Image Captioning

Team 6

Introduction

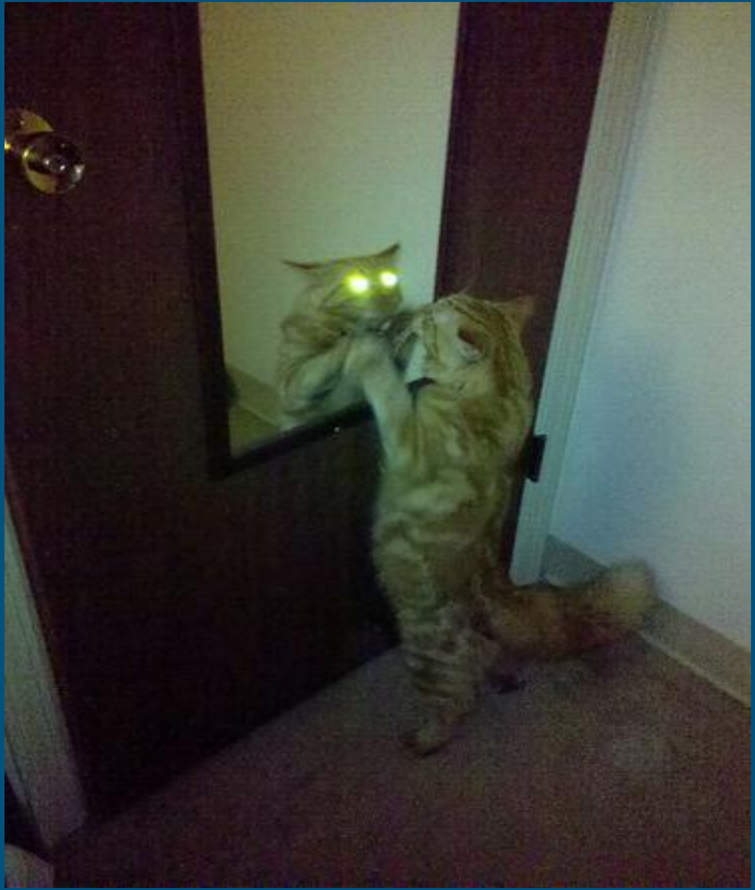
A quick glance at an image is sufficient for a human to point out and describe an immense amount of details about the visual scene. However, this remarkable ability has proven to be an elusive task for our visual recognition models. Generating descriptions for images has long been regarded as a challenging perception task,integrating vision, learning and language understanding. One not only needs to correctly recognize what appears in images but also incorporate knowledge of spatial relationships and interactions between objects. Even with this information, one then needs to generate a description that is relevant and grammatically correct

Dataset

MSCOCO - 2017

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

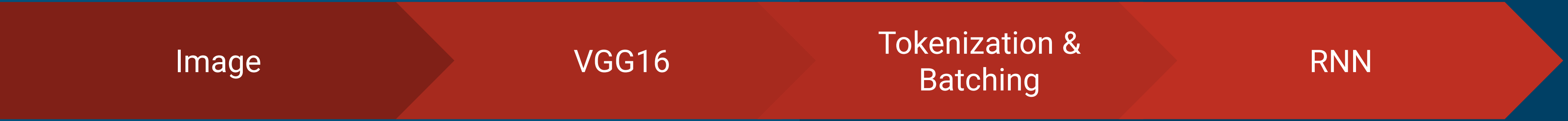


1. An orange cat with glowing eyes looking into a mirror.
2. A cat standing up in front of a mirror on the back of a door.
3. A cat looking at itself through a mirror.
4. A cat scared by its own reflection in the mirror.
5. A cat with shiny yellow eyes standing up to look at itself in a mirror.

Sample

Methodology

The ultimate goal of our model is to generate descriptions of image regions. During training, the input to our model is a set of images and their corresponding sentence descriptions. We first present a model that aligns sentence snippets to the visual regions that they describe through a multi-modal embedding. We then treat these correspondences as training data for a second, multi-modal Recurrent Neural Network model that learns to generate the snippets. Our multimodal Recurrent Neural Network generative model. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. START and END are special tokens.



Image

A Input Image resized to (224,224.3).

VGG16

Used pre-trained model of VGG16 on ImageNet Weights for object detection .

Tokenization & Batching

The Captions were converted it to numeric value without loss of any information so that the RNN model can work on it and Batch size of 512 was created to be forwarded to RNN model.

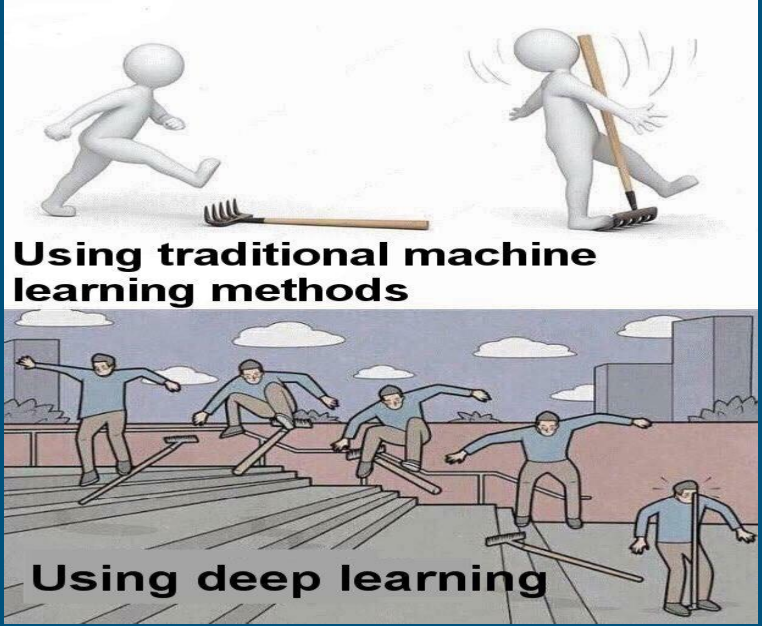
RNN

RNN Contains three staged layer of GRU. Which helps in producing a proper a complete sentences from VGG16 identified objects.

Object Detection

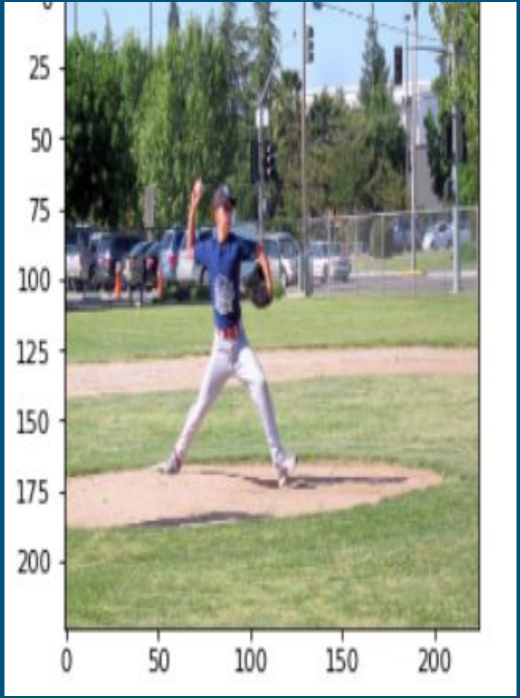
We used pre-trained VGG16 model on ImageNet for Objects Detection in Input Image.

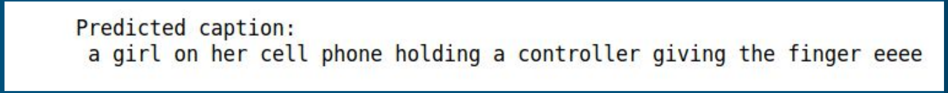
RNN



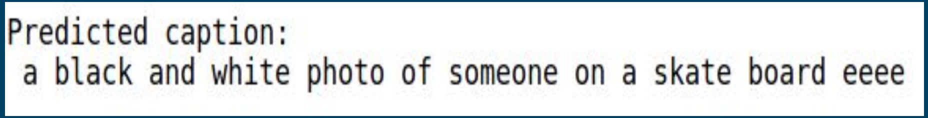
For text generation part we used RNN it takes in input the Image Object Detected features and Tokenized Caption for the particular match which leads to output of a full sentences explaining the image.

Results

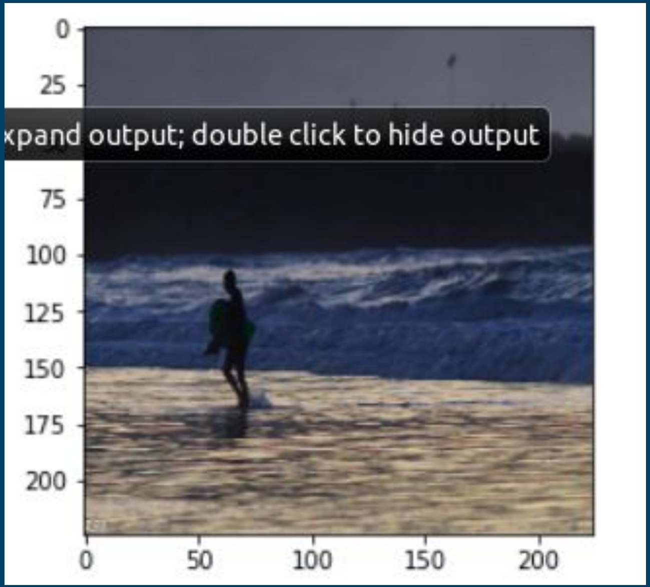




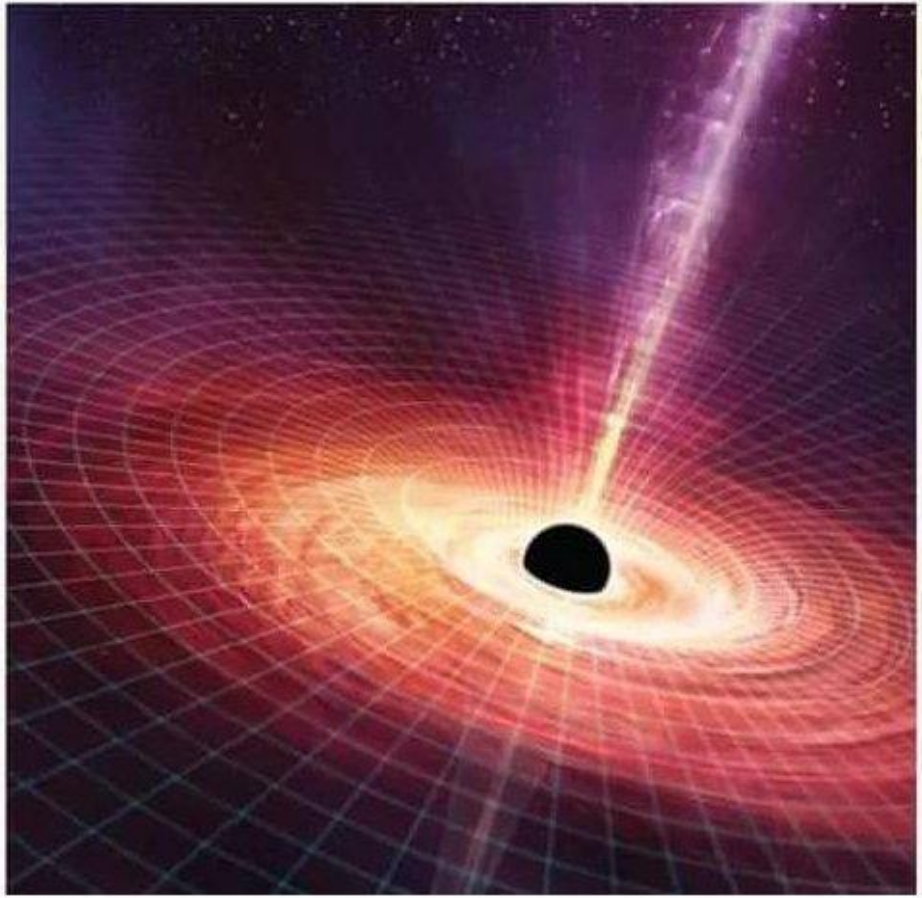
```
Predicted caption:
 a girl on her cell phone holding a controller giving the finger eeee
```



```
Predicted caption:
 a black and white photo of someone on a skate board eeee
```



Technique described in a paper



The same technique used in practice