IMAGE CAPTION GENERATOR USING CONVOLUTIONAL NEURAL NETWORK (CNN)

(ARTIFICIAL INTELLIGENCE PROJECT)

UNDER THE GUIDANCE OF
Ms. PALUCK ARORA
(ASSISTANT PROFESSOR)

SUBMITTED BY:

APARNA SINGH (102103414)

AKSHAY KHANNA (102103415)

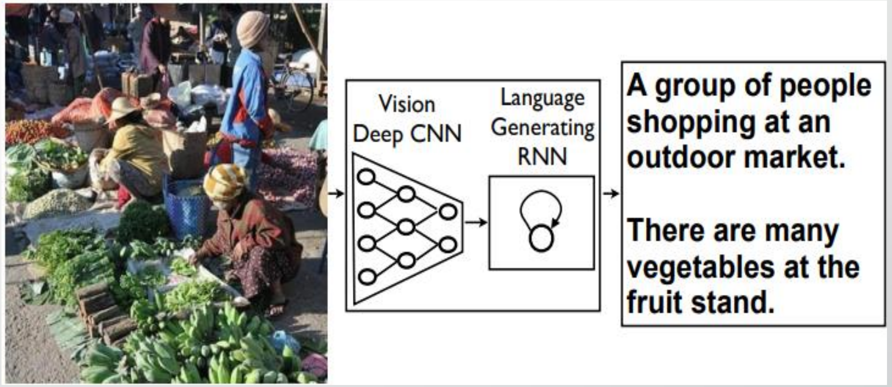
SARANYA (102103416)

ABSTRACT

- Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing.
- Here, we present a generative model that can be used to generate natural sentences describing an image.
- Experiments on datasets such as Flicker8k show the accuracy of the model and the fluency of the language it learns solely from image descriptions.

INTRODUCTION

- To describe the content of an image using properly formed English sentences is challenging, but it has great impact, for instance, helping visually impaired people understand the content of images properly.
- The description must express how these objects
  1. Relate to each other
  2. Relate to their attributes
  3. Relate to the activities they are involved in.

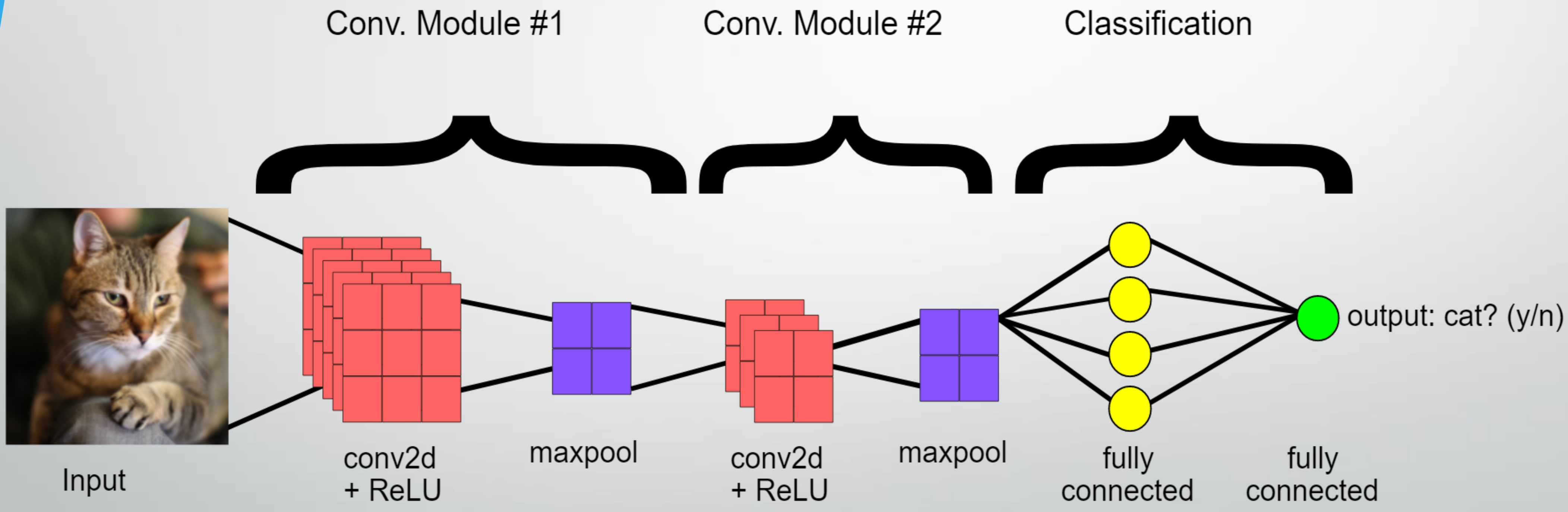

Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

- In our model, we use
  1. CNN for analysing and processing the image, and
  2. LSTM (a variety of RNN) for generating suitable captions for the same
- Both features are concatenated to predict the next word for the caption.
- We take help of BLEU score as a metric to evaluate the performance of our model

CNN (Convolutional Neural Network)

- CNN is a subfield of Deep learning used for the recognition and classification of images.
- It acts as an encoder and is used to process the data represented as a 2D matrix like images.
- It can deal with scaled, translated, and rotated imagery.
- It analyses the visual imagery by scanning them from left to right and top to bottom and extracting relevant features from that.
- Finally, it combines all the features for image classification.

BASIC ARCHITECTURE OF CONVOLUTIONAL NEURAL NETWORK (CNN)



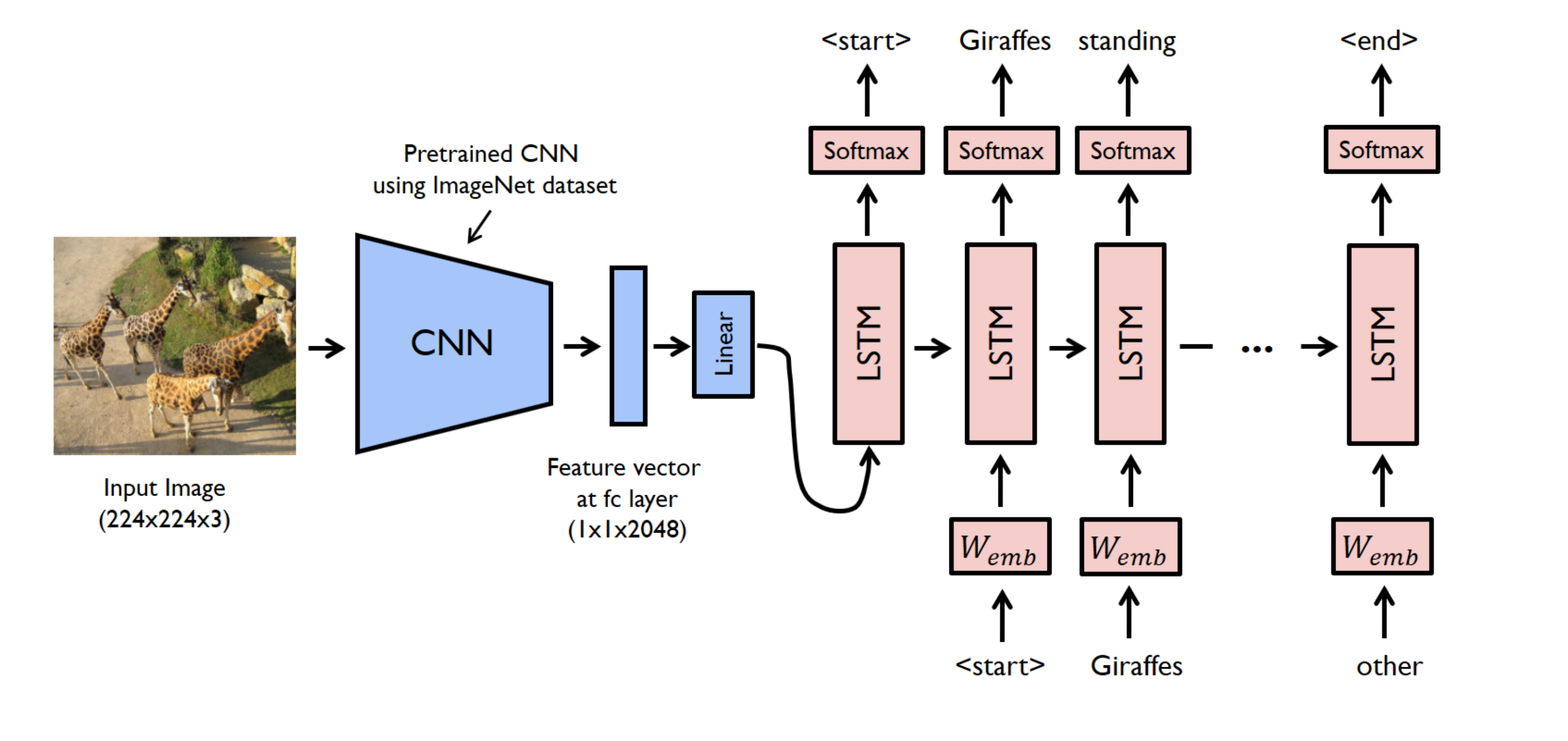
LSTM(LONG TERM SHORT MEMORY)

- Being a type of RNN (recurrent neural network), LSTM is capable of working with sequence prediction problems.
- Throughout the processing of inputs LSTM is used to carry out the relevant information and to discard non-relevant information using a forget gate.
- LSTM is way more effective and better compared to the traditional RNN as it overcomes the short term memory limitations of the RNN.
- It is mostly used for the next word prediction purposes, for instance, in Google search, system shows the next word based on the previous text.

Working of LSTM

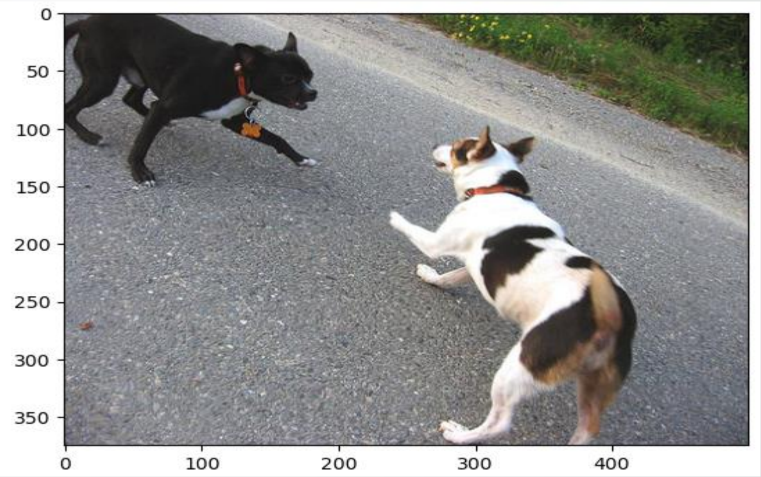
- The core of the LSTM model is a memory cell encoding knowledge at every time step of what inputs have been observed up to this step.
- The behaviour of the cell is controlled by “gates”.
- It can either keep a value from the gated layer if the gate is 1 or reject it if the gate is 0.

BASIC ARCHITECTURE OF LONG TERM SHORT MEMORY (LSTM)



DATASETS

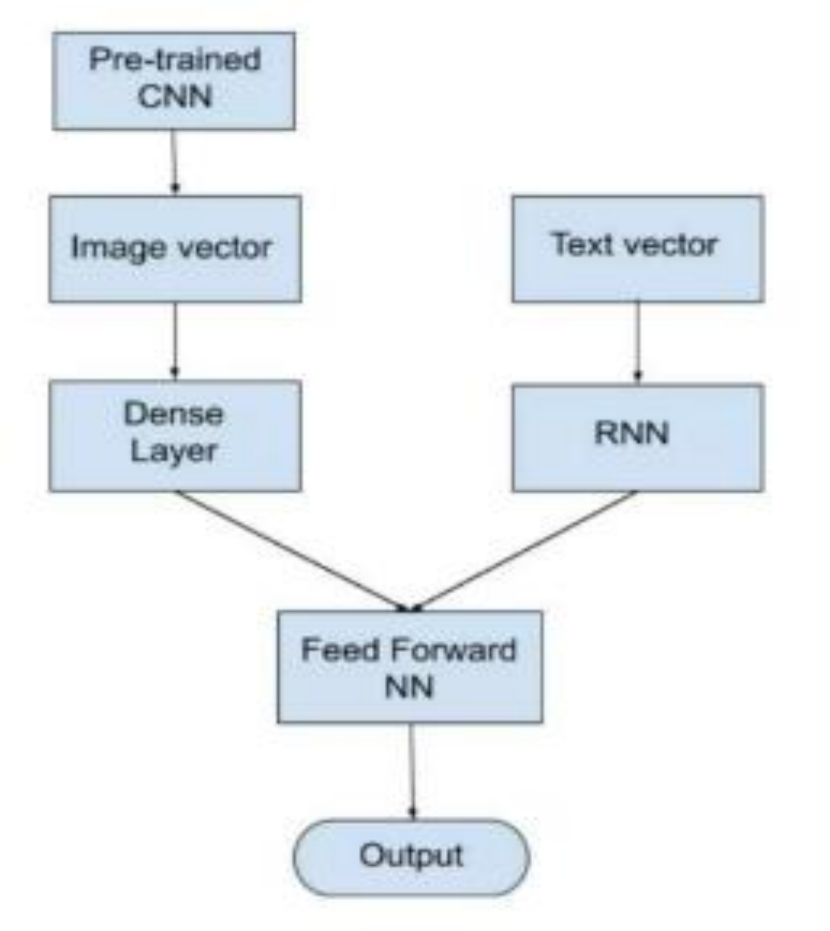
- We used Flicker8K dataset for this project.
- Flicker8K project contains a variety of images depicting scenes and situations.
- The dataset consists of 8091 images and each image has 5 corresponding descriptions.
- We split the data into 600, 1000 & 1000 images as training, validation and testing sets respectively.
- The images are of different dimensions.



- Black dog and spotted dog are fighting
- Black dog and tri-coloured dog playing with each other on the road
- Black dog and white dog with brown spots are staring at each other in the street
- Two dogs of different breed looking at each other on the road
- Two dogs on pavement moving toward each other

ARCHITECTURE

- We used an encoder decoder architecture.
- 2048 image vector is fed to a dense layer to generate 256 length image vector.
- 34 length word vector is fed to LSTM to output 256 length word vector.
- Decoder model adds both the encoder outputs and is fed to dense 256 layers.
- The last Dense layer will have as many nodes as the vocabulary size.
- The last softmax layer predicts the next word present in the output vocabulary.



EVALUATION METRICS

Bilingual Evaluation Understudy Score (BLEU)

- BLEU is a metric for evaluating a generated sentence to a reference sentence
- BLEU scores lie between 0 and 1.

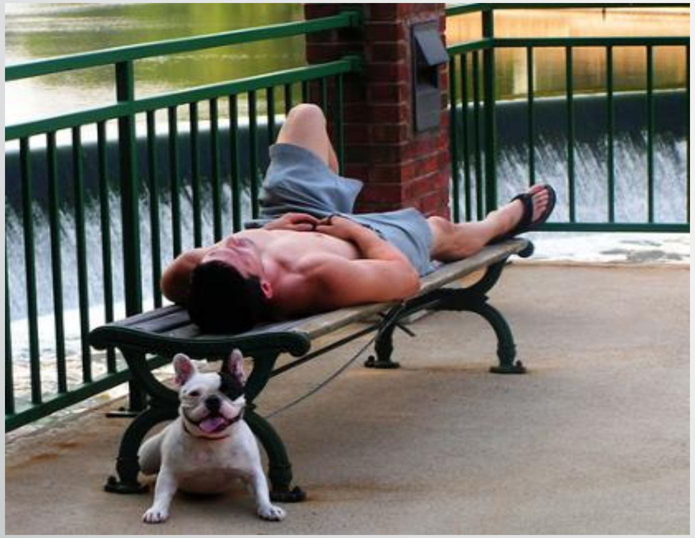
LSTM(Long Short Term Memory)

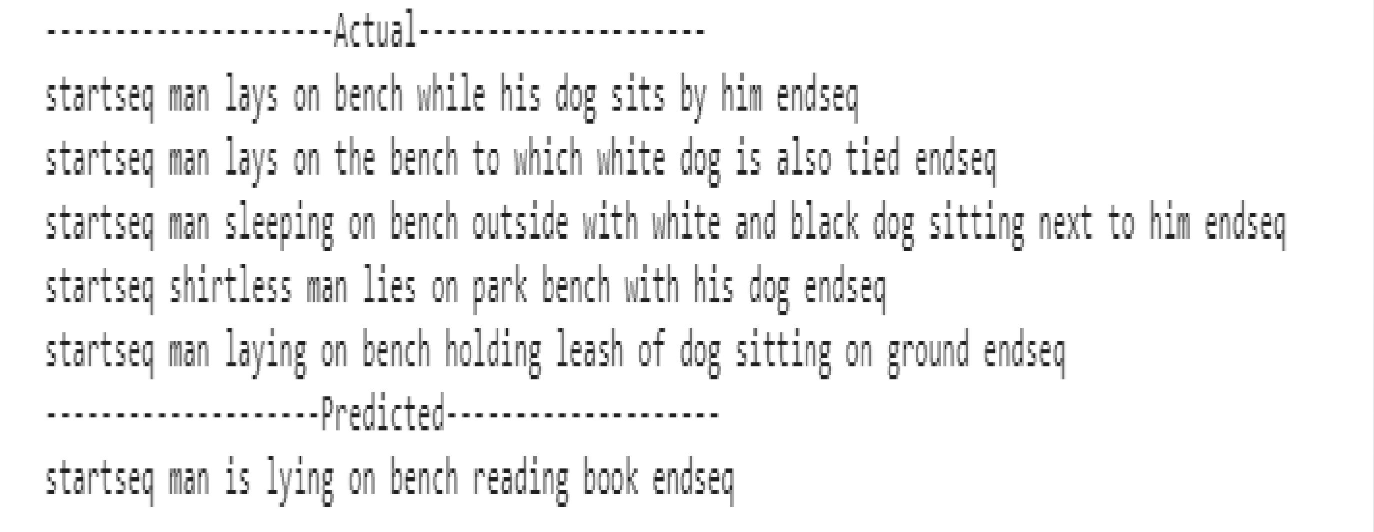
BLEU N-GRAM	SCORE
BLEU-1	0.572214
BLEU-2	0.339204
BLEU-3	0.237129
BLEU-4	0.116733

Simple RNN(Recurrent Neutral Network)

BLEU N-GRAM	SCORE
BLEU-1	0.364472
BLEU-2	0.181942
BLEU-3	0.103185
BLEU-4	0.085675

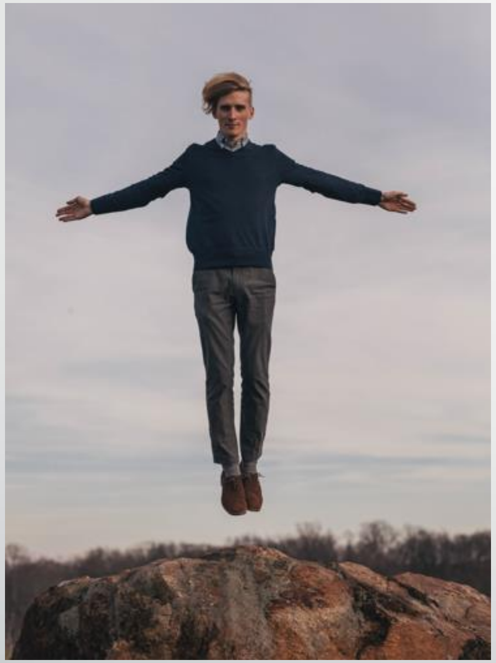
RESULTS



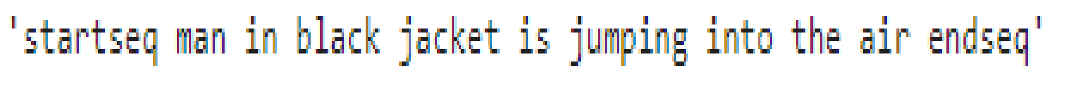


```
--------------------Actual--------------------
startseq man lays on bench while his dog sits by him endseq
startseq man lays on the bench to which white dog is also tied endseq
startseq man sleeping on bench outside with white and black dog sitting next to him endseq
startseq shirtless man lies on park bench with his dog endseq
startseq man laying on bench holding leash of dog sitting on ground endseq
--------------------Predicted--------------------
startseq man is lying on bench reading book endseq
```

TESTING WITH REAL IMAGE



Predicted Output :



```
'startseq man in black jacket is jumping into the air endseq'
```