Comparison of NASNetLarge and NASNetMobile as Encoder for Myanmar Image Caption Generation

San Pa Pa Aung
Natural Language Processing Lab
University of Computer
Studies,Yangon
Yangon,Myanmar
sanpapaaung@ucsy.edu.mm

Win Pa Pa
Natural Language Processing Lab
University of Computer
Studies,Yangon
Yangon,Myanmar
winpapa@ucsy.edu.mm

***Abstract*— The paper aims to generate the automatic Myanmar image description by recognizing the contents of images. Representing the information of an image is a complicated operation for machine. The combination of Computer Vision and Natural Language Processing are commonly utilized for tackling this issue. In this paper, we compared two popular convolutional network architecture-NASNetLarge and NASNetMobile as encoder for the same Myanmar image captioning system to investigate which approach is the best at feature extraction applied for caption generation in Myanmar language. The experiments are done on Myanmar image captions corpus that has 40460 sentences for 8k images and the system performance is measured based on the BLEU [4] scores.**

Keywords—NASNetLarget, NASNetMobile, Gated Recurrent Unit

I. INTRODUCTION

Image generation is a very critical research domain on a border between Computer Vision and Natural Language Processing. In order to generate precisely captions must comprehend not only what objects are existed in an image, but also connection between them. The image generation system can be utilized in a wide range of applications area. Nowadays, state-of-the-art techniques for image generation are particularly depend on an encoder-decoder framework. The aim of an encoder module is to convert the vector describing for input photo. The ambition of a decoder module is to produce a series of segmented caption for an image using previous features vector. An image consists of a huge number of contents. Currently, large amount of image is produced on social platform. Deep learning technique can be applied to spontaneously caption these images, therefore can institute where manual captioning performed. Furthermore, image description generation task can be designed as a machine translation task that transform image to a sequence of words in isolated language [4] [2]. In this work, the training is done on two different Myanmar image captions corpus [3] such word segmented corpus and syllable segmented corpus.

II. METHODOLOGY

In this paper, NASNetLarge [1] and NASNetMobile models of CNN are applied as encoder to examine the results achieved from each model. The vector including the output of the fully connected layer in individual model is fed to Gated Recurrent Unit (GRU) [5] as decoder to produce automatic descriptions in Myanmar language for any given image. The features vector is specified to be 4032 elements and the image pixels size is 331x331 for NASNetLarge. The 1056 elements feature vectors is specified and the image pixels size is 224x224 for NASNetMobile. The output layer of feature extraction models is eliminated because of the classification of the image is not needed. The main function of the GRU language model is for predicting image description in Myanmar language based on the previous feature vectors and entire vocabulary in our corpus.

III. RESULTS AND DISCUSSION

We applied four different hidden layer size (HLS) such as 128, 256, 512 and 1024, the comparison was done for both NASNetLarge with GRU and NASNetMobile with GRU on our two distinct Myanmar image captions corpus [3]. All investigations are run on NVIDIA Tesla K80 GPU. For the result of Table I, we found that the hidden layer size 1024 obtained highest BLEU-1 score of 64.79%, BLEU-2 score of 49.57%, BLEU-3 score of 40.88% and BLEU-4 score of 27.73 using NASNetLarge with GRU on word segmented corpus. In Table II, the hidden layer size 1024 achieved the highest BLEU-1 score of 69.88%, BLEU-2 score of 58.63, BLEU-3 score of 52.37% and BLEU-4 score of 39.79 using NASNetLarge with GRU on syllable segmented corpus compared with other hidden layer sizes.

TABLE I. VARIOUS HIDDEN LAYER SIZE USING NASNETLARGE WITH GRU ON WORD SEGMENTATION

HLS	BLEU-1	BLEU-2	BLEU-3	BLEU-4
128	63.88	48.49	39.63	26.26
256	64.67	49.69	41.17	27.72
512	64.21	49.06	40.38	27.01
1024	**64.79**	**49.57**	**40.88**	**27.73**

TABLE II. VARIOUS HIDDEN LAYER SIZE USING NASNETLARGE WITH GRU ON SYLLABLE SEGMENTATION

HLS	BLEU-1	BLEU-2	BLEU-3	BLEU-4
128	69.37	57.79	51.33	38.91
256	70.36	58.33	51.99	39.67
512	70.46	58.95	52.32	39.62
1024	**69.88**	**58.63**	**52.37**	**39.79**

TABLE III. VARIOUS HIDDEN LAYER SIZE USING NASNETMOBILE WITH GRU ON WORD SEGMENTATION

HLS	BLEU-1	BLEU-2	BLEU-3	BLEU-4
128	**64.39**	**48.47**	**39.37**	**26.14**
256	61.23	45.27	37.23	24.64
512	62.8	47.01	38.17	24.98
1024	62.53	46.67	37.93	25.12

TABLE IV. VARIOUS HIDDEN LAYER SIZE USING NASNETMOBILE WITH GRU ON SYLLABLE SEGMENTATION

HLS	BLEU-1	BLEU-2	BLEU-3	BLEU-4
128	66.82	54.68	48.13	35.62
256	67.96	56.20	49.85	37.02
512	**70.32**	**58.12**	**51.27**	**38.51**
1024	65.42	53.71	47.31	34.56

The more hidden layer size was set, the more training time will take, sometime we achieved the high performance.

In discussion of Table III, hidden layer size 128 obtained the highest BLEU-1 score of 64.39%, BLEU-2 score of 48.47, BLEU-3 score of 39.37% and BLEU-4 score of 26.14 using NASNetMobile with GRU on word segmented corpus compare to other hidden layer sizes. As can be seen in Table IV, hidden layer size 512 obtained the highest BLEU-1 score of 70.32%, BLEU-2 score of 58.12, BLEU-3 score of 51.27% and BLEU-4 score of 38.51 using NASNetMobile with GRU on syllable segmented corpus compare to other hidden layer sizes. According to the Table I, Table II, Table III and Table IV, NASNetLarge feature extraction model performs better than NASNetMobile feature extraction model as encoder. The best result is obtained from the combination of NASNetLarge and GRU on syllable segmented corpus. According to these various experimental results, we cannot be defined accurately which hidden layer size is good for the system. It can vary depended on the use of models and size of the dataset.

A. Experiments Results

In this section, we mainly focused on the predicted captions generated from NASNetLarge model as encoder and GRU network as decoder on word segmented corpus. Fig. 1 (a), the model can predict the gender accurately like “လူ တစ်ယောက်” (“A person”) and action “ရေ လှိုင်းစီး နေတယ်” (“is surfing the wave”). In Fig.1 (b), the generated caption quite effectively to identify the object gender and action such as “အမျိုးသား” (“A man”) and “တင်းနစ် ကစား နေတယ်” (“is playing tennis”).

(a) In English: A Person is surfing the wave

(b) In English: A man is playing tennis

Fig. 1. Generated Myanmar captions Using NASNetLarge and GRU

IV. CONCLUSION

In this work, two distinct image encoder such as NASNetLarge and NASNetMobile are compared for a Myanmar image captioning by measuring BLEU scores on two distinct segmented corpora. According to the results showed in Table I, Table II, Table III and Table IV, it can be summarized that NASNetLarge attained better evaluation results than NASNetMobile as an encoder for the Myanmar image description generation. Based on the evaluation results, we can be described that encoder performed a special role in image captioning and can be quite made better model without alternating a decoder structure.

REFERENCES

[1] B. Zoph and Q. V. Le, “Neural Architecture Search with Reinforcement Learning”, In International Conference on Learning Representations, 2017.

[2] H. Parikh, H. Sawant, B. Parmar, R. Shah, S. Chapaneri and D. Jayaswal, “Encoder-Decoder Architecture for Image Caption Generation”, 3rd International Conference on Communication System, Computing and IT Applications (CSCITA), 2020.

[3] S. P. P. Aung, W. P. Pa and T. L. Nwe, “Automatic Myanmar Image Captioning using CNN and LSTM-Based Language Model”, Proceeding of the 1st Joint SLTU and CCURL Workshop (SLTUCCURL 2020), pp. 139–143, 2020.

[4] V. Atliha and D. Sesok, "Comparison of VGG and ResNet used as Encoders for Image Captioning", 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 2020.

[5] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”, arXiv preprint arXiv:1406.1078, 2014.