Name: Fahad Ahmed Pranto

Id: 1612607042

Paper name: **TextMage: The Automated Bangla Caption Generator Based On Deep Learning.**

**Abstract**:

It is a significant task to produce text from a given image that contains a mixture of both computer vision and natural language processing sectors in order to interpret and describe a picture using a natural language. Fortunately, much of the current work has been performed on a single lingual context and on the same data collection. TextMage is a system capable of recognizing visual scenes belonging to the spatial context of Bangladesh and using its insight to reflects what it knows in Bengali. Therefore, they have trained a model with BanglaLekhaImageCaptions on their previously created and released dataset. This dataset includes 9,154, Images for each file, together with two descriptions.

**Introduction:**

In the area of intelligent machines, the generation of annotations for images automatically by combining computer vision and NLP is commonly known as the Image Captioning method. Automated image captioning through deep learning has gained considerable attention in recent years, resulting in the creation of different models capable of producing captions for images in various languages. Previous research focused on approaches that break the task into two sub-tasks by putting the findings together at the end, and fixed the problem. Throughout this paper, they developed an automated image definition scheme that follows the traditional method of extracting features from an image using a CNN and serving its previous output layer to an RNN producing language consisting of LSTM cells to generate a Bangla caption that explain the image.

**Literature review:**

They found that the majority of current systems, both contextual and lingual, have a Western bias. This paper's important aims and objectives are to establish a dataset that removes the current issue of bias.

**Methods:**

Firstly, the resizing part of our dataset is addressed. The entire dataset will be resized to 224 x 224 x 3 pixels. They used the features of VGG-16 where the pre-trained class of 1000 for the CNN pre-exists. Secondly, they were worried about the RNN and LSTM, typically used from given data images to establish textual definition. After this, to produce an error and reliability graph, both the CNN and RNN models were stapled and trained on the dataset again for 35 epochs.

**Dataset:**

In this project, the dataset used is named BanglaLekhaImageCaptions. There are 9,154 images in the dataset and all images are held in the same format. For each image in this dataset, they provided two descriptions in Bangla.

**Results:**

RNN[17], Flickr8K & Google NIC[2], and LRCN[11] models are based on the Flickr8K, Flickr30K, and MS COCO dataset study results showing the baseline outcome for that dataset. In addition, because they introduced a new task, they matched their benchmark outcome with the mentioned current work. Their model can create dynamic yield as our model figures out how to regulate the extent of the area and word inserting. In the CNN section of our dataset,they achieved remarkable results (0.758565) for the testing period and for the validation of 0.643476).

**Conclusion:**

In this paper, they have presented an automated image captioning system, TextMage, which can interpret and explain an image with a South Asian bias in Bangla. The model developed for TextMage was highly motivated by the first mutual model Show and Tell: A Neural Image Caption Generator model" from the perspective of architecture."