Image Captioning

Ansh Goyal 17BCE1278
Harshil Gupta 17BCE1112
Shekhar Gaur 17BCE11183

ABSTRACT

Image Captioning is the process of generating textual description of an image. It uses both Natural Language Processing and Computer Vision to generate the captions. The dataset will be in the form [image → Captions]. The dataset consists of input images and their corresponding output descriptions. Well, we can add “captioning photos” to the list of jobs robots will soon be able to do just as well as humans. After some training, the algorithm can describe the contents of a photo with staggering 94% accuracy. We will take a look at an interesting multi modal topic where we will combine both image and text processing to build a useful Deep Learning application, aka Image Captioning. Image Captioning refers to the process of generating textual description from an image – based on the objects and actions in the image.

INTRODUCTION

Image Captioning is the process of generating textual description of an image. It uses both Natural Language Processing and Computer Vision to generate the captions. Deep Learning is a very rampant field right now – with so many applications coming out day by day. And the best way to get deeper into Deep Learning is to get hands-on with it. We will take a look at an interesting multi modal topic where we will combine both image and text processing to build a useful Deep Learning application, aka Image Captioning. Image Captioning refers to the process of generating textual description from an image – based on the objects and actions in the image.

METHODS AND MATERIALS

The Convolutional Neural Network(CNN) can be thought of as an encoder. The input image is given to CNN to extract the features. The last hidden state of the CNN is connected to the Decoder.
The Decoder is a Recurrent Neural Network(RNN) which does language modelling up to the word level. The first time step receives the encoded output from the encoder and also the <START> vector. The image representation is provided to the first time step of the decoder.
Set x1 =<START> vector and compute the distribution over the first word y1. We sample a word from the distribution set its embedding vector as x2, and repeat this process until the <END> token is generated.

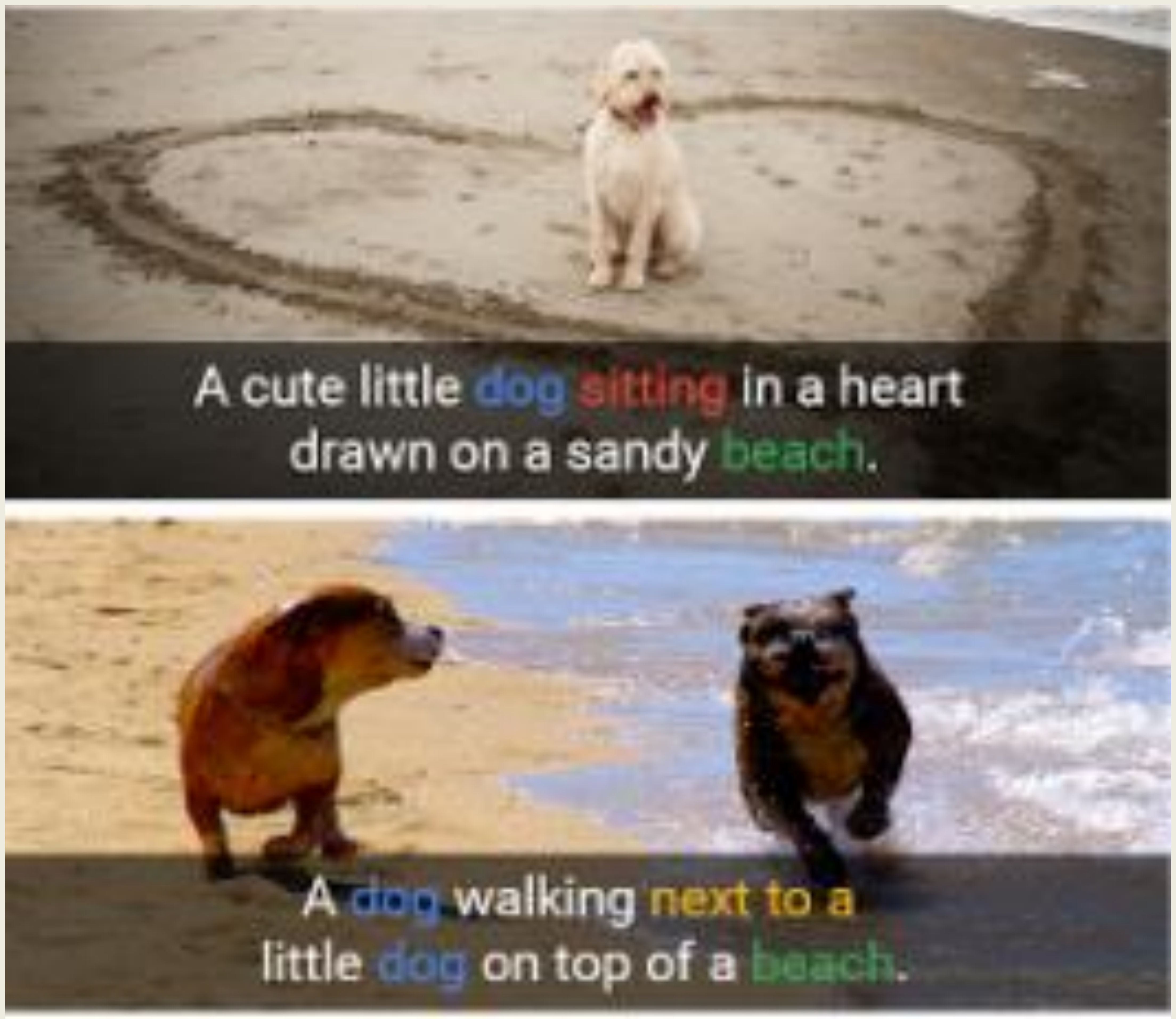
RESULTS

Image Captioning will provide a result with a description. For the preceding example, the result of image description would be a closeup of a white wall. This could be useful for generating content for a book or maybe helping the hearing or visually impaired. And another will be An open laptop computer sitting on top of a desk.
This is considerably more challenging as conventional neural networks are powerful, but they're not very compatible with sequential data. Sequential data is where we have data that's coming in an order and that order actually matters.

CONCLUSIONS

Most modern mobile phones are able to capture photographs, making it possible for the visually impaired to make images of their environments. These images can then be used to
generate descriptions that can be read out loud to the visually impaired, so that they can get a better sense of what is happening around them. When it comes to using image captioning in real world applications, most of the time only a few are mentioned such as hearing aid for the blind and content generation.

.


A cute little dog sitting in a heart drawn on a sandy beach.
A dog walking next to a little dog on top of a beach.