Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach

N. Komal Kumar[1], D. Vigneswari[2], A. Mohan[3], K. Laxman[4], J. Yuvaraj[5]
Department of Information Technology,
Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi, Chennai, Tamil Nadu, India.
[1]komalkumarnapa@gmail.com, [2]vigneswari121192@gmail.com, [3]anamalaimohan@gmail.com, [4]vinithkalaivanan3@gmail.com,
[5]yuvi0117@gmail.com

Abstract*—Image Caption Generator deals with generating captions for a given image. The semantic meaning in the image is captured and converted into a natural language. The capturing mechanism involves a tedious task that collaborates both image processing and computer vision. 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 using deep learning. Regional Object Detector (RODe) is used for the detection, recognition and generating captions. The proposed method focuses on deep learning to further improve upon the existing image caption generator system. Experiments are conducted on the Flickr 8k dataset using python language to demonstrate the proposed method.

Keywords*—Image, capturing, generator, regional, detector, deep learning

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 [1]. 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 progress made in the past such as the detection of objects from a given image, attribute classification, image classification, and classification of actions by 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 goal. 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. This paper aims at detection, recognition and generating captions using deep learning.

The paper is organized as follows. Section 2 describes the background study of the image caption generator, Section 3 deals with the proposed methodology, Section 4 deals with the experimental analysis and findings and finally concluded in Section 5.

II. BACKGROUND STUDY

This section describes the background study on image caption generators.

Image caption generator deals with generating caption for a given image. A. Kojima [2] used case structure, action hierarchy, and verb patterns to generate captions of human activities in a fixed environment. P. Hede [3] proposed a method for image caption generation, which involves a series of object names stored in a database called dictionaries, such method can generate captions for fixed image content, but the method fails to generate captions for real-world scenarios. An image caption generator system was proposed in [4], which uses deep neural networks to generate captions. A. Farhadi [5] proposed an information retrieval based image captioning system, where a score is generated for every object in an image and the score is compared with other images to generate captions. M. Hodosh [6] proposed a ranking based image captioning system, where the captions are generated with the help of sentence based image captioning ranking system. Y. Yang [7] proposed a sentence making strategy, which employs verbs, nouns, and prepositions for building a semantic sentence, the image is detected with the help of trained detectors and English corpus was used for the estimates of the image. R. Socher [8] proposed a decision tree based recursive neural networks to represent the visual meaning of the image. O. Vinyals [9] proposed a generative model that combines computer vision and machine learning to generate captions for a given image. Q. You et al. [10] proposed a model of semantic attention, which deals with the semantics stored in a hidden layer of neural networks and fusion them to gain more semantically sentence.

978-1-5386-9533-3/19/$31.00 ©2019 IEEE

III. PROPOSED METHODOLOGY

The Proposed methodology for generating captions with the detection and recognition of objects using deep learning 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.

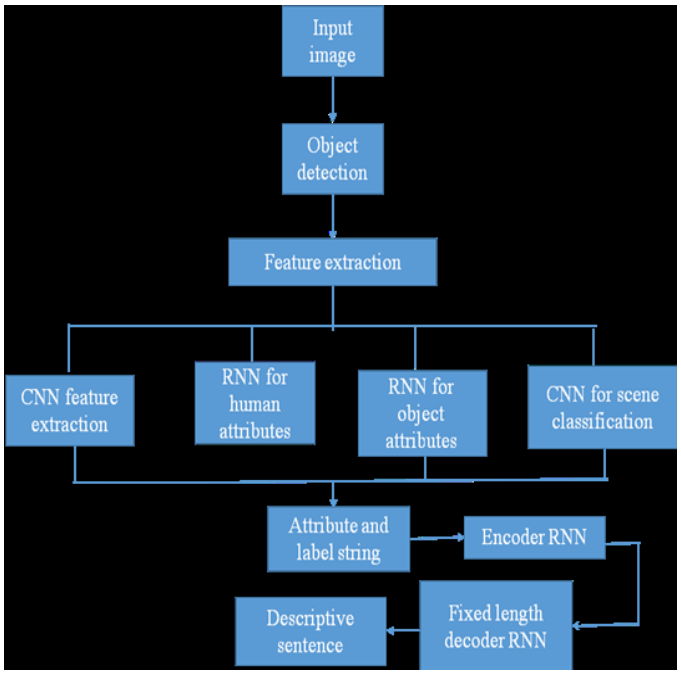

Fig.. 1. Proposed Methodology

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.

IV. 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 obtained from [11], from 8000 images, 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 generate_desc 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-3

```
# 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

V. CONCLUSION

In this paper, a deep learning method for image caption generation using neural networks is presented, the proposed method was applied to a Flickr 8k dataset. The proposed deep learning methodology generated captions with more descriptive meaning than the existing image caption generation generators. Applying hybrid image caption generator model can be developed in the future for more accurate captions.

REFERENCES

[1] L. Fei-Fei , A. Iyer , C. Koch , P. Perona. ,What do we perceive in a glance of a real-world scene? J. Vis. 7 (1) (2007) 1–29 .

[2] A. Kojima , T. Tamura , K. Fukunaga , Natural language description of human ac- tivities from video images based on concept hierarchy of actions, Int. Comput. Vis. 50 (2002) 171–184 .

[3] P. Hede , P. Moellic , J. Bourgeoys , M. Joint , C. Thomas ,Automatic generation of natural language descriptions for images, in: Proceedings of the Recherche Dinformation Assistee Par Ordinateur, 2004.

[4] J. Donahue , Y. Jia , O. Vinyals , J. Hoffman , N. Zhang , E. Tzeng , T. Darrell , DeCAF: a deep convolutional activation feature for generic visual recognition, in: Proceedings of The Thirty First International Conference on Machine Learning, 2014, pp. 647–655.

[5] A. Farhadi , M. Hejrati , M.A. Sadeghi , P. Young , C. Rashtchian , J. Hocken- maier , D. Forsyth , Every picture tells a story: Generating sentences from images, in: Proceedings of the European Conference on Computer Vision,,2010, pp. 15–29.

[6] M. Hodosh , P. Young , J. Hockenmaier , Framing image description as a rank- ing task: data, models and evaluation metrics, J. Artif. Intell. Res. 47 (2013) 853–899 .

[7] Y. Yang , C.L. Teo , H. Daume , Y. Aloimono , Corpus-guided sentence generation of natural images, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 4 4 4–454 .

[8] R. Socher , A. Karpathy , Q.V. Le , C.D. Manning , A.Y. Ng , Grounded composi- tional semantics for finding and describing images with sentences, TACL 2 (2014) 207–218 .

[9] O. Vinyals , A. Toshev , S. Bengio , D. Erhan , S3how and tell: a neural image cap- tion generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164 .

[10] Q. You , H. Jin , Z. Wang , C. Fang , J. Luo , Image captioning with semantic atten- tion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4651–4659 .

[11] M. Hodosh, P. Young and J. Hockenmaier (2013) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853-899