Automatic Image Caption Generator

Chun-Min Jen

Sep. 26th, 2020

Motivation -

What do you see from this photo?

1. A white dog in a grassy area
2. White dog with brown spots
3. A dog on grass and some pink flowers

can you write a computer program that takes an image as input and produces a relevant caption as output?

Applications (I)

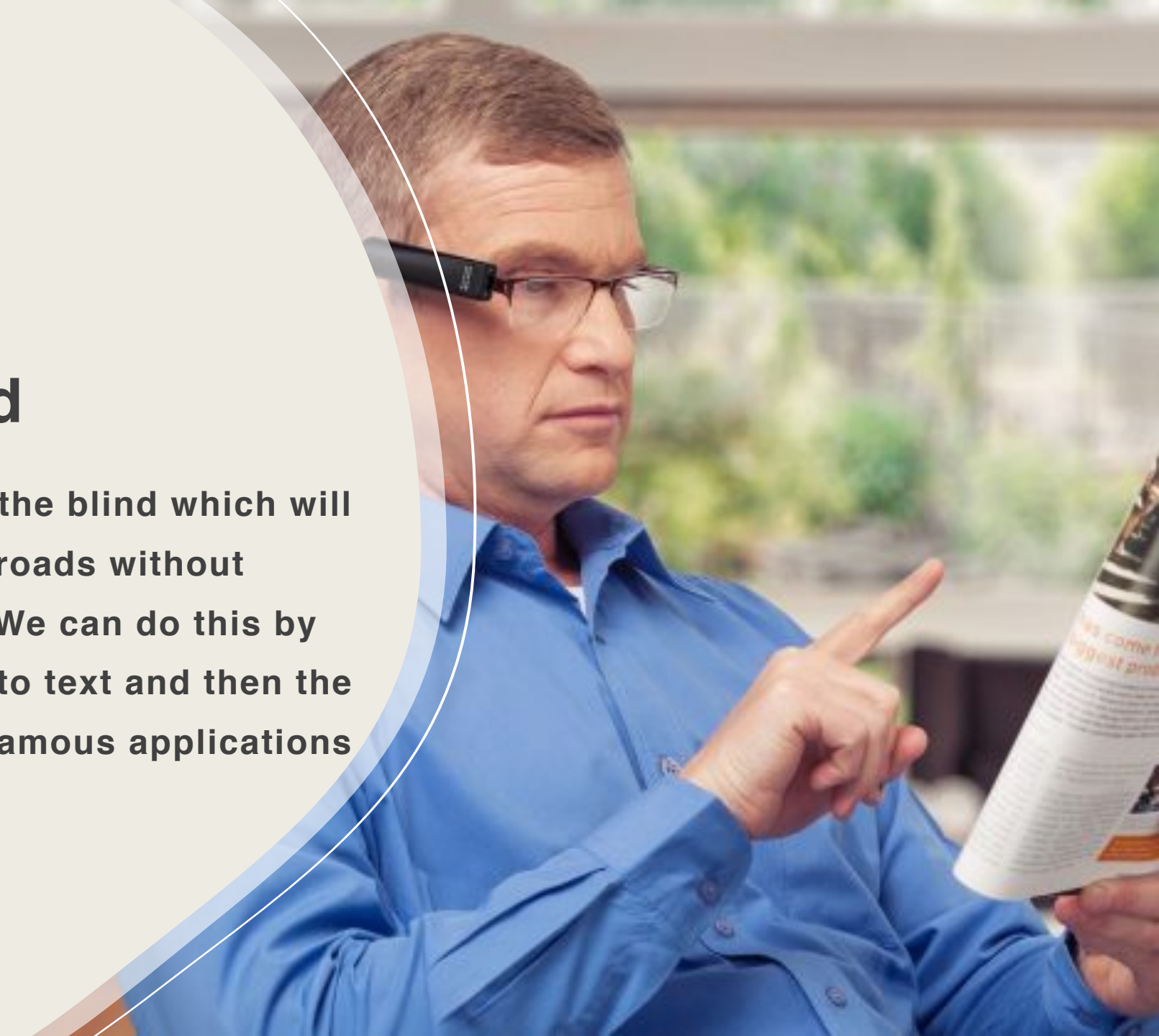
❖ Self-driving cars

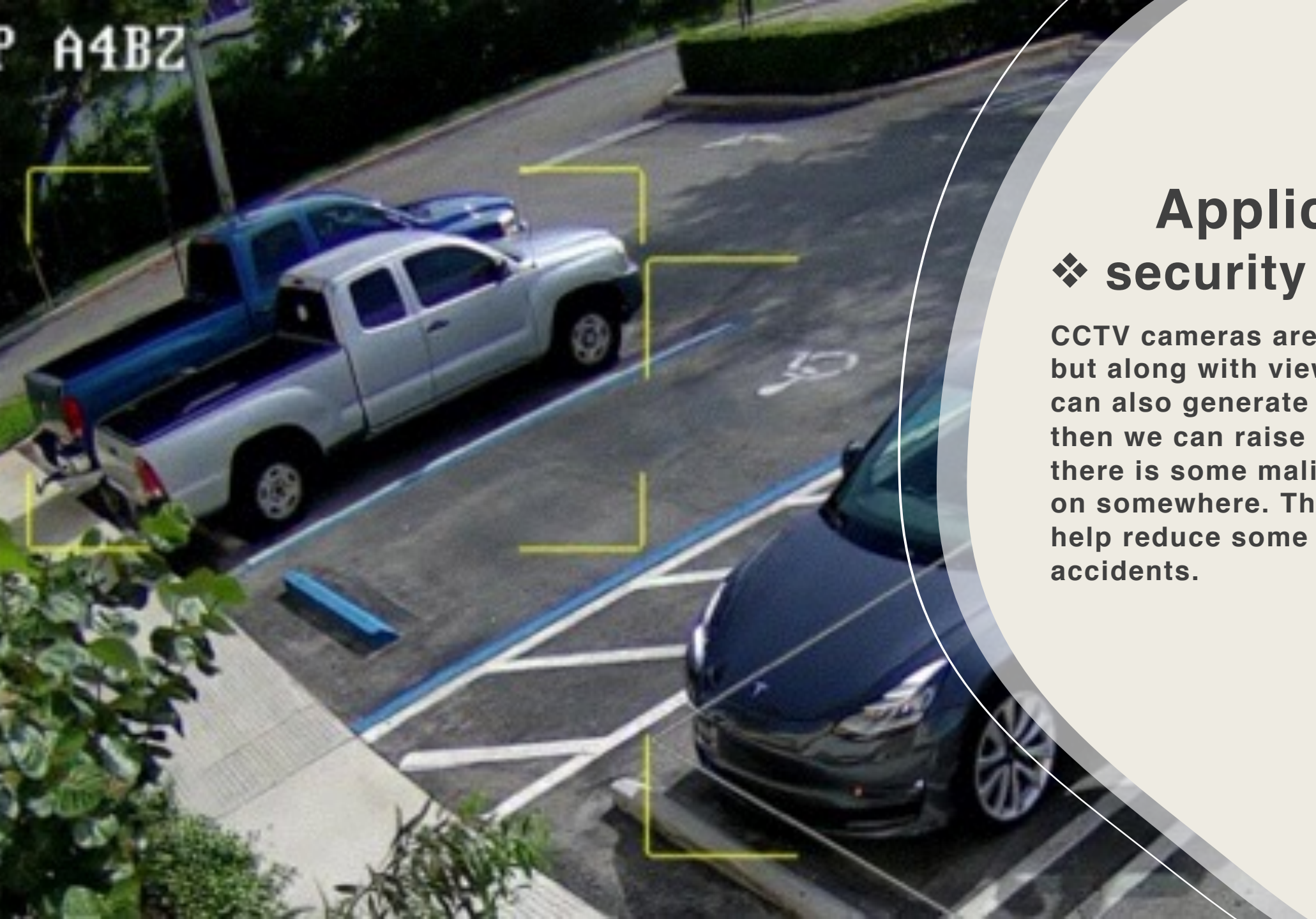
Automatic driving is one of the biggest challenges and if we can properly caption the scene around the car, it can give a boost to the self-driving system.

Applications (II)

❖ Aid to the blind

We can create a product for the blind which will guide them traveling on the roads without the support of anyone else. We can do this by first converting the scene into text and then the text to voice. Both are now famous applications of Deep Learning.



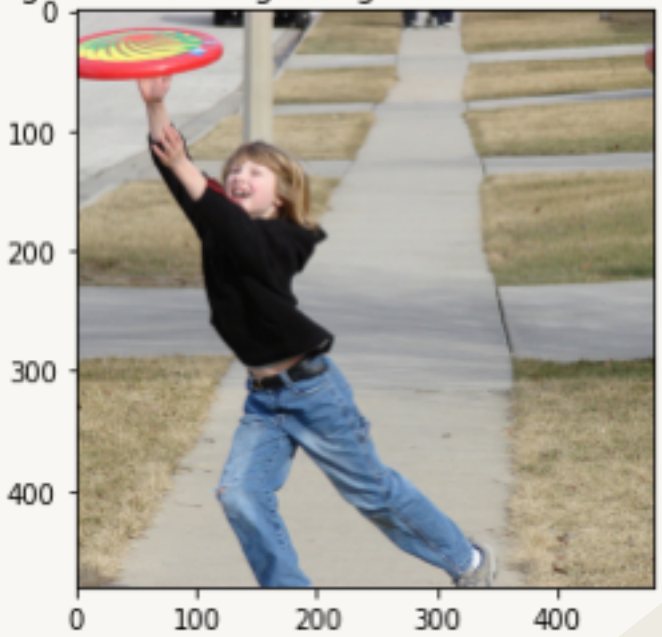


Applications (III)

❖ security camera

CCTV cameras are everywhere today, but along with viewing the world, if we can also generate relevant captions, then we can raise alarms as soon as there is some malicious activity going on somewhere. This could probably help reduce some crime and/or accidents.

A kid makes a side jump to catch a frisbee.
A child playing outside on the side walk with a frisbee
A child jumping up to catch a frisbee.
A child on a sidewalk catching a frisbee.
A young child catching a large red frisbee on a sidewalk.



A tree fallen down by a stop sign in front of a house.
A fallen down tree in front of a stop sign.
A tree has fallen down next to a stop sign.
A Stop sign is slightly covered up by a tree.
A fallen tree sitting next to a stop sign.

Data Collection

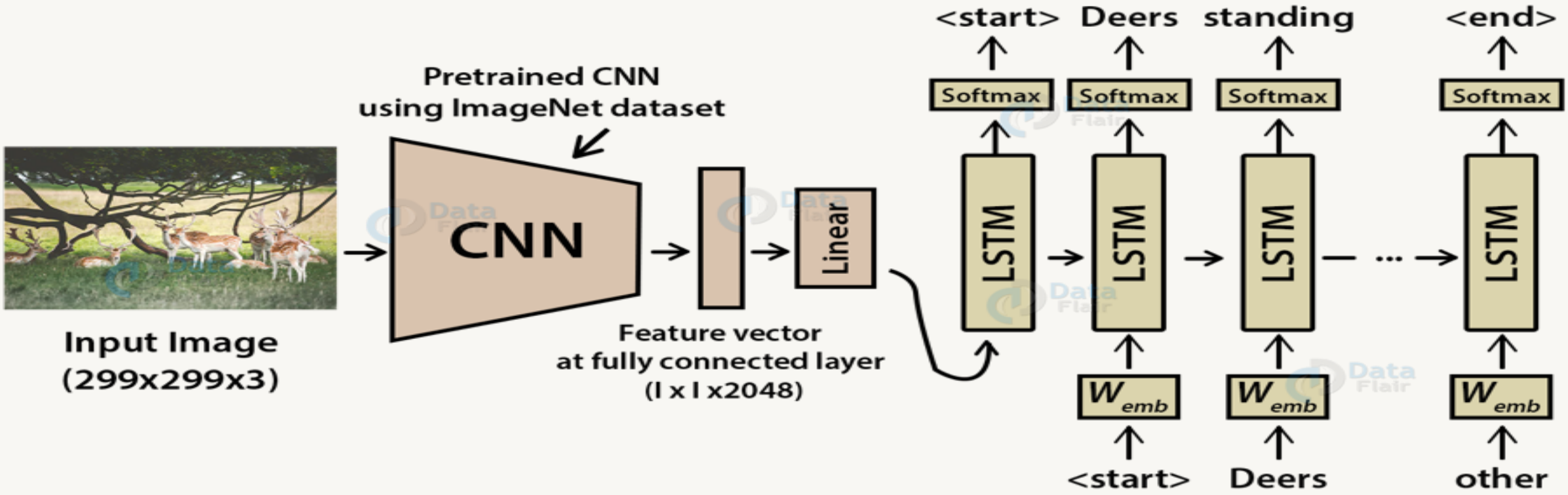
This dataset contains 123,287 images each with 5 captions (as we have already seen in the Introduction section that an image can have multiple captions, all being relevant simultaneously).

These images are bifurcated as follows:

- **Training Set — 82,783 images**
- **Validating Set — 40,504 images**

The caption files:

- **Training Set – 82,783 files**
- **Validating Set – 40,504 files**



Model Architecture

Recommendation - I

- **Evaluation of image embedding :**

Convolutional architectures which are tailored for a number of conventional problems in vision such as, image categories, fine-grained recognition, content-based retrieval, and various aspect of object recognition are needed.

1) **Image classification**
2) **Content-based image retrieval**
3) **Key-points regression**

Recommendation - II

- **Improvement on image embedding :**

Recalling the conventional sliding window, plus classifier approach culminating in Viola-Jones detector. Tracing the development of deep convolutional detectors up until recent days, we can consider R-CNN and single shot detector models.

1) **Sliding window detector**
2) **Employment of modern detector architecture, e.g. Region-based CNN (R-CNN)**

Recommendation - III

- **Inclusion of more features for image embedding :**

Considering video analysis and including material on optical flow estimation, visual object tracking, and action recognition. Motion is a central topic in video analysis, opening many possibilities for end-to-end learning of action patterns and object signatures.

1) **Object tracking**
2) **Action recognition**

Future Work

- A finer CNN model is needed
- Up-to-date RNN models, e.g. BERT and TRANSFORMER, can be considered

Thank you for your attention!

renjmindy@gmail.com

Appendix