Urdu Image Caption generation using Deep learning

Submitted by

Muhammad Osama Nusrat

Supervised by

Sir Shoaib Mehboob

MS-AI

FAST School of Computing

National University of Computer & Emerging Sciences

Islamabad, Pakistan.

May 28, 2022

ABSTRACT

In recent times deep learning has evolved as a major field in science and with each new day, some new innovations are been done in this field. A huge amount of research is going into different deep learning techniques Automatic image captioning is one of these techniques. Most of the research is in generating captions in English. Very little work has been done in generating captions in Urdu. This paper will focus on generating Urdu captions by using CNN (Exception) as an encoder and RNN (GRU) as a decoder with an attention model. I have made a dataset in Urdu comprising 1500 images that are related to a popular Flickr 8k dataset.

Chapter 1

Introduction

1.1 Background

We have seen captions under the images in newspapers magazines etc. Captions are an important part of the image and they describe what is happening in the image. They tell us the crux of the image. Automatic generation of captions is a new field and it has become more popular in the last decade. Basically, we use computer vision techniques and natural language processing techniques in image captioning. However, this process of captioning images is not as easy to process as it looks for humans we can just look at an image and visualize it in our mind and can give a good caption but doing the same task with the help of a machine is challenging. Techniques in deep learning like image classification and object detection are relatively Easy to achieve than image caption it is because in image caption the caption should be. According to the image what is happening or being shown in the image is precisely what I want to say is that there should be a connection this is the most tedious part of this research.

`

1.2 Motivation

The motivation for doing this project is that there is a gap for captioning images automatically in our national language Urdu. This has been done in various different languages such as English, French and Spanish, etc. so this is the reason which motivated me to do this project.

1.3 Problem Statement

The automatic generation of captions of images is an important problem in the domain of AI.
The problem statement is to generate caption of images in Urdu using deep learning and Nlp using the Flickr 8k dataset. 1500 images will be taken from 8k images and each image will be manually annotated in Urdu 5 times.

1.4 Contribution

My contribution is that Xception which is a CNN model has not been used previously as an encoder so I will use this CNN architecture for feature extraction. Moreover, I have used the Flickr 8k dataset and taken 1500 images from it, and manually caption them in Urdu. In short, there is a gap in image captioning in Urdu so I am filling the gap by doing research.

`

Chapter 2

Literature Review

I have read 6 papers related to my topic. In the first paper, the author proposed an image caption generating model with an encoder and decoder. The author used a pre-trained 16 layer vggnet convolutional neural network for feature extraction of the input image. It outputs a feature vector which is then fed to the recurrent neural network for decoding. Bidirectional long short-term memory (LSTM) networks are used as a decoder for generating captions for the input image. This model is then tested on three datasets which are Flickr8k, Flickr30k, and MSCOCO datasets respectively. Each image in the dataset is pre-annotated with 5 captions. The vocabulary of unique words is built by doing tokenization by removing all the words which occurred less than 5 times in the training set. The model showed the best results on the MS COCO dataset due to the greater no of training images. BLEU was used as an evaluation metric for performance.

In the second paper, the author has proposed an image caption generator using CNN and RNN. Pretrained models of VGG 16 and Vgg 19 are used as encoders for extracting features from the input images. LSTM is used as a decoder for generating captions of the images. 3 data sets are used which are Flickr8k, Flickr30k, and MSCOCO datasets respectively. Each image in the dataset is pre-annotated with 5 captions. 6000, 1000, and 1000 images were used in training, testing, and validation in the Flickr 8k dataset. On the contrary 28000, 1000, and 1000 images were used in the Flickr 30k dataset for training testing and validation while in MSCOCO 82k 40.5k and 40.7k images were used in training testing and validation. BLEU gave highest score on the MSCOCO dataset which was 0.67257.

`

In the 3rd paper, the author used the same generalized method of generating image captions with and without using the attention mechanism and compared the results of training images of Flickr 8k and 30k datasets. The author used single as well as combined image feature extraction models and checked the results on the given datasets. The datasets were split into training testing and validation sets. VGG-16 feature extraction model will give a one-dimensional feature vector with 4096 vector lengths and inceptionv3 will give a feature vector of length 2048. The performance of the model is being noted using BLEU score. It was observed that the model which uses attention

has a high BLEU score as compared to the models which did not use any attention mechanism.

In the 4th paper, author has proposed a transfer learning method for generating captions of the images. First of all the image is converted into vector form with the help of an encoder. Here GRU is taken as a decoder. Cross entropy loss is calculated. For better results, the Rms prop optimizer is used instead of Adams. In the proposed method attention mechanism is also added for better results. The author compared the results with and without using the attention model. InceptionV3 and inception-ResnetV2 are used as the CNN model. MS-COCO dataset is used and the images are trained at 2, 5, and 10 epochs respectively to compare and check the results.

In the 5th paper, the author proposed an image caption generator that uses a CNN model inception V3because it is cheap as compared to other CNN models. MS COCO dataset is used which has 82000 images that are unique with each image having 5 captions. 6000 images were used for training. The captions are broken down into words to form a dictionary and this process is called tokenization. Attention model is used so that it looks only at important features in the image

`

and generates captions. To generate captions GRU was used as a decoder. Other RNNs e.g. vanilla RNN is not used due to the vanishing gradient problems. GRU is used as a decoder because it is less complex and fast.

In the 6th paper, the author proposed a novel image captioning generator model that can do caption in Urdu. Flickr 8k dataset was used and 1k images of man dataset were taken from it and they were annotated manually in Urdu for training purposes. Each image was given 5 captions. They used resnet-101 as an encoder for feature extraction and they used attention before gated recurrent unit so that it only focuses on the important details of the image and then Gru will generate captions of the input image. As the dataset is small i.e. 1k images so gru is most suitable for this as it is fast than lstm. Finally they compared there results with other encoders like inception-v3. The results
were evaluated on BLEU scores.

`

Chapter 3

Materials and Methods

3.1 Overview

3.2 Dataset Creation

As there is no available dataset in Urdu. The only paper which has worked on Urdu image captioning have manually annotated 700 images from flickr8k dataset in Urdu. I will extend this work by taking 1500 images more and manually annotate them from flickr8k dataset**.**

3.2.1 Previous Datasets

There are several popular datasets which are used in image captioning such as flickr8k flickr30k and MS-COCO datasets. Flickr8k has 8000 images, flickr30k has 30000 images and MS-COCO dataset has 80,000 images respectively.

3.2.2 New Dataset

I have manually annotated 1500 images with 5 each image having 5 captions each and used it for my project. We have to create a vocabulary of distinct words of Urdu from all the captions.

3.2.3 Evaluation Datasets

I have manually annotated 1500 images with 5 each image having 5 captions each and used it for my project. We have to create a vocabulary of distinct words of Urdu from all the captions.

`

3.3 Methodology

Below is a block diagram of a simple image caption generator. It consists of an encoder which is a convolutional neural network. It is responsible for feature extraction of the input image. feature extraction means reducing the input image size by considering only the important parts of the image which can tell us the most details about the image. the final layer of CNN is a flattening layer that will generate a vector of 1D which will be Moreover we have to convert the input image to a vector so that it can be fed to RNN for generating captions of the image.

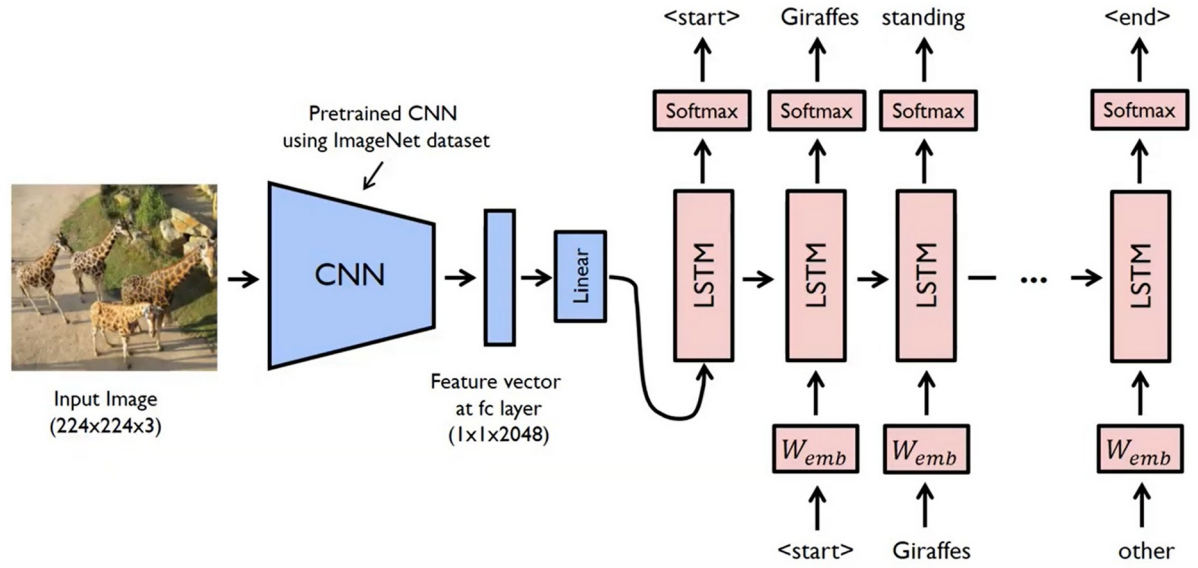


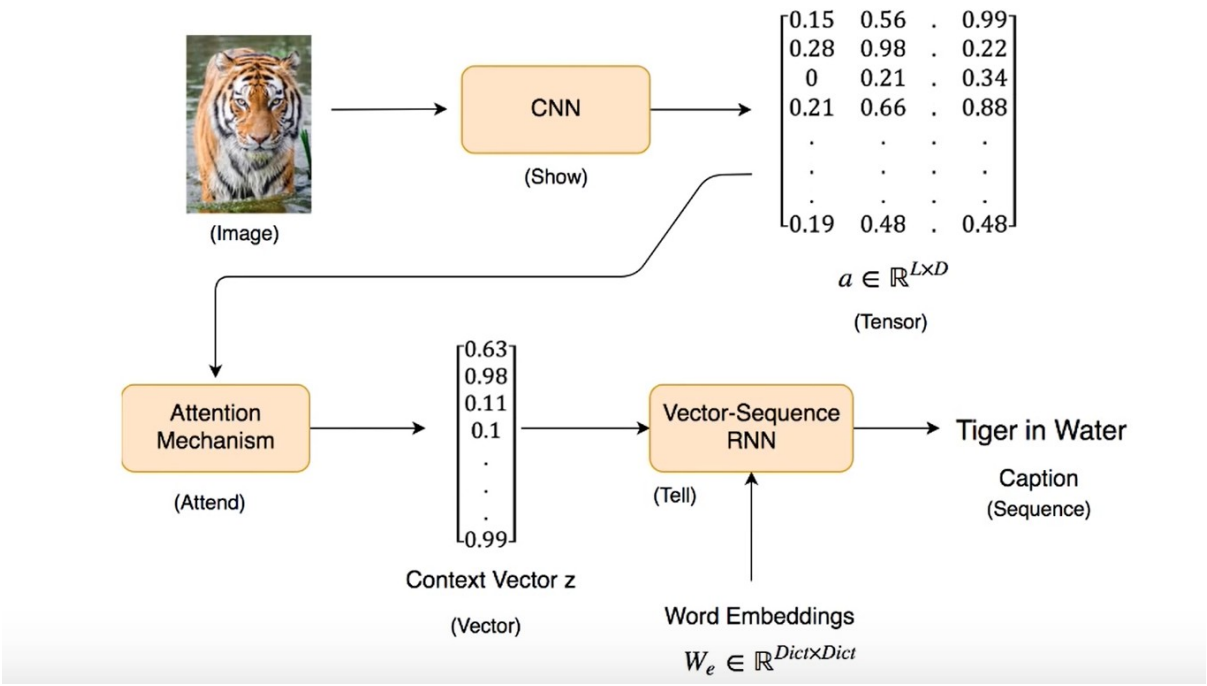
Image caption generator

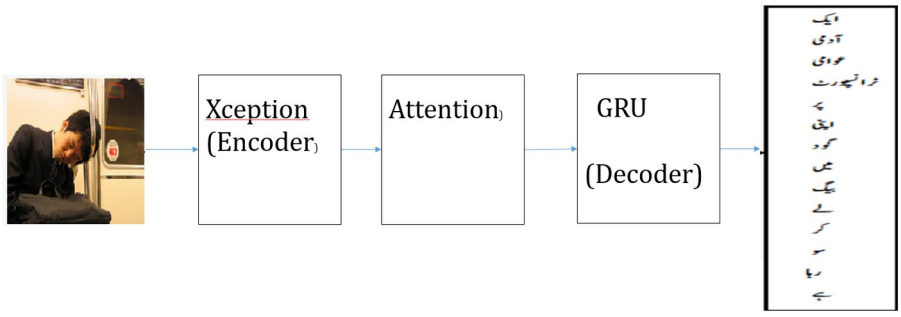
3.3.1 Proposed Model

The proposed model is a bit different from the above one. We will use the xception CNN model for feature extraction. Xception is used because its performance is better than inception, vgg, resnet, etc. because it is tested on the image net dataset where it gives more

`

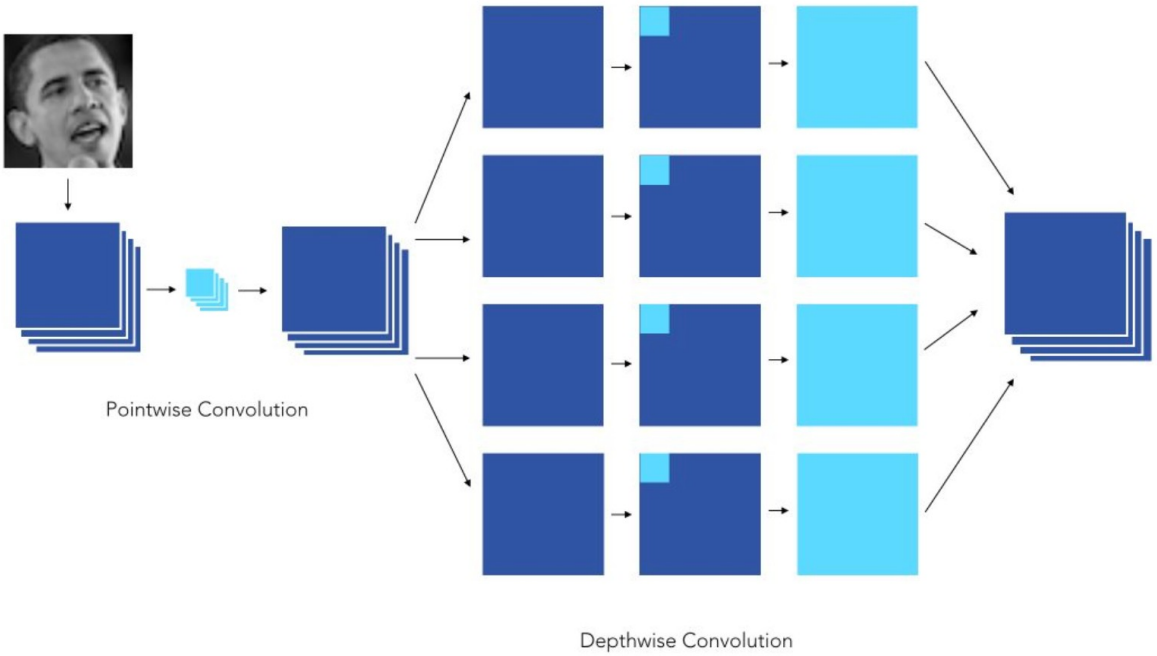
accuracy as compared to vgg-16, resnet152, and inception. The xception is an improved form of inception v3. In xception architecture depth-wise, separable convolution is modified to pointwise convolution followed by depth-wise convolution. The CNN will convert the input image to a feature vector and then it go to the attention model. Attention model is introduced so that it captures only the important details of the image and then pass it to a decoder which is GRU. The purpose of selecting GRU instead of lstm is that as the dataset is small so gru works faster than lstm that's why we choose gru instead of other rnn architecture. Gru will then decode the feature vector coming from attention and generate captions.





Proposed Image caption generator

Xception architecture

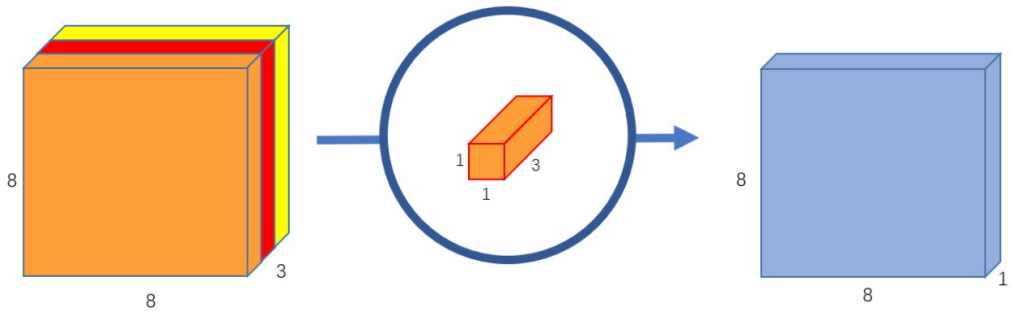


From the above figure, we can see that in xception pointwise convolution is followed by a depth-wise convolution.

Point-wise convolution

In pointwise convolution the filter is of size 1x1 with no of channels equal to the channels of the input image, It is then multiplied with the input image .

`



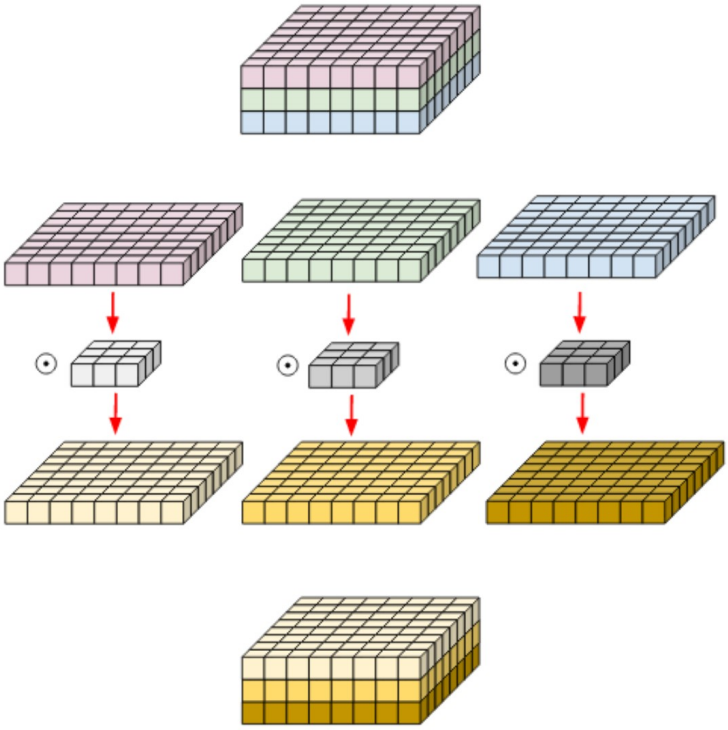
Point wise convolution

Depth wise convolution

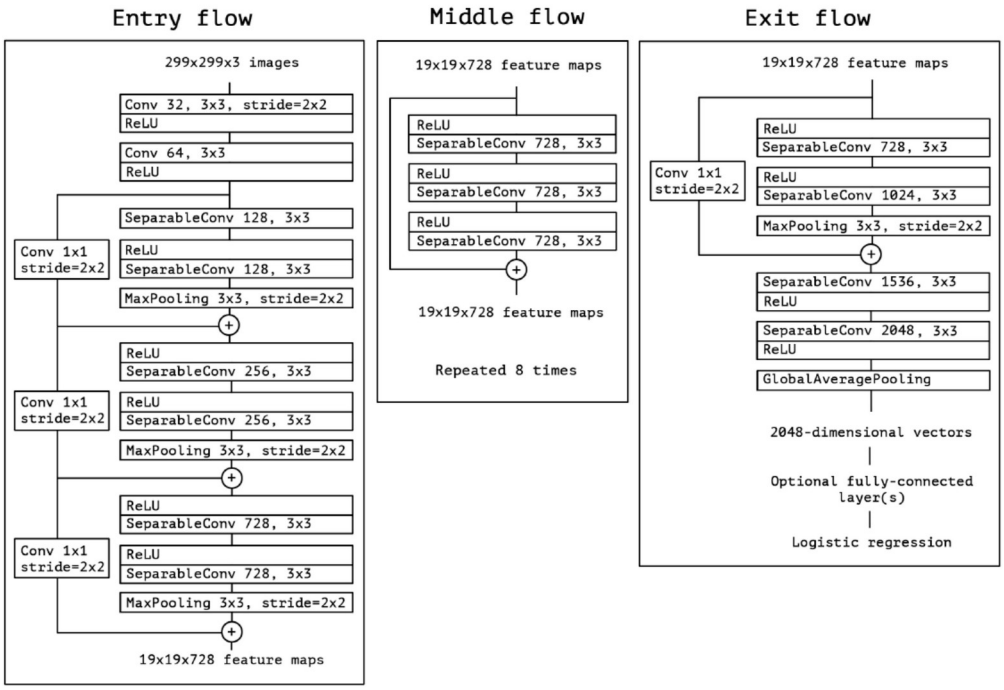
In depth wise convolution input image is broken down into its no of channels and similarly the filter is also broken down into its no of channels

For eg in case of a 6x6x3 image and a filter of 3x3x3. The image will be split into single channels as 6x6x1 and 6x6x1 and 6x6x1 and filter will also be split into 3x3x1, 3x3x1, 3x3x1 respectively. Now we will do convolution and convolve each single channel filter with single-channel image respectively one by one. After convolution, all the resulting feature maps will be stacked one by one over each other. It can also be seen in the figure below.

`



Depth wise convolution



`

Xception architecture

Chapter 4

Future Work & Conclusion

In this paper, we have proposed an Urdu image caption generator with Xception as a Convolutional Neural Network and a Gated recurrent unit as a decoder which will aid in generating captions. Lots of work has been done in image captioning but in Urdu language there is a lot of gap and work needs to be done in this language also. We used the flickr8k dataset and took 1500 images from it and manually annotated them. Xception was used because it gives good results than other CNN such vggnet , resnet and inceptionv3. In future we can extend this work by using a bigger dataset and we can also do some work on urdu grammar which can help in improving the results.

`

REFERENCES

Wang, C., Yang, H., Bartz, C., & Meinel, C. (2016, October). Image captioning with deep bidirectional LSTMs. In *Proceedings of the 24th ACM international conference on Multimedia* (pp. 988-997).

Amritkar, C., & Jabade, V. (2018, August). Image caption generation using deep learning technique. In *2018 fourth international conference on computing communication control and automation (ICCUBEA)* (pp. 1-4). IEEE.

Dang, T. X., Oh, A., Na, I. S., & Kim, S. H. (2019, January). The role of attention mechanism and multi-feature in image captioning. In *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing* (pp. 170-174).

Kesavan, V., Muley, V., & Kolhekar, M. (2019, October). Deep learning based automatic image caption generation. In *2019 Global Conference for Advancement in Technology (GCAT)* (pp. 1-6). IEEE.

Agrawal, V., Dhekane, S., Tuniya, N., & Vyas, V. (2021, July). Image Caption Generator Using Attention Mechanism. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

Ilahi, I., Zia, H. M. A., Ahsan, M. A., Tabassam, R., & Ahmed, A. (2020). Efficient Urdu Caption Generation using Attention based LSTM. *arXiv preprint arXiv:2008.01663*.

`