Navajyoti, International Journal of Multi-Disciplinary Research

Volume 6, Issue 1, August 2021

IMAGE CAPTION GENERATOR USING DEEP LEARNING

[1]Sudeshna Bose, [2]Dr. Shilpa Abhang

[1]Student, Dept. of MCA, Jyoti Nivas College

[2]Asst Prof, Dept. of MCA, Jyoti Nivas College

Abstract

Automatically generating caption for an image using deep neural network is a complicated task as it connects computer vision and natural language processing. While some people might argue that human accuracy is a function of training time it can be said with great confidence that automated classification models are at least as good as trained humans in classification problems. The ability of these models to analyze and describe complex images, however, is still an active area of research. Image description is a good starting point for imparting artificial intelligence to machines by allowing them to analyze and describe complex visual scenes. Image description is a good starting point for imparting artificial intelligence to machines by allowing them to analyze and describe complex visual scenes. Image Captioning is the process of generating textual description of an image.

In our proposed work, we present a multilayer neural network method closely related to the human visual system that automatically learns to describe the content of images. Our model consists of two sub-models, one is Convolutional Neural Network (CNN) that is Image based model which extracts the features of the image and another one is Recurrent Neural Network (RNN) it is a Language based model which translates the features & objects to a sentence.

Keywords: CNN, RNN, Image.

I. Introduction

In the past few years, deep neural network has made significant progress in image processing area, like image classification, object detection. However, tasks like image classification and object detection are far from the end of image understanding. One ultimate goal of image processing is deep image understanding to machines i.e. understanding the whole image scenario not individual objects. Image captioning follows the same path: by extracting the complete detail of individual object and their associated relationship from image. Finally, the system can automatically generate a neural sentence to describe the image. This problem is extremely important, as well as difficult because it connects two major artificial intelligence fields: computer vision and natural language processing. Computer vision enables the computers to see, identify and process images in the same way that human vision does, and then provide appropriate output. Natural Language Processing (NLP) is the analyzing, understanding and generating the languages that humans can understand.In the past years, supervised convolutional models have forever changed the computer vision and machine learning landscape. Due to the recent introduction of large supervised datasets and accelerated training models using Graphic Processing Units (GPUs), the traditional pairing of hand crafted low level vision features with complimentary classifiers have been bested by Convolutional Neural Networks (CNNs). Besides, a deep Recurrent Neural Network (RNN) based on Long Short-Term Memory (LSTM) units with attention mechanism for sentences generation. For training and testing Flickr8k dataset have been used.

II. PROPOSED WORK

In this project, I want to build a system that can generate an English sentence that describes objects, actions or events in an RGB image. The task of image captioning can be divided into two modules, one is Image based model, which acts as an encoder and extracts the features out of our image and we usually rely on a Convolutional Neural Network model for this purpose. Another is Language based model which acts as the decoder and translates the features & objects to a sentence and we use Recurrent Neural Network for this model. In Recurrent Neural Network I will use LSTM which refers Long Short Term Memory. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information. For training and testing Flickr8k dataset have been used, demonstrating state-of-the-art description results. Flilckr8K contains 8,000 images that are each paired with five different captions which provide clear descriptions of the salient entities and events.

Workflow

- Initially load each photo and collect the predicted features.
- Need to clean the description text. The descriptions are already tokenized and easy to work with. Clean the text by converting all words to lowercase, removing all punctuation, removing all words that are one character or less in length and remove all words with numbers in them. Once cleaned, we can save the descriptions to a new file.
- Train the data on all of the photos and captions in the training dataset.

Finally evaluate a model by generate captions for entirely new photographs in the test dataset.

III. METHODOLOGY

CNN

A Convolutional Neural Network (CNN) is a class of deep, feed-forward artificial neural networks that has successfully been applied to analysing visual imagery. CNN compares any image piece by piece and the pieces that it looks for in an image while detection are called as features. By finding rough feature matches in roughly, the same position in two images CNN gets trained. For all input images, we extract features, which is very well tuned for object detection. We obtained a 4096-Dimensional image feature vector that we reduce using Principal Component Analysis (PCA) to a 512-Dimensional image feature vector due to computational constraints. We feed these features into the first layer of our RNN or LSTM at the first iteration. Different layers of Convolution Neural Network are: Convolution Layer , Rectified Linear Unit Layer ,Pooling Layer ,Fully Connected Layer .

RNN

Recurrent Neural Network is a neural network with feedback, designed to model sequences of data, such as words (sequences of characters) or sentences (sequences of words). RNN maintains an internal hidden state that stores context information, i.e. information computed from past inputs. In its simplest formulation, given a sequence of inputs, RNN computes a sequence of outputs.In practice, RNNs suffer from vanishing and/or exploding gradients that is a problem where the backpropagation of an error signal over several iterations will diminish or explode quickly. The Long Short Term Memory (LSTM) models solved the gradient problem by replacing the traditional artificial neuron with a memory cell containing long and short term nonlinear capabilities. The LSTM's incredible power was first realized in the speech and natural language processing domains, and more recently to the annotation of image and videos. LSTMs are a natural fit for temporal sequences of varying lengths and can be trained using standard back propagation.

IV. IMPLEMENTATION

Environmental setup

Python SciPy environment installed ideally with Python 3.7.4, Keras is installed with TensorFlow, and other libraries like Pandas, NumPy, and Matplotlib is installed. Minimum of 6GB RAM is required.

Implementation Steps

Download and Install Anaconda and other Deep Learning Libraries

Flicker8k_Dataset – Dataset folder which contains 8091 images.
Flickr_8k_text – Dataset folder which contains text files and captions of images.The below files will be created by us while making the project.
Models – It will contain our trained models.
Descriptions.txt – This text file contains all image names and their captions after preprocessing.
Features.p – Pickle object that contains an image and their feature vector extracted from the Xception pre-trained CNN model.
Tokenizer.p – Contains tokens mapped with an index value.**Model.png** – Visual representation of dimensions of our project.
Testing_caption_generator.py – Python file for generating a caption of any image.
Training_caption_generator.ipynb – Jupyter notebook in which we train and build our image caption generator.

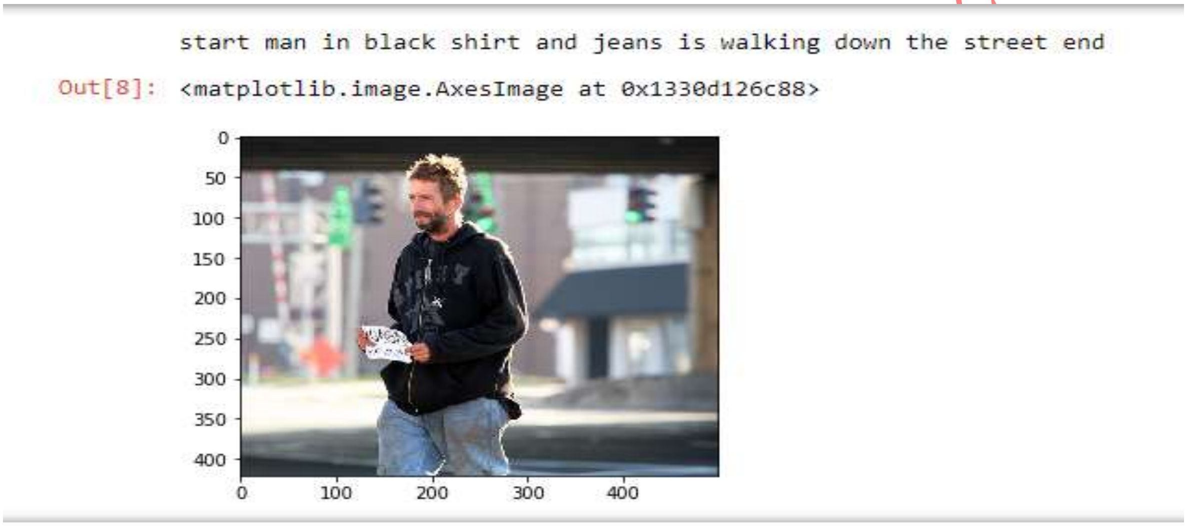
Download the Image and Caption Dataset

Prepare Text Data
Extracting the feature vector from all images
Loading dataset for Training the model
Tokenizing the vocabulary
Create Data generator
Create Data generator
Generate New Captions

V. TRAINING THE MODEL

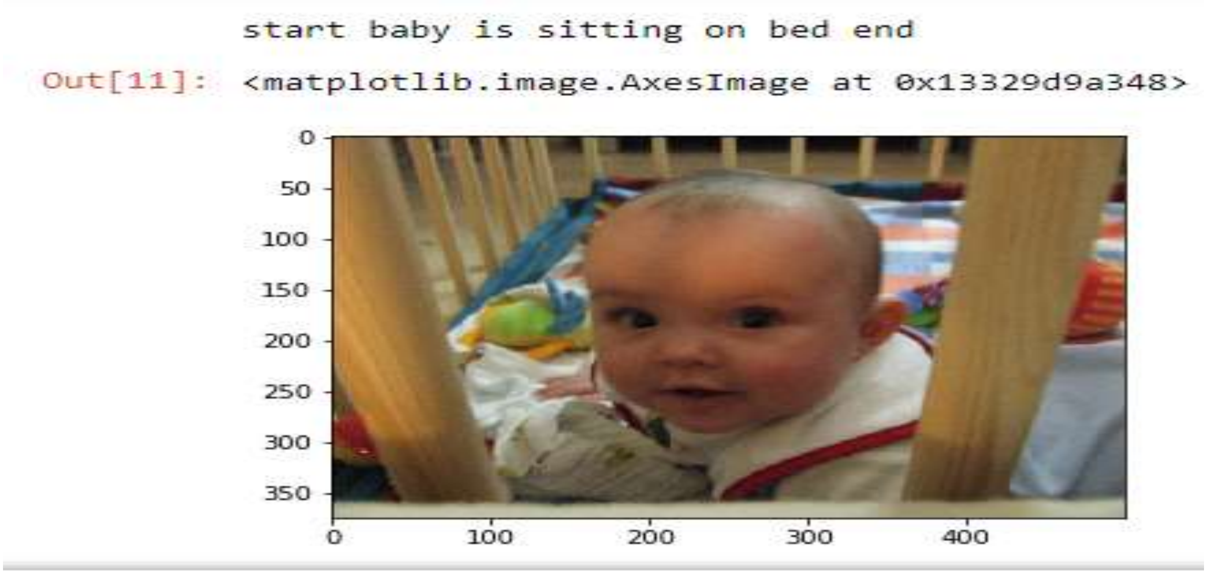
To train the model, 6000 training images needed by generating the input and output sequences in batches and fitting them to the model using model.fit_generator() method.

VI. RESULTS



```
         start man in black shirt and jeans is walking down the street end

Out[8]:  <matplotlib.image.AxesImage at 0x1330d126c88>
```



```
           start baby is sitting on bed end

Out[11]:   <matplotlib.image.AxesImage at 0x13329d9a348>
```

VII. CONCLUSION

In our proposed system, we present a multi-model Neural Network that automatically learns to describe the content of images. My model first extracts the information of objects and their spatial locations in an image, and then a deep recurrent neural network (RNN) generates a description sentence. Each word of the description is automatically aligned to different objects in the input image when it is generated. I trained my model with some image and text data set, which help my model to identify different object from input image.

VIII. REFERENCES

1. An Overview of Image Caption Generation Methods, Volume 2020 |Article ID 3062706 ,HaoranWang,YueZhang,and Xiaosheng Yu

2. Neural Image Caption Generation with Weighted Training and Reference, volume 11, Guiguang Ding, Minghai Chen, Sicheng Zhao, Hui Chen, Jungong Han & Qiang Liu Published: 08 August 2018

3. A Comprehensive Survey of Deep Learning for Image Captioning MD. ZAKIR HOSSAIN, Murdoch University, Australia FERDOUS SOHEL, Murdoch University, Australia MOHD FAIRUZ SHIRATUDDIN, Murdoch University, Australia HAMID LAGA, Murdoch University, Australia

4. Camera2Caption: A real-time image caption generator, June 2017,MD. ZAKIR HOSSAIN, FERDOUS SOHEL, MOHD FAIRUZ SHIRATUDDIN, HAMID LAGA

5. Show and tell: A neural image caption generator,June 2015, Conference: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

6. Image Caption Generator Based On Deep Neural Networks, JianhuiChen ,Wenqiang Dong ,Minchen Li Department CPSC 503 CS

7.RaffaellaBernardi, et al, *Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures*, 2016.

8. Akanksha P. Deshmukh, et al, *A Survey on Vision Based Approaches for Image Description*, 2003.