Name: Fahad Ahmed Pranto

Id: 1612607042

Paper name: Automatic Myanmar Image Captioning using CNN and LSTM-Based Language Model.

**Abstract:**

The framework for image captioning contains computer vision modules as well as natural language processing. The computer vision module is designed to identify salient objects or extract image characteristics, whereas the Natural Language Processing (NLP) module is designed to produce proper syntactic image captions. There really is no Myanmar language picture caption database. As part of the Flickr8k dataset in this ongoing project, the Myanmar image caption corpus is manually created. In addition, specifically for Myanmar image captioning, a generative merging model based on Convolutionary Neural Network (CNN) and Long-Short Term Memory (LSTM) is used.

**Introduction:**

A picture consists of various data, such as objects, features, scenes, and activities. With much less complexity, humans are able to create captions for pictures. Two tasks are involved in automatic image caption generation: identifying and understanding important objects in an image and explaining the correct link between these things. Image captioning represents a collection of two sub network, CNN for image detection of salient objects and LSTM for explaining relationship objects and decoding into phrases to perform these two tasks. The language of Myanmar is morphologically complex and has a shortage of annotated resources compared to English. Consequently, a corpus that is large enough to get the exact caption for Myanmar automatic image needs to be built.

**Methods:**

In this paper, they used CNN-LSTM based architecture. Two primary components form the architecture. The first is a framework for understanding images using CNN, and the second is a module for understanding text using LSTM. For image caption generation task, CNN is widely used because it has solved successfully for image annotation problems with high accuracy ( Aditya et al., 2019). Words or phrases are generated by the text comprehension portion based on the word embedding vector of the previous component. After all photos and all previous words have been seen, the language generation model is learned to determine each word in the description. For any specified sentence they add two additional signs for the start word and end word in Myanmar corpus which is signify the beginning and end of the sentence.

**Dataset:**

The flikr8k dataset is applied in the first Myanmar Image captioning. It consist 8092 images and five annotated English captions for every image. They picked only 3k images from the Flickr8k dataset with five annotated Myanmar captions for each image because of the limited time.

**Results:**

VGG16 and VGG19 produced captions are roughly identical and do not have any qualitative distinction. They focused primarily on the results produced with LSTM from VGG16. In their output, the significant features of the images are captured accurately and grammatically correct. Ultimately, it is the limits of their system and they want to emphasize the need for further work on the model.

**Conclusion:**

They developed the first image captioning corpus for the language of Myanmar, and manually reviewed and designed the explanations to complement captions and photographs in detail. The experimental results showed that the application of images captioning based on CNN and LSTM trained on their corpus will provide reasonable results. For Myanmar Image Captioning, such tiny corpus can help establish big corpora.