
Image
Captioning
With Machine Learning

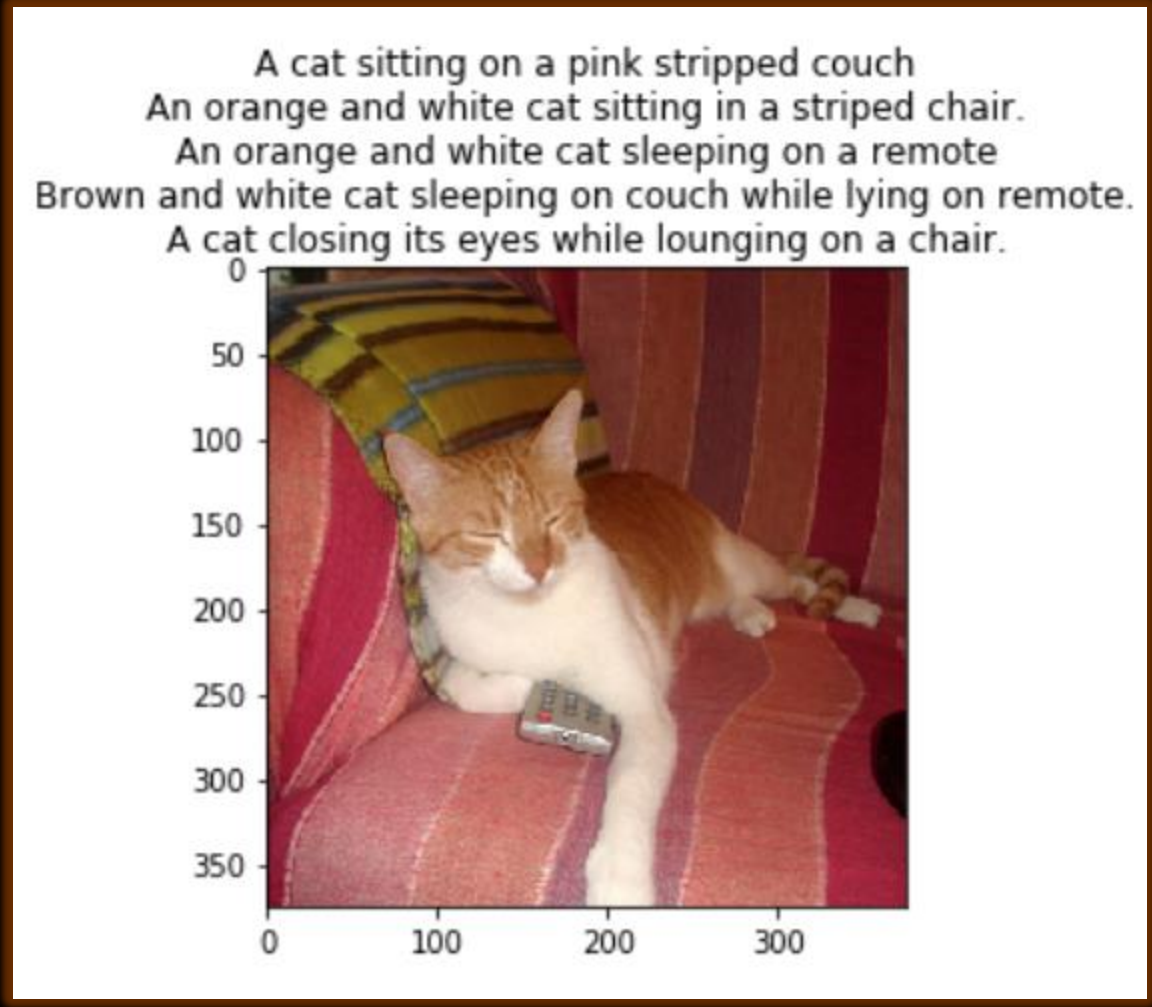
Our goal:

- To train a model that given an image, can automatically generate a caption for it

a dog is playing with a frisbee in the grass

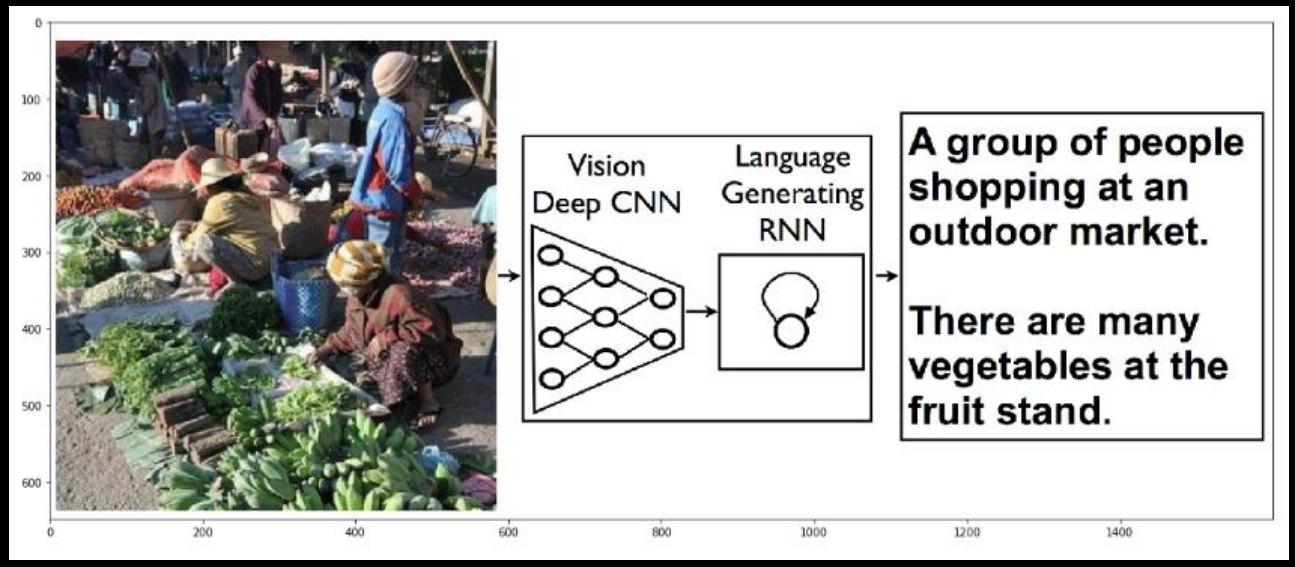
Our Data

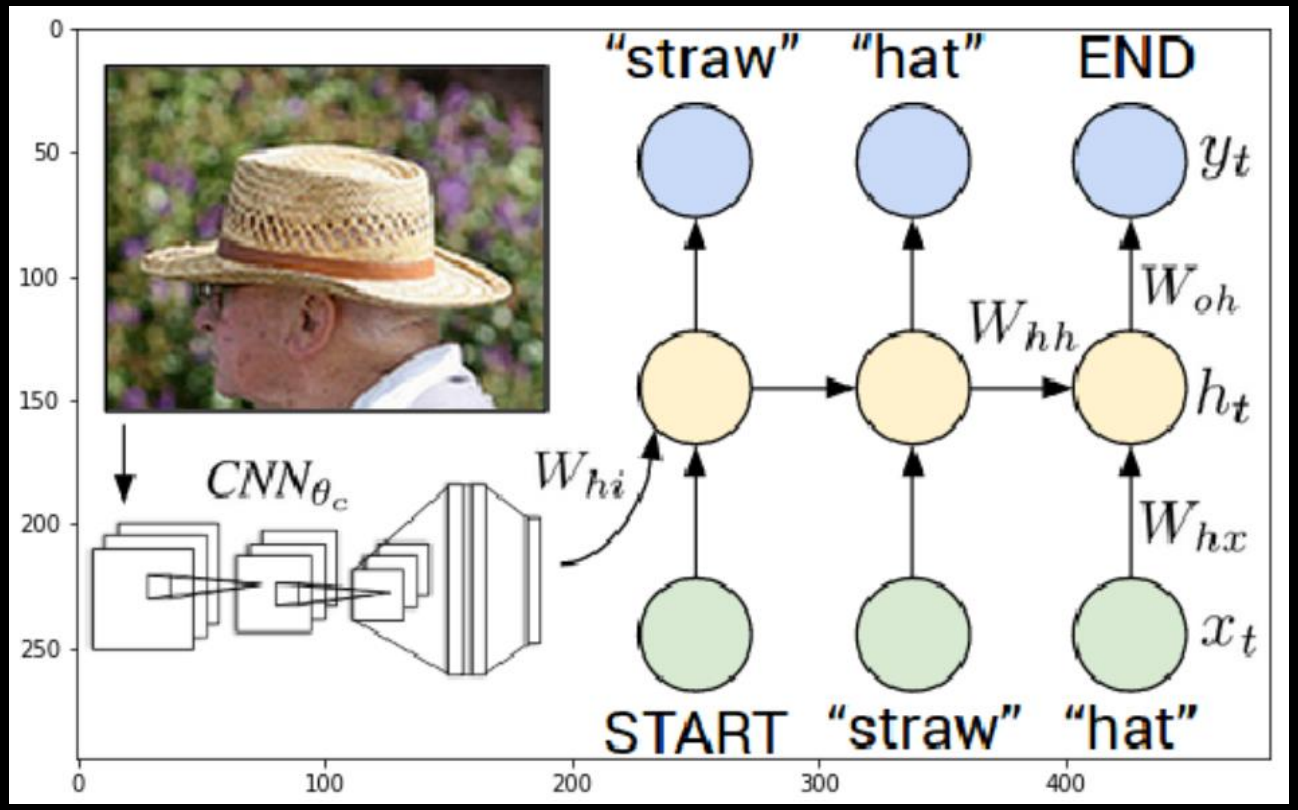
- We have 123,287 images, each paired with a string containing five different captions describing the contents of the image.
- We will split these data:
  - 82,783 for training
  - 40,504 for validation



Our Model

- Our model will take images and output captions
- We will start with a Convolutional Neural Network to take in the images
- We will end with a Recurrent Neural Network to produce the captions





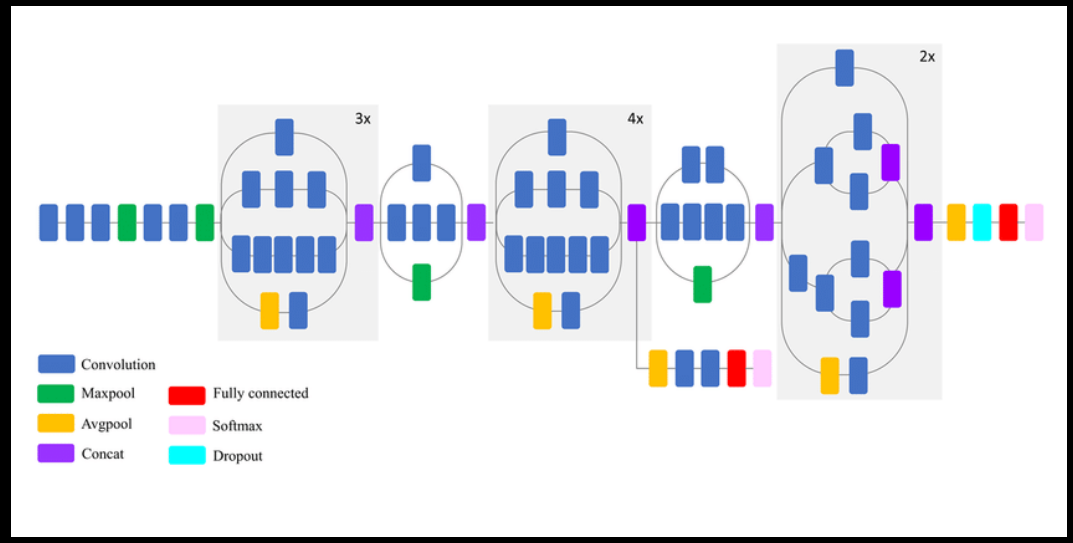
Model in More Detail

- The Convolutional Neural Network will reduce the image to a one-dimensional array of numbers
- This array will be used as the starting state for a Long-Short Term Memory (LSTM) Model
- The LSTM model will train on how closely it can approximate the captions paired with the image

Taking a Shortcut

- Training this model would take weeks if not months
- Google has a pre-trained model for the images: InceptionV3
- We can piggyback off InceptionV3 for the CNN portion of our model
- This means feeding weights from InceptionV3 into our RNN

InceptionV3



Basically, we will take everything from InceptionV3 except for the last layer (in this case, softmax)

How did it do?

a clock tower with a clock on top of it

a man riding on the back of a horse

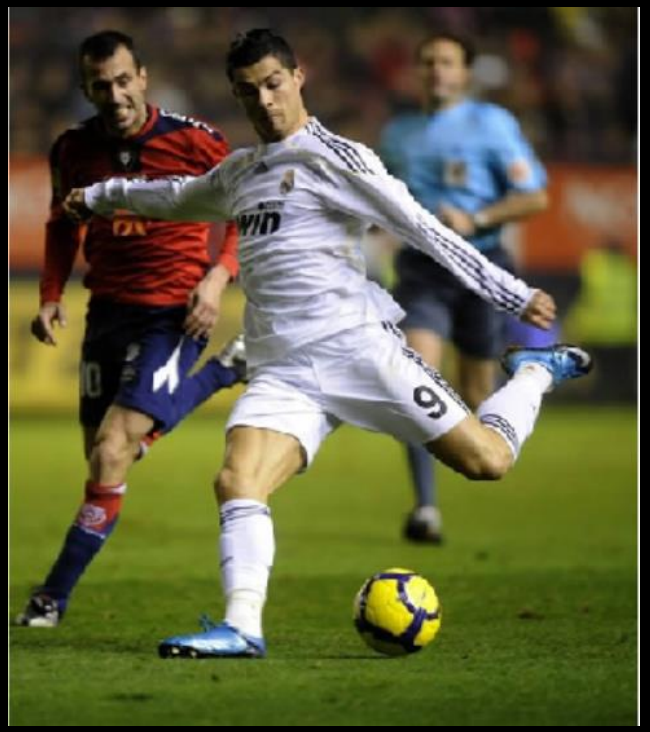
According to the model

- Accuracy for predicting the next word was 0.40629 for predicting the next word in a sentence

In Human Terms: Some Good Examples

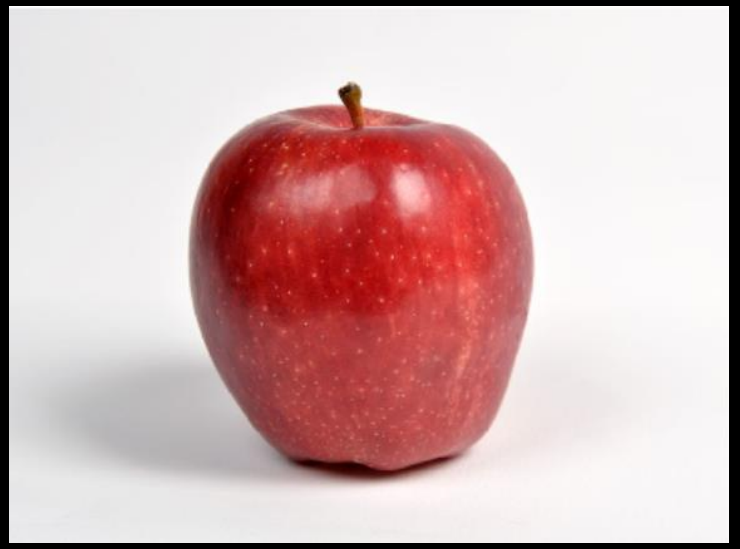
a bunch of apples sitting on a table

a woman in a cowboy hat holding a horse



A group of young men playing soccer on a field

In Human Terms: Some Bad Examples



a close up of a pair of orange scissors

a man is riding a surfboard on a wave

a person standing in the middle of a field

Uses and Consideration

- At this level, captions are at the level of a parlor trick
- However, this also shows how effective ML is at *classifying* images
- Some uses of image classification include:
  - Helping organize or search an unlabeled image archive
  - Autofocusing a camera on people, animals, or objects
  - Watching a nature camera for when a rare species appears