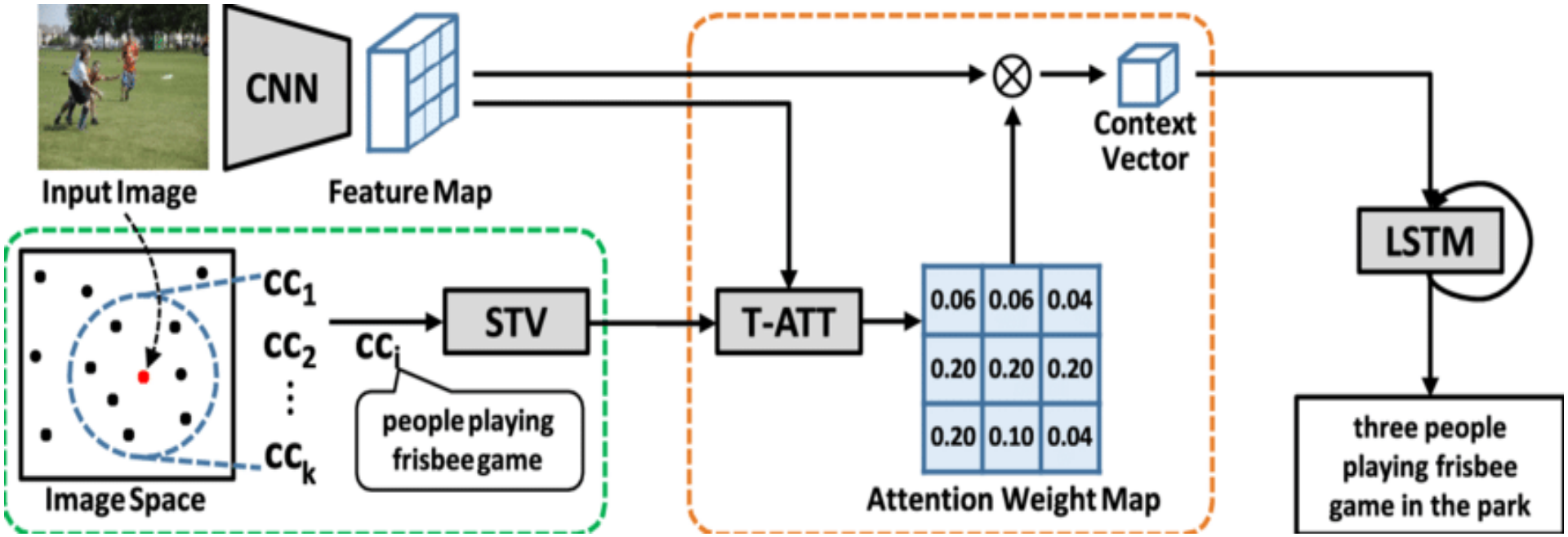
Image Caption Generator

Introduction

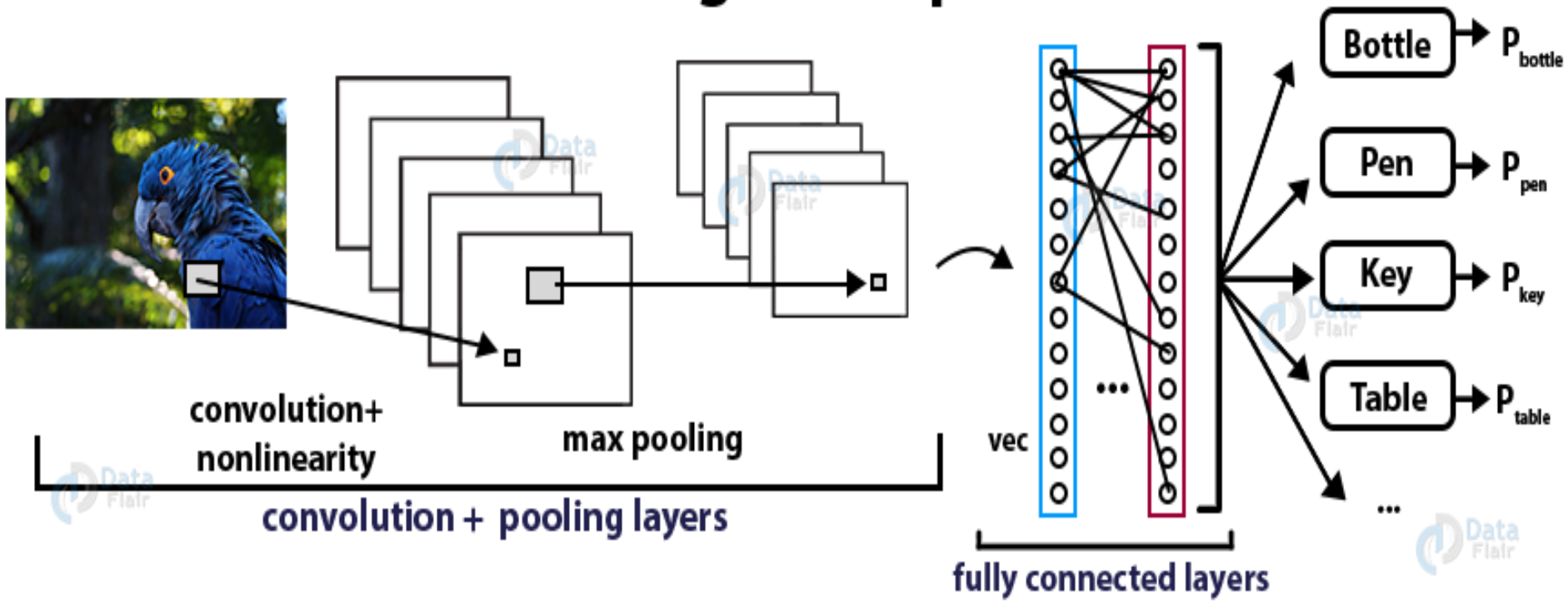
- Image caption Generator is a popular research area of Artificial Intelligence that deals with image understanding and a language description for that image. Generating well-formed sentences requires both syntactic and semantic understanding of the language. Being able to describe the content of an image using accurately formed sentences is a very challenging task, but it could also have a great impact, by helping visually impaired people better understand the content of images.
- Generating a caption for a given image is a challenging problem in the deep learning domain. In this article, we will use different techniques of computer vision and NLP to recognize the context of an image and describe them in a natural language like English. we will build a working model of the image caption generator by using CNN (Convolutional Neural Networks) and LSTM (Long short term memory) units.
  For training our model I'm using Flickr8K dataset. It consists of 8000 unique images and each image will be mapped to five different sentences which will describe the image.


Input Image
CNN
Feature Map
Context Vector
LSTM
CC1
CC2
CCk
CCi
STV
T-ATT
0.06 0.06 0.04
0.20 0.20 0.20
0.20 0.10 0.04
Attention Weight Map
people playing frisbee game
Image Space
three people playing frisbee game in the park

What is CNN?

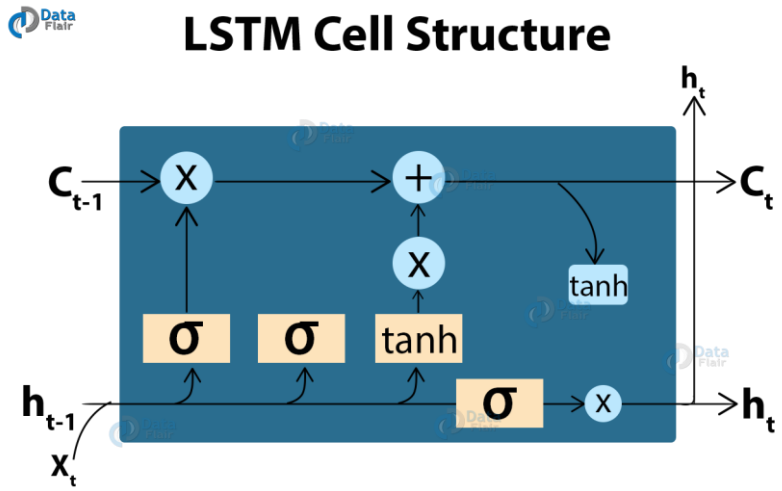
- Convolutional Neural networks are specialized deep neural networks which can process the data that has input shape like a 2D matrix. Images are easily represented as a 2D matrix and CNN is very useful in working with images.
- CNN is basically used for image classifications and identifying if an image is a bird, a plane or Superman, etc.
- It scans images from left to right and top to bottom to pull out important features from the image and combines the feature to classify images. It can handle the images that have been translated, rotated, scaled and changes in perspective.

Working of Deep CNN



What is LSTM?

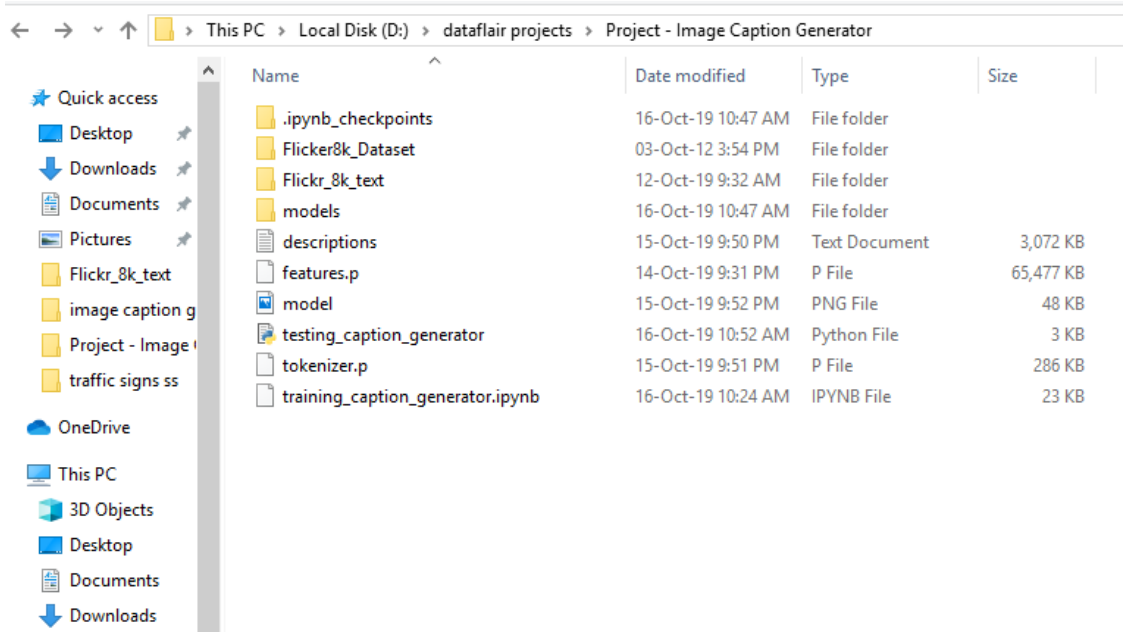
- LSTM stands for **Long short term memory**, they are a type of RNN (**recurrent neural network**) which is well suited for sequence prediction problems. Based on the previous text, we can predict what the next word will be. It has proven itself effective from the traditional RNN by overcoming the limitations of RNN which had short term memory. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information.



Understanding the Dataset

- A number of datasets are used for training, testing, and evaluation of the image captioning methods. The datasets differ in various perspectives such as the number of images, the number of captions per image, format of the captions, and image size. Three datasets: Flickr8k, Flickr30k, and MS COCO Dataset are popularly used.
- In the Flickr8k dataset, each image is associated with five different captions that describe the entities and events depicted in the image that were collected. By associating each image with multiple, independently produced sentences, the dataset captures some of the linguistic variety that can be used to describe the same image.
- Flickr8k is a good starting dataset as it is small in size and can be trained easily on low-end laptops/desktops

- Our dataset structure is as follows:-
- Flick8k/
  - Flick8k_Dataset/ :- contains the 8000 images
  - Flick8k_Text/
    - Flickr8k.token.txt:- contains the image id along with the 5 captions
    - Flickr8k.trainImages.txt:- contains the training image id’s
    - Flickr8k.testImages.txt:- contains the test image id’s



Resources and References

- Create Your Own Image Caption Generator using Keras! (analyticsvidhya.com)
- Image Caption Generator using Deep Learning on Flickr8K dataset – GeeksforGeeks
- Python based Project - Learn to Build Image Caption Generator with CNN & LSTM - DataFlair (data-flair.training)

Image links:

- https://data-flair.training/blogs/wp-content/uploads/sites/2/2019/11/working-of-Deep-CNN-Python-project.png
- https://data-flair.training/blogs/wp-content/uploads/sites/2/2019/11/LSTM-Cell-Structure-project-in-python.png
- https://data-flair.training/blogs/wp-content/uploads/sites/2/2019/11/structure-python-data-science-project.png
- https://artificialintelligence.oodles.io/wp-content/uploads/2020/04/Overall-architecture-for-image-captioning-with-text-guided-attention-Given-an-input.png