data and preprocess it.

4.2 Image Feature Extraction

The process of extracting pertinent features from an input image so that they can be used to create a natural language description of the image is known as image feature extraction. These characteristics may include specifics about the things, people, and scene shown in the picture, as well as information about the picture's lighting, color, and texture.

Once the characteristics have been extracted from the image, they may be fed into a model that creates natural language captions for images by learning to recognise the features in the image.

4.2.1 Inception v3

We have used Inception v3 for Image Feature Extraction, The Inception v3 model is based on the idea of "inception modules," which are neural network modules that carry out numerous parallel convolutions with various kernel sizes and concatenate their output to produce a more effective and expressive model. As a result of these and other advances, Inception v3 performs well on picture classification tasks. These developments include batch normalization, RMSprop optimization, and label smoothing.

4.3 Attention model

A neural network architecture called an attention model enables a model to choose to focus on particular input elements and give those elements more weight when making predictions. To enable the model to concentrate on particular elements of the input image when creating a caption, an attention model can be employed. This can be helpful when the image has several caption-relevant objects or scenarios and the model needs to be able to concentrate on various portions of the image at various times.

We extracted features from the input image using an Inception v3 feature extractor, and then we fed those features into an attention-based model. The attention-based model would then make use of the attributes to create a caption for the image, selectively focusing on various areas of the image as necessary.

4.4 Decoder model

Based on the information retrieved by the encoder model, the decoder model creates the final caption for the image. Long-term dependencies in the input data can be captured using recurrent neural network (RNN) layers such long short-term memory (LSTM), which are typical in this type of system.The final caption is output by the decoder model once it has finished generating words for the entire caption.

4.4.1 Long Short Term Memory Model

Language translation, text summarization, and the creation of image captions are just a few of the many natural language processing activities that frequently involve LSTMs. The LSTM creates a sequence of words one at a time by using a language model to predict the next word in the sequence given the previous words and the image features as input. The image features are extracted from the image by an encoder model (here , i.e., inception v3 model).

4.5 Encoder-Decoder Model

The integration of encoder model, attention model and decoder model results in an Encoder-Decoder model for our project. The output from encoder (feature extraction of images) is fed into the attention model that is there in the decoder model to give the significant attention for different features and the outputs of the attention model are fed into LSTM to guess the next

word in the sequence that is related to the different parts/features of the input image. The cell states and Hidden states of LSTM cell are updated and are connected to the attention model. By all this means, the predictions are carried out in the designed Encoder-Decoder model.

4.6 Evaluation Metrics

A key task in computer vision is learning how to automatically create captions that summarize the content of an image. Typically, measures like BLEU, METEOR, ROUGE, or CIDEr are used to evaluate image captioning algorithms. These metrics all primarily assess the degree of word overlapping between generated and reference captions. The recently developed SPICE measures how closely scene graphs made from the candidate and reference sentences resemble one another and exhibits improved agreement with human assessments.

4.6.1 BLUE scores

A variety of NLP measures have been created over time to address this issue. The Bleu Score is among the most well-known. Bleu Score stands for *Bilingual Evaluation Understudy.* It has numerous flaws and is far from ideal. However, it is easy to calculate, comprehend, and has a number of strong advantages. Despite having a wide range of alternatives, it remains one of the most popular measures. Its foundation is the notion that the more closely the predicted language resembles the human-generated target sentence, the more accurate it will be. Bleu Scores range from 0 to 1. A rating of 0.6 or 0.7 is regarded as ideal. Even two humans would probably come up with various sentence alternatives for a given issue and would infrequently find a perfect match.

Formula :

Precision n-gram = Number of correct predicted n-grams / Number of total predicted n-grams

For example, if

Candidate sentence: The guard arrived late because of the rain.
Reference Sentence: The guard arrived late because it was raining.

Model	Set of grams(correct grams)	Score
Unigram	“The”, “guard”, “arrived”, “late”, “because”	5/8
Bigram	“The guard”, “arrived late”, “late because”, “guard arrived”	4/7

CHAPTER 5

RESULTS

```
BLEU-1: 0.594435
BLEU-2: 0.440589
BLEU-3: 0.337126
BLEU-4: 0.197693
```

Fig - 5.1 : Bleu score -1,2,3,4 values for the designed model

Fig - 5.2 : original and predicted caption while training

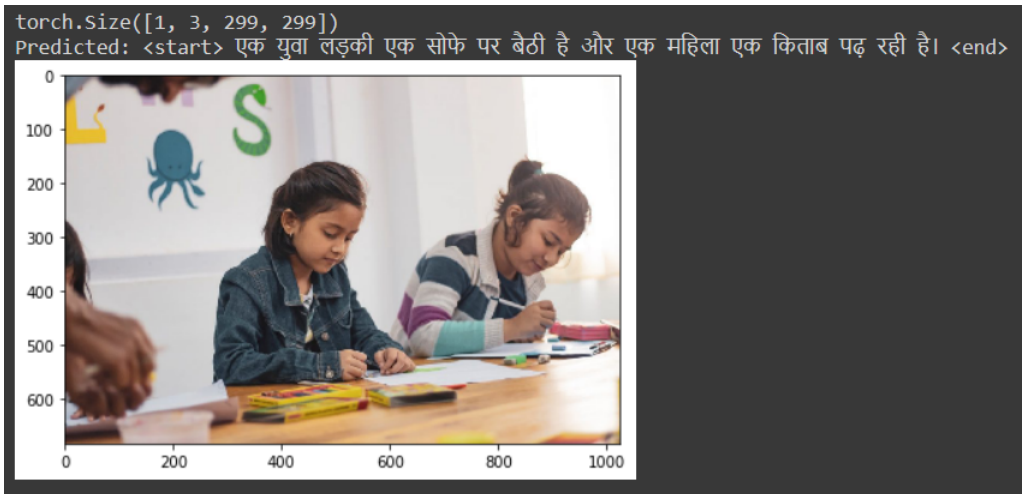

Fig - 5.3 : predicted output after validation

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

This project gave a Bleu-1 score of 0.594435 which is a better blue score in the field of image captioning.This project can be used to generate Hindi captions for all the indoor activity images.

- The data set can be further developed for indoor activities, and the algorithm can be implemented for the same.Using a bigger Dataset so that model has more data points to train on.
- Image captioning in the Hindi language can be done for video streaming by dividing the video into image frames.
- Feature extraction can be done using other models and can note down the observations.
- As English to Hindi conversion was done by a machine in Flickr8k. Doing so with the help of someone who speaks Hindi natively would definitely increase the results.
- We can also replace this whole sequential model (State-of-the-art model) with a transformer model which can be further developed by parallelism.

CHAPTER 7

SOURCE CODE

Source code uploaded in GitHub contains Dataset(contains images folder, captions with comma file, files that are generated during compilation of code), Deep learning source code.

Github Link (for reference): https://github.com/ANUPAMA0221/DeepLearningProject

CHAPTER 8

REFERENCES

- Deep learning approach for Image captioning in Hindi language,MSc Research Project Data Analytics Ankit Rathi.
- Comparison among Four Deep Learning Image Classification Algorithms in AI-based Diatom Test
- Image Caption Generator in Hindi Using Attention, Abhishek SETHI, Aditya JAIN, Chhavi DHIMAN, Electronics and Communication Engineering (ECE), Delhi Technological University (formerly DCE), Delhi, India
- Inception_v3 | PyTorch
- GitHub - rathiankit03/ImageCaptionHindi: Deep learning approach for Image Captioning in Hindi
- Understanding Encoders-Decoders with Attention Based Mechanism | DataX Journal