# Image Caption Generation System using Neural Network with Attention Mechanism

**Prof. N. G. Bhojne[1], Raj Jagtap[2], Rohit Jadhav[3], Rushikesh Jadhav[4], Akshay Jain[5]**

*[1]Professor Dept. of Computer Engineering Sinhgad College of Engineering Pune, India*
*[2,3,4,5]Dept. of Computer Engineering Sinhgad College of Engineering Pune, India*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** *Generating captions of an image automatically is a task very close to the scene understanding which is used to solve the computer vision challenges of determining which objects are in an image, and are also capable of capturing and expressing relationships between them in a natural language. It is also used in Content Based Image Retrieval (CBIR). Also it can be used to help visually impaired peoples to understand their surroundings. It represents a model based on a deep recurrent Neural Network that combines computer vision techniques that can be used to generate natural sentences describing an image. Model is divided into two parts Encoder and Decoder and Dataset used is Flickr8k. We are using Convolutional Neural Network (CNN) as encoder to extract features from images and Decoder as Long Short Term Memory (LSTM) to generates words describing image. Simultaneously using Attention Mechanism to provide more attention on details of every portion of image to generate more descriptive caption. To construct Optimal sentence from these words Optimal Beam Search is used. Further, generated sentence is being converted to audio which is found to help the visually impaired people. Thus, our system helps the user to get descriptive caption for the given input image.*

***Key Words***: computer vision, natural language processing

## 1. INTRODUCTION

Image captioning means automatically generating a caption for an image. To achieve the goal of image captioning, semantic information of images must be captured and expressed in natural languages. Humans are able to relatively easily describe the environments they are in. Given a picture, it's natural for a person to explain an immense amount of details about this image with a fast glance. Although great progress has been made in various computer vision tasks, such as object recognition, attribute classification, action classification, image classification and scene recognition, it is a relatively new task to let a computer use a human-like sentence to automatically describe an image that is forwarded to it. Using a computer to automatically generate a natural language description for an image, which is defined as image captioning, is challenging. Because it connects both research communities of computer vision and Language Processing, image captioning not only requires a deep understanding of the semantic contents of an image, but also must express the knowledge present in image to make it human-like sentence. Automatic image captioning has many application. It is used in image retrieval based on the contents present in image. It is also a base for video based image captioning.

## 2. RELATED WORK

Generating descriptions of natural language for images has been studied for a long time. In earlier approach the sentence potentials were generated and evaluated through complicated natural language processing technique [1]. The sentence is represented by computing the similarity between the sentence and the triplets which is generated during phase of mapping image space to meaning space. The major notable limitation of this method is that the triplet might mismatch together, such as bottle-walk-street. this triplet makes no sense at all. Extracting corresponding triplets was not that easy at before. In recent methods Encoder-Decoder architecture is used for generating captions, where recurrent neural networks are used as decoder for generating sentences. A similar approach is presented by vinayals et al [2] in show and tell. CNN is used as encoder and LSTM RNN is used as Decoder. Kelvin et al in their Show Attend and tell [3] proposed similar approach by adding Attention mechanism. Attention mechanism focuses on all parts of image adds more weight to important feature. Generating sentences from words obtained by decoder is important task in image captioning. Kelvin et al [3] were using greedy method for generating sentence. Beam search can be used for generating sentences from Decoder output. Beam search stops the generation of sentence when the end tag is generated. But the incomplete hypothesis may generate optimal sentence. This issue is overcome by using Optimal beam search [4] which continues searching until maxima is found. We are implementing Optimal beam search in our methodology to generate optimal sentence.

## 3. METHODOLOGY

Once the Encoder generates the encoded image, we transform the encoding to create the initial hidden state h for the LSTM Decoder.

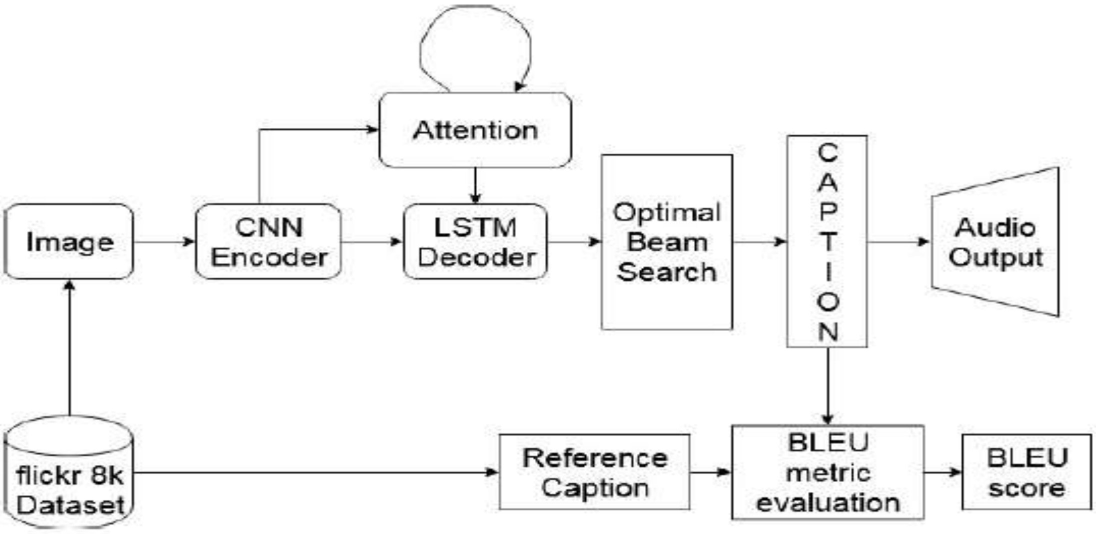

**Fig -1:** System Overview Diagram

At each decode step :-

1) The encoded image and the previous hidden state is used to generate weights for each pixel in the Attention network.

2) The previously generated token (word) and the weighted average of the encoding are fed to the LSTM Decoder to generate the next word. And at last, Optimal beam search will generate a sentence from tokens (output of LSTM). To check quality of generated caption BLEU score (Bilingual Evaluation Understudy) which is metric for evaluating a generated sentence to a reference sentence for a given image from testing dataset is used. We will be using Flickr8k dataset which contains 8000 images and for each image 5 captions are provided. Further, Datset will be splited into training (6000 images), development (1000 images), test (1000 images). The size of dataset is small (1.12 GB) so model can be trained easily on low-end systems.

### 3.1 CNN Encoder:

CNN acts as a feature extractor that compresses the information in the original image into a smaller representation. Since it encodes the content of the image into a smaller feature vector hence, this CNN is often called the encoder. The output of this phase will be a feature vector consisting of information about image. Then this feature vector is given as an input to Decoder. A feature vector is a one-dimensional matrix which is used to describe a feature of an image. They are often used to describe a whole image (Global feature) or a feature present at a location within the image space (local feature).

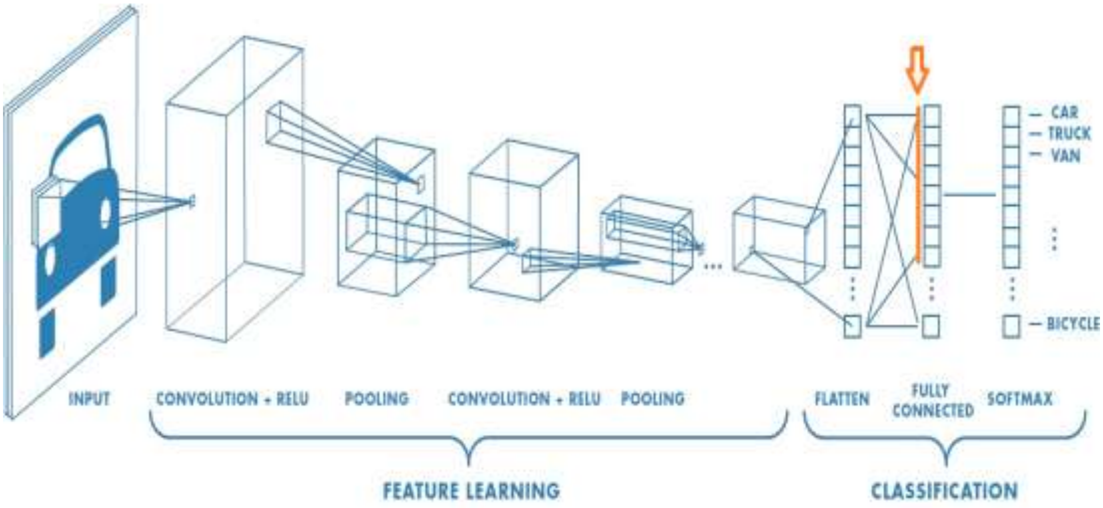

**Fig:2-** Feature Extraction from Image

They are shape, color, texture, objects present in image. There are various models available for this task: ResNets, VGG, inception, etc.

### 3.2 LSTM Decoder:

The Decoder's job is to look at the encoded image i.e. feature vector and turn it into natural language. Since it is generating a sequence, it would need to be a Recurrent Neural Network (RNN). We will use an LSTM. Each predicted word is employed to get subsequent word. Using these words, appropriate sentence is formed with help of Optimal beam search. Here, Softmax function will be used for prediction of word.

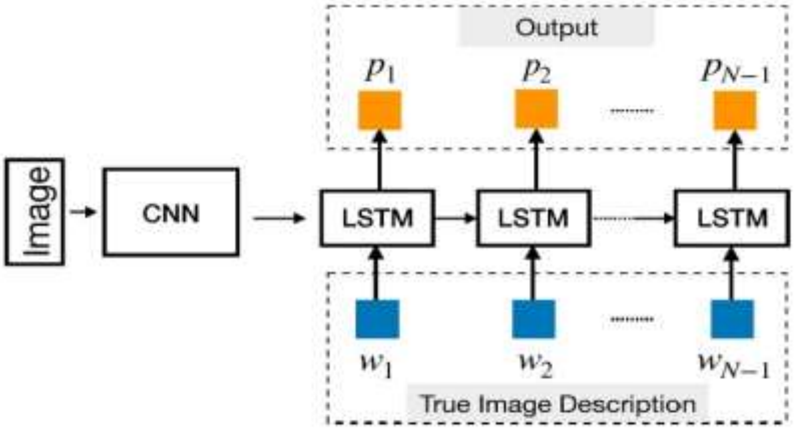

**Fig:3-** LSTM generating word Sequence

### 3.3 Attention:

The Attention network computes the weights with respect to important parts present in image. Intuitively, how would you estimate the importance of a certain part of an image? You have to notice every part in image, so you'll check out the image and choose what needs describing next. For example, after you mention a person, it's logical to declare the actions he is performing on some objects. This is exactly what the Attention mechanism does. It considers the sequence generated thus far, and attends to the part of the image that needs describing next. From fig. We can see that it takes two inputs: encoded image and previous output of decoder, to generate weighted encoded image.

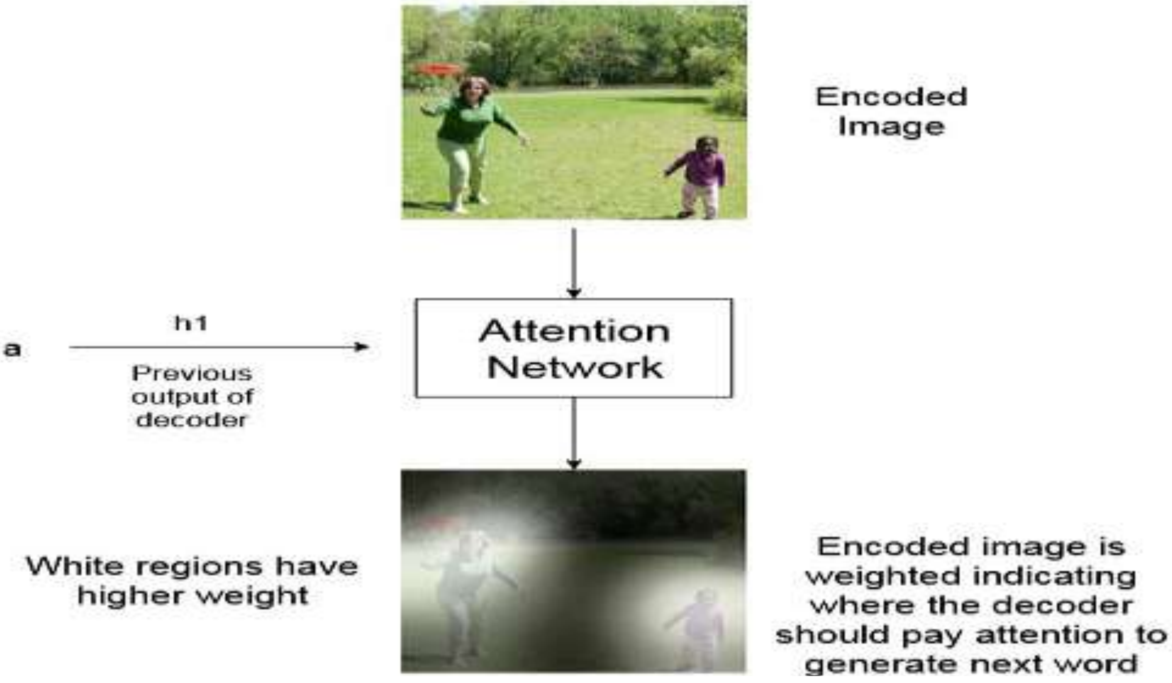
Encoded Image
a
h1
Previous output of decoder
Attention Network
White regions have higher weight
Encoded image is weighted indicating where the decoder should pay attention to generate next word

**Fig -4:** Attention Network

### 3.4 Optimal Beam Search:

Decoder generates words and their probability from the feature vector. Sentences are generated by using these words. Beam search is used for generating sentence from these words by selecting top words at given iteration. The probability of words is used to calculate the score of a sentence. When we get an end tag, a sentence having maximum score is selected as final output. Problem with this approach is, the incomplete sentences may be generated as an optimum sentence. The optimum beam search continues searching until it reaches a point of maxima. After that all sentences will have low score than that sentence. Beam search is a heuristic search algorithm where only the most promising nodes at each step of the search are retained for further branching. It is an optimization of best first search Optimal beam search expands nodes until it gets the sentence with highest score.

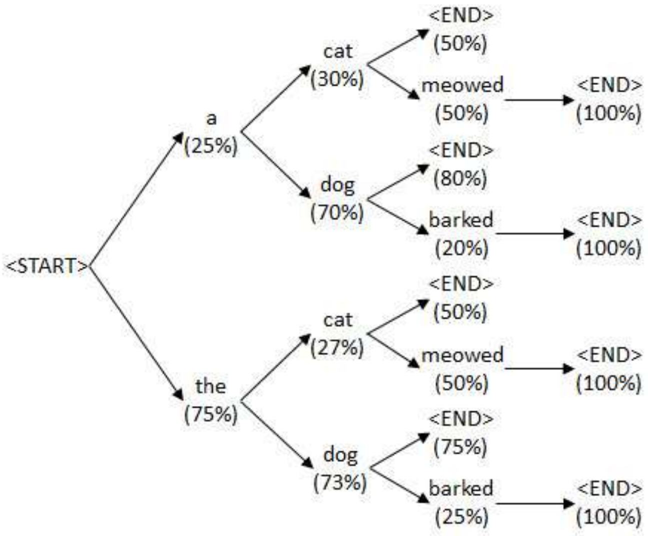

**Fig -5:**Beam Search

### 4. CONCLUSIONS

Thus, we can develop an image captioning system which will generate more descriptive and richer captions by using attention mechanism. And by using optimal beam search we can improve BLEU score.

### ACKNOWLEDGEMENT

We feel great pleasure in expressing our deepest sense of gratitude and sincere because of our guide Prof. N. G. Bhojne for his valuable guidance during the Project work, without which it would have been very difficult task. I feel always motivated and encouraged whenever I attend his meeting. I have no words to express my sincere thanks for valuable guidance, extreme assistance and cooperation extended to all the Staff Members of our Department too. This acknowledgement would be incomplete without expressing our special thanks to Prof. M. P. Wankhade (Head of the Department (Computer Engineering)) for his support during the work. Last but not least we would like to thank all the Teaching, Non-Teaching staff members of our Department, our parents and colleagues those who helped us directly or indirectly for completing the preliminary work for our project.

### REFERENCES

[1] Farhadi, M. Hejrati, M. A. Sadeghi, P. Young,C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV, 2010.

[2] Show and Tell: A Neural Image Caption Generator Oriol Vinyals ,Alexander Toshev ,Samy Bengio ,Dumitru Erhan [2015]

[3] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention Jimmy Lei Ba, Ryan Kiros, KyunghyunCho, Aaron Courville, Ruslan, Richard S. Zemel, Yoshua Bengio [2016]

[4] When to Finish? Optimal Beam Search for Neural Text Generation (modulo beam size) Liang Huang, Kai Zhao, Mingbo Ma.[2018].

[5] Pascanu, Razvan, Gulcehre, Caglar, Cho, Kyunghyun, and Bengio, Yoshua. How to construct deep recurrent neural networks.In ICLR, 2014.

[6] Tang, Yichuan, Srivastava, Nitish, and Salakhutdinov, Ruslan R.Learning generative models with visual attention. In NIPS, pp. 1808ˆa1816, 2014.

[7] Kulkarni, Girish, Premraj, Visruth, Ordonez, Vicente, Dhar, Sagnik,Li, Siming, Choi, Yejin, Berg, Alexander C, and Berg,Tamara L. Babytalk: Understanding and generating simple image descriptions. PAMI, IEEE Transactions on, 35(12):2891ˆa2903, 2013