

Department of Electronics and Communication Engineering
MNNIT Allahabad
Session 2021-2022

IMAGE CAPTION GENERATOR USING MACHINE LEARNING

Under Guidance of
Prof. V K SRIVASTAVA



Project Presented by Group of

Divas Gupta (20185116)

Priyanka Soni (20185053)

Dipesh Sharma Poudel (20185105)

Contents

Introduction

Motivation/uses

Dataset

Why 5 descriptions

Deep learning model

Image Based Model

Convolutional neural network

VGG-16 Model

Language Based Model

LSTM model

Model of Image Captioning

Flow chart of the process

Tech stacks

Text Pre-processing

Evaluation Metrics

Efficiency of Model

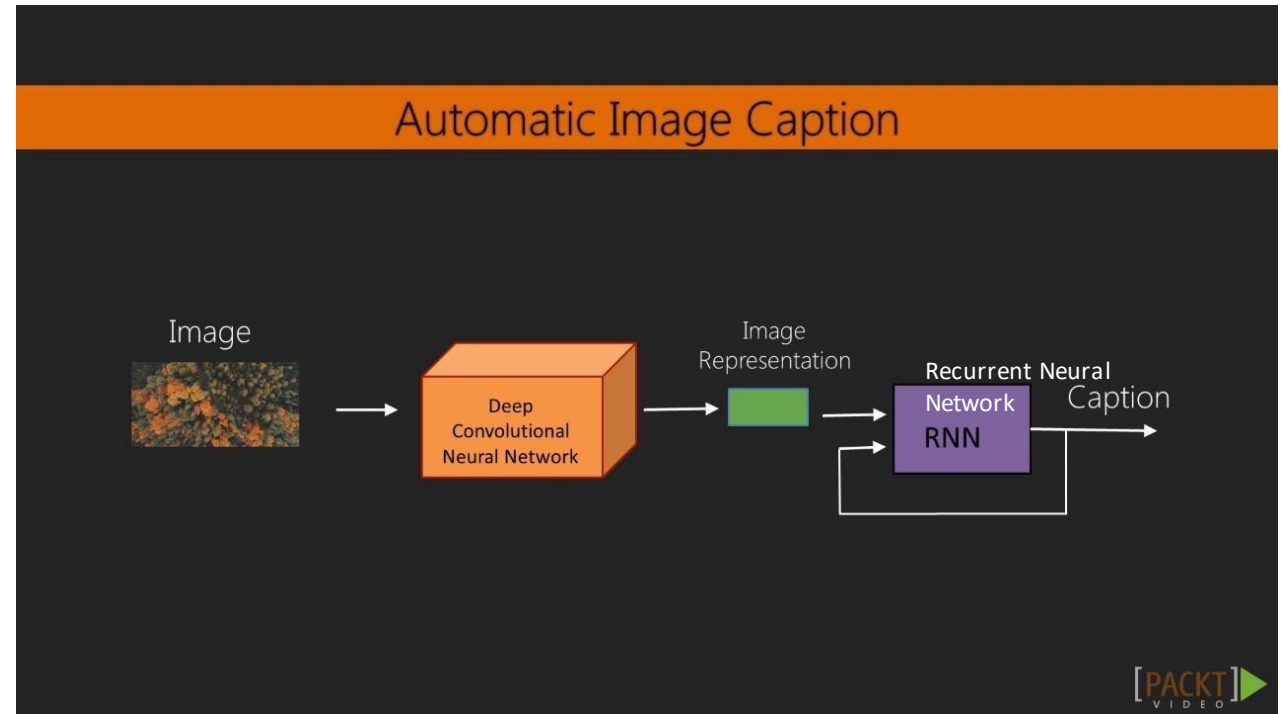
Results

Future Scope of Work

References

Introduction

- Automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines knowledge of computer vision and natural language processing.
- Goal is to generate a concise description of an input image.





MOTIVATION/USES



IMAGE SEARCH
ENGINE



GUIDANCE FOR
BLINDS



SELF DRIVING
CAR

DATASET (Flickr8k)

We are using Flickr8k Dataset for the project.

The dataset consists of 8000 images and each image has 5 corresponding descriptions.

We split the data into 6000, 1000 & 1000 images as training, validation and testing sets respectively.

WHY 5 DESCRIPTIONS?

- Well some will say "A white dog in grassy area" ,
- Some may say "white dog with brown spots" and
- Others might Say "A dog on grass and some pink flowers"



DEEP LEARNING MODEL

1. Image based model: We are using CNN to extract the features of an image.

2. Language based model: We are using LSTM to translates the features and objects extracted by image based model to a natural sentence.

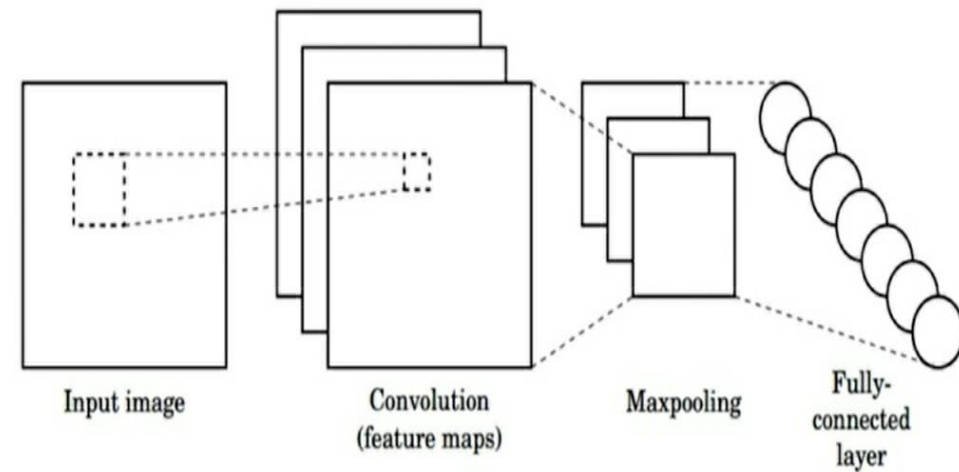
IMAGE BASED MODEL

Resized all images to a fixed size of 224×224 pixels.


Used pre-built keras VGG16 (3*3 layers , CNN based) 16 weighted layers model for features extraction of images.


Convolutional Neural Network


CNNs for Classification



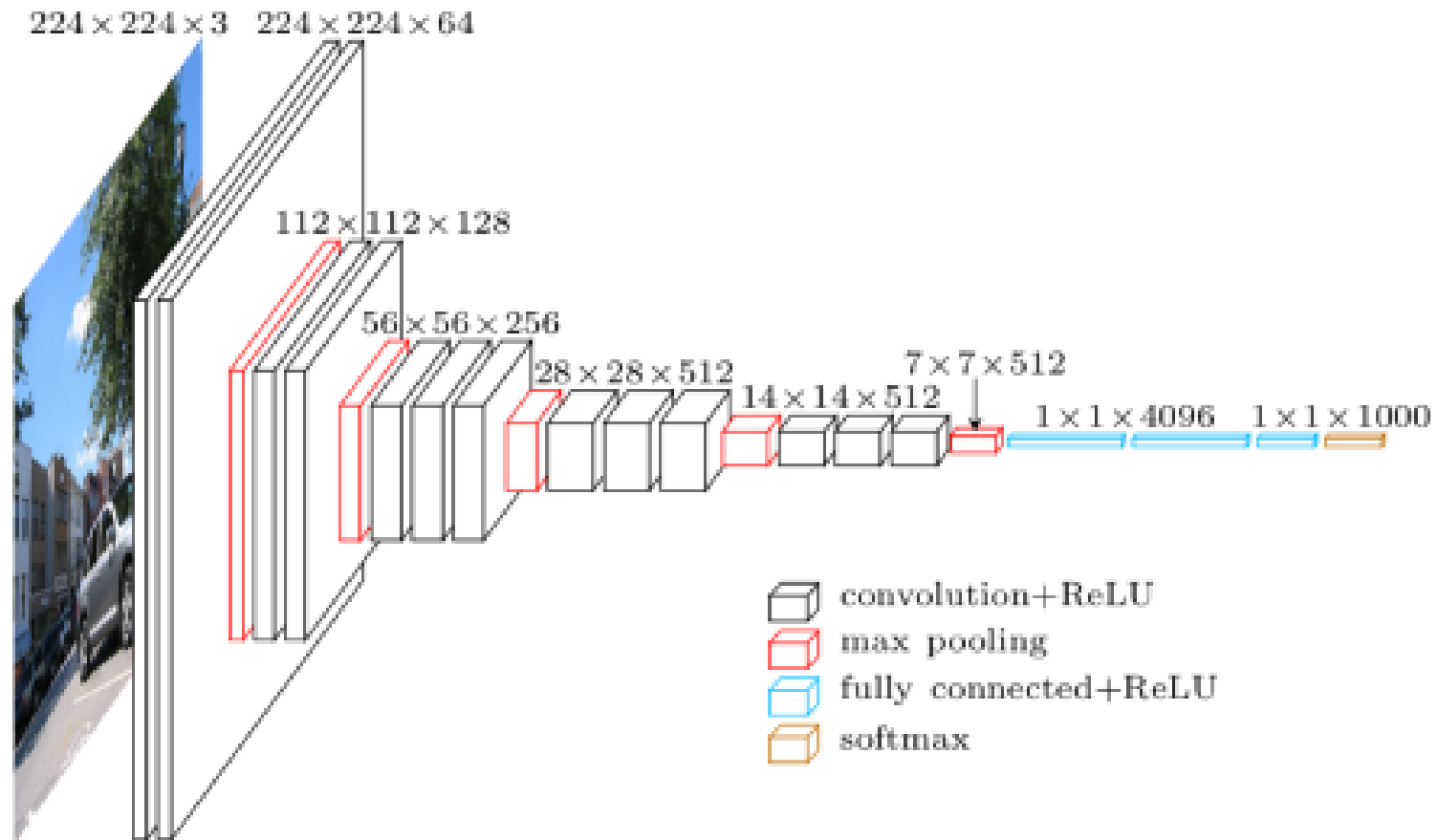
1. **Convolution:** Apply filters to generate feature maps.
2. **Non-linearity:** Often ReLU.
3. **Pooling:** Downsampling operation on each feature map.

 `tf.keras.layers.Conv2D`

 `tf.keras.activations.*`

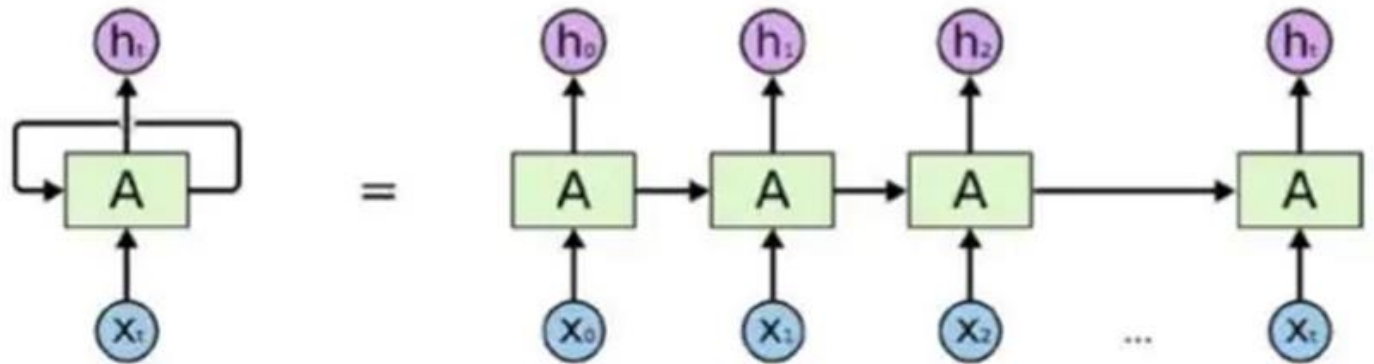
 `tf.keras.layers.MaxPool2D`

VGG16 Architecture



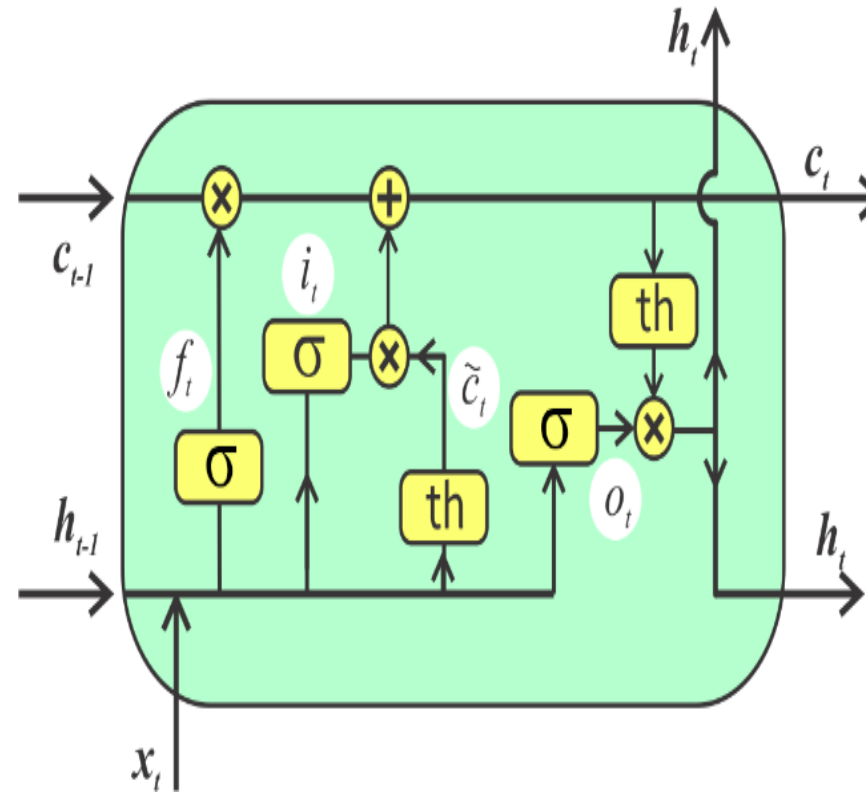
LANGUAGE BASED MODEL

- Language Generation is a serial task we generate words after word.
- This is well modeled by Recurrent Neural Cells: a neuron that uses itself over and over again to accept serial inputs, outputting each time a new value,



Long Short-term Memory (LSTM)

- A common LSTM unit is composed of a **cell**, an **input gate**, an **output gate** and a **forget gate**. The cell remembers values over arbitrary time intervals and the three *gates* regulate the flow of information into and out of the cell.

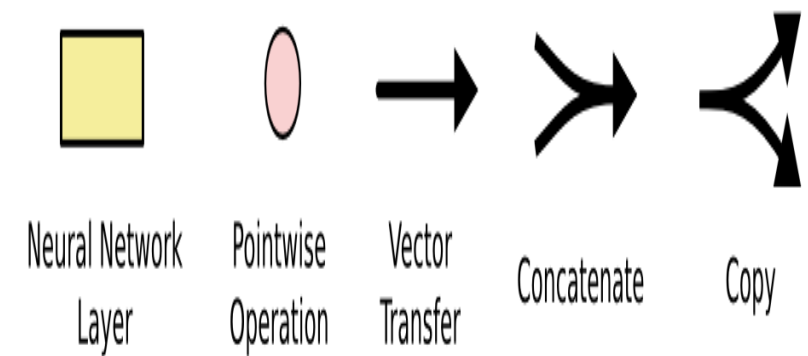
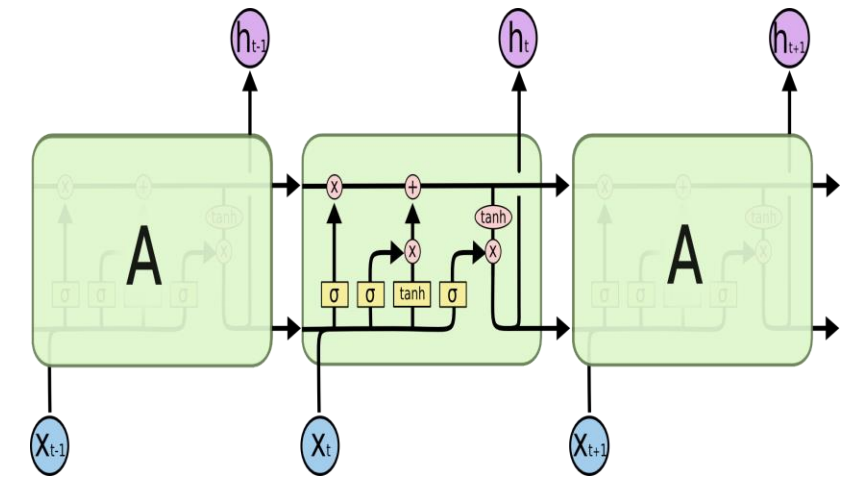


Legend:

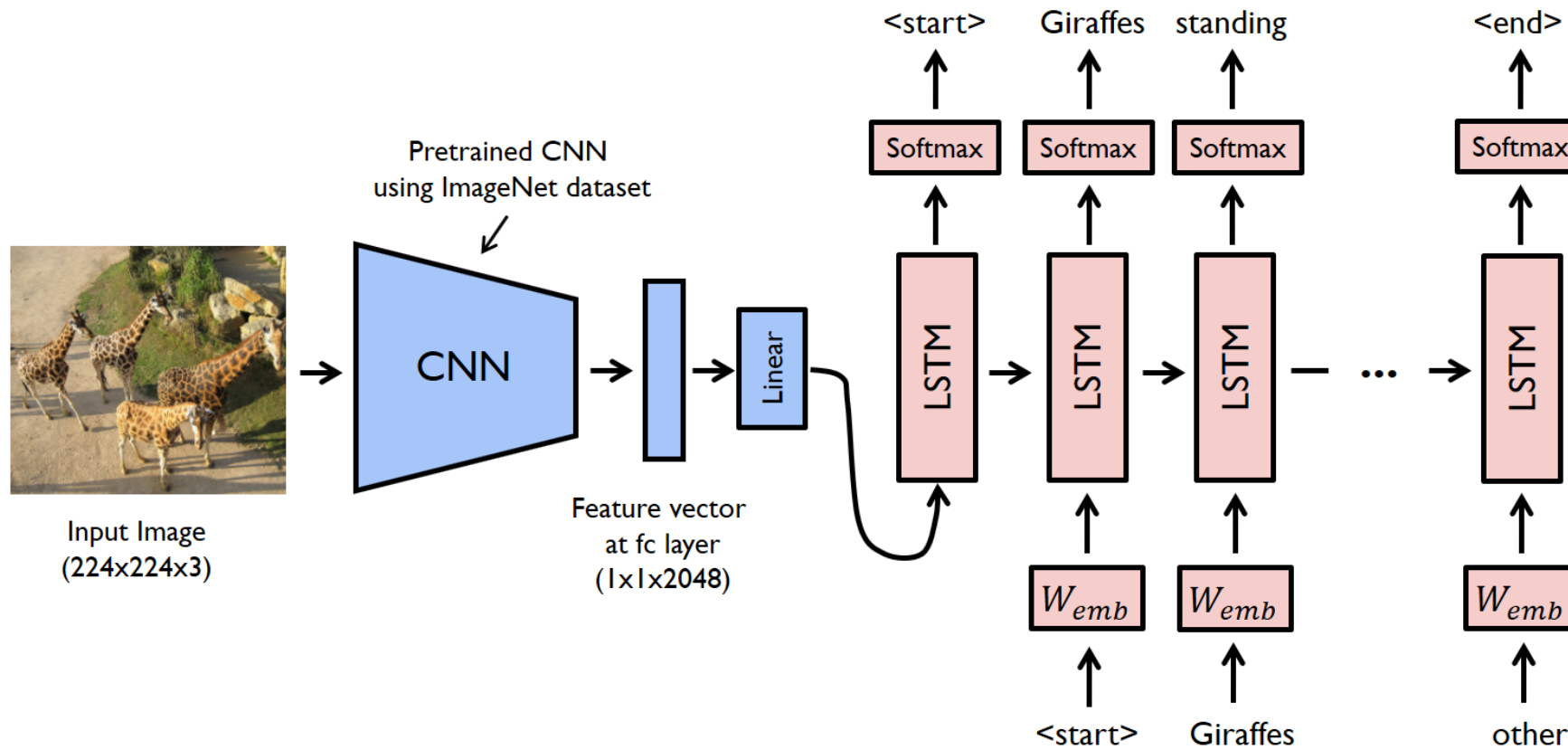
x_t input
 f_t forget gate
 i_t input gate
 \tilde{c}_t cell update
 c_t cell state
 o_t output gate
 h_t output

LSTMs: Key Concepts

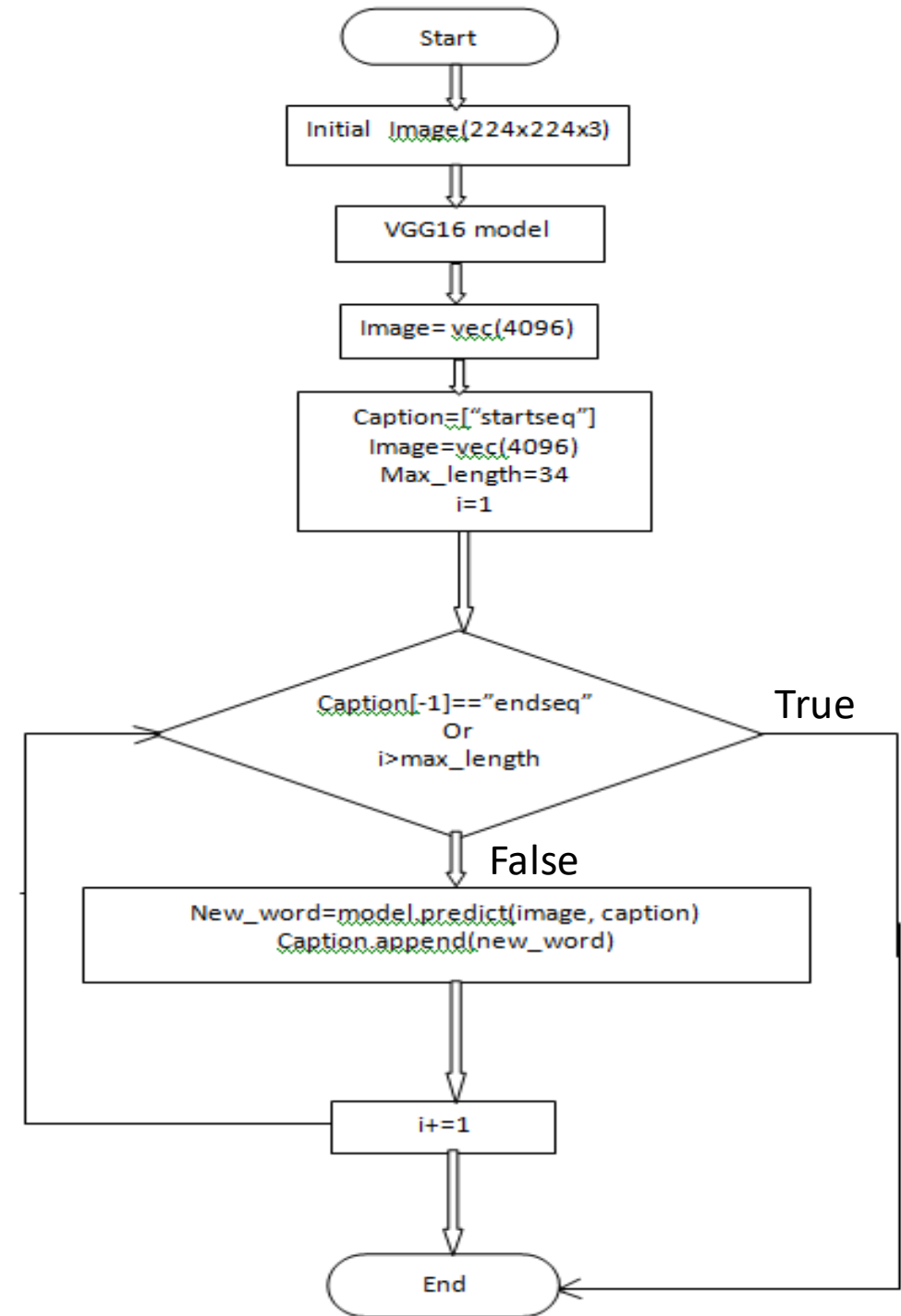
1. Maintain a **separate cell state** from what is outputted
2. Use **gates** to control the **flow of information**
 - **Forget** gate gets rid of irrelevant information
 - **Store** relevant information from current input
 - Selectively **update** cell state
 - **Output** gate returns a filtered version of the cell state
3. Backpropagation through time with **uninterrupted gradient flow**



MODEL OF IMAGE CAPTIONING



Flow chart of the Process



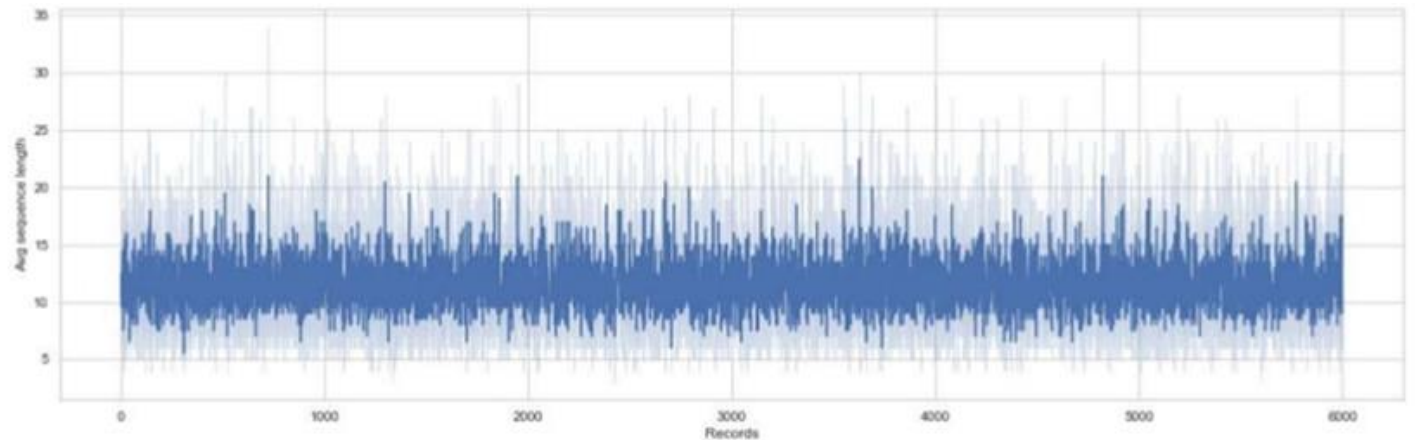
TRAINING LSTM MODEL

- For each image we trained the model by temporally injecting incremental sequences of the description.

Image	Partial Caption	Target Word
Image	startseq	a
Image	startseq a	young
Image	startseq a young	boy
.....
Image	startseq a young boy wearing a helmet and riding a bike in a park	endseq

TEXT PRE-PROCESSING

- Each description is tokenized and converted to lowercase.
- Removed alpha-numeric characters and punctuation marks.
- We use startseq and endseq as prefix and postfix for each caption respectively.
- Filtered out unique words from corpus and represented each word by a vector
- To generate a fixed length word vector we calculated the maximum length caption.



DESCRIPTIONS AFTER PRE-PROCESSING

File Edit Format Run Options Window Help

1000268201_693b08cb0e.jpg	child in pink dress is climbing up set of stairs in
1000268201_693b08cb0e.jpg	girl going into wooden building
1000268201_693b08cb0e.jpg	little girl climbing into wooden playhouse
1000268201_693b08cb0e.jpg	little girl climbing the stairs to her playhouse
1000268201_693b08cb0e.jpg	little girl in pink dress going into wooden cabin
1001773457_577c3a7d70.jpg	black dog and spotted dog are fighting
1001773457_577c3a7d70.jpg	black dog and tricolored dog playing with each other
1001773457_577c3a7d70.jpg	black dog and white dog with brown spots are staring
1001773457_577c3a7d70.jpg	two dogs of different breeds looking at each other
1001773457_577c3a7d70.jpg	two dogs on pavement moving toward each other
1002674143_1b742ab4b8.jpg	little girl covered in paint sits in front of paint
1002674143_1b742ab4b8.jpg	little girl is sitting in front of large painted red
1002674143_1b742ab4b8.jpg	small girl in the grass plays with fingerpaints in
1002674143_1b742ab4b8.jpg	there is girl with pigtails sitting in front of railing
1002674143_1b742ab4b8.jpg	young girl with pigtails painting outside in the grass
1003163366_44323f5815.jpg	man lays on bench while his dog sits by him

TECH STACKS

Python

TensorFlow

Keras (VGG16 model)

Numpy

EVALUTION METRICS

Bilingual Evaluation Understudy Score(BLEU)

- BLEU is a metric for evaluating a generated sentence to a reference sentence. BLEU Score lies between 0-1.
- Example:-

Actual sentence: *The quick brown fox jumped over the lazy dog.*

Predicted sentence: *The quick black fox run to the lazy dog.*

BLEU-1 Score: $6/9 = 0.667$ (*The, quick, fox, the, lazy, dog* => 6 out of 9)

BLEU-2 Score: $3/5 = 0.6$ (*The quick, the lazy, dog* => 3 out of 5)

Efficiency of model

Dataset: 6000

Descriptions: train=6000

Vocabulary Size: 7579

Description Length: 34

Dataset: 1000

Descriptions: test=1000

Photos: test=1000

BLEU-1: 0.529691

BLEU-2: 0.277107

BLEU-3: 0.182869

BLEU-4: 0.080318



RESULTS



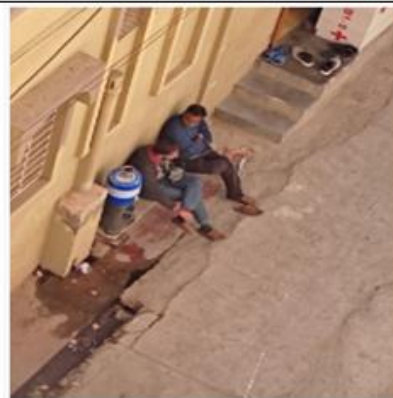
Result: man in red shirt is riding bike on the street



Result: man in black shirt is standing next to crowd of people



Result: dog is running through the grass



Result: man is sitting on the street

Future Scope of Work

We can combine a wide variety of training data in order to capture as many activities as possible since they are finite in number.

Use of an audio device to read the generated caption to help visually impaired to get better understanding of images in front of them.

We can also develop an android application for our project.

References

- [1] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, “Boosting image captioning with attributes,” in *Proceedings of the IEEE Conference on International Conference on Computer Vision*, pp. 4904–4912, Las Vegas, NV, USA, June 2016.
- [2] Karpathy, Andrej, and Li Fei-Fei. “Deep visual semantic alignments for generating image descriptions” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128-3137. 2015.
- [3] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, “Areas of attention for image captioning,” in *Proceedings of the IEEE Conference on International Conference on Computer Vision*, pp. 1251–1259, Venice, Italy, October 2017.
- [4] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. “Show and tell: A neural image caption generator.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156- 3164. 2015.
- [5] Donahue, Jeffrey, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. “Long-term recurrent convolutional networks for visual recognition and description.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625- 2634. 2015.

Thank You!