**Image Captioning System**

[Diagram

Description automatically generated](https://github.com/Gourang97/attention-based-image-captioning/blob/main/Result%20images/model_architecture.PNG)

**Idea**

We are trying to encounter the problem faced by the "classic" image captioning method using the Attention mechanism in the decoder. The attention mechanism will help the decoder to focus on relevant parts of the image. The decoder will only use specific parts of the image rather than conditioning on the entire hidden state h produced from the convolutional neural network. We can understand that there is one additional layer from the classic architecture, and this new layer makes the model as attention model. While predicting the next word while generating captions for an image, if we have previously predicted i words, the hidden state will be hi. Then the model will select the relevant part of the image using the attention mechanism which will be zi (which captures only relevant information from the image) and this will be go as an input to the LSTM. The LSTM then generates the next word and also passes on the information to the new hidden state hi+1 The model architecture is inspired by the Show, Attend and Tell paper, where an attention-based model was introduced to automatically learn and describe the content of images.

**Problem Statement**

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. Some of its application includes helping visually impaired people better understand the content of images on the web. One of the key challenge involves generating description that must capture not only the objects contained in an image, but also express how these objects relate to each other as well as their attributes and the activities they are involved in. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed in addition to visual understanding.

**Dataset**

We leveraged Flickr 8K dataset consisting 5 captions for each image to train our model. The dataset contains a total of 8092 images each with 5 captions, in total we have 40460 properly labelled captions.

[A picture containing graphical user interface

Description automatically generated](https://github.com/Gourang97/attention-based-image-captioning/blob/main/Result%20images/dataset.png)

**Dataset Preprocessing**

* We conducted following data Preprocessing steps
* Cleaned the captions by removing punctuations, single characters, and numeric values.
* Added start and end tags for every caption, so that model understands the start and end of each caption.
* Resized images to 224 X 224 followed by pixel normalization to suit our VGG16 image feature extraction model.
* Tokenized the captions (for example, by splitting on spaces) to obtain a vocabulary of unique words in the data.
* Padded all the sequence to be the same length as the longest one.

**Model Architecture**

We are using the Local attention based architecture for our model.Firstly, it produces the encoder hidden states, i,e encoder will produce hidden states for all the images in the input sequences. Then the alignment score is being calculated for each hidden state of the encoder and the decoder’s previous hidden state. These score are then combined and softmax is applied on them. To generate the contextual information, the softmaxed scores and the encoder hidden states are then combined to formulate a vector representation. This vector is then combined to the last decoder hidden state and fed into the RNN to produce a new word respectively. This complete procedure is recursive in nature and the stopping criteria is till the length of caption generated surpasses the maximum length.

[Diagram

Description automatically generated](https://github.com/Gourang97/attention-based-image-captioning/blob/main/Result%20images/attention.png)

**Steps Followed**

To simplify and formulate generalized approach we followed the below mentioned steps -

* We extracted the features from the lower convolutional layer of VGG16 giving us a vector with 512 output channels.
* This vector is then passed through the CNN Encoder which consists of a single fully connected layer followed by a dropout layer.
* The Recurrent Neural Network(here GRU), the takes in the image to predict the next word.
* Furthermore, the attention based model enables us to see what parts of the image the model focuses on as it generates a caption.

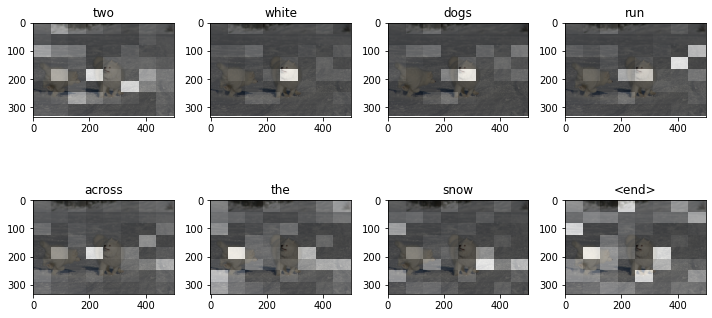
**Evaluation**

We are using greedy approach to evaluate the captions generated. The greedy approach computes the Maximum Likelihood Estimation (MLE) i.e.we select the word with the maximum logit value for a given output. We here greedily select the word which has the maximum probability. We are using BLEU(Bilingual Evaluation Understudy) Score, and evaluating the generated captions on our test set [5]. The BLEU score, will take the fraction of the tokens present in the predicted sentence to the ones that appears in the ground truth. It return a value between 0 and 1. If the value is closer to one that means that the prediction is very close to the ground truth.

**Results**

After performing the experiments as mentioned above, we were able to get significant results. We have plotted attention plots and are also observing the predicted captions, with respect to the original caption. We are also monitoring the BLEU score for the test images. We ran our model on several test images and plotted the attention plot, so as to observe which part of the image was focused upon while predicting a particular word in a caption. In attention plot in figure below,

[](https://github.com/Gourang97/attention-based-image-captioning/blob/main/Result%20images/snow_dog.PNG)

[](https://github.com/Gourang97/attention-based-image-captioning/blob/main/Result%20images/snow_attention.png)

We can observe that for every image the important feature is being highlighted and the predicted word is being also mentioned for that particular important feature. This clearly tells us that not every part of the image is important to predict a caption. Also, we can miss out on several minute details if we don’t use this kind of attention architecture.

In figure below -

[A picture containing graphical user interface

Description automatically generated](https://github.com/Gourang97/attention-based-image-captioning/blob/main/Result%20images/dog_attention.png)

the ground truth caption was "Two white dogs are playing in snow". The predicted caption is "Two white dogs run across the snow". The reported BLEU score is 61:47, which is very significant. If we compare our results to the ground truth result which is around 75. In figure, the ground truth caption was "Black dog with red collar is jumping in the water". The predicted caption is "black dog with red collar is jumping through the water" and the observed BLEU score is 67:32 which is also very significant. From the results of both the test images we observed that our model can reproduce significantly good captions. The captions are also semantically appropriate, and they report outstanding BLEU scores. The attention mechanism is highly utilized in recent years and is just the start of much more state-of-the-art systems. Our trained model shows good performance on Flicker 8k dataset using the BLEU metric and the captions generated are interpretable and well aligned with human intuitions. We also understood that the images used for testing should be semantically very close to the ones used in the training images. We can also alter the evaluation method i.e we could use beam search in order to generate better captions. The attention model successfully captures the important features from an image and generates semantically sound captions as well. We can further work on its improvement so as to improve on the BLEU scores and predict more closer ground truth captions for an image.