**Paper:** **(Oboyob) A sequential- semantic Bengali image captioning engine.**

Name: Milon Hossain

Id: 1610497042

**Abstract:** An active and demanding research subject in computer vision and natural language processing is understanding the meaning of the production of textual definition from an input image. In this article, by subsampling the computer translated data collection, they discuss a standard method for Bengali image caption generation. Later, for the state-of-art CNN-LSTM architecture-based models, they used many preprocessing approaches. The experiment is performed on the normal Flicker -8k dataset, along with some changes to fit it in the Bengali language. For more experimentation with 16 distinct models built in the entire training process, the training caption subsampled dataset is computed for both Bengali and English languages. With regards to multiple caption performance criteria, the qualified models for both languages are evaluated.

**Introduction:** To summarize the content of a picture in a way that tells the story without delving into unimportant information, an articulate picture definition is necessary. Text representation can help visually disabled individuals draw a realistic representation of an image in question. Multiple explanations can give an audience more viewpoints of how an image is viewed. These forms of sentence representation of an image are known as image captioning which deals with mainly two challenges. The first challenge is to identify the objects and second one is to create a correlation among the objects and sentence level description in the domain of natural language processing.

**Methodology:** An important approach is the top-down approaches which elects all the visual information through Recurrent Neural Network (RNN). The dramatic revolution in deep learning is incorporated by the Convolutional Neural Network (CNN) with RNN where CNN collects pictorial information from the image and RNN decodes into natural languages through incorporating vocabulary wise embedding from the language model and LSTM layers are used for contextual Bengali language. For evaluation a caption quality for more approaches “BLEU”, “METEOR”, “ROGUE” and “CIDER” also used here.

**Dataset:** For the entire experiment, Flicker -8k dataset is employed, which consists of total 8092 images, taken from Flicker. Corresponding image ids are split into training, validation and testing where 6000 are used for training, 1000 used for validation and 1000 for testing. Each image consists of 5 human annotated ground truth captions associated, resulting in 40,460 total sentences.

**Conclusion:** There are many systems for English language caption generating system. But this is the second search engine for Bengali languages which can generate caption when we will input an image inside the model. This system is little bit different from the others. Different approaches also have been discussed here as well.

**Reference:**

1. https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs179351