A Project Report

on

Image Captioning Using Neural Network

Bachelor of Technology

IN

COMPUTER ENGINEERING

BY

Chandresh Gupta
Enrolment No. : GJ6410

Ajay Gahlot
Enrolment No. : GI1448

Under the Guidance of

Mr. Asad Mohammed Khan
Mr. Tameem Ahmad
Department Of Computer Engineering

Zakir Husain College of Engineering & Technology
Aligarh Muslim University
Aligarh (India)-202002

March 2020

Dated …………..

Declaration

The work presented in project entitle "Image Captioning Using Neural Network" submitted to the Department of Computer Engineering, Zakir Husain College of Engineering and Technology, Aligarh Muslim University Aligarh, for the award of the degree of Bachelor of Technology in Computer Engineering, during the session 2019-20, is my original work. I have neither plagiarized nor submitted the same work for the award of any degree.

Date:
Place *Chandresh Gupta*

Md Adil Hameed

Ajay Gahlot

Dated :

Certificate

This is to certify that the Project Report entitled "Image Captioning Using Neural Network", being submitted by ***Chandresh Gupta, Md Adil Hameed and Ajay Gahlot****, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Engineering, during the session 2019-20, in the Department of Computer Engineering, Zakir Husain College of Engineering and Technology, Aligarh Muslim University Aligarh, is a record of candidate's own work carried out by him under my (our) supervision and guidance.*

Mr. Tameem Ahmad
Assistant Professor
Department of Computer Engineering
ZHCET, AMU, Aligarh

Mr. Asad Mohammed Khan
Assistant Professor
Department of Computer Engineering
ZHCET, AMU, Aligarh

Table of Contents

Abstract

Image captioning is a task which lies in the intersection of areas of object detection and natural language processing. We will be proposing a, model which will be utilizing both the areas of CV and NLP for the automatic generation of the captions of the given image. Model that we are going to propose mimics the human visual system that automatically describe image content. Main idea of our model is that rather than focusing on whole image it is better to focus on particular areas like the areas where objects are present in the image. Our model consists of two sub model, first sub model or an encoder consist of object detection part which is used to identify the object in the given image along with their spatial location and finally making annotation vector consist of object features and their spatial features. Second sub model or decoder consist of RNN based LSTM network along with attention network which produce a context vector based on annotation vector at a particular time and finally at each step LSTM takes input of attention network along with the other input to generate caption of a given image.

Acknowledgement

First of all, we want to thank almighty God, whose blessings helped us in achieving our project. We are very thankful to our supervisor "Mr. Asad Mohammed Khan" and "Mr. Tameem Ahmad" in guiding and helping us throughout our project. Our project would have not been completed if they haven't guided us during the course of the project.

We also want to express our gratitude to all other teachers for teaching us various different topics with which we were able to complete our project on time.

We are grateful to our families which supported us during our tough time, giving us motivation, encouragement as well as love, without their support this wouldn't seems possible.

At last, we also wants to thank our friends, classmates whose support, kind co-operation and motivation helped us in solving various problems that aroused during development of the project. We would like to extend our heartiest gratitude to all those who directly or indirectly helped us in completing the project.

Chandresh Gupta
Md Adil Hameed
Ajay Gahlot

Acronyms

BLEU	Bilingual Evaluation Understudy
CNN	Convolutional Neural Network
CV	Computer Vision
LSTM	Long Short Term Memory
NLP	Natural Language Processing
NN	Neural Network
RNN	Recurrent Neural Network

List of Figures

Chapter 1 : Introduction

1.1 Motivation

Past works are adjust in the decoding step, they more often than not utilize recurrent neural networks (RNN) based on long short-term memory (LSTM) units as a decoder. With respect to the encoding step is concern, the work is separated in to two noteworthy classes: CNN-RNN models and attention based models. CNN-RNN models represents an image as a single feature vector from the top layer of a pretrained convolutional neural network though, attention based models utilize the vector made up by the representations of image's sub regions as the source vector. The greatest disadvantage of CNN-RNN models is that they barely adjust diverse visual pieces of the information image to words in captions. Our model likewise pursues the Encoder-Decoder structure, however, it is absolutely different from the past models. Our inspiration in depicting a picture is to discover its contents, as opposed to concentrating on some futile regions related with it. Content which we are talking about in the image are objects which are present in the image. Along with the object in the image we will also combine their spatial location.

1.2 Objectives and Scopes

- We are going to propose a multimodal based image caption generation based on attention mechanism.
- In encoding part of our model, we utilize an object detection model to identify objects position in the image and each object region represented by a Convolutional feature vector.
- The decoder part utilizes task-explicit context to anticipate an attention distribution over the image areas. Attended vector is then processed as a weighted average of image feature over all areas.
- These attended feature vector along with other inputs is fed into LSTM network to predict the caption.
- In order to measure the performance, we will evaluate our model on Flickr8K image dataset using standard metrics like BLEU and compared results with previous state of arts object detection algorithm.

1.3 Organisation

This report is organized into 7 chapters.

- The first chapter gives the introduction and overview of the project.
- The second chapter then discusses the different network used in our model and previous work as well
- The third chapter discuss about the model of our project.
- The fourth chapter is about the whole implementation of the project.
- The fifth chapter is related to experimental setup.
- The sixth chapter shows the experimental results and all the outputs.
- Finally the last chapter provides conclusion and future work of this project.

Chapter 2 : Literature Review

In this section we provide overview on different networks we have used and relevant background on previous work on image caption generation and attention.

2.1 VGG16 as Feature Extractor

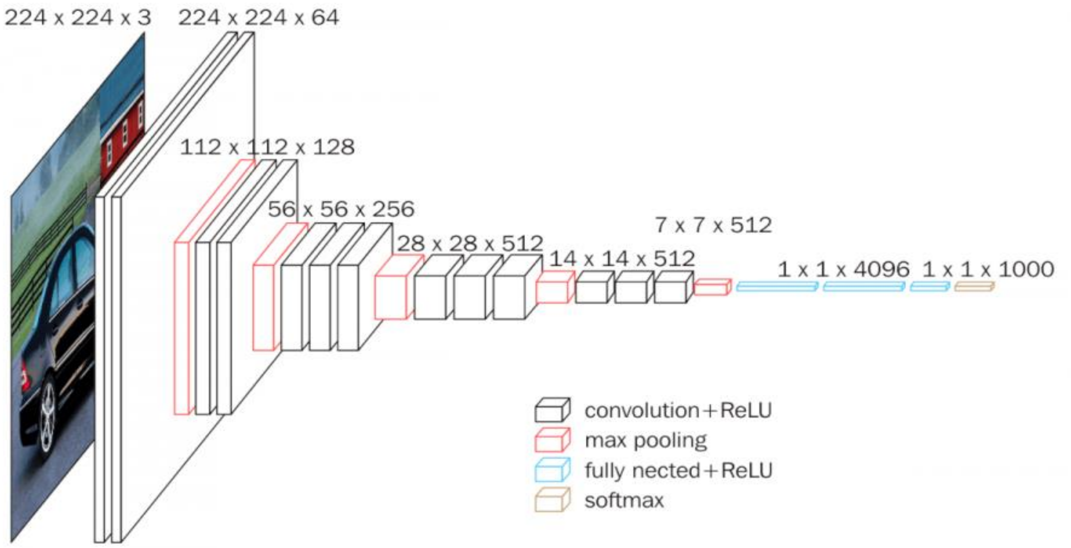


Figure 1. VGG16 architecture used as feature vector.

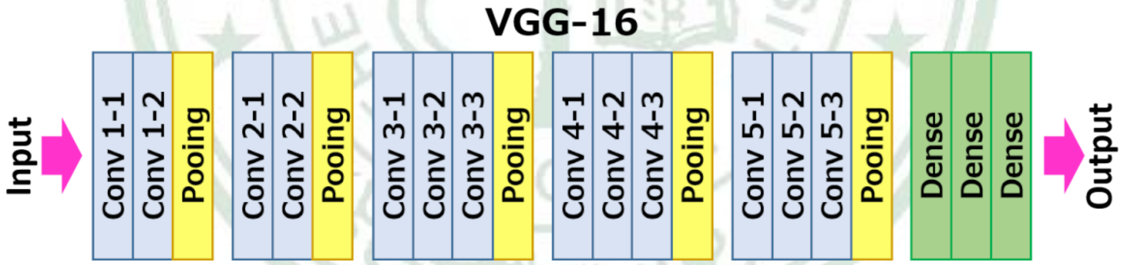


Figure 2. Layers of VGG16.

VGG16 is a convolution neural net (CNN) architecture which was used to win ILSVR (Imagenet) competition in 2014. It is considered to be one of the excellent vision model architecture till date. Most unique thing about VGG16 is that instead of having a large number of hyper-parameter they focused on having convolution layers of 3x3 filter with a stride 1 and always used same padding and maxpool layer of 2x2 filter of stride 2. It follows this arrangement of convolution and max pool layers consistently throughout the whole architecture. In the end it has 2 FC (fully connected layers) followed by a softmax layer.

2.2 InceptionV3 as Feature Extractor

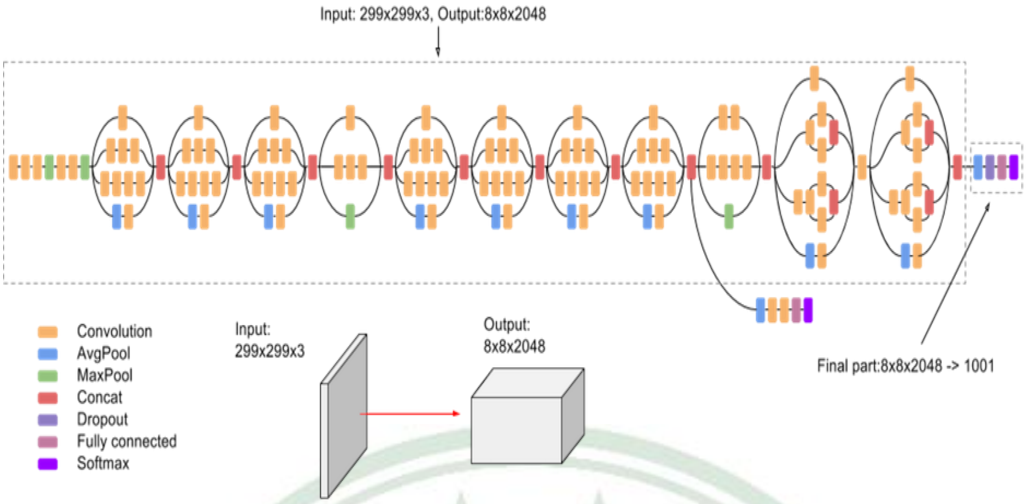


Figure 3. InceptionV3 Feature Extractor

Inception v3 is a widely-used image recognition model that has been shown to attain greater than 78.1% accuracy on the ImageNet dataset. The model is the culmination of many ideas developed by multiple researchers over the years. The model itself is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concats, dropouts, and fully connected layers. Batchnorm is used extensively throughout the model and applied to activation inputs. Loss is computed via Softmax.

2.3 RNN (Recurrent Neural Network)

Begin their reasoning starting from scratch. As you read this report, you see each word dependent on your comprehension of past words. You don't discard everything and begin thinking sans preparation once more. Conventional neural network can't do this, and it appears to be a noteworthy inadequacy. For instance, envision you need to classify what sort of occasion is occurring at each point in a motion picture. It's misty how a conventional neural system could utilize its thinking about past occasions in the film to educate later ones. This issue is address by RNN. RNN has loop in them allowing information to persist.

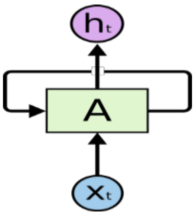


Figure 4. RNN with loop

A RNN can be thought of as more than one copies of the network, each transferring a message forward.

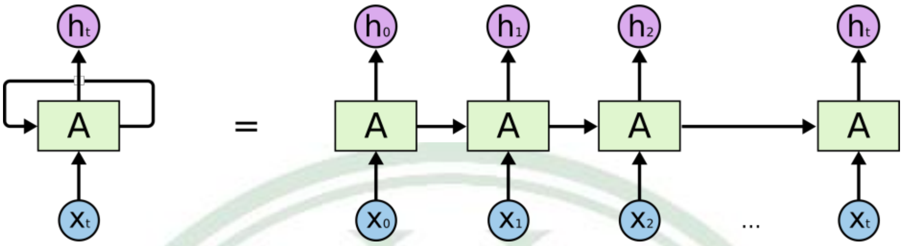


Figure 5. Unrolled RNN

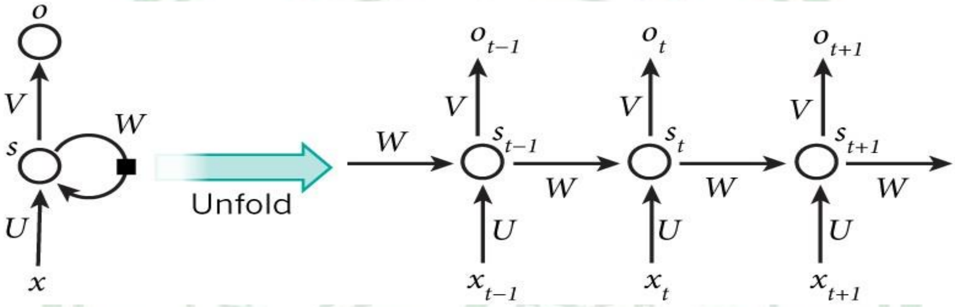


Figure 6. RNN with different weights and input and output at each step.

2.4 LSTM (Long Short Term Memory)

To avoid long term dependency LSTM are used. Their default behavior is remembering information for long time period. All RNNs has chain of repeating NN. This repeating NN has simple single tanh layer in RNNs. LSTM similar to RNN but neural network module has different structure. Instead of single layer like RNN it has 4 layers interacting in a different way.

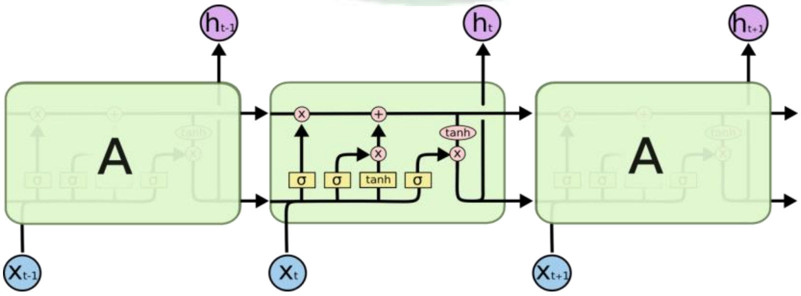


Figure 7. LSTM repeating module

2.5 Previous Work

In this section we provide previous background work that has been done in the field of image caption generation. The problem of generating natural language descriptions from visual data has long been studied in computer vision, but mainly for video.

Generally, the existing image captioning algorithms can be divided into three categories based on the way of sentence generation [8]: *template-based methods, transfer-based methods, and neural network-based methods.*

The ***template-based methods*** either use templates or design a language model, which fill in slots of a template based on co-occurrence relations gained from the corpus [11], conditional random field [12], or web-scale n-gram data [9]. Visual dependency representation is proposed to extract relationships among the objects [13]. The template-based methods are simple and intuitive, but are heavily hand designed and unexpressive, which are not flexible enough to generate meaningful sentences. Since these techniques achieve the image captioning task either by utilizing existing captions in the preparation set or depending on hard-coded language structures, the impediment of strategies received in early work is that they are not sufficiently adaptable. Therefore, expressiveness of produced descriptions by these strategies is, to a substantial degree, restricted.

The ***transfer-based methods*** are based on the retrieval approaches, which directly transfer the descriptions of the retrieved images to the query image. Some approaches took the input image as a query and selected a description in a joint image-sentence embedding space. Kuznets ova et al.[15,16] retrieved images that are similar to the input image, extracted segments from their captions, and organized these segments into a sentence. The generated sentences by the transfer based methods are often with correct grammar. However, these methods may misrecognize the visual content and cannot generate novel phrases or sentences, and thus are limited in image captioning. Notwithstanding, they indicate that we can take advantage of the images similar to the input image. This idea can be applied in other approaches, such as re-ranking candidate descriptions generated by other models and emotion distribution prediction. We also undertake this idea in our generation process.

The ***neural network-based methods*** come from the recent advantages in machine translation with the use of CNN and RNN. Ding et al. [1] proposed a multimodal layer to connect a deep CNN for images and a deep RNN for sentences, allowing the model to generate the next word given the input word and the image. Inspired by the encoder-decoder model [2] in machine translation, Vinyals et al. [3] used a deep CNN to encode the image instead of a RNN for sentences, and then used LSTM , a more powerful RNN, to decode the image vector to a sentence.

In this work we combine deep convolutional nets for image classification with recurrent networks for sequence modeling, to create a single network called as encoder-decoder that generates descriptions of images. The RNN is trained in the context of this single “end-to-end” network. The model is inspired by recent successes of sequence generation in machine translation and uses LSTM unit of RNN for the generation of captions for the given texts.

Chapter 3: Proposed Model

This section describe the detail of the proposed model, which consists of two main parts: encoding and decoding. The input to our model is a single image I, while the output is a descriptive sentence consists of K encoded words: $\boldsymbol{y} = \{y1, y2, y3, \dots, yC\}$, yi belongs to R^K ,Where K is the size of the vocabulary and C is the length of caption.

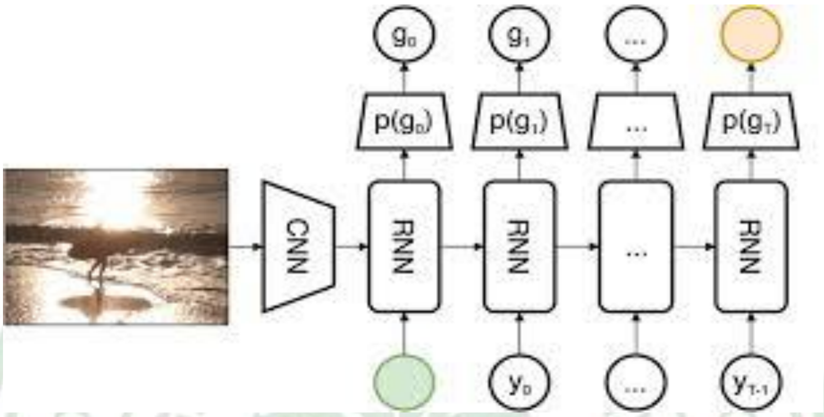


Figure 8. Model Architecture

In the encoding part we present a model that recognizes the input image and extract their features. All the information will be represented as a set of feature vectors referred as annotation vectors. The encoding part generates *L* annotation vectors, each of which is a D-dimensional representation corresponding to an object.

In the decoding part, we use a long short-term memory (LSTM) network that produces a caption using annotation vector by generating one word at every time step.

3.1 Encoder as CNN

In the previous couple of years, noteworthy advancement have been done in object detection. Everything began with "Rich feature hierarchies for accurate object detection and semantic segmentation" (R-CNN) in 2014, which utilized a algorithm called Selective Search to propose possible regions of interest and a standard Convolutional Neural Network (CNN) to classify and adjust them.

In our model we have used InceptionV3 CNN because it has better performance in terms of speed as well as accuracy as compared to other models.

As we referenced before, the initial step is utilizing a CNN pretrained for the classification task and utilizing the intermediate layer output i.e output of last convolutional layer. Each CNN layer produces abstractions dependent on the past data. The primary layers learn edges, the second discovers patterns in edges so as to actuate for more complex shapes. Finally we end up with a convolutional feature map which has spatial measurements much smaller than the first picture, yet more depth. The feature map width and height decline on account of the pooling applied between convolutional layers and the depth increments dependent on the quantity of filters the convolutional layer learns. In its depth, the convolutional feature map has encoded all the data for the image while keeping up the area of the "things" it has encoded with respect to the first image.

3.2 Decoder as RNN

Untill now, we have build CNN which convert the input image into its feature vector. Now these image feature vector along with the partial caption acts as input to the repetitive unit of RNN and predict the next word in the sequence of the caption.

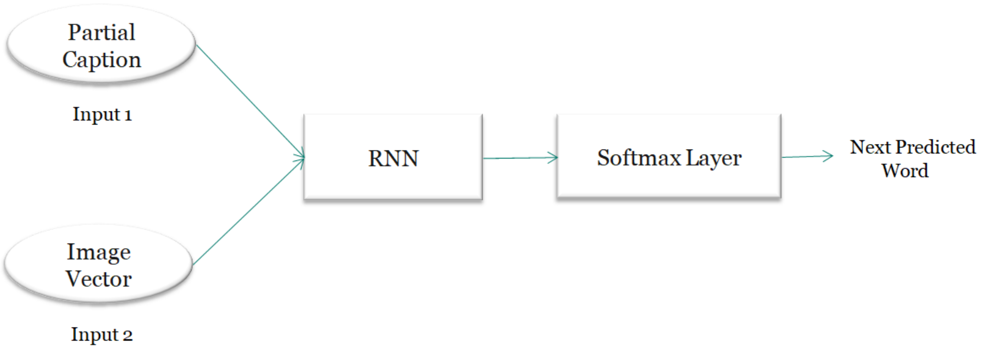


Figure 9. RNN Data Flow Model

Thus, we propose to directly maximize the probability of the correct description given the image by using the following formulation:

$$\theta^{\star} = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta)$$

where θ are the parameters of our model, I is an image, and S its correct transcription.

Since S represents any sentence, its length is unbounded. Thus, it is common to apply the chain rule to model the joint probability over S0,...,SN, where N is the length of this particular example as

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

It is natural to model p(St|I,S0,...,St−1) with a Recurrent Neural Network (RNN), where the variable number of words we condition upon up to t−1 is expressed by a fixed length hidden state or memory ht. This memory is updated after seeing a new input xt by using a non-linear function f:

$$h_{t+1} = f(h_t, x_t)$$

Now for the implementation of function f we have used LSTM unit of RNN because it has the ability to deal with the vanishing gradients problem.

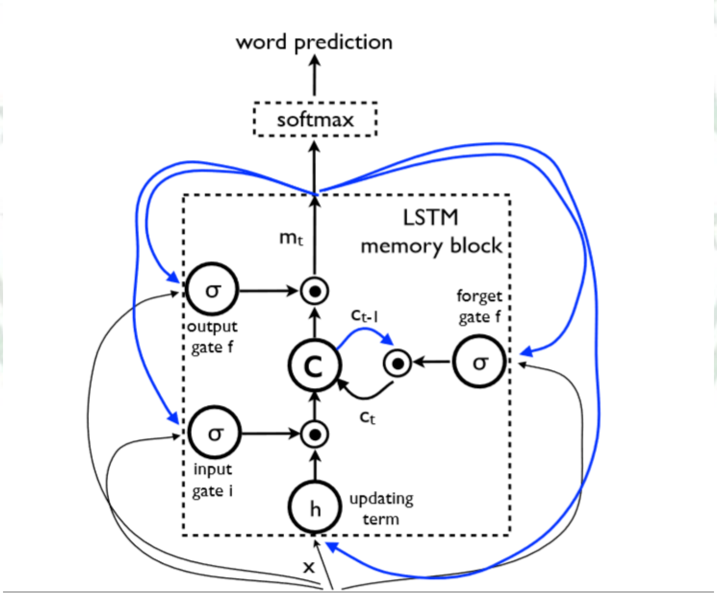


Figure 10. Description of LSTM Network

The core of the LSTM model is a memory cell c encoding knowledge at every time step of what inputs have been observed up to this step. The behavior of the cell is controlled by "gates"–layers which are applied multiplicatively and thus can either keep a value from the gated layer if the gate is 1 or zero this value if the gate is 0. In particular, three gates are being used which control whether to forget the current cell value (forget gate f), if it should read its input (input gate i) and whether to output the new cell value (output gate o). The definition of the gates and cell update and output are as follows :

$$
\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1}) \\
f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1}) \\
o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1}) \\
m_t &= o_t \odot c_t \\
p_{t+1} &= \text{Softmax}(m_t)
\end{aligned}
$$

Now we get the probability of the word which is going to be predicted by the model. In this way we use probabilistic measure for the prediction of the word in the generation of the caption.

Chapter 4: Implementation

4.1 Dataset

We will be using Flickr8K image dataset for image captioning. In this images are classified as 6000 in training set, 1000 in validation set and 1000 in testing set and each image is associated with five caption each.

4.2 Data Preparation

This section describes the preprocessing of images and caption before giving input to the network.

4.1.1 Image

Since we are going to use the pre trained object detector module, we need to process the images into the form pre trained object detector is accustomed to. We will resize the image of Flickr8K dataset to 256 x 256 for uniformity. And image input to the object detector must be float tensor of dimension N, 3,256,256 where N is the batch size.

4.1.2 Caption

Each image contains five captions associated with it. We need to clean these captions i.e.

4.1.2.1 We will convert word to lowercase in all captions.

4.1.2.2 We will remove punctuation in all captions.

4.1.2.3 Remove word from the caption which contains number in them.

Our main objective is to get the vocabulary which is small and the same time expressive also.

Captions are the one which we give input to the decoder as well as the output we get is also the captions as at each step word generated depends upon the previous step word generated.

In order to generate the first word we will going to give starting word as input i.e. <start>. And the last word predicted should be <end>, when we get the <end> as output of decoder we came to know that we need decoding to be stop.

We will also going to create the word map i.e. mapping for each word in the corpus to the unique index, it should also include <start> and <end> tokens.

4.3 Feature Extraction

- Using VGG16 and InceptionV3 pre-trained model, we will going to detect the object present in the image.
- VGG16 has accuracy of 90.1% and InceptionV3 model has accuracy of 93.7% as defines by documentation of Keras which is a Python library.
- Firstly we remove the last layer from the model which is softmax layer because we do not want the classification probability, we just want the feature vector corresponding to the input image.
- We get the feature vector of length 2048 and we save all the feature vector of training images in dictionary which maps image to its corresponding feature vector.
- We have also save all the feature vector in the .pkl file.
- Now this feature vector will act as the input to the LSTM unit along with the partial caption which will generate the full caption corresponding to that image.

4.4 Decoding using LSTM

- Decoder's task is to analyze the encoded image and produce the caption word by word.
- Since it repeat itself again and again, so we will used LSTM which is a recurrent neural network.
- In decoder, we simply take the mean of encoded image over all the pixels and fed this as the input to the first hidden state and produce the caption.
- Each predicted word in the output is used to predict the next word in the sequence to produce full caption.

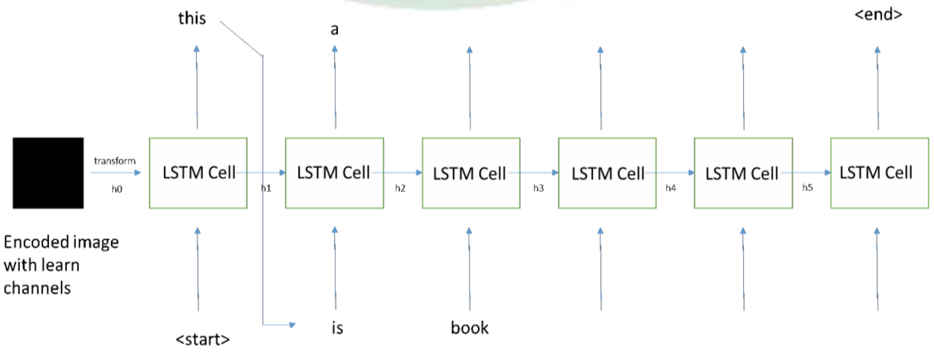


Figure 11. Diagrammatic Representation of decoder

4.5 Greedy Search

- Greedy Search is used to select the next predicted word from the generated probability distribution.
- The model basically generates the probability distribution of all the words in the vocabulary.
- Model greedily selects the most likely word with highest probability.
- This is called **Maximum Likelihood Estimation (MLE)** or **Greedy Search.**

4.6 BLEU Evaluation Metrics

In order to evaluate the BLEU (bilingual evaluation understudy) metric compare produce hypothesis with reference sentences in n- grams. In order to determine BLEU-1 hypothesis is compared with reference sentence in unigram, while for evaluation of BLEU-2, matching is done with bigram and so on. A greatest order of four is experimentally resolved to get the best relationship with human decisions.

Chapter 5: Experimental Setup

5.1 System Specification and library used

- **Language used :** Python 3.6.6
- **Libraries used :** Numpy 1.15.4
Sklearn 0.19.1
Kears
NLTK 3.3
Matplotlib 3.0.2
- **Platform used :** Google Cloud Colab
- **GPU:** GeForce GTX 1080 Ti/ PCIe/ SSE2

Chapter 6 : Results

We have used BLEU Evaluation Metrics for analyzing the performance of our model.

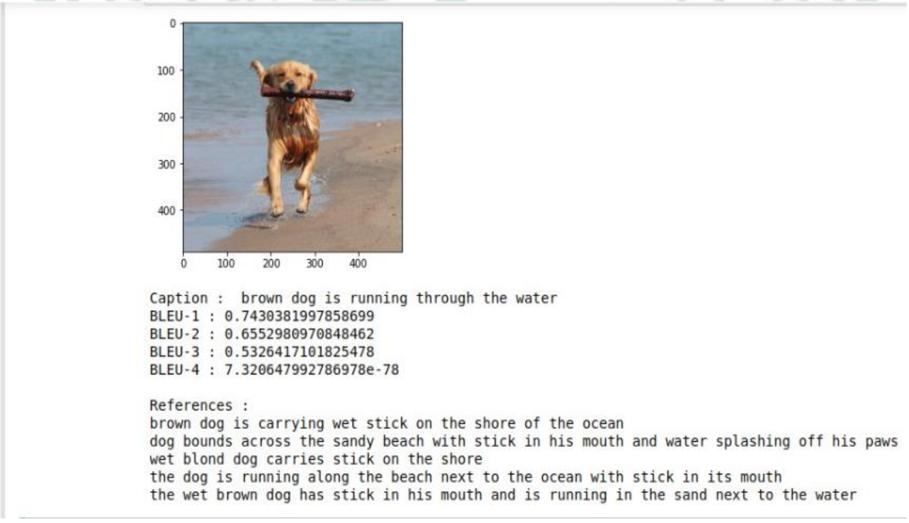
BLEU Score	Model Performance
BLEU-1	0.58
BLEU-2	0.30
BLEU-3	0.18
BLEU-4	0.08

Some of the results have been shown in below figures.

We can see that the model generates the sentence which describes the image quite well. But it sometimes makes some mistakes in generation of sentences but it is pretty obvious as humans also make some mistakes.

Above shown results can be improved if we train big network on larger dataset.

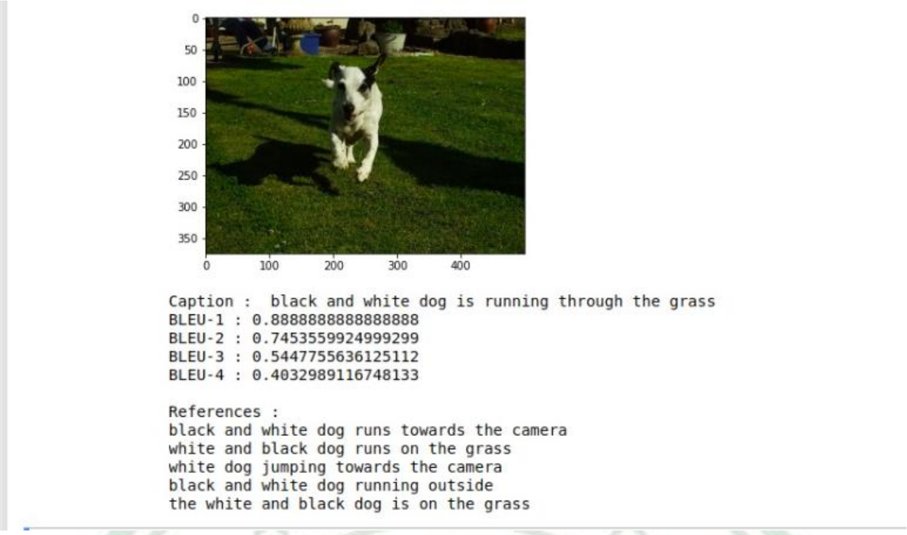
Output 1 :



```
Caption :  brown dog is running through the water
BLEU-1 : 0.7430381997858699
BLEU-2 : 0.6552980970848462
BLEU-3 : 0.5326417101825478
BLEU-4 : 7.320647992786978e-78

References :
brown dog is carrying wet stick on the shore of the ocean
dog bounds across the sandy beach with stick in his mouth and water splashing off his paws
wet blond dog carries stick on the shore
the dog is running along the beach next to the ocean with stick in its mouth
the wet brown dog has stick in his mouth and is running in the sand next to the water
```

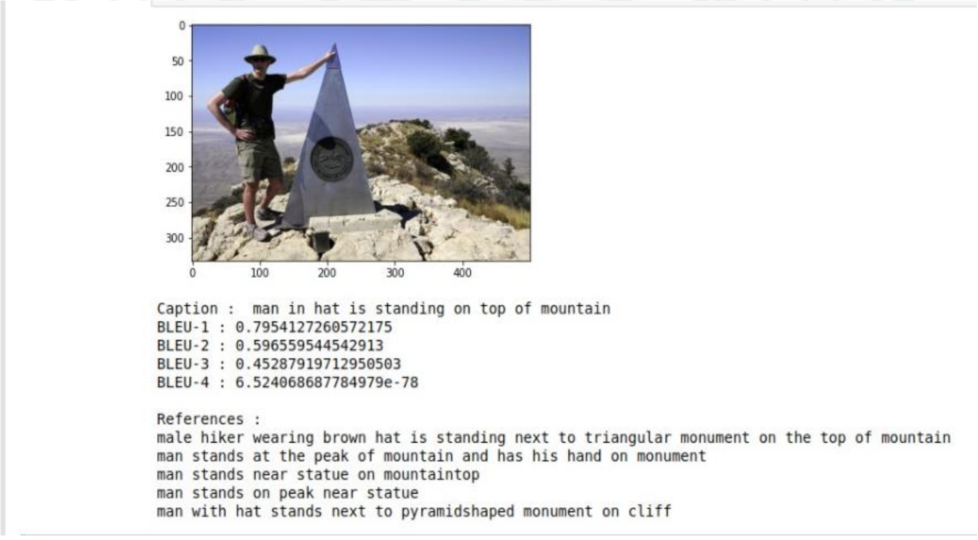
Output 2 :



```
Caption :  black and white dog is running through the grass
BLEU-1 : 0.8888888888888888
BLEU-2 : 0.7453559924999299
BLEU-3 : 0.5447755636125112
BLEU-4 : 0.4032989116748133

References :
black and white dog runs towards the camera
white and black dog runs on the grass
white dog jumping towards the camera
black and white dog running outside
the white and black dog is on the grass
```

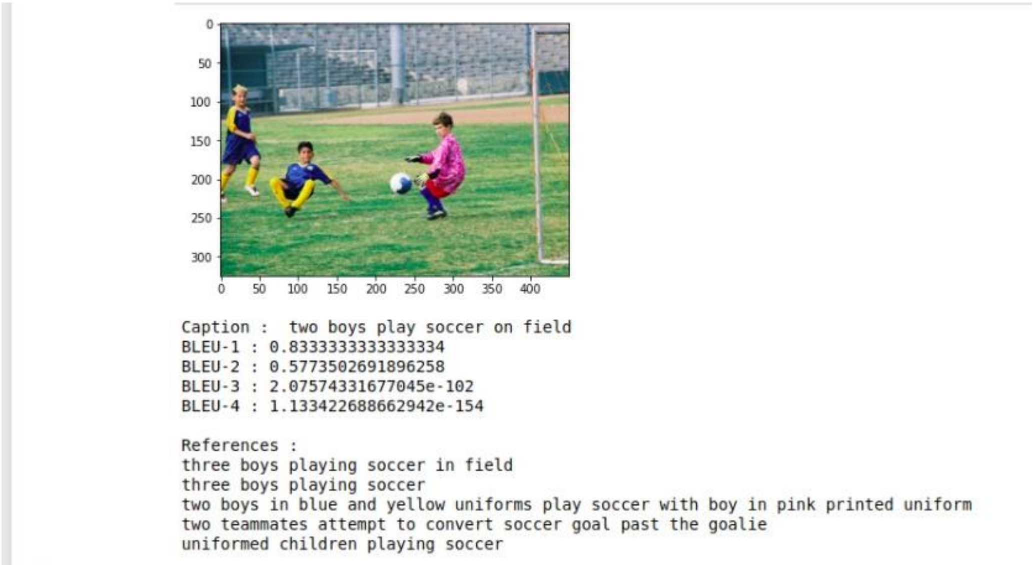
Output 3 :



```
Caption :  man in hat is standing on top of mountain
BLEU-1 : 0.7954127260572175
BLEU-2 : 0.596559544542913
BLEU-3 : 0.45287919712950503
BLEU-4 : 6.524068687784979e-78

References :
male hiker wearing brown hat is standing next to triangular monument on the top of mountain
man stands at the peak of mountain and has his hand on monument
man stands near statue on mountaintop
man stands on peak near statue
man with hat stands next to pyramidshaped monument on cliff
```

Output 4 :



```
Caption :  two boys play soccer on field
BLEU-1 : 0.8333333333333334
BLEU-2 : 0.5773502691896258
BLEU-3 : 2.07574331677045e-102
BLEU-4 : 1.133422688662942e-154

References :
three boys playing soccer in field
three boys playing soccer
two boys in blue and yellow uniforms play soccer with boy in pink printed uniform
two teammates attempt to convert soccer goal past the goalie
uniformed children playing soccer
```

Chapter 7 : Conclusions and Future Work

7.1 Conclusions

We proposed an image encoder model which will convert image to the required feature vector. Now these feature vector can be used for the generation of image description. In this overall work, we combine "Image Labeling" and "Automatic Machine Translation" into an end-to-end hybrid neural network system. The developed model is capable to autonomously view an image and generate a reasonable description in natural language with reasonable accuracy and naturalness. Further extension of the present model can be in regard to increasing additional CNN layers or increasing/implementing pre-training, which could improve the accuracy of the predictions. These applications can be used by common people in the form of image search like Amazon Image Search and Google Image Search. In future research work similar models can be used which rather than directly using CNN features will be using features related to more specific regions of the image.

7.2 Future Work

- Rather than just using VGG16 and InceptionV3 model as feature extractor, we can also use image encoders and their results can be compared to show which one will going to give better results.
- Reverse image captioning model can be generated following the similar network i.e. network that generate similar images given the description of image.
- Neural network based video captioning can also be done following similar patterns but employing different algorithms which can work on videos easily and effectively.
- Much better results can be get by using better attention mechanism that is by using hard attention mechanism than using soft attention mechanism and also other variants of attention mechanism which combines idea of both hard and soft attention mechanism

References

[1] Ding, Guiguang, et al. "Neural image caption generation with weighted training and reference." *Cognitive Computation* (2018): 1-15.

[2] Singha, Rahul, and Aayush Sharma. “Image captioning using Deep Neural Networks.” (*ACM Transactions on Intelligent Systems and Technology (TIST)* May 2018).

[3] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015).

[4] Pu, Yunchen, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. "Variational autoencoder for deep learning of images, labels and captions." In *Advances in neural information processing systems*, pp. 2352-2360. 2016.

[5] He, Xiaodong, and Li Deng. *"Deep learning for image-to-text generation: A technical overview." IEEE Signal Processing Magazine 34.6 (2017): 109-116.*

[6] Hossain, MD Zakir, et al. "A comprehensive survey of deep learning for image captioning." *ACM Computing Surveys (CSUR)* 51.6 (2019): 1-36.

[7] Wang, Cheng, et al. "Image captioning with deep bidirectional LSTMs." *Proceedings of the 24th ACM international conference on Multimedia. 2016.*

[8] Jia X, Gavves E, Fernando B, Tuytelaars T. Guiding the long short term memory model for image caption generation. In: *IEEE international conference on computer vision, pp 2407–2415 2015.*

[9] Li S, Kulkarni G, Berg T, Berg A, Choi Y. “Composing simple image descriptions using web-scale n-grams”. In: *The SIGNLL conference on computational natural language learning, pp 220– 228. 2011.*

[10] Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing. "Convolutional image captioning." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018..*

[11] Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D. “Every picture tells a story: generating sentences from images.” *In: European conference on computer vision, pp 15–29. 2010.*

*[12]*Kulkarni G, Premraj V, Dhar S, Li S, Choi Y, Berg A, Berg T. Baby talk: “understanding and generating simple image descriptions.” *In: IEEE conference on computer vision and pattern recognition, pp 1601–1608. 2011.*

*[13]*Elliott D, Keller F. “Image description using visual dependency representations.” *In: Conference on empirical methods on natural language processing, pp 1292–1302. 2013.*

*[14]*K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: “A method for automatic evaluation of machine translation.” *In ACL, 2002.*

*[15]*Kuznetsova P, Ordonez V, Berg A, Berg T, Choi Y. “Collective generation of natural image descriptions”. In: *Annual meeting of the association for computational linguistics, pp 359–368. 2012.*

*[16]*Kuznetsova P, Ordonez V, Berg T, Choi Y. Treetalk: “composition and compression of trees for image descriptions.: *Trans Assoc Comput Linguist. 2014;2(10):351–62.*