ImageSpeak : Generating Captions from Pixels

PRESENTED BY

- Shivani (21BCS6285)
- Ashmandeep Kaur (21BCS6284)
- Tarushi Sandeep Gupta (21BCS6280)

Introduction

Image captioning is the task of generating descriptive and appropriate sentences of a given image.

An image caption generator helps computers understand pictures and write descriptions about them, almost like a human would.

The image caption generator is like a teamwork between two parts of artificial intelligence – one that looks at pictures (computer vision) and the other that understands language (natural language processing).

Project scope

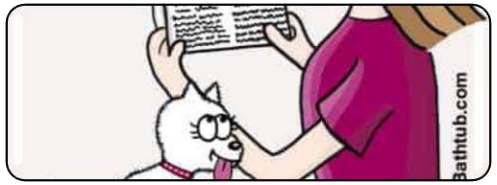


Image caption generators can automatically generate captions for images used in articles, blogs, and websites. This helps in enhancing the content's accessibility and engagement, as well as improving its search engine optimization.

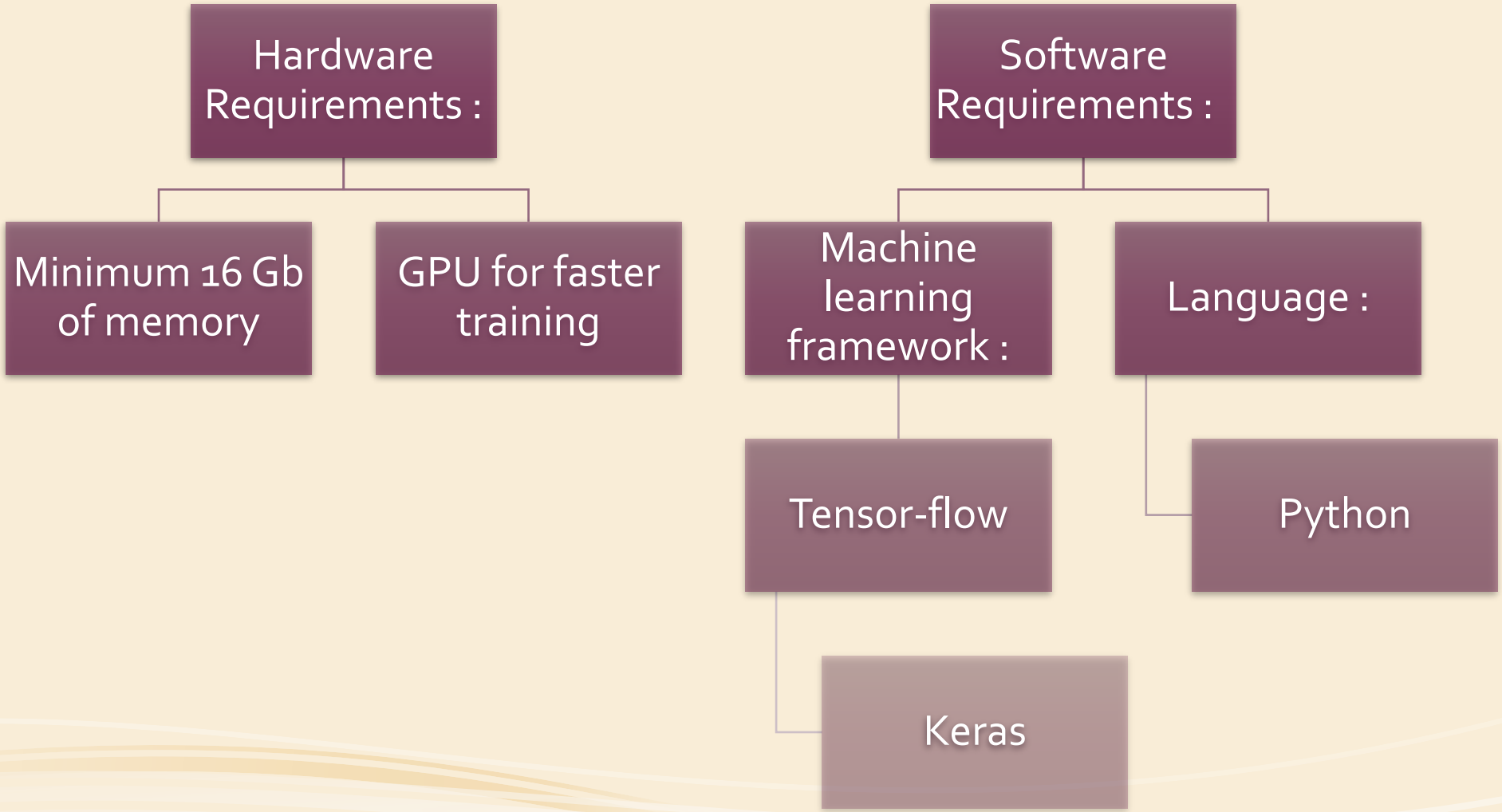
Online shopping can be enhanced by generating detailed captions for product images, providing potential buyers with more information and a better understanding of products.

With image captions, search engines can better understand the content of images, improving image search results and making it easier to find relevant visuals.

Image captioning can be integrated into AI assistants to provide more informative and contextually relevant responses, particularly when dealing with image-related queries.

On social media platforms, image caption generators can generate creative and contextually relevant captions for images, making posts more engaging and shareable.

Requirements



Data we worked on

There are many open-source datasets available for this problem, like : -

COCO stands for Common Objects and Contexts, and it contains a large variety of images.

- Flickr 8k (containing 8k images),
- Flickr 30k (containing 30k images),
- MSCOCO (containing 180k images), etc.

a tan dog is playing in the grass
a tan dog is playing with a red ball in the grass
a tan dog with a red collar is running in the grass

a yellow dog runs through the grass
a yellow dog is running through the grass
a brown dog is running through the grass

a group of people stand in front of a building
a group of people stand in front of a white building
a group of people stand in front of a large building

a man and a woman walking on a sidewalk
a man and a woman stand on a balcony
a man and a woman standing on the ground

Methodology

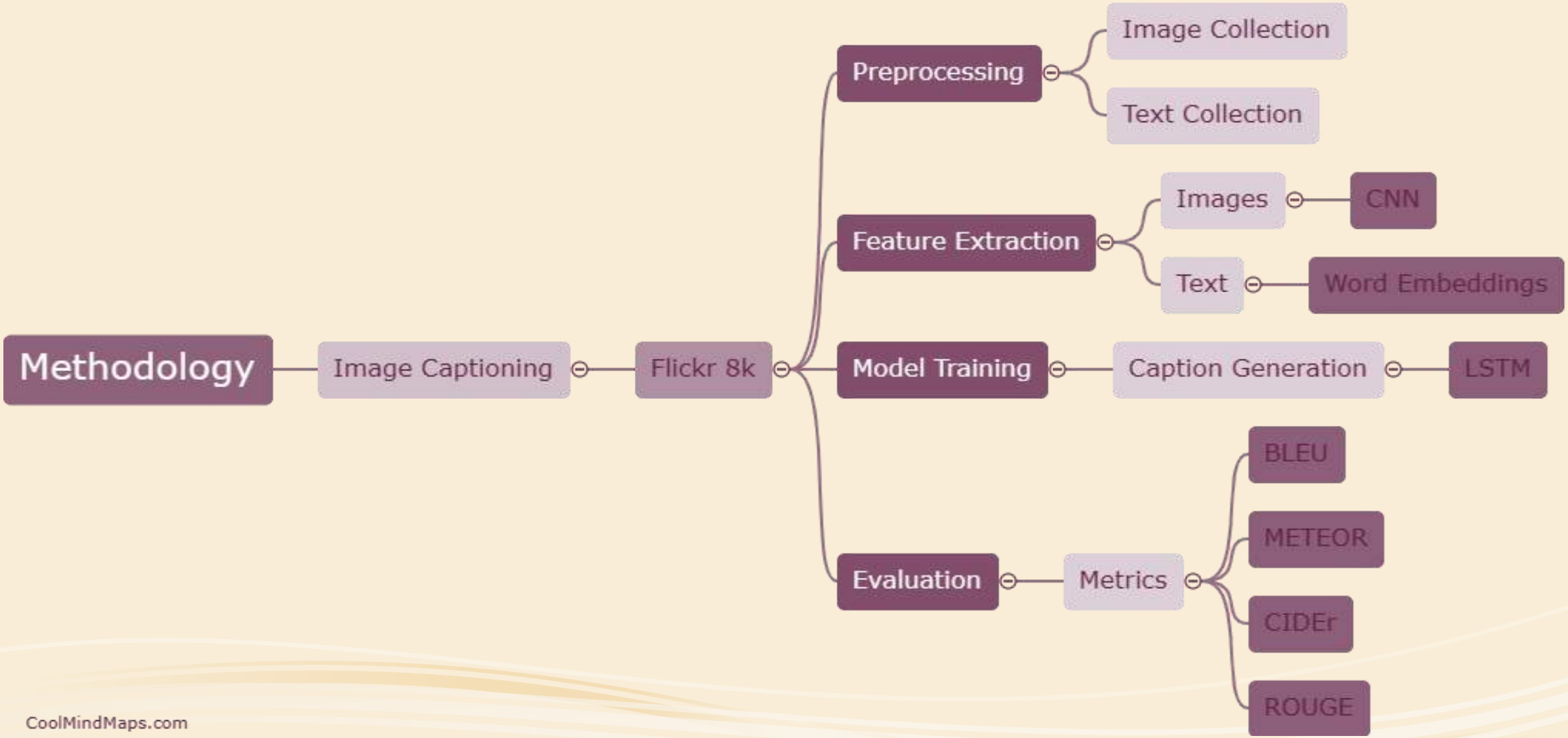
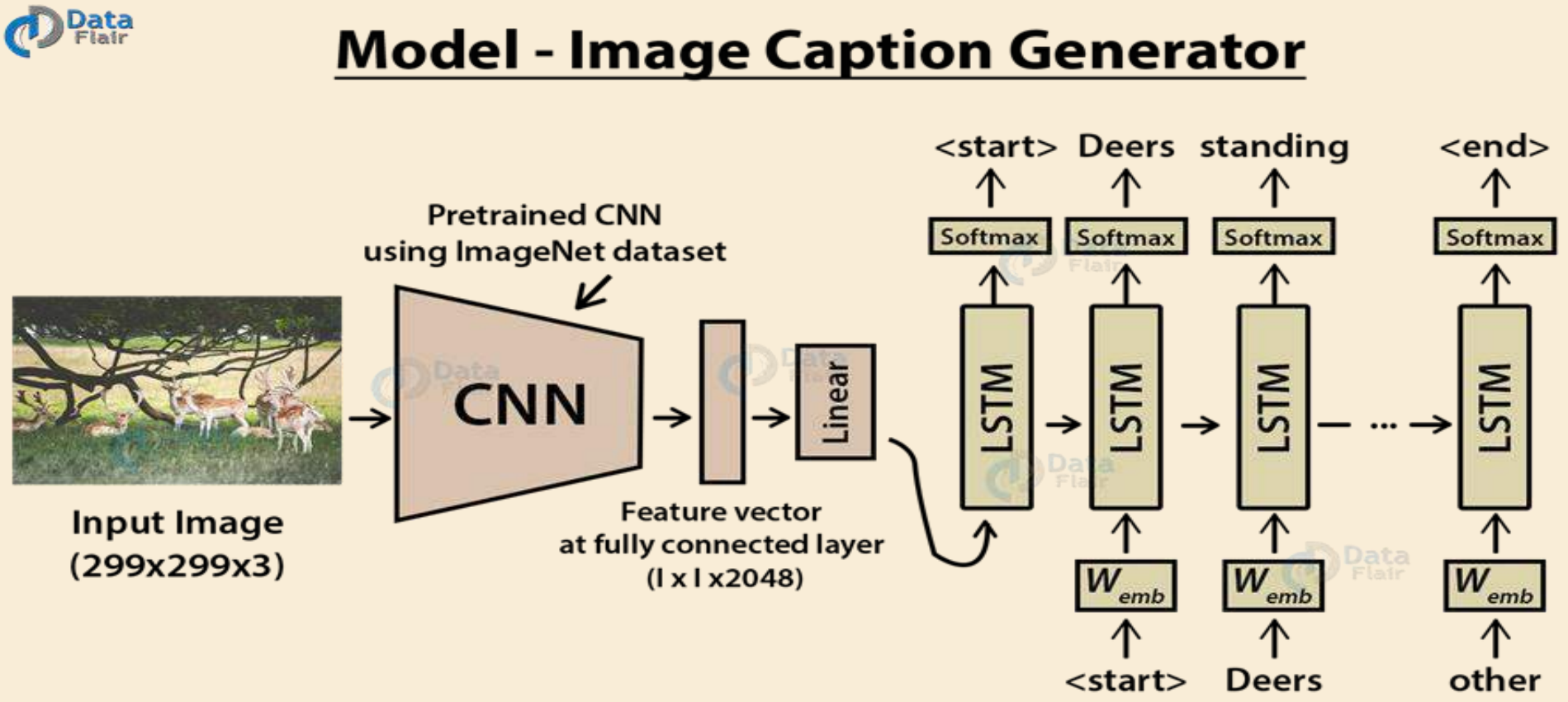


Image Caption Generator Model

1. CNN
2. LSTM

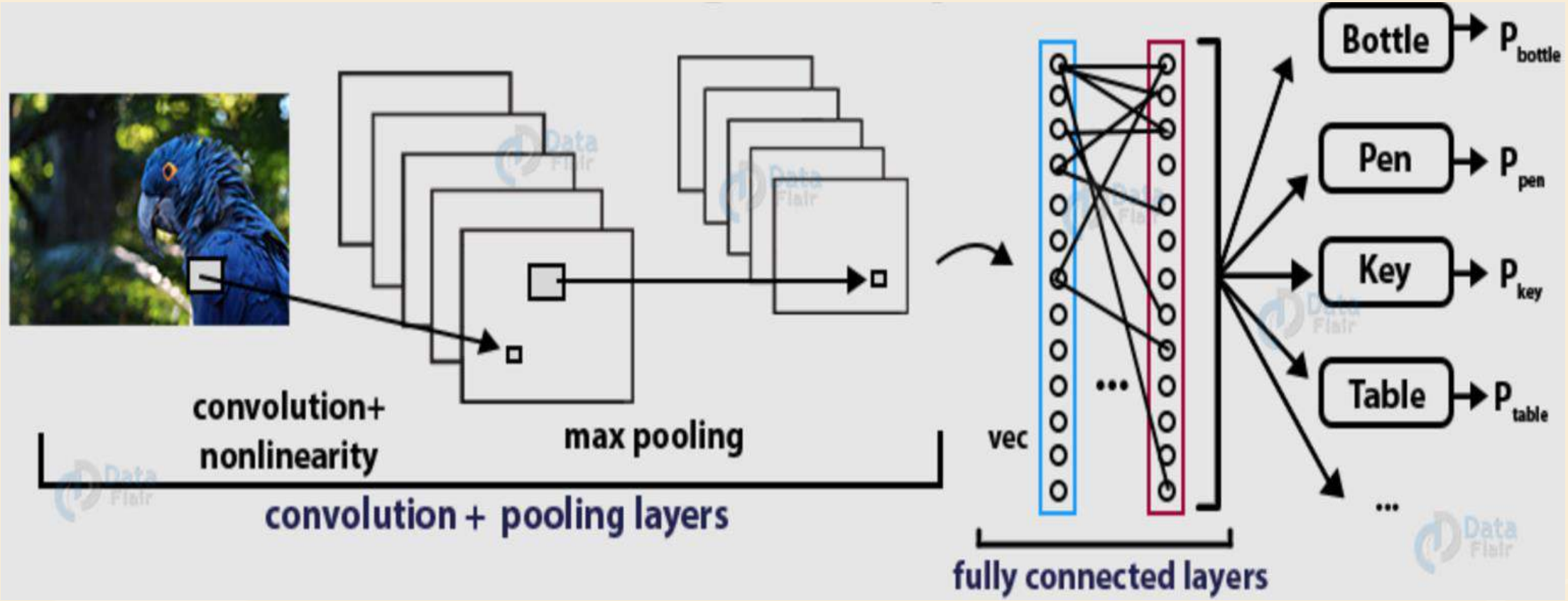
Design Flow


Model - Image Caption Generator
Pretrained CNN
using ImageNet dataset
CNN
Linear
Input Image
(299x299x3)
Feature vector
at fully connected layer
(I x I x2048)
<start>
Deers
standing
<end>
Softmax
Softmax
Softmax
Softmax
LSTM
LSTM
LSTM
LSTM
W emb
W emb
W emb
<start>
Deers
other

What is CNN?

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

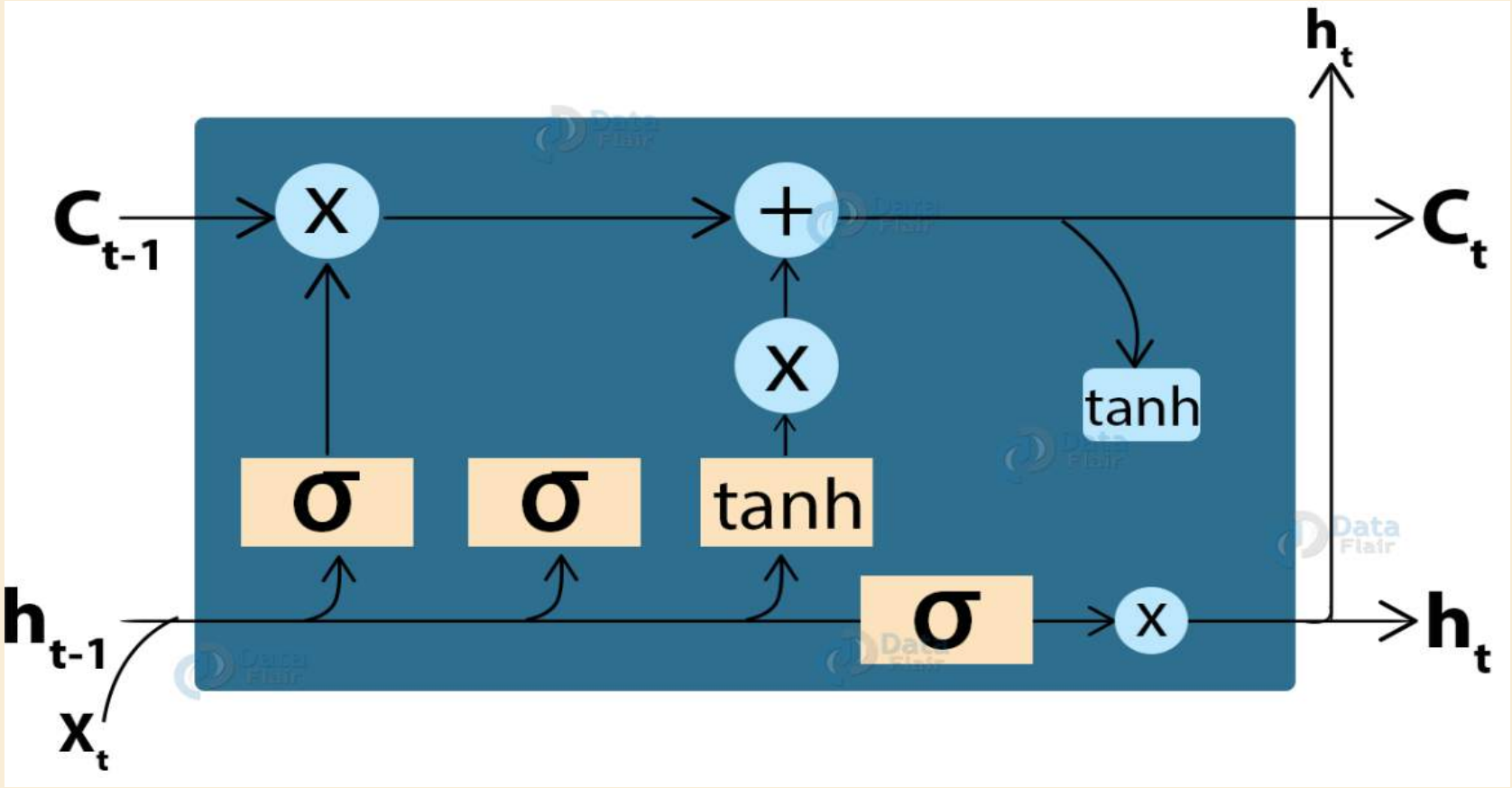
Working on deep CNN



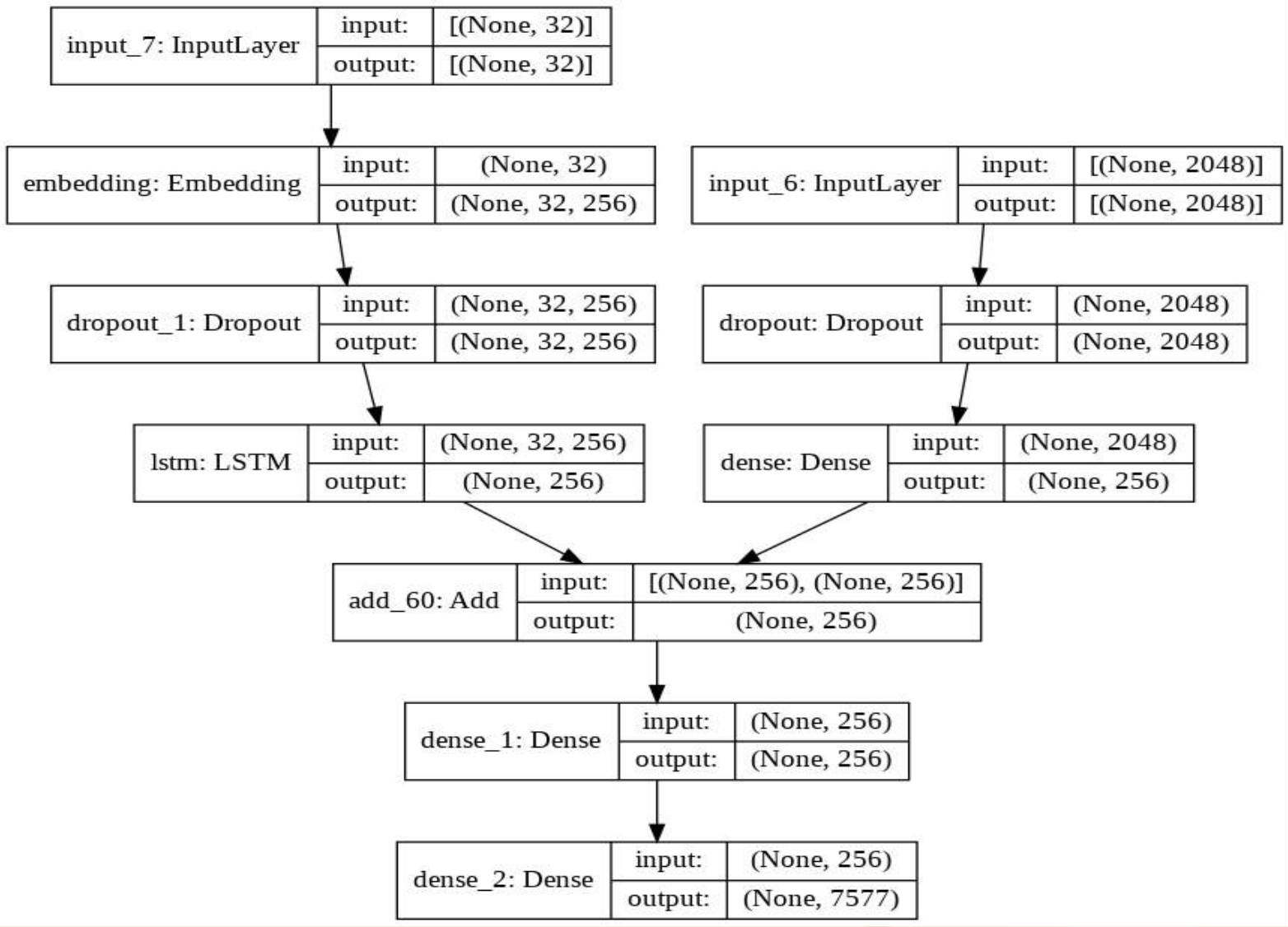
What is LSTM?

LSTM stands for Long short-term memory, they are a type of RNN, that is well suited for sequence prediction problems. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information.

LSTM cell structure



Visual representation of the final model



Result

```
from PIL import Image
img = Image.open('/content/drive/MyDrive/ML/Flicker8k_Datase
img
```

```
pciBusID: 0000:00:04.0 name: Tesla T4 computeCapability: 7.5
coreClock: 1.59GHz coreCount: 40 deviceMemorySize: 14.75GiB devi
2021-06-25 22:56:00.968023: I tensorflow/stream_executor/cuda/cu
2021-06-25 22:56:00.968415: I tensorflow/stream_executor/cuda/cu
2021-06-25 22:56:00.968773: I tensorflow/core/common_runtime/gpu
2021-06-25 22:56:00.968819: I tensorflow/stream_executor/platfor
2021-06-25 22:56:01.556718: I tensorflow/core/common_runtime/gpu
2021-06-25 22:56:01.556777: I tensorflow/core/common_runtime/gpu
2021-06-25 22:56:01.556797: I tensorflow/core/common_runtime/gpu
2021-06-25 22:56:01.556963: I tensorflow/stream_executor/cuda/cu
2021-06-25 22:56:01.557421: I tensorflow/stream_executor/cuda/cu
2021-06-25 22:56:01.557859: I tensorflow/stream_executor/cuda/cu
2021-06-25 22:56:01.558223: W tensorflow/core/common_runtime/gpu
2021-06-25 22:56:01.558287: I tensorflow/core/common_runtime/gpu
2021-06-25 22:56:03.862377: I tensorflow/compiler/mlir/mlir_grap
2021-06-25 22:56:03.862834: I tensorflow/core/platform/profile_u
2021-06-25 22:56:14.028765: I tensorflow/stream_executor/platfor
2021-06-25 22:56:14.526717: I tensorflow/stream_executor/cuda/cu
2021-06-25 22:56:15.375889: I tensorflow/stream_executor/platfor
2021-06-25 22:56:15.851255: I tensorflow/stream_executor/platfor


start man is climbing up the side of cliff end
```

Challenges in captioning images from Flickr 8k

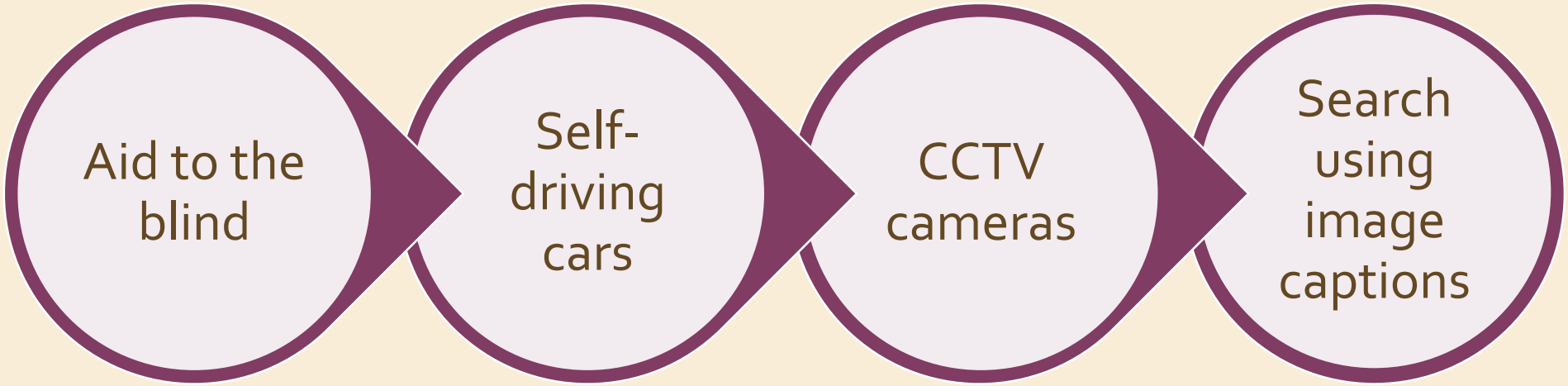
- Large dataset (over 8,000 images) requires significant processing power and time.

- Diverse image types (landscapes, objects, people) need adaptable captioning models.

- Ambiguous and subjective images demand contextual understanding.

- Generating descriptive, concise, and linguistically correct captions is complex.

Real Life Use Cases Of Image Captions



Future Scope

Real-Time Processing: Develop models and techniques that enable real-time caption generation for live video streams, making it applicable for applications like video captioning and live event coverage.

Explainable AI: Develop techniques to provide explanations for the generated captions, enabling users to understand how the model arrived at its descriptions.

Emotion and Sentiment Analysis: Incorporate emotion and sentiment analysis into the image caption generator to generate captions that not only describe the visual content but also convey the emotional tone of the scene.

Thank You