International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 25th Jun 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue= Spl Issue 05

DOI: 10.48047/IJIEMR/V11/SPL ISSUE 05/23

Title IMAGE CAPTION GENERATION USING DEEP LEARNINGVolume 11, SPL ISSUE 05, Pages: 151-156

Paper Authors
Mr. Y. Nagendra Kumar, SD. Beebi, N. Jahnavi, R. Anil Kumar, S. Isaac Roshan

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IMAGE CAPTION GENERATION USING DEEP LEARNING

Mr. Y. Nagendra Kumar[1], SD. Beebi[2], N. Jahnavi[3], R. Anil Kumar[4], S. Isaac Roshan[5]

[1] Assistant Professor, Dept. Of CSE, [2]18ME1A05A0, [3]18ME1A0575,
[4]18ME1A0594,[5]18ME1A05A5.
Ramachandra College of Engineering, A.P., India
nagendrayakkala@rcee.ac.in, sayedbeebi2001@gmail.com,
jahnavi555n@gmail.com, anilkumar10rayavarapu@gmail.com, isaacroshanism@gmail.com

Abstract:
In this work, we use CNN and LSTM to learn about the image caption. Image caption generation is a system that recognises the image's connection in English using natural language processing and computer vision standards. We cautiously examine a number of important principles of photograph captioning and its familiar methods in this research study. Picture captioning is the process of creating a description for an image. It necessitates identifying the relevant objects in an image, their qualities, and the relationships between them. For the purposes of this work, we discuss the Keras library, NumPy, and Jupiter notebooks. We also discuss the Flickr dataset and the CNN algorithm for photo classification.

Keywords- CNN, LSTM, image captioning, deep learning, python, OpenCV.

Introduction:

We see a lot of images in the news every day, social media, and in the environment newspapers. Humans have the ability to perceive patterns merely the photos themselves, but machines on the other hand, require images to be taught, after which it would generate the automatically captioned photos. Image captioning could be useful in a variety of situations, such as assisting the visionless person with text-to-speech via real-time feedback about the situation over a camera feed, and improving social medical leisure by reorganising captions for photographs in social feeds as well as messages to speech.

In the generation of virtual world, conventional artwork of writing is being changed via way of means of virtual artwork. Digital artwork refers to types of expression and transmission of artwork shape with virtual shape. Relying on cutting-edge technology and era is the exclusive traits of the virtual manifestation. Traditional artwork refers back to the artwork shape that's created earlier than the virtual artwork. From the recipient to analyse, it may absolutely be divided into visible artwork, audio artwork, audio-visible artwork and audio-visible imaginary artwork, which incorporates literature, painting, sculpture, architecture, music, dance, drama and different works of artwork. Digital artwork and conventional artwork are interrelated and interdependent. Social improvement isn't a people`s will, however the wishes of human lifestyles are the primary riding pressure anyway. The equal state of affairs occurs in artwork.

3. Implementation Techniques:

CNN-

Revolutionary CNN is useful for working with photos because neural systems are

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

specific important neural systems that can provide information with an information shape, such as a 2D lattice. It analyses images from the left to the right corner and all the way through to extract major highlights and consolidate the element to define images. It can handle images that have been interpreted, rotated, scaled, and changed. The neural system is a sophisticated learning calculator that examines the data and assigns meaning to the picture's numerous components / protests, and it is recognised by each other.

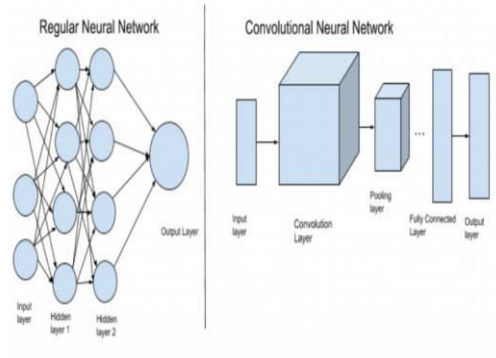


Fig.1 CNN Architecture

When compared to other order calculations, ConvNet requires very less pre-handling. Despite the fact that channels are hand-designed in rudimentary techniques, ConvNets is capable of learning these channels/highlights with proper preparation. The curving system's shape is inspired by the way the visual cortex is organised and is similar to the neural network design found within the human brain. Singular neurons only respond to upgrades in a small area of the observable field called as open field. The collection of such fields encompasses all visual areas.

CNN: Architecture - When it comes to interpreting huge photos and videos, a pure primitive neural network, in which all neurons in one layer merge with all neurons in the next layer, is inefficient. The range of restriction using an acceptable neural system for a standard size picture with many picture pieces called pixels and 3-tone colours (RGB i.e., red colour, green colour, blue colour) will be in the thousands, which can lead to overfitting.

CNN uses a 3D arrangement in which each adjustment of neurons breaks down a little area or "highlight" of picture to constrain effective quantities of constraints & recognition of the neural system on significant pieces of picture. Instead of all neurons skipping to the next brain layer, each group of neurons spends a significant amount of time differentiating one aspect of the picture, such as a nose, left ear, mouth, or leg. The final result is a point of scope, demonstrating how plausible each of the skills is chosen as a member of the class.

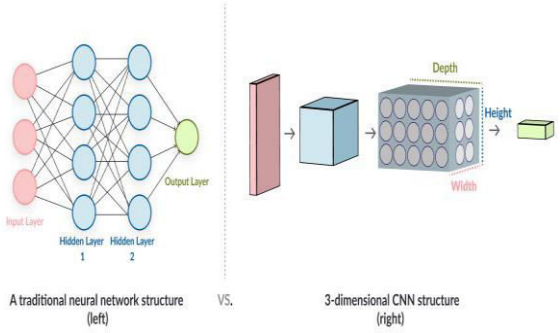
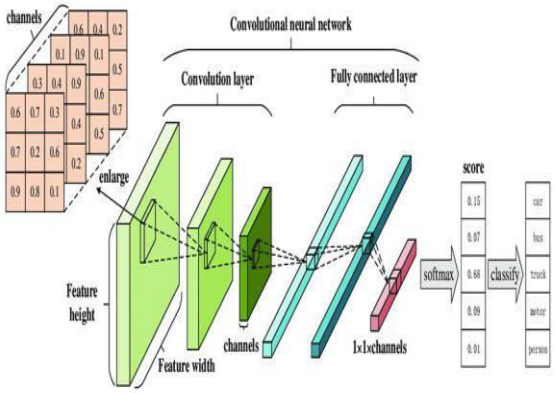


Fig.2 Working of CNN

How does CNN work?
As previously discussed, a fully connected neural network, in which every input in the preceding layers is connected to every input in the following layers, is useful for the task at hand. Along these lines, CNN suggests that the neurons in a cell may be connected to a specific cell area before it, rather than all the neurons in the same way.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org



This helps reduce the complexity of neurons Gain network and less computing power. when for each new computer, under the standard image Use a number for each pixel. We are generally Compare two images checking pixel values All pixels. This technique is useful for comparison Only two identical photos, If the images to be compared are different, the comparison will fail. CNN makes image comparisons little by little. piece.

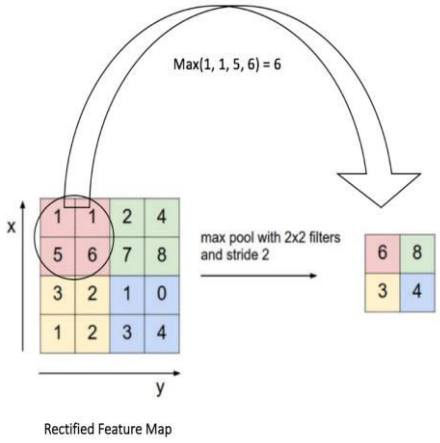


Fig.3 Feature map of CNN picture

The main reasons for using the CNN algorithm are: This is the only algorithm for taking pictures as input, based on the input image Draw a feature map. That is, classify each pixel Based on similarities and differences. CNN Pixels are categorized and a matrix is created It is called a feature map. There is one feature map A collection of similar pixels placed elsewhere Category. These matrices are Find the essence of the problem in the input image.

More about CNN –
There is total 3 types of layers in CNN model.

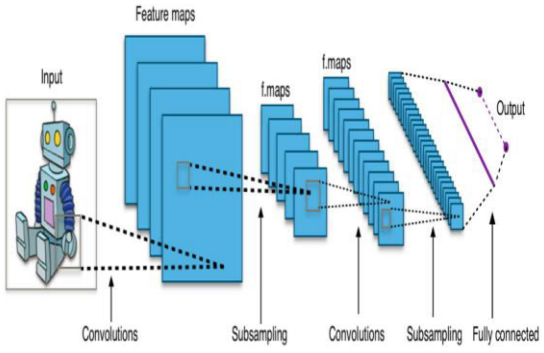
1. Convolutional

2. Pooling

3. Fully connected

The input image is read through the CNN in the first layer, and a feature map is built on top of that. It acts as an input to the feature map from there. The Pooling layer is one of the following layers. The feature map is divided into sections by the pooling layer. This layer adds depth to the feature map to evaluate the context of the photograph by finding the extra easy components order to find the most important information regarding the image.

The first and second layers, Convolutional and Pooling, are practised numerous times depending on the image in order to obtain densed information about the image. These two layers combine to provide an extra dense feature map. The last layer, Fully Connected, makes use of this deep feature map. This layer is responsible for classification. It arranges the pixels according to their similarities and differences. Classification is carried out to an extreme degree in order to extract the essence of the image and aid in the identification of objects, people, and things.



These layers aid CNN in locating and identifying visual elements. The image of fixed length inputs is turned into fixed size

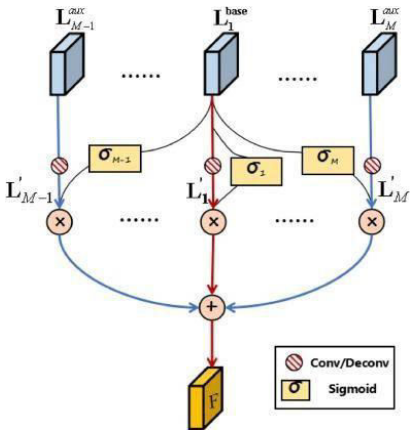
outputs by extracting important aspects from the image.

Computer Vision-
In the field of medical sciences, image analysis is done entirely by CNNs. With this, the inner structure of the body may be easily studied.

It's been used in mobile phones for a variety of purposes, including determining a person's age and unlocking the phone by scrutinising a photo taken with the camera.

Origin of LSTM-
In 1997, two German academics, Sepp Hochreiter and Jurgen Schmidhuber, were the first to investigate LSTM. Long short-term memory is referred to as LSTM. LSTM is a vital component in the Deep Learning discipline of recurrent neural networks. The unique feature of LSTM is that it not only maintains the input data, but it can also make predictions about forthcoming datasets using its own data. This LSTM network saves the input for a specific time period and then predicts or assigns future values to it. This is the key reason why LSTM is preferred over regular RNN.

LSTM (Long Short-Term Memory)-
LSTMs are a type of RNN that has the capacity to hold more data values than RNNs. LSTMs are frequently used in today's world. LSTMs are built in such a way that they can overcome the problem of Vanishing Gradients. Errors are always present in LSTM. And as a result of these mistakes, LSTM continues to examine the data values over time. It also makes learning backpropagation simple.



In theory, as shown in the picture above, the LSTM stores data in numerous gates before processing it and sending the output to the final gate. When we talk about RNNs, we used to think that they just passed the data straight to the final gate. The entire network can influence the data from these gates in LSTM in a variety of ways, including storing and reviewing the data. The gates in the LSTM are capable of making autonomous decisions about facts and data.Furthermore, by opening and closing the gates, these gates are capable of making decisions on their own.

LSTM Architecture: -
The LSTM design is fairly simple; it comprises of three major gates that store data for a longer amount of time and assist in solving problems that RNNs could not. The three principal gates covered by the LSTM are:

1.Forget gate – the fundamental function of the forget gate is to filter the data, i.e. to remove any data that will not be needed in the future to complete a certain task.This gate is in charge of the LSTM's overall performance and data optimization.

2.Input gate — the LSTM begins with this gate, i.e., the input gate. This gate receives input from the user and passes it on to other gates.

3.Output gate – This gate is in charge of properly displaying the desired outcome.

Python-
Python is a computer programming language often used to build websites and software, automate tasks, and conduct data analysis. Python is a general-purpose language, meaning it can be used to create a variety of different programs and isn't specialized for any specific problems.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Open CV-

OpenCV (Open-Source Computer Vision Library) is an open-source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in the commercial products. Ever desired to attract your creativeness through simply waiving your finger in air. Here we can discover ways to construct an Air Canvas that may draw something on it through simply shooting the movement of a colored marker with camera. Here a colored item at tip of finger is used because the marker. We may be the use of the pc imaginative and prescient strategies of OpenCV to construct this project. The favoured language is python because of its exhaustive libraries and smooth to apply syntax however knowledge the fundamentals it could be carried out in any OpenCV supported language. Here Colour Detection and monitoring is used with a purpose to gain the objective. The coloration marker in detected and a masks is produced. It in particular attention on device gaining knowledge of area for correct results. Machine gaining knowledge of is part of Artificial intelligence that is used for the look at of algorithms.

This makes the consumer to have an interactive surroundings wherein the consumer can draw some thing he desires with the aid of using deciding on his required colours from the displayed ones. So, we finish that Virtual Sketch is advanced the use of the library NumPy and in Open CV wherein we've many libraries and set of rules in constructed which makes the interfaces greater energetic whilst the use of . We used python as, it have many in-built libraries and lots of modules which constitute the creativeness truly while used at the side of OpenCV in addition to its morphological processes.

Problem Statement

Existing system-

In previous model it is based on a deep learning neural network that consists of a vision CNN. It generates complete sentences as an output captions or descriptive sentences. By using VGG16 model it can automatically describe the content of a picture using properly formed English sentences. When the image is given by the user which is out of dataset it was advantageous in naming the objects in an image but it could not tell us the relationship among them when the image is not clear.

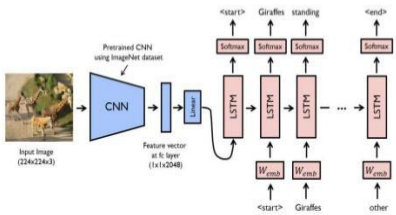
Proposed System-

In this we are differentiating the image given as input whether it is valid or not. In the existing one it was accepting the invalid input and generating the false captions. So, we are proposing a system to differentiate the blur image and clear image. The caption will be generated if the image is valid or else it generates an invalid message.

For blur detection we will be using Open CV in this the blur detection will be done with the help of threshold value. If the variance falls below a pre-defined threshold, then the image is considered *blurry*; otherwise, the image is *not blurry*. In this the Laplacian operator is used for identifying the clearness of image.

Image Caption Generation Model: (for input from dataset)- We'll combine the two architectures to create an image caption generation model. CNN-LSTM model is another name for it. So, to get the caption for the input photographs, we'll use these two architectures. CNN was utilised to extract the key features from the input image. LSTM has been used to store and process the data or features from the CNN model, as well as to assist in the production of a good caption for the image.

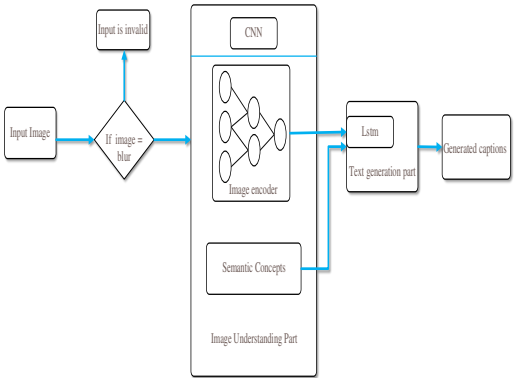
Model



Model (user image)-

In this model we will be using the CNN and LSTM for feature extraction and sentence

generation. And it will contain a initial condition block for blurness detection. Here user gives his own image for caption generation.



Conclusion-

The CNN-LSTM model was created with the intention of creating captions for the input images. This concept can be used to a wide range of situations. In this we looked into the CNN and LSTM models and finally the model creates captions for the data picture.

References-

[1] Abhaya Agarwal and Alon Lavie,
Meteor, m-bleu, and m-ter: are evaluation metrics for machine translation output that have a high connection with human ranks. In Third Workshop on Statistical Machine Translation Proceedings.115–118, Association for Computational Linguistics.
[2] Aker, Ahmet, and Gaizauskas, Robert.
Using dependency relational patterns to generate image descriptions. In Proceedings of the Association for Computational Linguistics' 48th Annual Meeting. 1250–1258, Association for Computational Linguistics.
[3] Basura Fernando, Mark Johnson, and Stephen Gould are among the cast members. 2016. Spice: Evaluation of semantic propositional image captions. In the European Computer Vision Conference 382–398. Springer.
[4] Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang.
[5] Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, and others 2016. Deep compositional captioning: describing novel object categories in the absence of matched training data. In IEEE Conference on Computer Vision and Pattern Recognition Proceedings.
[6] Yoshua Bengio, Dzmitry Bahdanau, and Kyunghyun Cho 2015. By learning to align and translate together, neural machines can translate. Learning Representations International Conference (ICLR).
[7] Shan An and Shuang Bai, 2018. Automatic Image Caption Generation: A Survey Neurocomputing. Vol. 0, No. 0, Article 0 of ACM Computing Surveys. Date of Acceptance: October 2018. 0:30 Hossain and colleagues.