**IMAGE CAPTIONING**

Image Captioning can be defined in simple words as "**Automatically generating the textual description using an artificial intelligence based system for an image".** The goal of such a system is to convert a given input image into the Natural Language Description.

For Example



The above example is sufficient to understand image captioning.

Caption generation is a challenging artificial intelligence problem where a textual description must be generated for a given photograph.

It requires both methods from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order.

This is also a very active area of research and an interesting multi modal topic where combination of both image and text processing is used to build a useful Deep Learning application, aka Image Captioning.

**Applications of Image captioning**

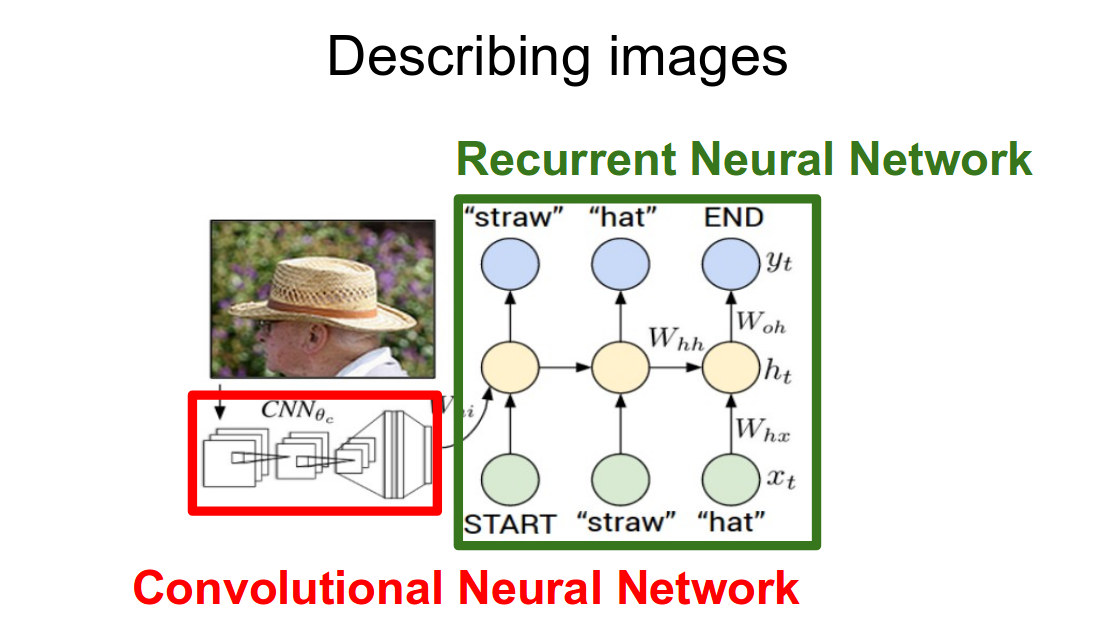
1. Probably can be used in the applications where text is used mostly and with the use of this we can infer a image in form of text.
2. NLP is used extensively in the market now-a-days. For example, summarizing or gaining insights from a large corpus of text. In a same way, we can use this to get insights from images as well.
3. We can build a 360-degree metastore and make use of it in a wide variety of business like making searches by user more efficient on an e-commerce platform based on metadata of products, other may be some other things like recommendations and all. One such application is here: https://www.sophist.io/
4. A slightly long term but yes is a use case where we can describe like what happen in a given video segment.
5. Can be used to give something back to mankind for visually impaired people.

and many more.

**Proposed Methodology**

The task of image captioning can be divided into two modules logically – one is an **image-based model** – which extracts the features and nuances out of our image, and the other is a **language-based model** – which translates the features and objects given by our image-based model to a natural sentence.

For our image-based model (viz encoder) – we usually rely on a Convolutional Neural Network model. And for our language-based model (viz decoder) – we rely on a Recurrent Neural Network. The image below summarizes the approach given above.



This architecture heavily employees **"Encoder-Decoder"** framework where encoder part is implemented using the convnet architecture and decoder part is implemented using LSTM cells based recurrent nets.

For the convnet part we can employ transfer learning and can use some models from ImageNet and then Fine tune them as per our requirement.

This kind of system is known as **attention-based system,** which automatically learns to describe the content of an image in words. We can train this model in a standard manner using standard backprop technique and stochastically by maximising a variationally lower bound.

For implementing the model, we can use the COCO – Common Object in Context dataset sponsored by CVDF, Microsoft, Facebook or we can use the Flikr8K dataset by University of Illinois.

I found these two papers quite handy:

1. [Show and Tell: A Neural Image Caption Generator](https://github.com/mayankskb/Research_Papers/blob/master/Applications%20-%20NN/Show%20and%20Tell:%20A%20Neural%20Image%20Caption%20Generator.pdf)
2. [Show, Attend and Tell: Neural Image Caption Generation with Visual image](https://github.com/mayankskb/Research_Papers/blob/master/Applications%20-%20NN/Show%2C%20Attend%20and%20Tell:%20Neural%20Image%20Caption%20Generation%20with%20Visual%20image.pdf)