International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Detecting and Describing the Real World Objects from the Given Images

Leo Francis M, Anjan S Chavan, Dr. Zafar Ali Khan

Presidency University,India
DOI: https://doi.org/10.55248/gengpi.2022.3.6.18

ABSTRACT

In this paper we are gone see how to describe the physical appearance or characteristics of the real world objects to visually impaired persons, from the captured images and how to find the exact size of the objects, and to calculate the distance of the object from the device or the user. Another aim of this model is to give the exact information of the images to the users.

INTRODUCTION

Image captioning is the process where the AI model involves in describing the image into textual or human readable form. The main aim of image captioning is to automatically generate descriptions for the images. Image captioning involves in both computer vision and NLP (Natural Language Processing). It is a end-to-end and sequence-to-sequence problem. Here we are gone use CNN (Convolutional Neural Network) to process the images and RNN (Recurrent Neural Network) for generating text. CNN is used to classify the images, label them and detect the objects in the images. NLP is a computer program that can understand human language which can be obtained by using RNN.

EXISTING SYSTEM OR RELATED WORKS

There are many researches going on related to object detection and image captioning.Among them the most famous work is Google lens.This system just takes the images as input and shows some related information based on the image.Google lens gives information only for the images of animals or flowers and other living things.But when it comes to describing those images or the objects in the images it is not good enough.For example: if we scan/upload an image of a laptop or a bag it will show some AD'S and 'shopping websites' based on the image.

CURRENT WORK AND PRELIMINARY RESULTS

The Internet is a pool of information with never-ending results for the query searched. When it comes to searching with images, the whole searching aspect will differ. First, we need to process the image and then compare them with the existing images from the database. This might sound simple to hear but its not an easy task to process the images and convert them into convolutional layers.
Here we are going to detect the objects or the scenes from the given image. It can be done by using CNN or YOLO (You Only Look Once).

Look at the above image if we are asked to describe the above image, we can describe it in several ways but what about our model how it's gone select or choose the right statement for the image…????? This has to be done parallelly while we are looking at the image even the correct statement or description has to be framed. So, the first part will be done by CNN and the second one will be done by RNN.

Now let's look how to evaluate the model. We can use BLEU (Bilingual Evaluation Understudy) score, it evaluates the generated description with the referenced description, for correct match the score will be 1.0 and for correct mismatch the score will be 0.0.

In our network topology we are gone use "encoder" and "decoder", Encoder is dealing with the CNN where the input image will be given to it and extracts the features from the image, Decoder goes with the RNN, this receives the encoded image from the encoder and does the language developing in the word level.

CONCLUSION AND FUTURE WORK

In this paper until now we have seen how we are using CNN and RNN to build our model. These two are good enough to detect an object and describe them, there is another algorithm called YOLO (You Only Look Once), this algorithm is CNN based it detects multiple objects at a single iteration so we can use this algorithm instead of CNN in future.

In future we are planning to add few more features like calculating the size of the objects in the images and the distance of the object from the user. Here our main aim is to describe the things to visually impaired people.

REFERENCES

[1].https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/
[2].https://pjreddie.com/darknet/yolo/
[3].https://machine-learning.paperspace.com/wiki/convolutional-neural-network-cnn#:~:text=A%20Convolutional%20Neural%20Network%20(CNN,a%20subset%20of%20machine%20learning.
[4].https://builtin.com/data-science/recurrent-neural-networks-and-lstm
[5].https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks
[6].https://towardsdatascience.com/a-guide-to-image-captioning-e9fd5517f350
[7].https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2
[8].https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP#:~:text=Natural%20language%20processing%20(NLP)%20is,in%20the%20field%20of%20linguistics.
[9].https://www.ibm.com/cloud/learn/natural-language-processing
[10].https://www.javatpoint.com/nlp