IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

IMAGE CAPTION GENERATOR USING MACHIN LEARNING

Mrs. Nisha Patil, Nishant Dabhade, Sachin Gaund, Utsav Chaudhary, Omkar Govardhane

Information Technology

Sandip University, Nashik, India

Abstract: **The paper basically describes about the anomaly of demonstrating the mathematical methods and the algorithm required for their recognition. With having the straight intent of new and advance algorithm for the identification of various images and heterogenous data processing. This paper mainly focuses on image caption generator, in simple terms it is a data generator by using the input as an image. This paper consists of different approaches to a straight-line pattern sentence formation using different machine learning techniques. Several machine learning algorithms like Multiplayer Perception, Support Vector Machine, Convolutional Neural Network, and many more. The main purpose or the main objective of the following paper is that how the simple image having a content gets converted into the simple English sentence which can be readable to any of the simple English learner. The paper shows how different algorithm shows different output with different accuracy.**

Keywords**: Image recognition, machine learning, machine learning algorithm, neural network, classification algorithms.**

1.INTRODUCTION

From centuries, the mode of communication is getting change to communicate with each other. From starting with the transformation of data using birds and moving in today's digital world the modes have changed a lot. Handwritten notes pass the most important aspect in passing the data and now the internet to pass the same role. But still there is some leakage in data transfer due to internal problems and extracting data from images plays a major role in todays world to transfer the data. In this research paper we are going through how all the data is been is extracted from a image and get converted into simple English sentence so that each one can get it easily. We have use R-CNN with VGG16 Module to extract the information from the image. The main purpose of creating this Image caption generator is that it involves computer vision and natural language processing concept to recognize the context of the image into simple natural language like English. We know that Convolutional Neural Networks are used for image classification and object localisation. Recurrent neural network is mainly used for text generation.

Human Beings usually describe a sense of creating or owning the scene using natural language, which is concise and compact. However, machine vision system describes the sentence by using the scene of taking image which is a 2-dimension array. The main idea is to map the image and the sentence with each other and form a grouping of it to perform the sentence formation task. Proposed a more general Long-term Recurrent Convolutional Network (LCRN) method. LCRN model is such a model which not only deal with one to many but it also deals with many to one model for its task creation. This work is based on the LCRN method.

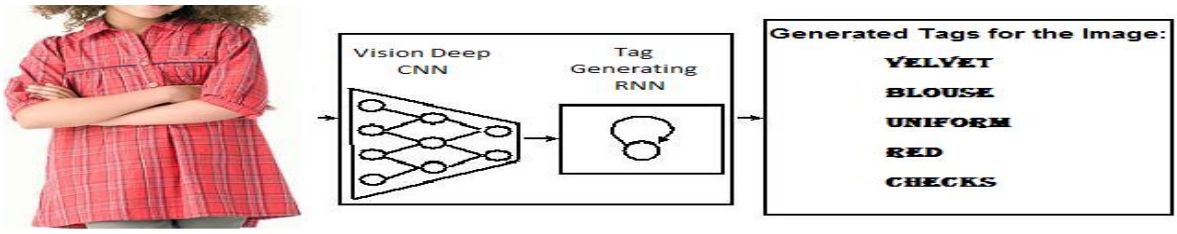


Figure: Image caption generation pipeline. The framework consists of CNN followed by RNN. It generates an English sentence from an input image.

ILLITRATURE REVIEW

In this section of our paper, we will highlight some of the below mentioned paper which is used as a reference to depict what different techniques were used by different researchers in the field of data extraction.

Zeeshan Khan, Sandeep Kumar and Anurag Jain put forward a paper based on Content Based Image Classification using Machine Learning Approach, in which they mentioned different techniques like KNN, DT and SVM which are used for image classification and present a detailed comparative analysis of the above techniques. They came to a conclusion that SVM performs better results as compared to the other techniques but finds out SVM still faces some problem related to feature outlier and core problem.

Salvador España-Boquerón, Maria J. C. B., Jorge G. M. and Francisco Z. M. [4], this paper outlines the hybrid Hidden Markov Model (HMM) is used to conceive the unconstrained offline handwritten texts. The main characteristics of the recognition systems is to produce a new way in the form of pre-processing and recognition which are both based on ANNs. The pre-processing is used to clean the images and to enhance the non-uniform slant and slope correction. Whereas the recognition is used to estimate the emission probabilities.

Karpathy (Karpathy and Fei-Fei) proposed a visual-semantic alignment (VSA) method. The method generates descriptions of different regions of an image in the form of words or sentences. Technically, the method replaces the CNN with Region-based convolutional Networks (RCNN) so that the extracted visual features are aligned to particular regions of the image. The experiment shows that the generated descriptions significantly outperform retrieval baselines on both full images and on a new dataset of region-level annotations. This method generates more diverse and accurate descriptions than the whole image method such as LRCN and NIC. The limitation is that the method consists of two separate models. This method is further developed to dense captioning (Johnson et al., 2016) and image-based question and answering system

DESCRIPTION OF A PROBLEM:

TASK: In this project we want to build a system that can generate simple English sentence from an image of RGB input. Let us consider the following equation

$$S=f(I)$$

Where *I* is a RGB image and *S* is a simple English sentence. *F* is a function that we are going to learn in this project by performing the various task.

We have used flicker 8k dataset at VGG16 to complete this project as this dataset has a good number of trained image and we just have to trained the remaining images. Flicker 8K dataset has 6k images as trained with the following sentence to perform a task and remaining 2k images we are going to trained in this following project.

Requirement Analysis:

Flickr8k_Dataset.zip (1 Gigabyte) An archive of all photographs.

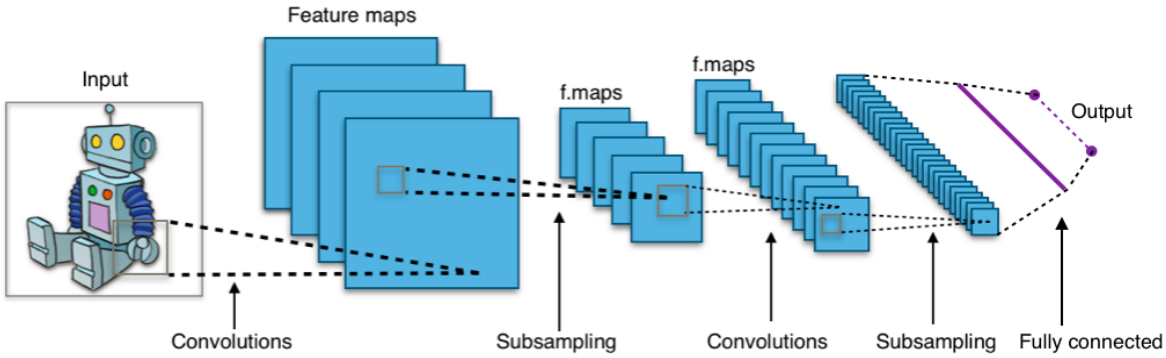
Flickr8k_text.zip (2.2 Megabytes) An archive of all text descriptions for photographs

Flickr8k_Dataset: Contains 8092 photographs in JPEG format.

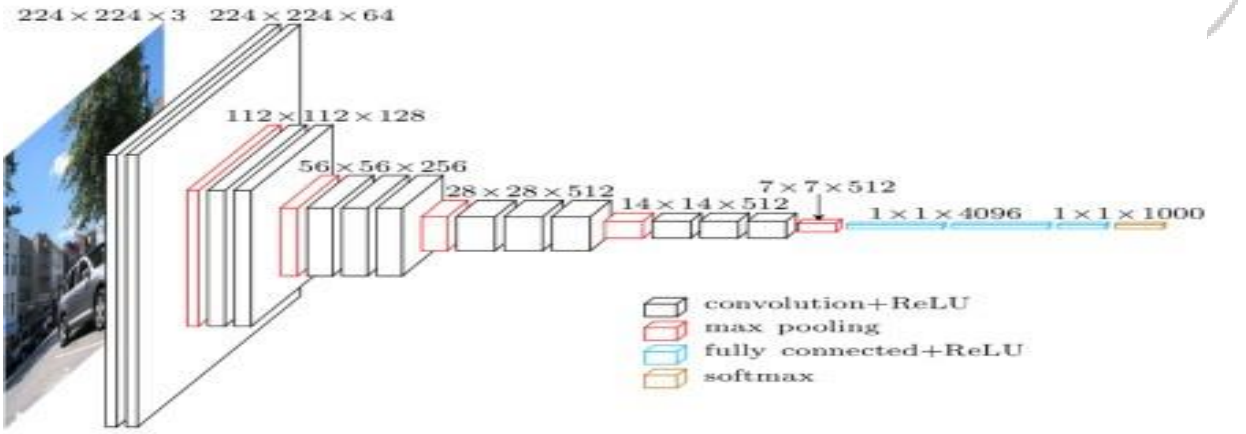
Flickr8k_text: Contains a number of files containing different sources of descriptions for the photographs

CONVOLUTIONAL NEURAL NETWORK (PRINCIPLE)

In this project convolutional neural network perform the most important part to convert the image into simple English sentence. CNNs have broken the mold and ascended the throne to become the state-of-the-art computer vision technique. Among the different types of neural networks (others include recurrent neural networks (RNN), long short term memory (LSTM), artificial neural networks (ANN), etc.), CNNs are easily the most popular. These convolutional neural network models are ubiquitous in the image data space. They work phenomenally well on computer vision tasks like image classification, object detection, image recognition, etc. The big idea behind CNNs is that a local understanding of an image is good enough. The practical benefit is that having fewer parameters greatly improves the time it takes to learn as well as reduces the amount of data required to train the model. Instead of a fully connected network of weights from each pixel, a CNN has just enough weights to look at a small patch of the image. It's like reading a book by using a magnifying glass; eventually, you read the whole page, but you look at only a small patch of the page at any given time.



One can see the image has been inputted the convolution layer extract the information from the image it gets converted into the feature maps and subsampling is been done further f. maps does and the same process is being carried out number of times until the fully connected layer does not exist and we get an output.



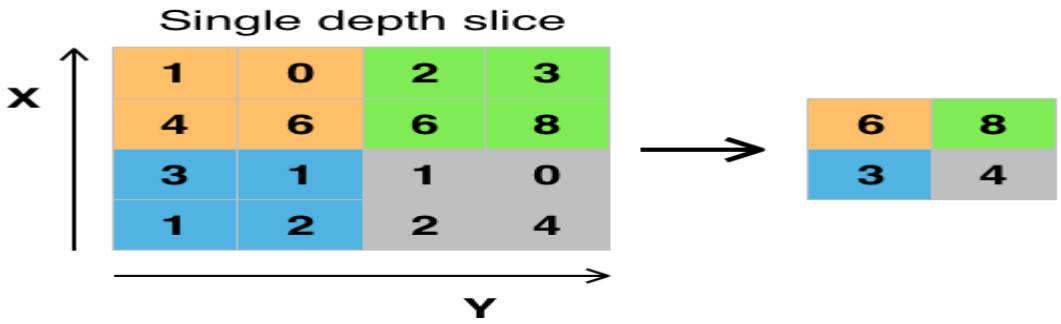
A CNN architecture is formed by a stack of distinct layers that transform the input volume into an output volume through a differentiable function. A few distinct types of layers are commonly used. These are further discussed below.

CONVOLUTIONAL LAYER

The convolutional layer is the core building block of a CNN and the project Image Caption Generator. This layer extract all the important context from image with the help of RELU.

POOLING LAYER

Pooling mainly works on the non-linear functions. Pooling mainly consist of two types max and min pooling, but it always works on max-pooling. Max pooling means let us considered the (fig 2) , you can see the number of boxes with same colour. Max pooling will work on it and remove the maximum amount of needed data from the boxes.



RELU LAYER

Relu is a rectified linear image which works on the liner equation of the image. Were the derivate of 0 and 1 is been sent to SoftMax to convert it into the simple English sentence.

Recurrent neural network

To prevent the vanishing problem, the long short-term memory (LSTM) method is used as the RNN component. A simplified LSTM updates for time step t given inputs xt, ht−1, and ct

$$it = \sigma (Wxixt + Whiht{-}1 + bi)$$

$$ft = \sigma (Wxfxt + Whfht{-}1 + bf)$$

$$ot = \sigma (Wxoxt + Whoht{-}1 + bo)$$

$$gt = \varphi (Wxcxt + Whcht{-}1 + bc)$$

$$ct = ft$$

$$ct{-}1 + it$$

$$ht = ot$$

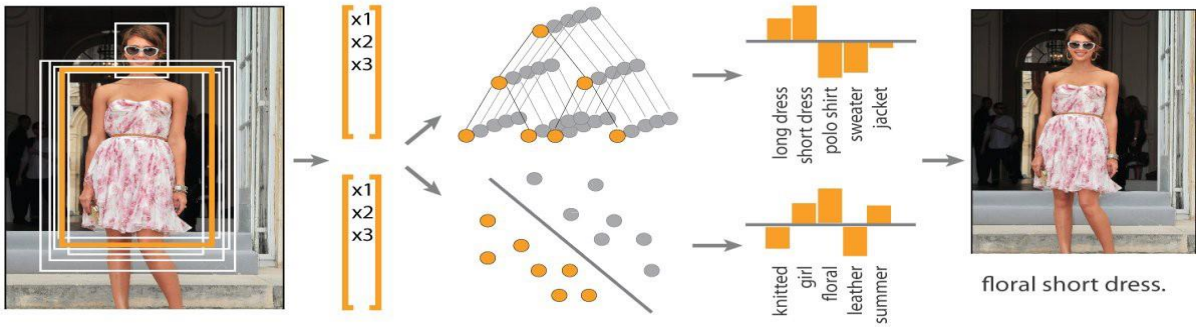
where $\sigma(x) = (1 + e - x) - 1$ and $\varphi(x) = 2\sigma(2x) - 1$. In addition to a hidden unit $ht \in RN$, the LSTM includes an input gate $it \in RN$, forget gate $ft \in RN$, output gate $ot \in RN$, input modulation gate $gt \in RN$, and memory cell $ct \in RN$. These additional cells enable the LSTM to learn extremely complex and long-term temporal dynamics. Additional depth can be added to LSTMs by stacking them on top of each other.

SENTENCE GENERATION

The output of LSTM is the probability of each word in the vocabulary. Basically, beam search is used to generate sentences. Beam search is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set.

IMPLEMENTATION

As we want to stay with the architecture that we have of an CNN we have an image of 224*224 and we nulled it into the smaller and cropped its size. As we know the same image will get cropped multiple times the CNN knows the actual image with its actual size and then they will be cropped up to the SoftMax layer doesn't exits and hence in fully connected layer we will gain an output. Pre-processing Because we want to keep the architecture of the CNN, the input image is randomly cropped to the size of 224 *224. As a result, only part of the images is used in training at particular iteration. Because one image will be cropped multiple times in the training, the CNN can probably see the whole image in the training (once for part of the image). However, the method only sees part of the image in the testing except the dense cropping is also used. For the sentences, the method first creates a vocabulary only from the training captions and removes lower frequency words (less than 5). Then, words are represented by one-hot vectors.



Given that fact, the complete image classification pipeline can be formalized.

Future Prospects:

For future work, we propose the following four possible improvements:

An image is often rich in content. The model should be able to generate description sentences corresponding to multiple main objects for images with multiple target objects, instead of just describing a single target object.

For corpus description languages of different languages, a general image description system capable of handling multiple languages should be developed.

Evaluating the result of natural language generation systems is a difficult problem. The best way to evaluate the quality of automatically generated texts is subjective assessment by linguists, which is hard to achieve

CONCLUSION:

In this overview, we have compiled all aspects of the image caption generation task, discussed the model framework proposed in recent years to solve the description task, focused on the algorithmic essence of different attention mechanisms, and summarized how the attention mechanism is applied. We summarize the large datasets and evaluation criteria commonly used in practice. In this advanced Python project, we have implemented a CNN-RNN model by building an image caption generator. Some key points to note are that our model depends on the data, so, it cannot predict the words that are out of its vocabulary. We used a small dataset consisting of 8000 images. For production-level models, we need to train on datasets larger than 100,000 images which can produce better accuracy models. 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:

REFERENCE:

[1] Zeeshan Khan, Sandeep Kumar, Anurag Jain, "A Review of Content Based Image Classification using Machine Learning Approach", International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume2 Number-3 Issue-5 September-2012, pp 55-60.

[2] K. Gaurav and Bhatia P. K., "Analytical Review of Pre-processing Techniques for Offline Handwritten Character Recognition", 2nd International Conference on Emerging Trends in Engineering & Management, ICETEM, 2013.

[4] Batyrkhan Omarov, Young Im Cho," Machine Learning based Pattern Recognition and Classification Framework Development", Department of Computer Engineering, Gachon University, Seoul, 1342, Republic of Korea.

[5] Sultan Alhusain and Simon Coupland, Robert John, Maria Kavanagh," Towards Machine Learning Based Design Pattern Recognition".

[6] Anuj Dutt, AashiDutt," Handwritten Digit Recognition Using Deep Learning", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 6, Issue 7, July 2017, ISSN: 2278 – 1323.