

1. Introduction

1.1 Overview

Image Captioning refers to the process of generating textual description from an image – based on the objects and actions in the image. The ability to recognize image features and generate accurate, syntactically reasonable text descriptions is important for many tasks in computer vision. In recent years, with the rapid development of artificial intelligence, image caption has gradually attracted the attention of many researchers in the field of artificial intelligence and has become an interesting and arduous task. Image caption, automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing. The application of image caption is extensive and significant, for example, the realization of human-computer interaction.

For example,

- Self-driving cars — Automatic driving is one of the biggest challenges and if we can properly caption the scene around the car, it can give a boost to the self-driving system.
- Aid to the blind — We can create a product for the blind which will guide them travelling on the roads without the support of anyone else.

Although image caption can be applied to image retrieval, video caption, and video movement and the variety of image caption systems are available today, experimental results show that this task still has better performance systems and improvement. It mainly faces the following three challenges: first, how to generate complete natural language sentences like a human being; second, how to make the generated sentence grammatically correct; and third, how to make the caption semantics as clear as possible and consistent with the given image content.

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.

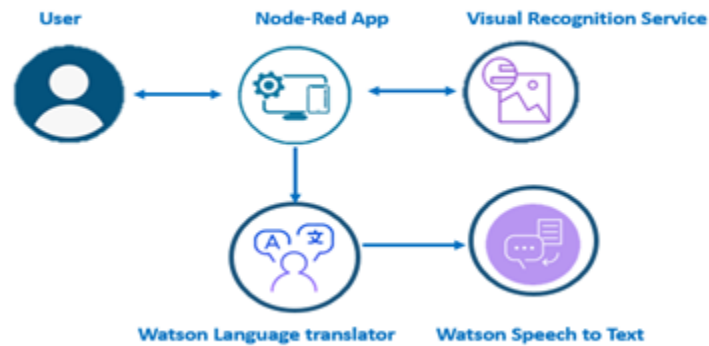
And for our language based model (viz decoder) – we rely on a Recurrent Neural Network.

1.2 Purpose

We can do this by first converting the scene into text and then the text to voice. Both are now famous applications of Deep Learning. CCTV cameras are everywhere today, but along with viewing the world, if we can also generate relevant captions, then we can raise alarms as soon as there is some malicious activity going on somewhere. This could probably help reduce some crime and/or accidents.

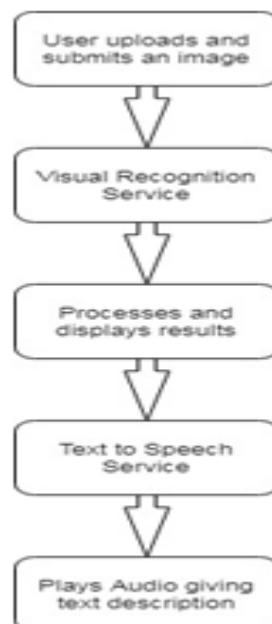
2 Proposed Solution

This project aims at building an application which takes input as image analyses it and generates the captions in the form of speech. To achieve this, we will be using IBM Services like node-red service to build a web UI where user uploads a picture. This picture is analyzed by visual recognition service and the analyzed description is then converted in to text to speech service using text to speech service.



3. Theoretical Analysis

3.1 Block Diagram



3.2 Hardware/Software designing

- OS: Windows 7+
- Platform: IBM Watson Studio
- Services: Visual Recognition Service/Text to Speech/language translator
- UI Services: NodeRED

4. Results


Caption

Watson Visual Recognition

Caption Generator

Choose file

Browse

 Selected Image

Submit

Results

Caption

Watson Visual Recognition

Caption Generator

Choose file

Browse



Submit

Results

Watson thinks this picture contains a reddish orange color flowering plant.

Class	Confidence
florist shop	0.76
shop	0.76
retail store	0.76
building	0.76
dahlia	0.625
flower	0.886
flowering plant	0.887
plant	0.887

Caption

Choose file

Browse



Submit

Class	Confidence
florist shop	0.76
shop	0.76
retail store	0.76
building	0.76
dahlia	0.625
flower	0.886
flowering plant	0.887
plant	0.887
florist's chrysanthemum	0.556
Barberton daisy	0.5
nature	0.79
reddish orange color	0.966
dark red color	0.74

5. Applications

The main implication of image captioning is automating the job of some person who interprets the image (in many different fields).

1. It will be useful in cases/fields where text is most used and with the use of this, you can infer/generate text from images. As in, use the information directly from any particular image in a textual format automatically.
2. There are many NLP applications right now, which extract insights/summary from a given text data or an essay etc. The same benefits can be obtained by people who would benefit from automated insights from images.
3. A slightly (not-so) long term use case would definitely be, explaining what happens in a video, frame by frame.
4. It may serve as a huge help for visually impaired people. Lots of applications can be developed in that space.
5. Social Media. Platforms like facebook can infer directly from the image, where you are (beach, cafe etc), what you wear (color) and more importantly what you're doing also (in a way).
6. Self-driving cars — Automatic driving is one of the biggest challenges and if we can properly caption the scene around the car, it can give a boost to the self-driving system.

6. Conclusion

In this project, an artificial Intelligence model is developed that first converting the scene/image into text and then the text to voice. To achieve this, IBM Services like node-red service is used to build a web UI where user uploads a picture. This picture is analyzed by visual recognition service and the analyzed description is then converted in to text to speech service using text to speech service.

7. Bibliography

1. https://www.ripublication.com/ijaer18/ijaerv13n9_102.pdf
2. https://www.researchgate.net/post/Any_ideas_on_more_applications_of_image_captioning
3. <https://www.hindawi.com/journals/cin/2020/3062706/>
4. <https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning>