International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

Deep Learning Based Image Caption Generator

Manish Raypurkar[1], Abhishek Supe[2], Pratik Bhumkar[3], Pravin Borse[4], Dr. Shabnam Sayyad[5]

Department of Computer Science, All India Shri Shivaji Memorial Society's COE, Pune-1
Savitribai Phule Pune University, Pune, Maharashtra, India

Abstract - *Automatically describing the content of images using natural languages is a fundamental and challenging task. It has great potential impact. For example, it could help visually impaired people better understand the content of images on the web. Also, it could provide more accurate and compact information of images/videos in scenarios such as image sharing in social network or video surveillance systems. The framework consists of a convolutional neural network (CNN) followed by a recurrent neural network (RNN). By learning knowledge from image and caption pairs, the method can generate image captions that are usually semantically descriptive and grammatically correct. Human beings usually describe a scene using natural languages which are concise and compact. However, machine vision systems describes the scene by taking an image which is a two dimension arrays. The idea is mapping the image and captions to the same space and learning a mapping from the image to the sentences.*

Key Words: **NUERAL NETWORKS, CNN, RNN, OBJECT, LSTM, NLP.**

1. INTRODUCTION

Image captioning models typically follow an encoder-decoder architecture which uses abstract image feature vectors as input to the encoder and generate caption. Generating a natural language description from images is an important problem at the section of computer vision, natural language processing, artificial intelligence and image processing. Image caption, automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing. The application of image caption is extensive and significant, for example, the realization of human-computer interaction. This summarizes the related methods and focuses on the attention mechanism, which plays an important role in computer vision and is recently widely used in image caption generation tasks. Further more,this project model highlights some open challenges in the image caption task.

2. RELATED WORK

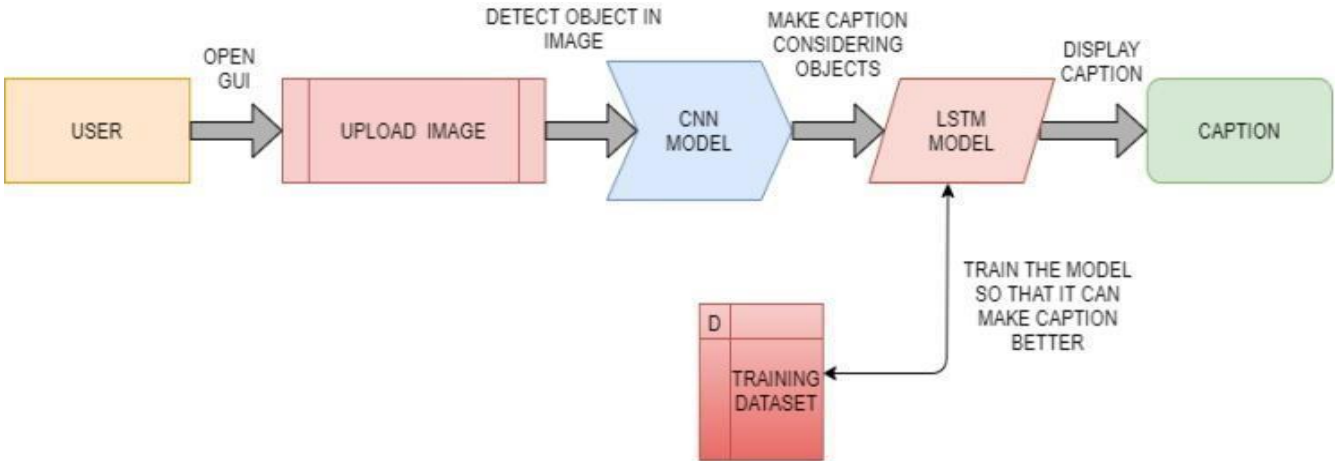
We introduce a synthesized output generator which localize and describe objects, attributes, and relationship in an image, in a natural language form. So, to make our image caption generator model, we will be merging these architectures. It is also called a CNN-RNN model.

- CNN is used for extracting features from the image.
- LSTM will use the information from CNN to help generate a description of the image.
  - CNN- Convolutional Neural networks are specialized deep neural networks which can process the data that has input shape like a 2D matrix. Images are easily represented as a 2D matrix and CNN is very useful in working with images. CNN is basically used for image classifications and identifying if an image is a bird, a plane or Superman, etc. It scans images from left to right and top to bottom to pull out important features from the image and combines the feature to classify images. It can handle the images that have been translated, rotated, scaled and changes in perspective.

  - LSTM

    LSTM stands for **Long short term memory**, they are a type of RNN (**recurrent neural network**) which is well suited for sequence prediction problems. Based on the previous text, we can predict what the next word will be. It has proven itself effective from the traditional RNN by overcoming the limitations of RNN which had short term memory. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information.

- DFD Diagram-



3. PROPOSED ARCHITECTURE

We started by adopting an encoder-decoder architecture that incorporates visual attention mechanism to generate image captioning. The encoder part is CNN based and the decoder one uses the visual attention module. The proposed architecture is illustrated within following fig. Suppose {S0,..., ST-1} is a sequence of words in a sentence of length T, the model aims to directly maximize the probability of the correct description given an image.

- *The Encoder part*

Under the encoder-decoder framework for image captioning, a CNN can produce a rich representation of the input image by embedding it to a fixed length vector representation. Many different CNN can be used, e.g., VGG, Inception V3, ResNet. In this paper, we use Inception V3 model created by Google Research as encoder. This model was pre-trained on ImageNet dataset where it was the first runner up for image classification in ILSVRC 2015. We have removed the last layer of the model as it is used for classification. We have pre-processed images with the Inception V3 model and have extracted features.

Thus, the optimization problem can be formulated by

$$\theta = \arg\max_{\theta} \sum \log p(S_{i\,i} I ; \theta) \quad (1)$$

Where 8 represents the parameters of the model, Ii is an image and S is the generated description. Probability of the created by Google Research as encoder. This model was pre-trained on ImageNet dataset where it was the first runner up for image classification in ILSVRC 2015. We have removed the last layer of the model as it is used for Classification. We have pre-processed images with the Inception V3 model and have extracted features.

The extractor produces L vectors, each of which is a D-dimensional representation corresponding to a part of the image:

$$a = \{a_{\,i}, \ldots, a_{\,L}\}, a \in {}^{D} \quad (2)$$

Global image feature can be obtained by:

$$a\,g = \frac{1}{L} \sum ai \quad (3)$$

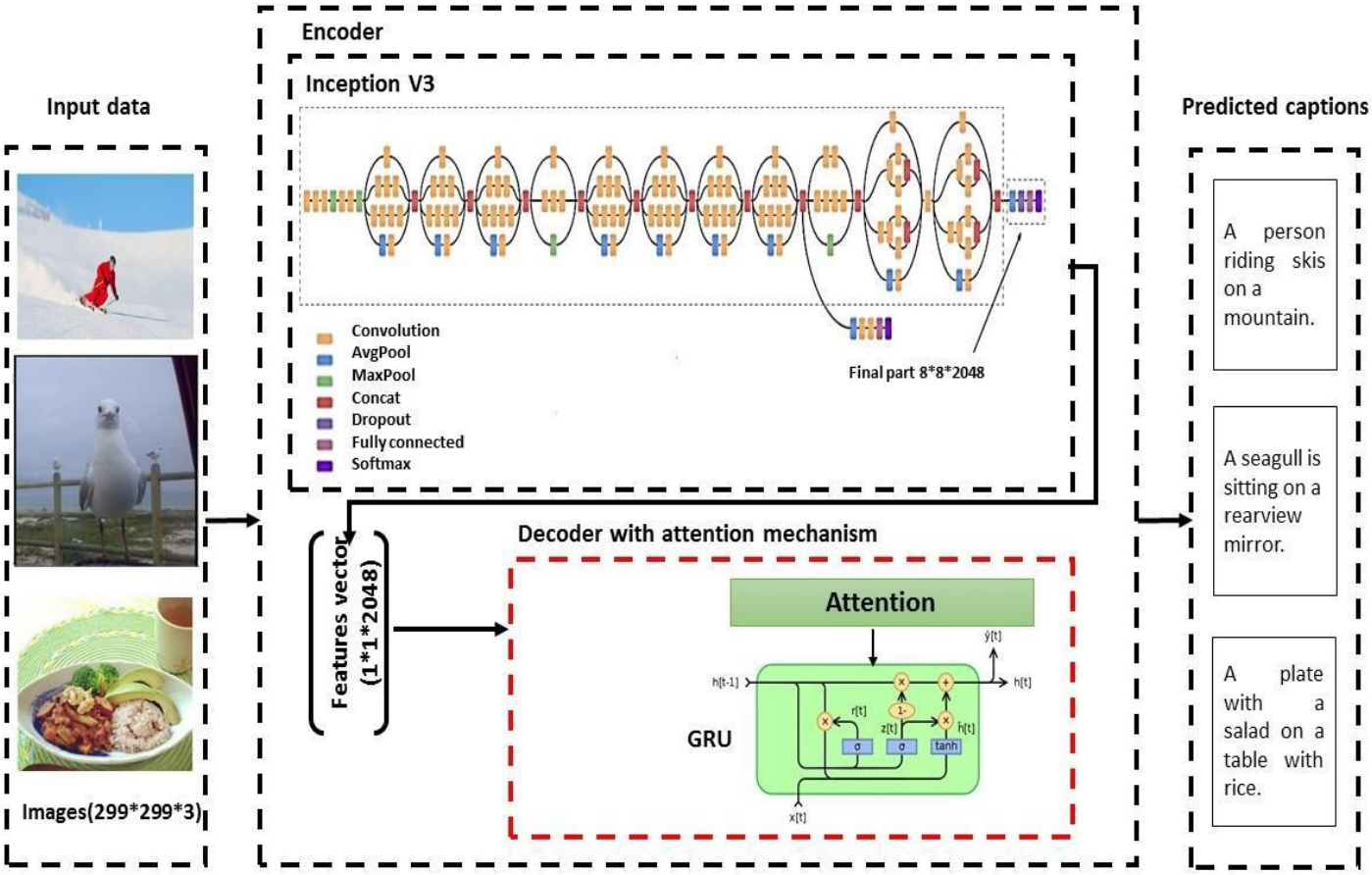
Image Vector and Global Image Vector can be obtained by using a single layer perceptron with rectifier activation function:

$$v_i = \mathrm{Re}\, LU\,(W_a a_i) \quad (4)$$

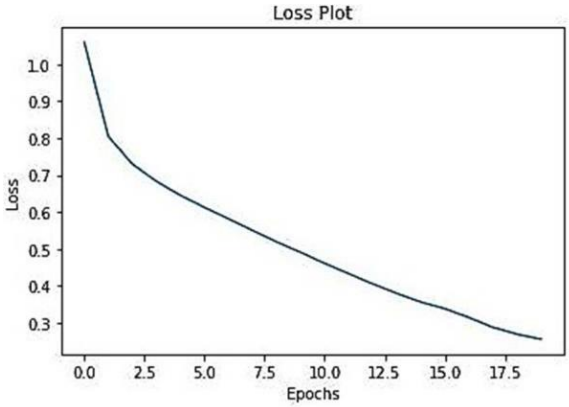
transformed spatial image feature form is V =

$$v_g = \mathrm{Re}\, LU\,(W_g\, a_g) \quad (5)$$

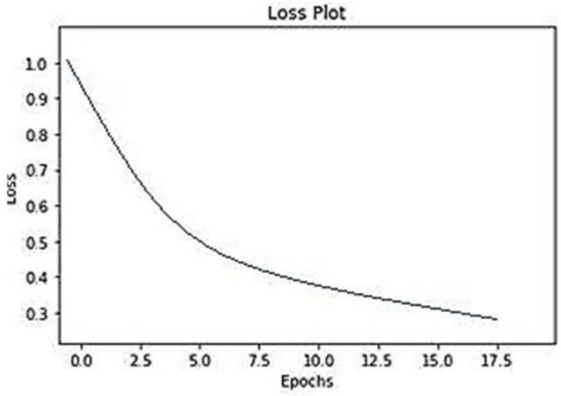
[v1, ..., vL]



Loss presentation without ADAM-



Loss Presentation with ADAM-



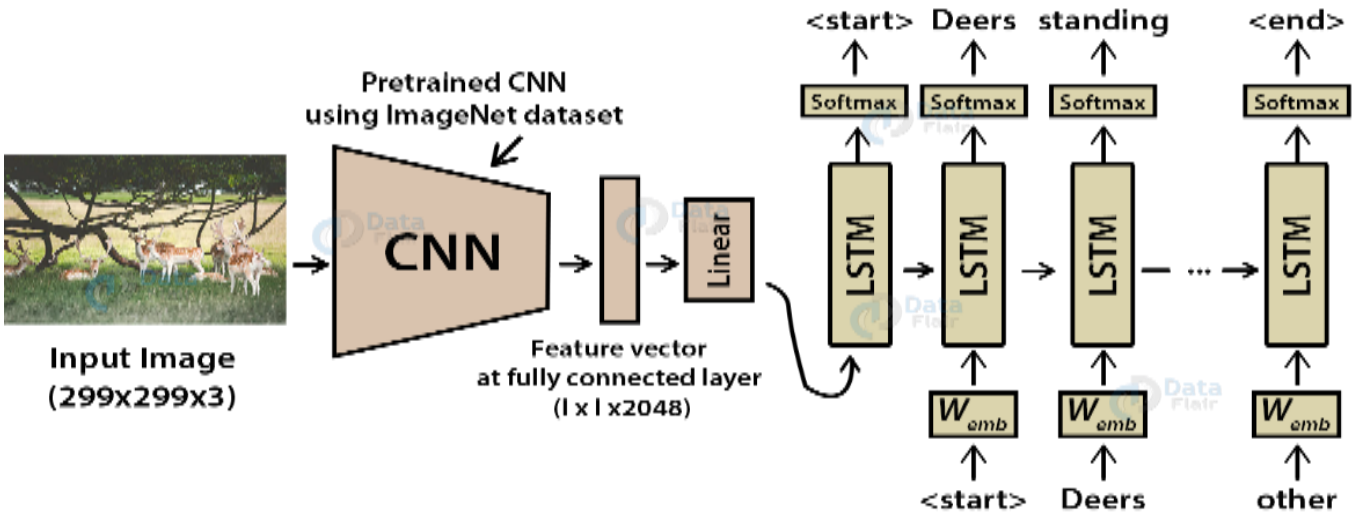
Attention mechanism

For image captioning, attention tends to focus on specific regions in the image while generating descriptions.

At time t, based on the hidden state, the decoder would attend to the specific regions of the image and compute context vector using the spatial image features from a convolution layer of a CNN

$ct = g(v, ht)$ (6)

We feed V and h_t through a single layer neural network



followed by a soft-max function to generate the attention distribution over the k regions of the image

$$Z_t = W_h^T \tanh(W v_v + (W_g h_t)1^T \quad (7)$$

$$\alpha_t = soft\max(Z_t) \quad (8)$$

Where 1 k is a vector with all elements set to 1. Wv,Wg ∈R LXD and Wh∈ R L are parameters to be learnt. α ∈ R L is the attention weight over features in V. Based on the attention distribution, the context vector ct can be obtained

$ct = \sum(\alpha ti\ vti)$ (9)

- *The Decoder part-*

Given the image representations, a decoder is employed to translate the image into natural sentences. A decoder is a RNN which are typically implemented using either LSTM or GRU.

Here we have used GRU as a decoder which has a simpler structure than LSTM.

Also, unlike RNN, GRU does not suffer from the vanishing gradient problem.

Xt is the input vector and we obtain it by concatenating the word embedding vector, Wt, global image feature vector, vg, and context vector, ct, to get the input v is number of vectors and D is size of each vector. The transformed spatial image feature form is

V = [v1, ..., vL]

$Xt = [Wt ; vg ; ct]$ (10)

4. HELFUL HINTS

4.1 Image Captioning Datasets:

- For the image caption generator, we will be using the Flickr_8K dataset. There are also other big datasets like Flickr_30K and MSCOCO dataset but it can take weeks just to train the network so we will be using a small Flickr8k dataset.
- The advantage of a huge dataset is that we can build better models.
- Flickr8k dataset contains a variety of images depicting scenes and situations.
- The dataset consists of 8000 images and each image has 5 corresponding descriptions.
- We split the data into 6000, 1000, & 1000 images as training, validation and testing sets respectively.
- The images are of different dimensions.

4.2 The summary of a few recent works for Image Caption

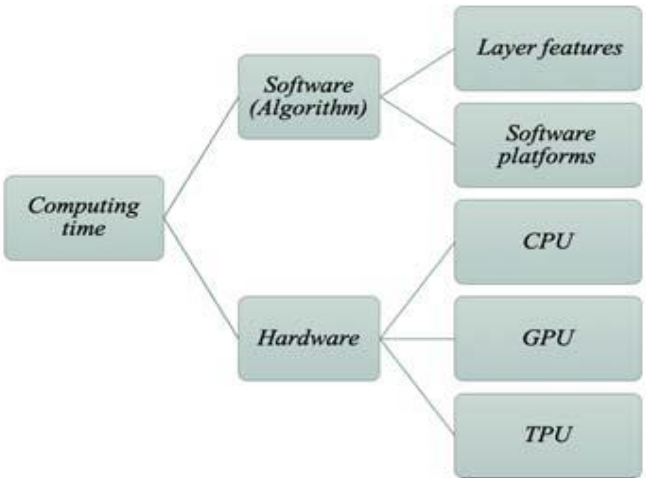
No.	Author (s)	Title of Paper	Year of Publication	Proposed Methodology and dataset used	Conclusion/Findings
1	Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra, Nand Kumar Bansode	Camera2Caption: A real-time image caption generator	(2017) International Conference on Computational Intelligence in Data Science(ICCIDS)	Proposed Deep Learning Based Advanced Technique Deep Reinforcement Learning that's led by Computer Vision and machines translation based on deep learning model. Dataset used in this model is MS-COCO.	The proposed model based on deep learning, well optimize and perform in real time environment (mobile devices) and produce high quality captions by using help of tensorFlow by google.
2	Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares	Image Captioning: Transforming Objects into Words.	June 2019 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.	Proposed Object Relation Transformer model, focuses on spatial relationship between objects of images through used of faster R-CNN with ResNet-101. Mainly focuses on Improve the relationship between objects. Dataset used in this model is MS-COCO with Pycharm IDE.	The proposed model encodes positions and size and relationship between detected objects in images and extracted features by building upon the bottom-up and topdown image captioning approach and CNN.
3	R. Subash	Automatic Image Captioning Using Convolution Neural Networks and LSTM	November 2019 Journal of Physics Conference Series 1362:012096	Proposed Deep Learning based Convolution Neural Networks and Natural Language Processing (NLP) Techniques reasonable sentences are framed and inscriptions are produced.dataset used in this model is MS-COCO.	Proposed model having convolution neural network whose output is paired to Long Short Term Memory network which helps us generate descriptive captions for the image. Also model don't require huge dataset to produce caption of images.
4	B.Krishnakumar,K.Kousalya, S.Gokul,R.Karthikeyan, D.Kaviyarasu	Image Caption Generator Using Deep Learning	(IJAST)Vol.29 NO.3s(2020).	Proposed Deep Learning based Convolution Neural Networks to identify objects in the images using OpenCv. Detected Images converted into audio using GTTP and then converted to text using to Long Short Term Memory network.They used Pre-trained model VGG16 as a baseline model.	Proposed Model successfully trained to generate captions of images using CNN technique, model is depends on data and used small data set.The model generate caption by using Keras Framework used in Jupyter notebook and also conclude keras has strong support for multiple GPU's.
5	Seung-Ho Han, Ho-Jin Choi	Domain-Specific Image Caption Generator with Semantic Ontology	(2020) IEEE International Conference on Big Data and Smart Computing (BigComp)	Proposed model uses domain specific image caption generator to overcome the limitation of open dataset MSCOCO which include general images. Firstly model uses objects and attribute information of images and then reconstruct generated caption using **Semantic Ontology.** dataset used in model is MS-COCO	The Proposed model provide natural language description for given specific-domain. Model generates captions of images using visual and semantic attention. Replacing specific words in captions with domain-specific words. For eg The general word "MENS" replace with "WORKERS" in image of "GROUP OF PEOPLE/MENS WEARING HELMETS AND STANDS IN A ROADS"

5.THE REQUIRED PLATFORM FOR IMPLEMENTATION

Deep Learning has dramatically improved the accuracy of image recognition. Image recognition is considered to be one of the most challenging problems in image science.

5.1 Software Requirement

- Tensor-flow: Tensor-Flow is an end-to-end open source platform for machine learning. Tensor-flow is developed by Google and has integrated the most common units in deep learning frameworks. It supports many up-to-date networks such as CNN and RNN with different settings. Tensor-flow is designed for remarkable flexibility, portability, and high efficiency of equipped hardware.



- PyTorch: PyTorch is a Python-based scientific computing package that serves two purposes: as a replacement for NumPy to use the power of GPUs and as a deep learning research platform that provides maximum flexibility and speed.

- Keras: Keras is a high-level neural network API, written in Python and capable of running on top of Tensor-flow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation. Being able to go from idea to result with the least possible delay is the key to doing good research. Keras allows for easy and fast prototyping (through user friend lines modularity, and extensibility).Keras supports both convolutional networks and recurrent networks, as well as a combination of both.

5.2 Hardware Requirement

The science and methodology behind deep learning have been in existence for decades. In recent years, however, there has been a significant acceleration in the utilization of deep learning due to an increasing abundance of digital data and the involvement of the powerful hardware.

- GPU- Compared to CPU, the performance of matrix multiplication on Graphics Processing Unit is significantly better. With GPU computing resources, all the deep learning tools mentioned achieve much higher speedup when compared to their CPU-only versions GPUs have become the platform of choice for training large, complex Neural Network based systems because of their ability to accelerate the systems.
- TPU- Tensor Processing Unit (Domain-Specific Architecture) is a custom chip that has been deployed in Google data centers since 2015. DNNs are dominated by tensors, so the architects created instructions that operate on tensors of data rather than one data element per instruction. To reduce the time of deployment, TPU was designed to be a coprocessor on the PCI Express I/O bus rather than be tightly integrated with a CPU, allowing it to plug into existing servers just as a GPU does. The goal was to run whole inference models in the TPU to reduce I/O between the TPU and the host CPU. Minimalism is a virtue of domain-specific processors.

CONCLUSION

Our model based on multi label classification using fast Text and CNN, is useful in detecting and extracting objects from image and generate caption according to the provided datasets. We have presented multiple approaches for Image caption Generator like (Convolution neural network, Long short-term memory, Recurrent Neural Network)

REFERENCES

[1] Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra and Nand Kumar Bansode (2017) International Conference on Computational Intelligence in Data Science(ICCIDS) : Camera2Caption: A Real-Time Image Caption Generator
[2] Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares (june 2019) : Image Captioning: Transforming Objects intoWords.
[3] R. Subash November 2019 Journal of Physics Conference Series 1362:012096 : Automatic Image Captioning Using Convolution Neural Networks and LSTM.
[4]B.Krishnakumar,K.Kousalya, S.Gokul,R.Karthikeyan, D.Kaviyarasu (IJAST)Vol.29 NO.3s(2020) :Image Caption Generation Using Deep Learning.
[5] Seung-Ho Han, Ho-Jin Choi (2020) IEEE International Conference on Big Data and Smart Computing (BigComp) : Domain-Specific Image Caption Generator with Semantic Ontology