**Automatic Myanmar Image Captioning using CNN and LSTM-Based Language**

**Model**

An image consists of several information such as the objects, attributes, scenes and activities. Humans are capable of generating captions for images with much less difficulty. However, automatic caption generation for a given image is a very challenging task for machine (Yang et al., 2018). Automatic image caption generation involves two tasks: 1) recognizing and understanding significant objects in an image and 2) describing the proper relationship between these objects. To perform these two tasks, image captioning uses a combination of two subnetworks, CNN for salient object detection in images and LSTM for understanding relationship objects and decoding into sentences.

The image caption generation is mainly split in retrievablebased approaches and constructive-based approaches. The first category is used in the earlier attempts to solve image captioning which has the problem as a retrieval task. A database is constructed based on image features extraction and caption generation for given images and then the most appropriate sentence is extracted (Jacob et al., 2015). This approach is not effective to describe novel captions and the caption generation is restricted to the features size of the images and the database size.

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). We have trained and tested two different models for feature extraction of images datasets. The two models have different capabilities in extracting features of images and the input image size of both models are 224× 224× 3 and the convolutional feature size of VGG is 4096.

The Flickr8k1 dataset (Khumaisu et al., 2018 ; Micah et al., 2013) is applied in the first Myanmar Image Captioning task. It contains 8092 images and five annotated English captions for each image. Due to the limited time, we selected only 3k images of the Flickr8k dataset with five annotated Myanmar captions for each image. We constructed Myanmar image captions corpus in two different ways: 1) Automatic translation from English descriptions and 2) Direct image descriptions with Myanmar language.

We created the first corpus of image captioning for Myanmar language, and manually checked and built the descriptions in detail to match captions and images. Convolutional Neural Network based on Visual Geometry Group (VGG) OxfordNet CNN and single hidden layer LSTM model were applied for Myanmar automatic image caption generation in this work. The experimental results showed that applying CNN and LSTM based image captioning trained on our corpus can give acceptable performance.