Image Caption Generator with Visual Attention Model

Abdul Rahman
PES1UG19CS006

Abhishek D
PES1UG19CS020

Pratham S Golechha
PES1UG19CS346

***Abstract*— Caption generation does not only include detecting an object in the image but also describes the interaction between various objects present in the images and their respective actions.**
In this paper, we will be discussing the various models available in order to generate captions on images and also describe about the architecture we have implemented which basically contains an Encoder and Decoder model along with Visual Attention models to generate captions.The validation of the captions is performed on a dataset which is a combination of Flickr 8k and COCO images.

I. INTRODUCTION

Image interpretation has become significantly important as social media images and various other types of informational images are always utilised in our daily lives. Since images are one the most unstructured form of data it makes it harder for us to search, index, and query images. To solve this we introduce a solution of image captioning using a visual attention model.

The major challenge that arises in the image captioning models is that the model must be powerful enough to solve the challenges arising in the field of computer vision, like determining which objects are in an image, but they must also be able to capture and demonstrate their relationships in a decoded natural language. It is a big challege for machine learning models to percieve things the way we humans do and being able to exactly capure all the spatial level and object level feature dependencies.

In this paper, we design approaches to caption generation that attempt to incorporate a form of attention inspired from The work of Neural machine translation in recent paper by *Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua* using the Bahanadou's attention model.

For Dataset we have prepared our own which is the combination of Flickr8k and COCO-MS which has a total of 98967 images which approximately 5 captions per image.

II. LITERATURE REVIEW

In this section we provide relevant background on previous work on image caption generation and attention. Recently, several methods have been proposed for generating image descriptions.

B. Deep Learning Model to Generate Captions Using Computer Vision & Machine Translation.

This work aims to detect different objects found in an image, recognize the relationships between those objects and generate captions.It also uses a technique called Transfer Learning with the help of Xception model.The model is based on a deep learning neural network that consists of a vision CNN followed by a language generating RNN. It generates complete sentences as an output.

They use a CNN + LSTM to take an image as input and output a caption. An "encoder" RNN maps the source sentence (which is of variable length) and transforms it into a fixed-length vector representation. A "decoder" RNN which ultimately generates the final meaningful sentence as a prediction, but they have replaced this RNN with a deep CNN. Since deep CNN can produce a rich representation of the input image by embedding it to a fixed-length vector - by first pre-training it for an image classification task and using the last hidden layer as an input to the RNN decoder that generates sentences.

C. *An Overview of Image Caption Generation Methods*

This paper discusses various methods used for Image Captioning using Deep Learning Techniques. The work done in this paper shows that there are basically two categories in Image captioning models i.e., [1] a method based on a statistical probability language model to generate handcraft features and [2] a neural network model based on an encoder-decoder language model to extract deep features. The first method used in this paper first analyzes the image, detects the object, and then generates a caption. Words are detected by applying a convolutional neural network (CNN) to the image area and integrating the information with MIL. Later it transforms the Image caption generation model into an optimization problem and searches for the most likely sequence of the words to describe the image.

The ability to focus on specific features of the image and drop out the unimportant features is known as attention. The work done in this paper also deals with various Attention Mechanism, their achievements, and also their application in image caption generation. Some of the mentioned attention mechanism algorithms are Hard Attention, Local Attention, Global Attention, Scaled Dot-Product Attention, and so on. The figure below shows a glimpse of two attention models.

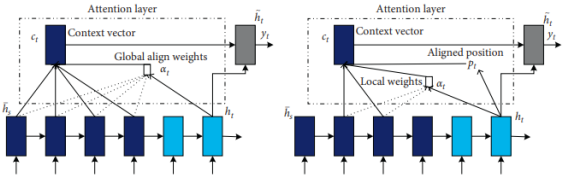


FIGURE 4: (a) Global attention model and (b) local attention model.

Along with this various evaluation metrics like BLEU, METEOR, ROGUE and SPICE.

III. DATA PREPROCESSING

For preprocessing, We chose the flickr8k dataset which consists of 8092 images and each image consists of approximately 5 captions then we also downloaded the COCO-MS dataset which consists of 90875 images and approximately 5 captions per image, so then we took the image path for each dataset elements in the images dataset from both flickr8k and coco-ms dataset as the key and value as a list of all the captions associated with it. From this dictionary of key-value pairs, we shuffled randomly and selected the first 16000 images to train. We first convert all the caption sentences to lower case, we then add a <start> and an <end> token to every caption to have the encoder-decoder model learn the start and end of the sentence captions being generated, and we resize every image to (299,299) and we used the resnet_v2.preprocess_input from Tensorflow to preprocess the input images as we were using ResNet101 for feature extraction. We then tokenize each of the captions using keras. layers.TextVectorization and we set the maximum caption size to be generated to 50 and we used regex replace to remove special characters from the captions, we then pick the most common words from the vocabulary after making the caption words unique and we select the top 5000 words for prediction in our model. We then covert the words to indices and add a <UNK> tag for the words that fall out of the range of the common vocabulary words. We then create training and validation sets using an 80-20 split randomly.

IV. TRAINING DETAILS

For our paper we propose two different models which is Image captioning using GRU based decoder and Image captioning using LSTM based decoder, further details about which is mentioned below.

- *MODELS*

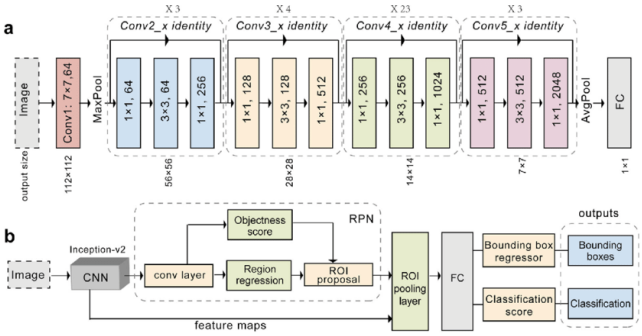
A. Image Caption Generation Using Resnet-101 Feature Extraction Model and LSTM-based decoder.

The model leverages exemplar images to map captions related to captured features and dependencies.

For feature extraction, we use the resnet-based feature extraction convolutional neural network.

ResNet-101 is a convolutional neural network that is 101 layers deep. For the training purpose, we use a pre-trained version of the network trained on more than a million images from the ImageNet database. The pretrained network can easily classify various images upto 1000 object categories. As a result, the network has learned rich feature representations for a wide range of images.

ResNet101 Architeture



For feature extraction, we remove the last fc layers and output layers since our aim is solely to extract features. The model has a total of 42,658,176, of which 42,552,832 are trainable and 105,344 are Non-trainable.

In our ResNet101 feature extraction model, we set the last layer as a hidden layer and the generated feature is of dimension (100, 2048)

For the Encoder, Our model takes a single raw image and generates a caption y encoded as a sequence of 1-of-K encoded words.

We use a convolutional neural network in order to extract a set of feature vectors which we refer to as annotation vectors. The extractor produces L vectors, each of which is a D-dimensional representation corresponding to a part of the image.

In order to obtain a correspondence between the feature vectors and portions of the 2-D image, we extract features from a lower convolutional layer, unlike previous work which instead used a fully connected layer. This allows the decoder to selectively focus on certain parts of an image by selecting a subset of all the feature vectors.

For the Decoder,

We use a long short-term memory (LSTM) network that produces a caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words.

In simple terms, the context vector $z\hat{}t$ is a dynamic representation of the relevant part of the image input at time t. We define a mechanism φ that computes $z\hat{}t$ from the annotation vectors ai , i = 1, . . . , L, corresponding to the features extracted at different image locations. For each location i, the mechanism generates a positive weight αi which can be interpreted either as the probability that location i is the right place to focus for producing the next word or as the relative importance to give to location i in blending the ai's together. The weight αi of each annotation vector a_i is computed by an attention model fatt for which we use a multilayer perceptron conditioned on the

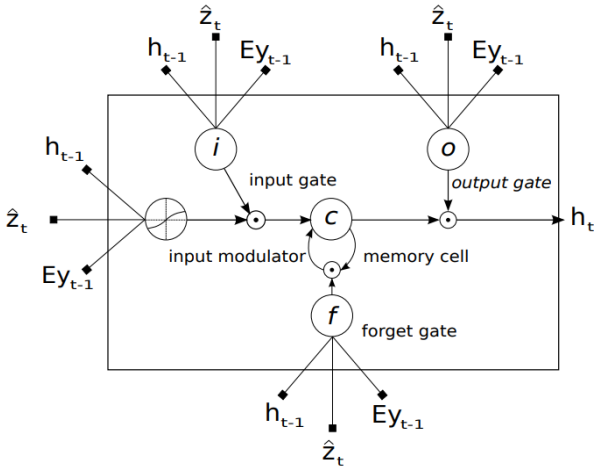
previous hidden state s_{t-1}. For the attention model we represent the global attention model which is the Bahdanau attention model inspired from neural machine translation.For emphasis, we note that the hidden state varies as the output RNN advances in its output sequence: "where" the network looks next depends on the sequence of words that has already been generated.

$$e_{ti} = f_{att}(a_i , s_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}$$

Once the weights are computed, the context vector zˆt is computed by

where φ is a function that returns a single vector given the set of annotation vectors and their corresponding weights. The initial memory state and hidden state of the LSTM are predicted by an average of the annotation vectors fed



Attention Model,

We modelled an overall global attention mechanism inspired from Neural machine translation in recent paper by *Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua* namely Bahdanau attention. The decoder takes the encoder features and passes it to the attention model together with the previous hidden decoder state, st−1

This generates an attention score:

$$e_{t,i} = a(s_{t-1}, f_i)$$

The weights are calculated using two dense layers of size 512 , one for the hidden state s_{t-1} and the other for the encoded features.

$$a(s_{t-1}, f_i) = V^T \tanh(W_1(f_i) + W_2(s_{t-1}))$$

Here,

V is a weight vector which is a one dimension dense layer.

The model is parametrized as a feedforward neural network, and jointly trained with the remaining system components.

Subsequently, a softmax function is applied to each attention score to obtain the corresponding weight value:

$$\alpha_{t,i} = \text{softmax}(e_{t,i})$$

The application of the softmax function essentially normalizes the annotation values to a range between 0 and 1 and, hence, the resulting weights can be considered as probability values.

Each probability (or weight) value reflects how important f_i and s_{t-1} are in generating the next state s_t ,and the next output, y_t .

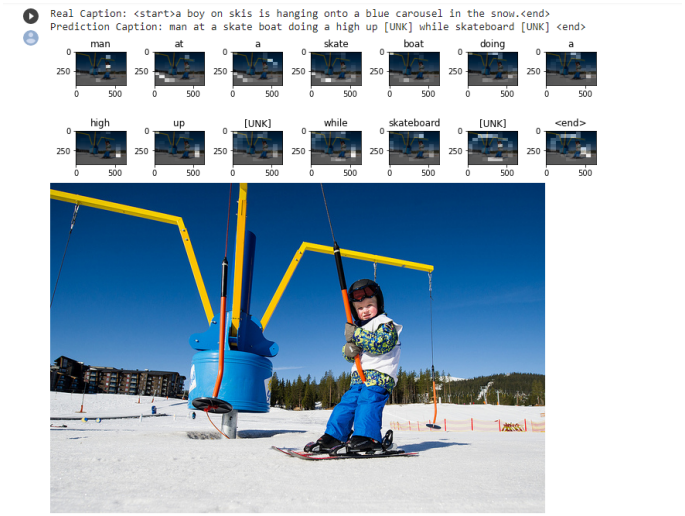
V. EVALUTATION OF CERTAIN IMAGES

Image modelling attention on different words also shown

Test image [1]

Real Caption: <start>a boy on skis is hanging onto a blue carousel in the snow.<end>

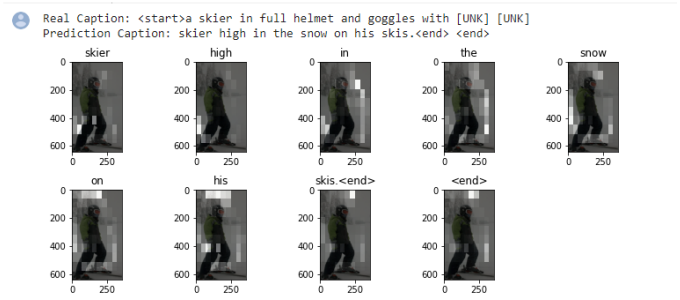
Prediction Caption: man at a skate boat doing a high up [UNK] while skateboard [UNK] <end>



Test image [2]

Real Caption: <start>a skier in full helmet and goggles with [UNK] [UNK]

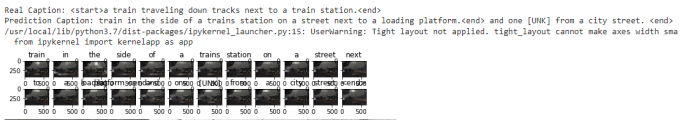
Prediction Caption: skier high in the snow on his skis.<end> <end>



Test image [3]

Real Caption: <start>a train traveling down tracks next to a train station.<end>

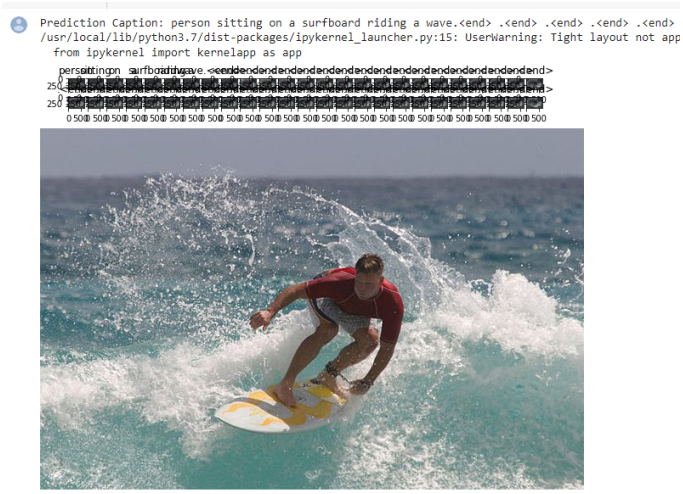
Prediction Caption: train in the side of a trains station on a street next to a loading platform.<end> and one [UNK] from a city street. <end>



Test image [4]

For test data outside the dataset:

Prediction Caption: person sitting on a surfboard riding a wave.<end> .<end> .<end> .<end> .<end> .<end> .<end>



VI. ACKNOWLEDGMENTS

We place on record our gratitude to Professor Uma and the entire Topics in Deep Learning course team, for motivating us and providing us with this opportunity to get hands-on experience in the domain. This project has proven to be a great experience in learning how various models work with various type of data and what processing techniques to be used to keep the model up and running to get the required results. We hope to work on the project in the future to further our findings and establish them fully.

REFERENCES

[1] Haoran Wang , 1 Yue Zhang,1 and Xiaosheng Yu 2, “An Overview of Image Caption Generation Methods” , Research Article Published 8 January 2020.

[2] Kelvin Xu, Jimmy Lei Ba†,Ryan Kiros, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention