Detection and Recognition of Object for Image Captioning

Prashant Yadav, Vishal Vishwakarma, Aakash Tiwari, *Student, SLRTCE,* Komal Champanerkar, *Mentor, SLRTCE*

***Abstract*—As one of the most intelligent beings on the planet, we are equipped with the most powerful visual and language system as it is easy for us to extract visual information from a given image and transform it into proper linguistic description. Image Caption Generator deals with generating captions for a given image. The capturing mechanism involves a tiring task that collaborates both computer vision and image processing. The mechanism must detect and establish relationships between objects, people, and animals. The aim of this paper is to detect, recognize and generate worthwhile captions for a given image. We use transfer learning CNN on sentences, and extract image representation with Neural Networks.**

Keywords— Image caption, Neural Networks, Computer Vision, Natural language Processing.

I. INTRODUCTION

The basic ability of human beings is the tendency to describe an image with an ample amount of information about it by just a quick glance. Creating a computer system to simulate the abilities of human beings is a long-time researcher goal in the fields of machine learning and artificial intelligence. There are several research progresses made in the past such as the detection of objects from a given image, attribute classification, image classification, and classification of actions by goal human beings.

Making a computer system to detect the image and produce a description using natural language processing is an exigent task, which is called an image caption generator system. Generating a caption for an image involves various tasks such as understanding the higher levels of semantics and describing the semantics in a sentence by which human can understand. In order to understand the higher levels of semantics, the computer system must learn the relationships between the objects in a given image. Usually, communication in human beings occurs with the help of natural language, so developing a system that produces descriptions that can be understandable by human beings is a challenging.

There are several steps to generate captions, such as understanding visual representation of objects, establishing relationships among the objects and generating captions both linguistically and semantically correct. Considerable research has been done over the past few years for solving the image captioning problem. Most contemporary solutions involve the use of an encoder-decoder mechanism in the form of a Convolutional Neural Network.

This paper indents to provide a thorough survey of the various factors that constitute towards the development of the system. The paper is organized as follows. Section 3 describes the background study of the image caption generator, Section 4 deals with the proposed methodology, Section 5 deals with the experimental analysis and findings and finally concluded in Section 7.

II. PROBLEM STATEMENT

Generating a caption from a given image has its own set of challenges such as perfect exposure and light to stress the image, well-defined captions, etc. These factors have a significant impact on the accuracy of the result. To overcome these challenges our proposed system uses Convolutional Neural Network (CNN) for the feature extractions and Long Short-Term Memory (LSTM) to generate the captions from the extracted features. Also, it is important to cover an extensive range of objects or datasets for model training to achieve higher accuracy for providing a human like ability to read the image and process the event happening in it, which can prove to be very useful for future technologies, security and many other applications.

III. LITERATURE REVIEW

Here we will elaborate the aspects like the literature survey of the project and about the existing ones being used or been developed in the market. which we took the motivation from and thus decided to go ahead with the project. Literature review aids us study the past innovations related to the system we aim to build and also help ameliorate them. Understanding the literature review assures the better implementation of a project by minimizing discrepancies.

Related work

[1] Deep Learning is used to detection and recognition of objects in image. Convolution Neural Network (CNN) approach is used for feature extraction. [2] Here Deep neural models can generate an image caption. But there is an issue in LSTM, the next predicted word of the caption depends mainly on the last predicted word, rather than the image content. To overcome this, a modified LSTM cell with an additional gate(read-only-unit) can result in more accurate captions. The Loss Value of LSTM was 2.85 while loss value of RNN, LSTM cell and read only unit is 1.85. [3] Using Zero-Shot Learning. It uses Long-Short Term Memory (LSTM) that uses a word embedding trained on external corpora as well as training data captions. This system comprises of 3 modules. The first is the image caption generator, the second module is an object detection algorithm and the third module is a semantic word embedding. [4] We use transfer learning CNN on sentences, and then extracting image representation with deep Fisher kernel. In Fisher kernel, all the extracted activations are aggregated to Fisher vectors. Finally, the vectors are pooled to a finale vector with MPP also we focused on better representation of an image to outperform existing caption generation models. [5] In this paper, the sentence embedding should be able to capture

details from various descriptions means that it should scan from left to right and top to bottom. This determines the robustness of understanding image detail.

IV. PROPOSED METHODOLOGY

The Proposed methodology for generating captions with the detection and recognition of objects using Neural Network is shown in Fig. 1. It consists of object detection, feature extraction, Convolution Neural Network (CNN) for feature extraction and for scene classification, Recurrent Neural Network (RNN) for human and objects attributes, RNN encoder and a fixed length RNN decoder system.

The steps for generating captions with object detection and feature extraction using neural networks are as follows.

Step 1: Object detection

In this step, the objects in the input image are detected using R-CNN region proposal approach.

Step 2: Feature Extraction

In this step, the features in the image are extracted using principal component analysis using NumPy. CNN is used for scene classification and RNN is used for detecting objects and human attributes.

Step 3: Creating attributes

In this step, the features extracted by the neural networks were used to define the attributes with its label strings.

Step 4: Encoder and Decoder

In this step, the label strings were subjected to an encoder RNN for encoding the label strings to a proper format, and the resultant variable length string is subjected to a fixed length decoder for converting to a fixed length descriptive sentence.

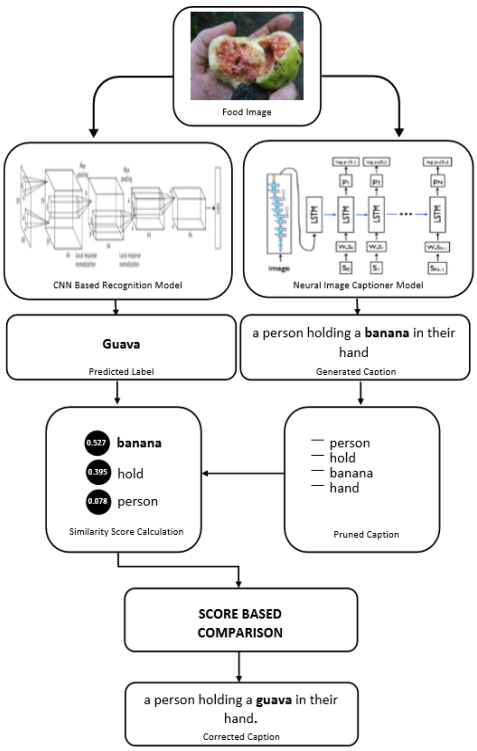

Fig. 1.

V. EXPERIMENTAL ANALYSIS

The aim of this paper is to propose a deep learning method for generating captions using neural networks. The dataset details have been described in this section. The experimental evaluation of the proposed methodology is done by Flickr 8k dataset, just for simplicity, only three images were subjected to the proposed methodology and the results were obtained. Fig. 2 represents the input image to which the caption needs to be generated. Fig. 3 describes the caption generation process, first, the input image is subject to the feature extraction using the feature extraction command to extract the features in that image and captions are generated using the generated command which takes parameters such as model, tokenizer, input image and length as input. The proposed model accurately generated a caption that a dog running through the water for the Fig. 2. The model is also evaluated with Fig. 4 and Fig. 6, the model accurately generated caption as shown in Fig. 5 and Fig. 7.

Fig. 2. Input Image–1

```
In [26]: # load and prepare the photograph
         photo = extract_features('example.jpg')
         # generate description
         description = generate_desc(model, tokenizer, photo, max_length)
         print(description)

         startseq two dogs are running through the water endseq
```

Fig. 3. Output: Caption generated

Fig. 4. Input Image-2

```
# load and prepare the photograph
photo = extract_features('ex2.jpeg')
# generate description
description = generate_desc(model, tokenizer, photo, max_length)
print(description)

startseq two children are playing with soccer ball in the grass endseq
```

Fig. 5. Output: Caption generated

Fig. 6. Input Image-2

```
# load and prepare the photograph
photo = extract_features('2.jpg')
# generate description
description = generate_desc(model, tokenizer, photo, max_length)
print(description)
```

```
startseq man is climbing down each mountain endseq
```

Fig. 7. Output: Caption generated

VI. ADVANTAGE

Captioned image can help visually impaired people understand the content of the image through text to voice conversion.

Categorization of Image based on the object detected in the image.

The trained model can be used in other help understand unidentified or unknown objects.

Can help in security field by identifying objects in restricted areas.

A well-trained model can also help in medical field in detecting tumour and other diseases as well.

VII. CONCLUSION

In this paper, the proposed a system which will help generate caption for the given input image. This model can be very useful in other technological fields capable of providing accurate results and monitoring purpose. The proposed deep learning methodology generated captions with more descriptive meaning than the existing image caption generation generators. It is important to train the model with a lot of datasets and well-defined captions to ensure a wide range of features are detected correctly. Also, the extracted features can help individual understand the image better and help categorize them based on the features.

REFERENCES

[1] N. Komal Kumar, D. Vigneswari, A. Mohan, K. Laxman, J. Yuvaraj, “Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach.”, 8728-516 ©2019 IEEE.

[2] Aghasi Poghosyan, Aghasi Poghosyan, “Long Short-Term Memory with Read-only Unit in Neural Image Caption Generator.” 8312-163 ©2017 IEEE.

[3] Mirza Muhammad Ali Baig, Mian Ihtisham Shah, Muhammad Abdullah Wajahat, Nauman Zafar and Omar Arif, “Image Caption Generator with Novel Object injection”, 8615-810 ©2018 IEEE.

[4]] Dong-Jin Kim, Donggeun Yoo, Bonggeun Sim, In So Kweon, “Sentence Learning on Deep Convolutional Networks for Image Caption Generation”, 7625-747 ©2016 IEEE.

[5] Hao Dong, Jingqing Zhang, Douglas McIlwraith, Yike Guo, “I2T2I: LEARNING TEXT TO IMAGE SYNTHESIS WITH TEXTUAL DATA AUGMENTATION” , 8296-635 ©2017 IEEE.