JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Image Caption Generator (Deep Learning Approach Using RNN & CNN)

Shreyas N. Kale
Department of Information Technology
AISSMS Institute of Information Technology
Pune, India
kaleshreyas45@gmail.com

Mr. Riyaz Jamadar
Assistant Professor
Department of Information Technology
AISSMS Institute of Information Technology
Pune, India
riyaz.jamadar@aissmsioit.org

Ajinkya H. Jagadale
Department of Information Technology
AISSMS Institute of Information Technology
Pune, India
ajinkyajagdale21@gmail.com

Adil Khan
Department of Information Technology
AISSMS Institute of Information Technology
Pune, India
workwithadil4@gmail.com

Rishabh Kunwar
Department of Information Technology
AISSMS Institute of Information Technology
Pune, India
rishabhkunwar11@gmail.com

Abstract- **Generating proper, accurate content or caption of images deals with generating captions for a various query images. The semantic meaning within the image is captured and transformed into a proper content. It is going to be a tedious task that collaborates both object detection, image identification and computer vision. OpenCV library is used to recognize and detect the object properly. The mechanism should detect and create relationships between objects, people, and animals. The project target includes object detection, object recognition using the computer vision and generating captions after processing entire image. For detection, recognition and creating various different captions for various different images, Regional Object Detector (RODe) is taken into consideration. The proposed methodology mainly focuses on deep learning to further improve upon the prevailing image caption generator system. Experiments are conducted on the Flickr 8k dataset using python language to demonstrate the selected method.**

Keywords—image caption, RNN, CNN

I. INTRODUCTION

A large amount of data is stored in a picture . Everyday huge image data is generated on social media and observatories. Deep learning are often wont to automatically annotate these images, thus replacing the manual annotations done. this may greatly reduce the human error as well because the reports by removing the necessity for human intervention. The generation of captions from images has various practical benefits, starting from aiding the visually impaired, to enabling the automated , cost-saving labelling of the many images uploaded to the web a day , recommendations in editing applications, beneficial in virtual assistants, for indexing of images, for visually challenged people, for social media, and a number of other natural language processing applications. the sector brings together state-of-the-art models in tongue Processing and Computer Vision, two of the main _elds in Artificial Intelligence. one among the challenges is availability of large number of images with their associated text ever expanding internet. However, most of this data is noisy and
hence it can't be directly utilized in image captioning model. For training a picture caption generation model, a huge data set with properly available annotated image is required. In this paper, we decide to demonstrate a system that generates contextual description about objects in images. Given a picture , break it right down to extract the various objects, actions, attributes and generate a meaningful sentence for the image.

Table 1 Literature Survey

SR NO.	TITLE	METHODOLOGY	CONCLUSION
1	**Detection and Recognition of objects in image caption generator system : A Deep Learning Approach**	Input image is detected using R-CNN and then features are extracted using NumPy followed by scene classification by CNN. Extracted features are used to define the attributes with its label strings and this string is passed to an encoder for encoding it in a proper format.	**The proposed deep learning methodology generated captions with more descriptive meaning than the existing image caption generation generators using unsupervised learning.**
2	**Retrieving images with generated textual descriptions**	Using the encoded vector of the generated description, the similarity of any two images could be measured calculating the distance between their generated encoded vector and retrieval of images is done by using word mover distance	**We have presented a semantic image retrieval method based on generated textual descriptions which attempt to explore the high level semantic content incorporated in the generated descriptions.**
3	**Domain specific image caption generator with semantic ontology.**	There are two parts one is image caption generator and other caption re-constructor and image caption generator generates common caption using mainly object detector and attributes predicter and this common caption then reconstructs using caption reconstructor to generate domain specific caption to that image	**This approach is going to generate caption using semantic ontology technique.**
4	**Deep Learning based Automatic Image Caption Generation**	We propose a transfer learning approach to generate automated captions for any given image. Encoder used is pretrained VGG16 model. This model makes use of a RNN which encodes the variable length input into a fixed dimensional vector and uses this representation to \decode" it to the desired output	**In this paper, we have presented a semantic image retrieval method based on generated textual descriptions which attempt to explore the high level semantic content incorporated in the generated descriptions.**
		sentence.	
5	**A Novel Convolutional Neural Network-Gated Recurrent Unit approach for Image Captioning**	The image features were computed by VGG16 model and then features to our model. The test data is cleaned by making unique photo, description and converted in to words. At last a dictionary of images and description saved and fitted in our model	**Our model after training generates caption with any given image with accuracy of 82.3 percent. Our model can be used for various real time image captioning applications like language modeling, image indexing , etc.**
6	**New Image Captioning Based On Text Summarization Using Image As Query**	It consist of 4 parts 1)Text Encoder, 2)image encoder 3)Decoder, 4)Attention Mechanism. The text encoder is used to encode text and image encoder is used encode image and get converted into the vector representation. The RNN Model is used for both encoding and decoding content from the image and generates caption	**News image captioning is different from generic image captioning in that news image captions contain more detailed information such as entity names and events than general image captions do, and detailed information is usually contained in news text but not in news images**
7	**A New CNN RNN framework for remote sensing image captioning**	Generate multiple captions for target image using beam search algorithm and choose best option among multiple generated captions on the basis of the lexical similarity with reference captions of similar images from the archive	**This model is really promising. We can develop more about CNN-RNN framework in future more sophisticated lexical similarity**
8	**Encoder-Decoder Architecture for Image Caption Generation**	It is based on a Convolutional Neural Network, which is used as an encoder to process the images and then relevant captions are generated using the Gated Recurrent Unit. Then BLEU algorithm checks the quality of the generated captions and chooses best one.	**Using captioning for videos, we can also provide real time information for visually impaired people, since a particular image can only give the caption at one specific time.**
9	**Semantic Descriptions of High Resolution Remote Sensing Images**	To relieve the limitation, a completely unique captioning task is proposed and a completely unique framework is proposed to unravel the novel task. The proposed framework uses semantic embedding to live the image	**Multi-sentence captioning task is proposed considering the complex objects distribution of remote sensing images. We propose a novel method called CSMLF to generate five sentences for a**

		representation and therefore the sentence representation. The captioning performance is improved by a proposed sentence representation	**given remote sensing image..**

II. METHODOLOGY

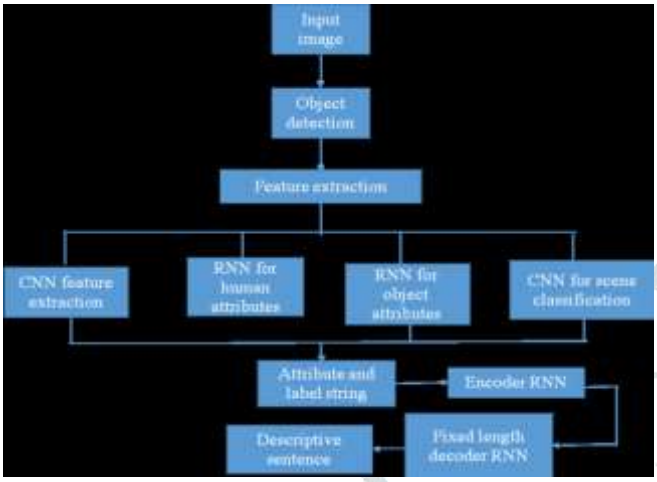

Fig.1 Methodology

There are various algorithms and method to generate proper caption for target image. In this, we are going to talk about RETRIEVING IMAGES WITH GENERATED TEXTUAL DESCRIPTIONS. As per name suggest using this methodology we can also retrieve small sub images from target image as well as we can also generate captions for our query image. This methodology contains mainly three parts: 1) image textual description generation; 2) textual description encoding; and 3) image retrieval using the generated textual descriptions. But we will talk only first two parts as per our project name.

A. Image textual caption generator Proposed system:

For The main work of image description generation is to generate natural description of the content of an image. For the text generation during this work, we resort to the long STM (LSTM) which may be a extraordinary case of the recurrent neural networks (RNN).confirmation.

- It have been huge success in usual process of (NLP) _eld in word recognition task using recurrent neural network. Caption generation is recognised on the basis of human thoughts where the identification of latest words depends on the previous ones.
- The main trademark of the RNN is that the aforementioned property is satisfied by means of feedback loops which make the information to persist through the network.
- When the prediction of a replacement word is said to a faraway previous information, RNN suffers the long term dependency.
- To notice this problem in the Long Short Term Memory concept is introduced. There is a cell state which allows the unchanged owing of data through the network and three gates which are wont to control the knowledge ow through the cell.
- There is a condition for image content to identify word to word meaning. For that, we fetch the image feature employing a pre-traine convolutional neural network (CNN). especially , we use the ResNet50 model.
- Using one-hot encoding having dimension of the vocabulary size, the words (composing the sentences) are encoded then imported to an embedding layer that's ready to explore their semantic content.
- The individual word embedding is called sentence representation. The word embedding (composing the sentences) are given as input to the LSTM that stores and learns the semantic temporal context of words through its recurrent layers.
- The final output of the LSTM is added with image features during a `multimodal' feedforward layer to get textual descriptions of the content of a picture . At inference stage we input the image to the model and acquire the generated description of its content.

B. Sentence Encoding:

- Vector of numbers using two different recent word embedding techniques take the input of each word of the generated information. That two techniques are: word2vec and GloVe Both techniques are based on co-occurrence of words so as to require under consideration context during a text represented by the adjacent words.
- The word2vec is built on a feed-forward neural network using two predictive models, nonstop bag of words (CBOW) and skipgram model to find out the embedding of the words. • CBOW model attempts to expect a word given its context, while skip-gram attempts to predict the context from a given word. During this work we use fast Text , a faster version of word2vec which takes under consideration the word morphology.
- This system is predicated on the skip-gram model and each word is represented as a sum of its n-gram character vectors.

III. DESIGN

A. *HARDWARE*

i) Processor
ii) RAM(MIN - 4GB)

B. *SOFTWARE*

i) Open CV
ii) Python
iii) Tenserflow

IV. SYSTEM ARCHITECTURE

The first one generates textual descriptions of the content of the RS images combining a convolutional neural network (CNN) and a recurrent neural network (RNN) to extract the features of the pictures and to generate the descriptions of their content, respectively. The second step encodes the semantic content of the generated descriptions using word embedding techniques ready to produce semantically rich word vectors. The third step retrieves the foremost similar images with reference to the query image by measuring the similarity between the encoded generated textual descriptions of the query image and the people of the archive.

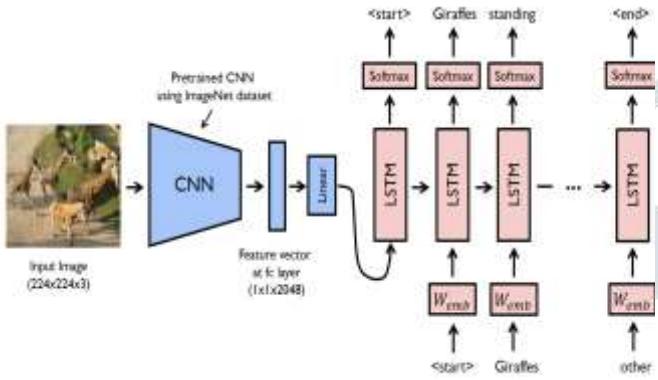

Fig.2 System Architecture

V. CONCLUSION

- In this paper, we implemented the image retrieval mechanism for generating proper captions according to various kind of images.
- A comparison is formed between using the important content and therefore the generated content for RS image retrieval purpose, from which we will notice that there's a mean gap of 0.3 in terms of mean BLEU score.
- So we concluded that to reduce this mean gap and to improve this current system, the technology can be implemented called caption generation block in future.

VI. REFERENCE

[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". 2019 IEEE.

[2] Genc Hoxha, Farid Melgani, Bengim Demir \Retrieving images with generated textual descriptions". 2019 IEEE.

[3] Seung-Ho Han and Ho-Jin Choi \Domain speci_c image caption generator with semantic ontology". 2020 IEEE.

[4] Varsha Kesavan, Vaidehi Muley, Megha Kolhekar \Deep Learning based Automatic Image Caption Generation". 2019 IEEE.

[5] Sarthak Singh Rawat, Kartikeyan Singh Rawat, Rahul Nijhawan \A Novel Convolutional Neural Network-Gated Recurrent Unit approach for Image Captioning". 2020 IEEE.

[6] Jingqiang Chen, Hai Zhuge \News Image Captioning Based On Text Summarization Using Image As Query". 2019 IEEE

[7] Harshit Parikh, Harsh Sawant, Bhautik Parmar, Rahul Shah, Santosh Chapaneri, Deepak Jayaswal \Encoder-Decoder Architecture for Image Caption Generation". 2020 IEEE.

[8] Adela Puscasiu, Alexandra Fanca, Dan-Ioan Gota, Honoriu Valean \Automated image caption". 2020 IEEE.

[9] BinqiangWang , Xiaoqiang Lu , Senior Member, IEEE, Xiangtao Zheng , and Xuelong Li 'Semantic Descriptions of High-Resolution Remote Sensing Images' 2019 IEEE.

[10] Harald Scheidl, Stefan Fiel, Robert Sablatnig Computer Vision Lab TU Wien \Word Beam Search: A Connectionist Temporal Classi_cation Decoding Algorithm" 2018 IEEE