IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

IJARSCT
Impact Factor: 6.252

Volume 2, Issue 3, April 2022

Review on Image Caption Generation

Aishwarya Mark[1], Sakshi Adokar[2], Vageshwari Pandit[3], Rutuja Hambarde[4], Prof. Swapnil Patil[5]

Students, Department of Computer Science and Engineering[1,2,3,4]
Faculty, Department of Computer Science and Engineering[5]
SKN College of Engineering, Pune, Maharashtra, India

Abstract: *With the rapid development of Deep learning, AI along with Computer Vision and Natural Language processing Image caption has become an interesting and complex task. Image caption generation is the process of generating textual description of the given image and it is a challenging task because it consists of apprehension of objects. If the machine will be programmed to accurately describe an image or environment like human vision, it will be highly beneficial for robotic vision, business and many more***.** *In order to generate an effective description of the image, the machine needs to detect, recognize objects as well as understand the scene type or location, object properties, their relationships and their interactions with each other. In this paper, we focus on advanced image captioning techniques such as CNN (Convolutional Neural Network)-LSTM(Long Short Term Memory) to generate meaningful captions. and the advantages and limitations of each method are discussed.*

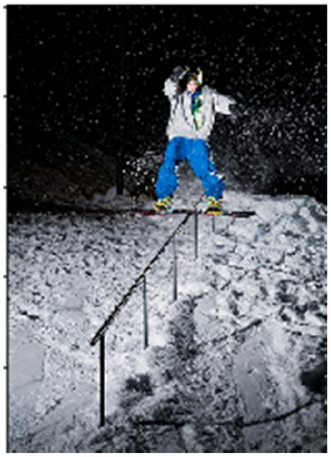
Keywords: AI, Deep learning, CNN, LSTM

I. Introduction

Every day we encounter images in many ways; e.g., the Internet, news articles, document diagrams and advertisements. Humans usually find it easy to interpret these images and give a textual description. However, if machines need to give a textual description of an image, the machines need to understand the semantic and the context of the image. A long-standing goal in the field of Artificial Intelligence is to enable machines to see and understand the images of our surrounding. Most photo posts on social networks like Facebook and Instagram hardly contain any description or caption. Hence, lots of opinions and emotions are conveyed through visual content only. Today, social networks have grown to be one of the most important sources for people to acquire information on all aspects of their lives. Social media images provide a potentially rich source for understanding public opinions/sentiments. Such an understanding of images may in turn benefit or even enable many real-world applications such as advertisement, product based recommendation, marketing and health-care.

In the past few years, computer vision in image processing field has made significant progress, like image classification [1] and object detection [2]. Due to this, it has become possible to automatically generate one or more sentences to understand the visual content of an image, which is the problem known as Image Captioning. Generation of complete natural image descriptions automatically has many potential impacts, such as titles attached to news images, information associated with medical images, text-based image retrieval, information accessed for blind users, human-robot interaction. These applications in image captioning have important theoretical and practical research value. Therefore, image captioning is a more complex but also a meaningful task in the age of artificial intelligence**.** Mimicking the human ability of providing descriptions for images by a machine is itself an impressive step along the line of Artificial Intelligence. The major challenge of this task is to capture how objects relate to each other in the image and to express them in a natural language. Basically, this model takes image as input and gives caption for it. With the advancement of the technology the efficiency of image caption generation is also increasing. The organizational structure of this paper is as follows. The second part focuses on Literature survey and third part mainly introduces and analyzes the CNN and LSTM model of image captioning and its design ideas. finally the proposed model for image caption generation and conclusion.

DOI: 10.48175/IJARSCT-3247



Input image

Output: Man is skateboarding on ramp

II. CNN AND LSTM

This section mainly introduces the theoretical concepts of CNN and RNN for image caption generation.

2.1 CNN for Extracting Features

Image captioning techniques are mainly categorized into two types, based on the template method and another is based on the encoder decoder structure. A Convolutional Neural Network or CNN is a deep learning neural network which is specifically designed for processing structured arrays of data such as images. Convolutional neural network is really good at identifying the key features and patterns in the input image such as lines, circles, even eyes and faces. Another property of CNN is that it can directly work on raw image without preprocessing makes it so powerful. It has many applications such as Photo and Video recognition, Image classification, medical image analysis, Computer vision, Natural language processing (NLP)etc.

The mathematical function of convolution is a special kind of linear operation in which two functions are multiplied to produce a third function that expresses how the shape of one function is modified by the other which is denoted by the word "Convolution" in Convolution Neural Network. A convolutional neural network is a feed-forward neural network, often with up to 20 or 30 layers. With three or four convolution layers it is possible to recognize handwritten digits and with 25 layers it is possible to distinguish human faces. The Basic layers of a CNN architecture are Convolution Layer, Pooling Layer & Fully Connected Layer. Addition to these three layers, there are two more important parameters which are the dropout layer and the activation function .

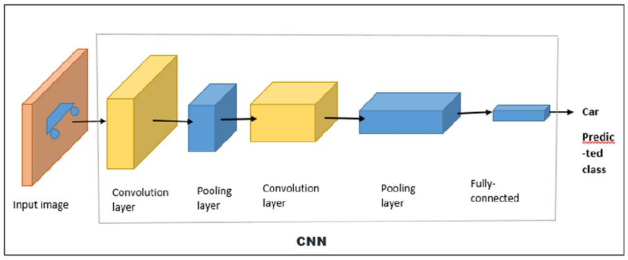


Figure 1: CNN Architecture

Convolutional Layer is the first layer that mainly extracts the various features & characteristics from the input images. In this layer, convolution mathematical operation is carried out between the input image and a filter of a specific size MxM. The dot product is taken between the filter and the sections of the input image with respect to the size of the filter by sliding the filter across the input image (MxM).The Feature map is the outcome, and it consist of information of the image such as its corners and edges. This feature map is then given as input to further layers, which learn a range of other features from the input image.

A Pooling Layer is usually applied after the Convolutional Layer. The main goal of this layer is to reduce the size of the collapsed feature map to reduce computational costs. This is achieved by reducing the connections between layers and independently operating on each feature map. There are different types of Pooling procedures, depending on the mechanism used.

Fully connected (Fc) layers are mainly used to consist of weight and bias with neurons, and are used to connect neurons between two different layers. These layers are usually placed in front of the start layer and form the last layer of the CNN architecture. In this, the input image from the previous layers are flattened and fed to the FC layer. The flattened vector then passes through some additional FC layers where the mathematical functions operations usually occurs. The classification process begins at this stage.

Dropout, when all the features are connected to the FC layer, it can cause over fitting in the training dataset. Over fitting occurs when a specific model works so well on the training data causing a negative impact in the model's performance when used on a new data. To conquer this problem, a dropout layer is applied in which some neurons are dropped from the neural network in the course of training process resulting in reduced size of the model. On providing a dropout of 0.3, 30% of the nodes are dropped out randomly from the neural network.

Finally, one of the most important parameters of the CNN model is the activation function. They are mainly used to learn and approximate any type of continuous and complex relationship between variables of the network. In simple words, it determine which information of the model to deliver forward and which ones should not at the end of the network. It adds non-linearity to the network. There are several commonly used activation functions such as the ReLU, Softmax, tanH and the Sigmoid functions. Each of these functions have a specific purpose. For a binary classification CNN model, sigmoid and softmax functions are recommended for a multi-class classification, generally softmax is used.

2.2 LSTM for Generating Captions

The main drawback with RNN was that vanishing/exploding gradient effect could occur, if the sequence is very large or if neural network has more than one hidden layer due to back propagation. To overcome these issues Long Short Term Memory (LSTM) was developed. LSTM is a type of RNN architecture that addresses the vanishing/exploding gradients and allows learning of long term dependencies. LSTM has risen to prominence with state-of-the-art performance in speech recognition, language modeling, translation, image captioning. LSTM can preserve information for longer periods when compared to RNN. It mainly uses long-term memories (info collected long time back) and short-term memories (info that is collected a few timestamps back) along with current event to generate a new modified long-term memory. It is done by trying to to "remember" all the past knowledge that the network seen so far and by "forgetting" irrelevant data. In simple words at each time step, it will filter the memory which needs to be passed to the next time step.

A. LSTM Architecture

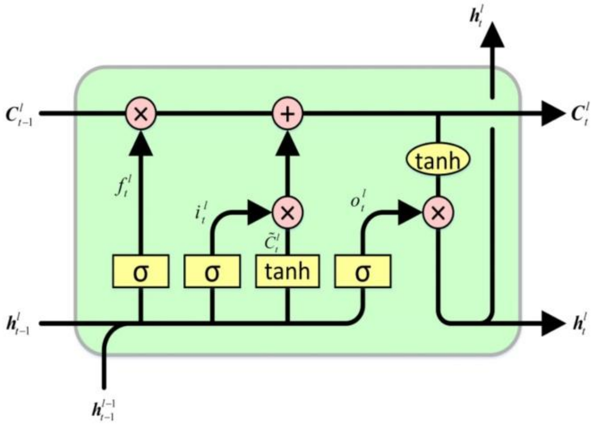


Figure 2: LSTM Architecture

There are mainly two outputs from one LSTM unit that are "Ct" and "ht". The hidden state ht is the short-term memory that is obtained from the immediately previous steps and vector Ct is the Cell state which is responsible for storing the long-term memory events. LSTMs will make use of a mechanism called gates to add and remove certain information into this cell state. LSTM Network mainly consists of four different gates for different purposes as described below:-

1. Forget Gate: It determines which information from the previous data should be discarded.
2. Input Gate: It determines what information can be written onto the Cell State from current input.
3. Remember Gate: It is used to modulate the information that the Input gate will write onto the Internal State Cell.
4. Output Gate: It determines what output(next Hidden State) to generate from the current Internal Cell

Working of an LSTM Recurrent Unit:

It firstly takes input from the current input, the previous hidden state, and the previous internal cell state. Then it Calculates the values of the four different gates by:

1. **By** calculating the parameterized vectors for the current input and the previous hidden state by element-wise multiplication with the concerned vector with the respective weights for each gate.
2. Applying the respective activation function for each gate element-wise on the parameterized vectors.
3. Then Calculate the current internal cell state by first calculating the element-wise multiplication vector of the input gate and the input modulation gate, then calculate the element-wise multiplication vector of the forget gate and the previous internal cell state and then adding the two vectors.

   c_{t} = i\odot g + f\odot c_{t-1}
4. Lastly Calculate the current hidden state by first taking the element-wise hyperbolic tangent of the current internal cell state vector and then performing element-wise multiplication with the output gate. Some of the drawbacks of LSTMs are longer training times, large memory requirements, unable to parallel training, etc.

III. Proposed Model

Our model includes use of deep learning for image captioning. We are using two techniques mainly CNN and LSTM for image classification. So, to make our image caption generator model, we will be merging these architectures i.e. CNN-LSTM model. It is also known as encoder-decoder model**.** The neural network-based image captioning methods work as just simple end to end manner. These methods are very similar to the encoder-decoder framework-based neural machine translation. In this network, global image features are extracted from the hidden activations of CNN and then fed them into an LSTM to generate a sequence of words.

The CNN-LSTM architecture is built by using CNN layers for feature extraction on input data combined with LSTMs to support sequence prediction. This model is specifically designed for sequence prediction problems with spatial inputs, like images or videos. They are widely used in Activity Recognition, Image Description, Video Description and many more. CNN-LSTMs are generally used when their inputs have spatial structure, such as the 2D structure or pixels in an image or the 1D structure of words in a sentence, paragraph, or document and also have a temporal structure in their input such as the order of images in a video or words in text, or require the generation of output with temporal structure such as words in a textual description.

As we are using LSTM over RNN, we are introducing more & more controlling knobs, which controls the flow and mixing of Inputs as per trained Weights. And thus, bringing in more flexibility in controlling the outputs. So, LSTM gives us the most Control-ability and thus, Better Results

- CNN-RNN Model: Objection detection using CNN: CNN provides optimistic results for object detection and will be best suited for image captioning.
- RNN-LSTM for generating captions: RNN-LSTM will be used to generate meaningful captions from the image and object detection features. The input will be object detection and the output will be caption for the particular image.

In past few years image captioning has made significant improvement. The neural image caption generator gives a beneficial framework for learning to map from various images to human-level image captions. Neural networks can handle all of the issues by generating suitable, expressive and highly fluent caption using tensorflow and algorithms. The content-based image retrieval efficiency can be enhanced by text description of the images, the expanding application scope of visual understanding in the fields of science, security, defense and other fields, which has wide application prospect. This Image Captioning deep learning model is very useful to inspect the large amount of unstructured and unlabeled data to detect the patterns in those images for guiding the Self driving cars, for building the software to guide blind people.

DOI: 10.48175/IJARSCT-3247

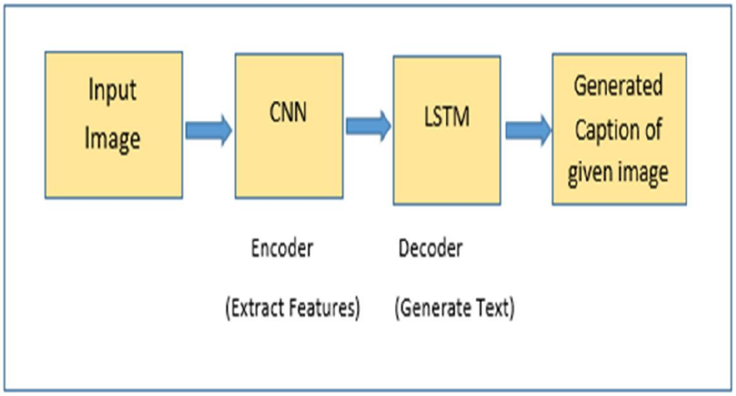


Fig 3. A block diagram of simple Encoder- Decoder

Architecture - based image captioning

IV. Dataset

We are using the Flickr8k dataset as a standard benchmark dataset for the sentence description of the image. This dataset consists of 8000 images with five captions for each image. Each caption provides a clear description of entities and events present in the image. The dataset represents a diversity of scenarios and events and doesn't have images of well-known people and places so that dataset won't be more generic. It is divided into parts as 6000 images in the training dataset, 1000 images in the development dataset, and 1000 images in the test dataset.

The advantages of using this dataset for this project are:

- Single image is mapped for multiple captions to make the model generic and avoid overfitting the model.
- Various categories of training images can make the image captioning model work for multiple categories of images and hence can make the model more robust.

V. Conclusion

In this paper, we have implemented a deep learning approach for generating captions for the images. Our described model is based upon a CNN feature extraction model that encode an image into a vector representation, followed by LSTM decoder model that can generates corresponding sentences based on the image features learned .

References

[1]. Philip Kinghorn, Li Zang, “a region based image caption generator with refined descriptions” , Elsiver B V, 6 july 2017, Ling Shao University Northumbria New castle NE1,United Kingdom.

[2]. Priyanka Raut, Rushali A Deshmukh, “An Advanced Image Captioning using combination of CNN and LSTM”, Turkish Journal of Computer and Mathematics Education, 05 April 2021, Savitribai Phule Pune Univresity, faculty, Maharhatra/India.

[3]. Shuang Liu, Liang Bai, Yanli Hu and Haoran Wang ,“Image Captioning Based on Deep Neural Networks”, MATEC Web of Conferences ,2018 ,College of Systems Engineering, National University of Defense Technology,410073 Changsha, China.

[4]. Raj kadam , Uday Kumbhar , Onkar Gulik , Dr Makrand Shahade, “Object Detection and Automatic Image Captioning Using Tensorflow”, International Journal of Future Generation Communication and Networking2020, Scholar, Department Of Computer Engineering, JSPM’s RSCOE Pune.

[5]. Priyanka Kalena , Aromal Nair,Nishi Malde, Saurabh Parkar “Visual Image Caption Generator Using Deep Learning”, ICAST-2019,K.J Somaiya College Of Engineering, Mumbai.

DOI: 10.48175/IJARSCT-3247