International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 10 | Oct 2021 www.irjet.net p-ISSN: 2395-0072

Image Caption Generator Using Deep Learning

Shubham Patil[1], Bhagesh Patil[2], Ankit Shewale[3], Sachin Bandal[4], Bhakti Patil[5]

[1-5]Department of Computer Science, All India Shri Shivaji Memorial Society's COE, Pune-1
Savitribai Phule Pune University, Pune, Maharashtra, India

Abstract - *Using natural languages to automatically describe the content of photographs is a fundamental and difficult task. It has a lot of potential. It could, for example, aid visually handicapped people in comprehending the content of web images. It could also deliver more accurate and concise image/video information in settings like social media image sharing or video surveillance systems. A convolutional neural network (CNN) is followed by a recurrent neural network in the framework (RNN). The approach generates image captions that are usually semantically meaningful and grammatically correct by gaining information from image and caption pairs. Natural languages are used by humans to describe scenes because they are brief and compact. Machine vision systems, on the other hand, characterise the scene by capturing an image that is a two-dimensional array. The concept is to combine the image and captions into one place and then learn a mapping from the visual to the phrases.*

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

1.INTRODUCTION

Encoder-decoder architectures are commonly used in image captioning models, which employ abstract image feature vectors as input to the encoder and output caption. In the fields of computer vision, natural language processing, artificial intelligence, and image processing, generating a natural language description from photographs is a challenging challenge. Image caption is a key aspect of scene understanding, which combines computer vision and natural language processing knowledge to automatically generate natural language descriptions based on the content observed in an image. The use of image captions is widespread and important, for example, in the development of human-computer interfaces. This summarizes the parallel approaches and focuses on attention-grabbing, which plays an important role in computer recognition and has recently been widely used in the production of image captions jobs. In addition, this project model highlights some of the open challenges in the work of image descriptions.

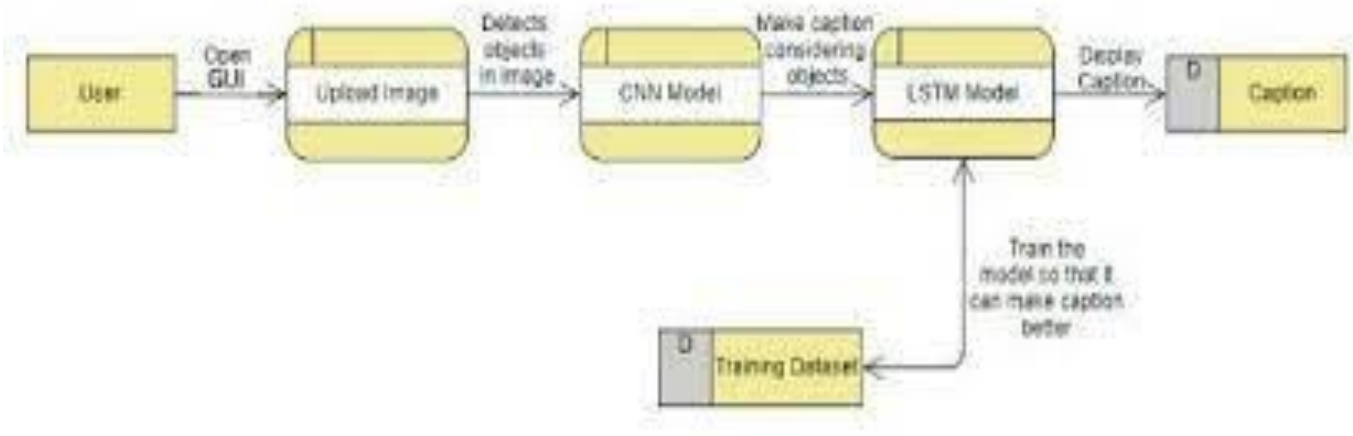
2. RELATED WORK

We introduce a synthesized output generator that creates space and describes objects, attributes, and relationships in an image, in the form of natural language. So, to make our own photo caption model, we will incorporate this art. Also called the CNN-RNN model.

- CNN is used to extract features from images.
- LSTM will use information from CNN to help generate image definition.
    - CNN- Convolutional Neural Networks with special deep neural networks that can process input data as a 2D matrix. Images are easily represented as 2D matrix and CNN is very useful for working with images. CNN is used for image classification and to determine whether the image is a bird, a plane or a Superman, etc. Scan images from left to right and top to bottom to extract key features from the image and integrate the feature to separate images. It can handle translated, rotated, zoomed and shifted images in view.
    - LSTM

      LSTM represents long-term memory, they are a type of RNN (repetitive neural network) that is well suited for sequence predictive problems. Based on the previous text, we can predict what the next world will be. It has proven to be effective from the traditional RNN restrictions that had temporary memory. LSTM can make the correct information when processing input and through the forget gateway, discarding incorrect information.

- DFD Diagram-



3. PROPOSED ARCHITECTURE

We started by adopting an encoder-decoder design that includes a visual approach to generating image captions. Part of the encoder is based on CNN and the decoder uses the viewing module. The proposed structure is shown within the fig following. Suppose {S0, ..., ST-1} word order in the sentence T, T model aims to maximize the potential of the appropriate meaning given to the image.

- *The Encoder Part*

Under the encoder-decoder framework for image captioning, CNN can generate rich representation of the input image by embedding it in a long, consistent vector image. Many different CNNs can be used, e.g. VGG, Inception V3,ResNet. In this paper, we use the Inception V3 model created by Google Research as an encoder. This model was previously trained in the ImageNet database where it became the first image editing runner at ILSVRC 2015. We have removed the last layer of the model as it is used for partition. We have pre-processed images with the Inception V3 model and we have released the features.

Therefore, the problem of efficiency may be caused by

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

Where 8 represents the model parameters, ii the image and S is the generated definition. Opportunities created by Google Research as an encoder. This model was previously created in the first imageNet database where it became the first runner of image editing a ILSVRC 2015. We removed the last layer of the model as it used for partition. We have pre-processed images with the Inception V3 model and we have released the features. The output generates vectors L, each of which is a D-expression corresponding to part of the image:

$$a = \{a_i, ..., a\ L\},\ a \in^{D} \quad (2)$$

The global image feature can be accessed by:

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

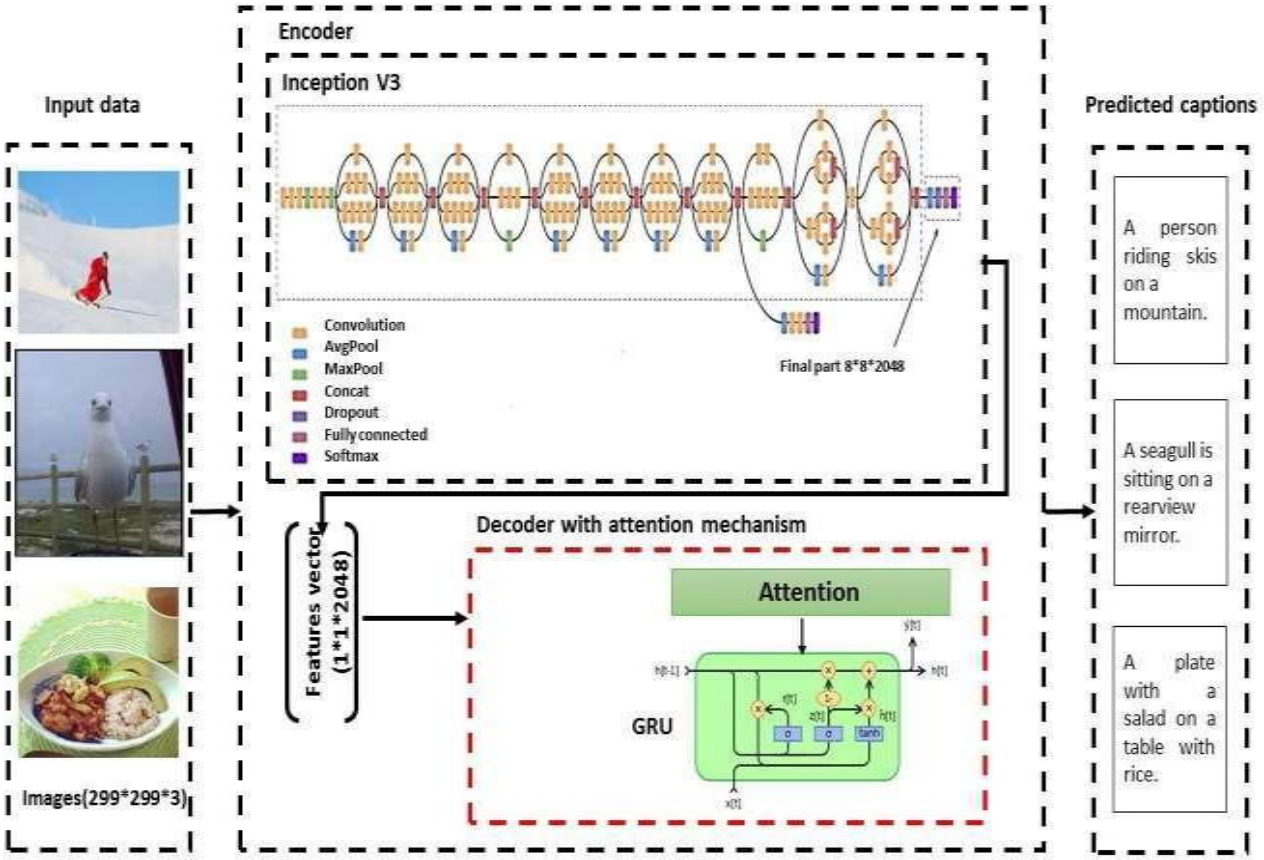
Image Vector and Global Image Vector can be accessed using a single-layer perceptron with a corrective startup function:

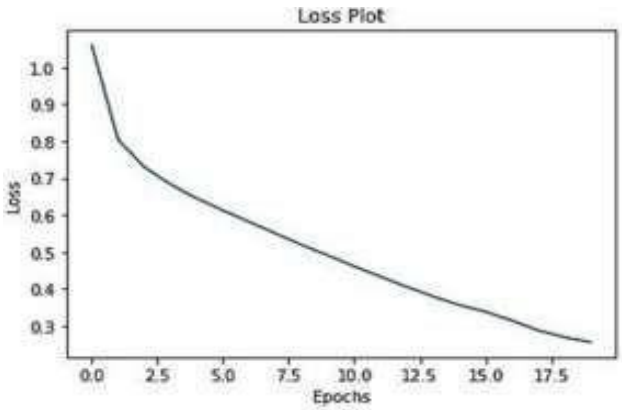
$$vi = Re\ LU\ (W_a a_i) \quad (4)$$

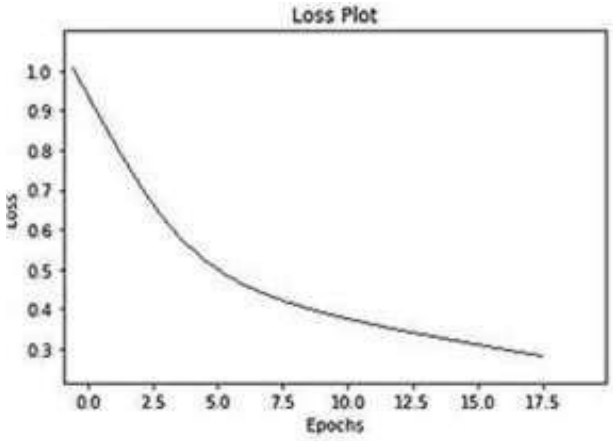
the converted form factor image form is

$$V = vg = Re\ LU\ (W_g\ a_g) \quad (5)$$

[v1, ..., vL]







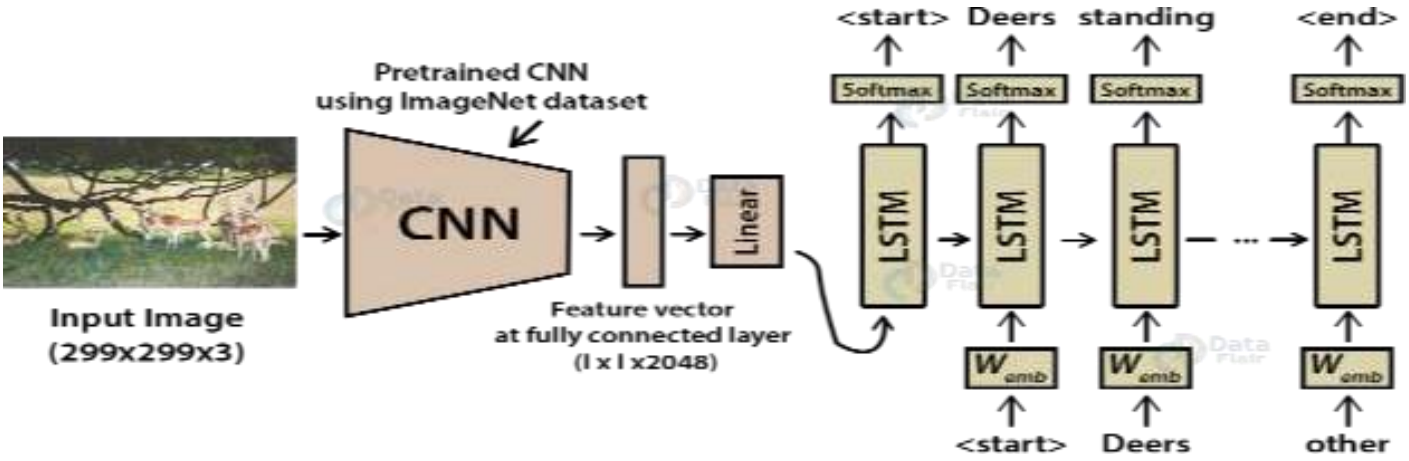
Attention mechanism:

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

At time t, based on the hidden state, thedecoder 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.

$$c_t = g(v, h_t) \quad (6)$$

We feed V and ht through a single layerneural network followed by a softmax function to generate the attention distribution over the k regions of the image.



$$Z\,t = W\,h^{T}\ \tanh(W\,v\,v + (W\,g\,h\,t)1^{T} \qquad (7)$$

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

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

ct = ∑ (αti vti) (9)

- *The Decoder Part-*

Given the image representations, adecoder is employed to translate the image into natural sentences. A decoderis a RNN which are typically implementedusing either LSTM or GRU.

Here we have used GRU as a decoder with a simpler structure than LSTM. And, unlike RNN, GRU does not suffer from the problem of gradient disappearance.

Xt is the input vector and we get it by combining the embedded vector name, Wt, global image element vector, vg, and context vector, ct, input finder iv is the number of vectors and D in size for each vector. Status feature changed by

V = [v1, ..., vL]

4. 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	Grishma Sharma, Priyanka Kalena, Nishi Malde, Aromal Nair, Saurabh Parker.	Visual Image Caption Generator Using Deep Learning	(2019) International Conference on Advances in Sciences & Technology (ICAST)	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 detection. Image recognition is considered to be one of the mostchallenging issues in image science.

5.1 SOFTWARE REQUIREMENT

- TensorFlow: TensorFlow is an end-to-end open source platform for machine learning. TensorFlow 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. TensorFlow is designed for remarkable flexibility, portability and high efficiency of quipped hardware.

- PyTorch: PyTorch is a Python-based scientific computing package that serves two purposes: a 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 TensorFlow, 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 friendliness 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 based systems because of their ability to accelerate the systems.

- TPU – TPU- Tensor Processing Unit (Domain-Specific Architecture) is a custom chip that has been deployed in Google statistics facilities seeing that 2015. DNNs are ruled through tensors, so the architects created commands that perform on tensors of statistics as opposed to one statistics detail consistent with instruction. To lesson the time of deployment, TPU changed into designed to be a coprocessor at the PCI Express I/O bus as opposed to be tightly included with a CPU, permitting it to plug into present servers simply as a GPU does. The aim changed into to run complete interface fashions withinside the TPU to lesson I/O among the TPU and the host CPU. Minimalism is a distinctive feature of domain – particular processors.

6. CONCLUSIONS

Our approach, which is based on multi label classification utilising quick Text and CNN, is effective for recognising and extracting objects from images, as well as generating captions depending on the datasets provided. We've discussed a variety of ways for Image Caption Generator, including

(Recurrent Neural Network, Convolutional Neural Network, Long Short-Term Memory).

7. REFERENCES

[1] Grishma Sharma, Priyanka Kalena, Nishi Malde, Aromal Nair, Saurabh Parkar(2019) International Conference on Advances in Sciences & Technology (ICAST): Visual Image Caption Generator Using Deep Learning.

[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 Ontolog.