Image Caption Generator

Team

Khan Shayaan Shakeel - 3118030

Masalawala Murtaza Shabbir - 3118033

Qazi Faizan Ahmed - 3118039

Guided By

Er. Nafeesa Mapari

Problem Statement

- The problem statement is to generate detailed caption of an Image and further convert the caption into an Audio.

- The objective is to make an efficient system for describing an image in detail. Through this objective one can aim to find numerous uses in real life, like:

    - It can help visually impaired people to visualize an image and guide them at any given point of time.

    - The Model can be extended for videos to generate captions for the video.

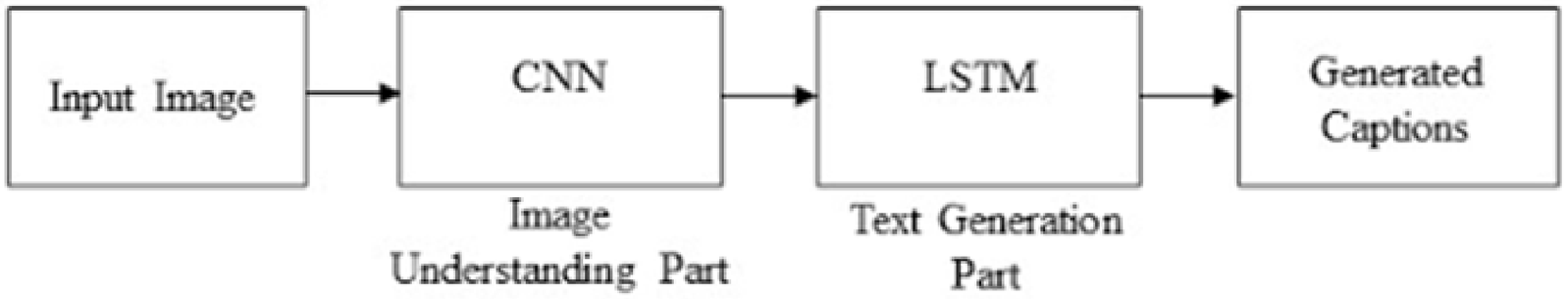
Introduction

- Automatically generating captions of an image is a task, that is very close to the heart of the Scene Understanding, which is one of the primary goals of Computer Vision.

- Not only the caption generation models should be powerful enough, to solve the Computer Vision challenges of determining objects, they should also be capable of capturing and expressing their relationship in a Natural Language.

- Recent work has significantly improved the quality of caption generation using combination of CNN to obtain vectorial representation of images and RNN to decode those representations into Natural Language Sentences

Block Diagram



Convolutional Neural Network

- Convolutional Neural networks are specialized deep neural networks which can process the data that has input shape like a 2D matrix. Images are easily represented as 2D matrix and CNN is very useful in working with these images.

- CNN is basically used for various image classifications and identifying if an image is a bird, a plane or Superman etc.

- It scans images from left to right or top to bottom to pull out important features from the image and combine these features to classify the images. It can handle the images that have been translated, rotated, scaled and have undergone any changes in perspective.

Long Short Term Memory

- LSTM stands for Long short term memory, they are a type of RNN (recurrent neural network) which is well suited for sequence prediction problems.

- Based on the previous text, we can predict what the next word will be like. It has proven itself effective from the traditional RNN by overcoming the limitations of RNN which originally had a short term memory.

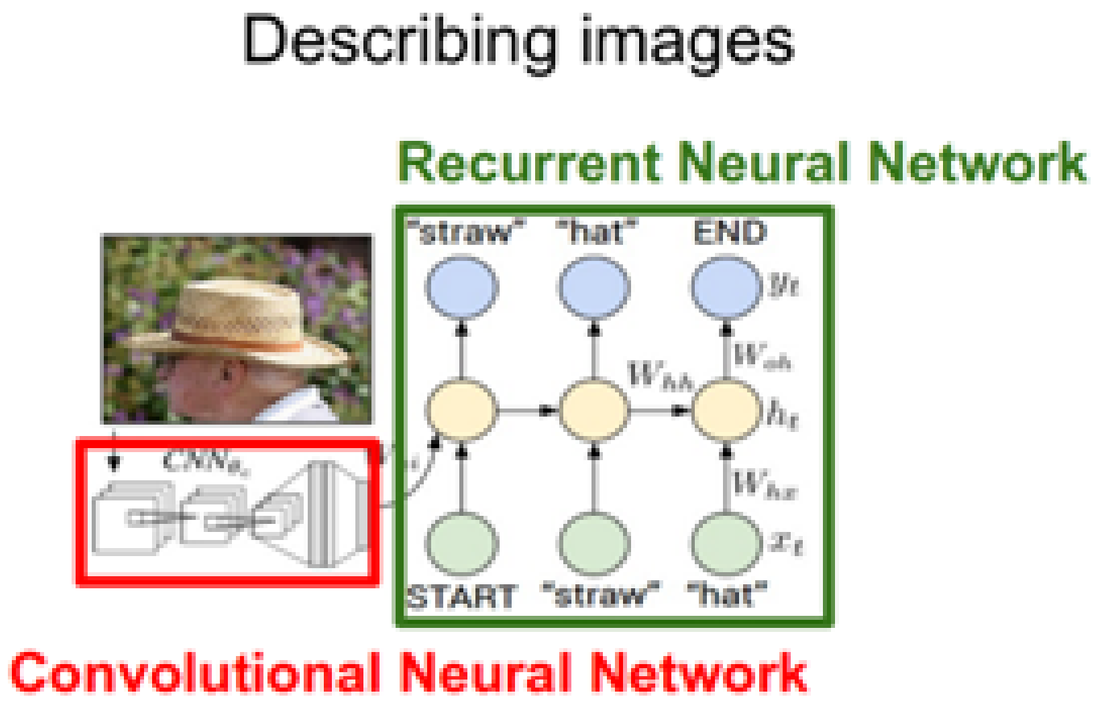
- LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information.

Image Captioning

The task of image captioning can be divided into two modules logically:

1. Image based model – Extracts features from images

2. Language based model - which translates extracted features and objects to sentences.



Dataset

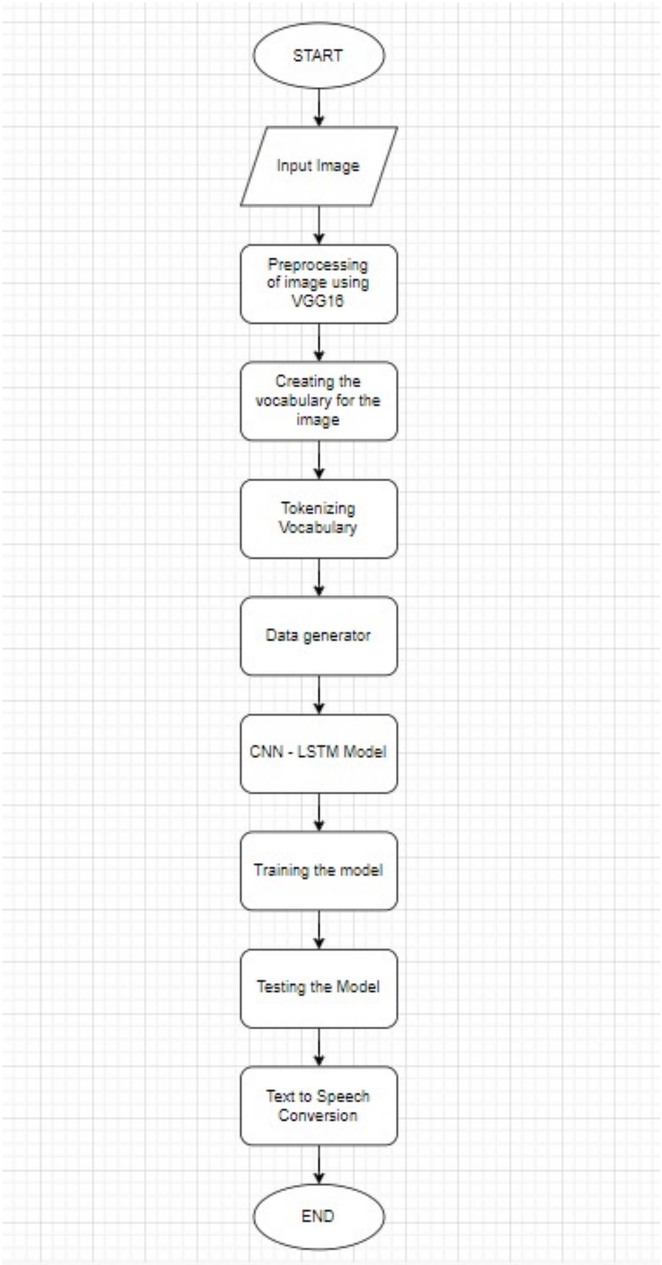
The Flickr8k dataset contains two directories:

- Flickr8k_Dataset: Contains 8092 photographs in JPEG format.

- Flickr8k_text: Contains a number of files containing different sources of descriptions for the photographs. Flickr_8k_text folder contains file Flickr8k.token, which is the main file of our dataset that contains names of images and their respective captions separated by newline(“\n”).

The image dataset is divided into 6000 images for training, 1000 images for validation and 1000 images for testing.

Major Steps


START
Input Image
Preprocessing of image using VGG16
Creating the vocabulary for the image
Tokenizing Vocabulary
Data generator
CNN - LSTM Model
Training the model
Testing the Model
Text to Speech Conversion
END

Output

IMAGE CAPTION GENERATOR

Home

Caption Your Image with the help of Deep Learning

GET STARTED

127.0.0.1:5000

IMAGE CAPTION GENERATOR

Home

Caption Image

Choose Files | img1.jpg

Or Drag It Here.

SUBMIT

IMAGE CAPTION GENERATOR

Generated Text:
two children playing soccer

0:00 / 0:02

Conclusion

During the project timeline, we have conducted requirement gathering and all the research required to be done. This includes researching on various approaches for generating captions efficiently, finding relevant and recent research papers related to the topic and eventually discovering the modules that were required to be further converted to obtained captions into an audio format. We have successfully created a web application having a single web page which sends request to the backend web server, created in Flask and successfully responds with a sample caption and audio.

References

1. V. Julakanti, “Image Caption Generator using CNN-LSTM Deep Neural Network“, International Journal for Research in Applied Science and Engineering Technology, vol. 9, no., pp. 2968-2974, 2021.

2. P. Mathur, A. Gill, A. Yadav, A. Mishra, and N. K. Bansode, “ Camera2Caption: A real-time image caption generator “, International Conference on Computational Intelligence in Data Science(ICCIDS), 2017.

3. S. Amirian, K. Rasheed, T. Taha and H. Arabnia, “Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap“, IEEE Access, vol. 8, pp. 218386-218400, 2020.

4. S. Shukla, S. Dubey, A. Pandey, V. Mishra, M. Awasthi and V. Bhardwaj, “Image Caption Generator Using Neural Networks“, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp. 01-07, 2021.

5. M. Panicker, V. Upadhayay, G. Sethi and V. Mathur, “Image Caption Generator“, International Journal of Innovative Technology and Exploring Engineering, vol. 10, no. 3, pp. 87-92, 2021.

6. J. Karan Garg and Kavita Saxena, “Image to Caption Generator“, International Journal for Modern Trends in Science and Technology, vol. 6, no. 12, pp. 181-185, 2020.

7. Xu Zhao, Kai-Hsiang Lin, Yun Fu, Yuxiao Hu, Yuncai Liu and T. Huang, “Text From Corners: A Novel Approach to Detect Text and Caption in Videos“, IEEE Transactions on Image Processing, vol. 20, no. 3, pp. 790-799, 2011.

8. Y. Huang, J. Chen, W. Ouyang, W. Wan and Y. Xue, “Image Captioning With End-to-End Attribute Detection and Subsequent Attributes Prediction“, IEEE Transactions on Image Processing, vol. 29, pp. 4013-4026, 2020.

9. S.-H. Han and H.-J. Choi, “Domain-Specific Image Caption Generator with Semantic Ontology,” IEEE International Conference on Big Data and Smart Computing (BigComp), 2020

10. M. Wang, L. Song, X. Yang, and C. Luo, “A parallel-fusion RNN-LSTM architecture for image caption generation,” IEEE International Conference on Image Processing (ICIP), 2016.