Improving Myanmar Image Caption Generation Using NASNetLarge and Bi-directional LSTM

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

Tin Lay Nwe
Visual Intelligence Department
Institute for Infocomm Research
Singapore
tlnma@i2r.a-star.edu.sg

***Abstract*— The main objective of this paper is to improve the automatic Myanmar captions by learning the contents of images using NASNetLarge and Bi-LSTM model. Describing the contents of an image is a complex task for machine without human intervention. Computer Vision and Natural Language Processing are widely used to tackle this problem. This paper proposed a deep learning-based Myanmar image captioning system which used a NASNetLarge feature extraction model of CNN as an encoder and a deep Recurrent Neural Network (RNN) with Bi-directional Long Short-Term Memory (LSTM) as a decoder. For corpus construction, we created and annotated the Myanmar image captions corpus (consists of over 40k Myanmar sentences), which is based on Flickr8k dataset. Furthermore, two different types of segmentations such as word segmentation level and syllable segmentation level are studied in text preprocessing step. In this work, the proposed Bi-directional LSTM model is compared with LSTM, GRU as well as the baseline model. Experiments on the updated dataset is presented that all of our models using syllable segmentation give higher and comparable BLEU scores than word segmentation for Myanmar image captioning system. NASNetLarge with Bi-directional LSTM model using syllable segmentation approach achieved the highest BLEU-4 score 40.05% which is 12.5% better than word segmentation in this work and 15.67% BLEU-4 score better than our previous work.**

Keywords—NASNetLarget, Recurrent Neural Network, Long Short-Term Memory, Gated Recurrent Unit

I. INTRODUCTION

Automatic description generation connects Computer Vision and Natural Language Processing which are two major fields in Artificial Intelligence. In order to produce qualitative image description must understand not only what objects are presented in an image, but also relationships between them. The image description generation system can be applied in a wide range of practical tasks, such as image search on the web, human-computer interaction, social media platform, help to visually-impaired people and several other natural language processing applications [6].

Today, state of the art models for image description generations are usually based on an encoder-decoder framework. The goal of an encoder part is to transform the vector representing for given images. The objective of a decoder part is to predict a sequence of token generating an image utilizing that features vector [7]. An image stored a large amount of information. Every day, a lot of image data is generated on social media. Deep learning methods can be used to automatically annotate these images, thus can be replaced in place of manual annotations done. This will be greatly reduced the human error as well as the efforts. Moreover, image caption generation task can be considered as a machine translation task that is used to transform from input image to a sequence of words with some language.

For Myanmar language, VGG16 and LSTM-based language model (around 15k Myanmar sentences) is found publicly [5]. However, any segmenter is not used for text preprocessing step in this publication. Therefore, in this work, we developed the more Myanmar image captions sentences (over 40k sentences) and also investigated how segmentation level affects the Myanmar image captioning system performance. More than that, different deep learning models such as NASNetLarge with GRU, NASNetLarge with LSTM, VGG16 with Bi-LSTM and NASNetLarge with Bi-LSTM are compared with baseline model[5]. The experimental results reveal that syllable level segmentation can give significantly better performance for Myanmar image description compared with the result of previous published work in [5] and state of the art model[20].The key contributions of this paper are described as follows:

1) NASNetLarge feature extraction model and a multi-task Bi-LSTM language generation model are applied for Myanmar image caption generation, that can accurately identify the objects in the images and also generate grammatically correct sentences with relevant images.

2) The proposed model is evaluated on Myanmar image captions corpus, which has over fourty-thousand captions. The system performance significantly improved over previous work. The proposed model achieved an accuracy of 27.55% BLEU-4 score on word level and 40.05% on syllabel level compared with the baseline model of 24.38% BLEU-4 score.

The rest of paper is organized as follows, existing work flow of image captioning system is described in section II. Section III describes the encoder-decoder framework and the process flow of our system is explained in section IV. Section V reports details of dataset and experimental setups. Result and analysis are presented in section VI. Conclusion and future improvements are concluded in section VII.

II. RELATED WORK

In the literature, there exists two types of approaches for generation image captioning task, the first one is the top-down approach [11][15] and the second one is the bottom-up approach [10].

The top-down approach starts from the input image and converts it into words while the bottom up approach first comes up with words which describe the various aspects of an image and combines them to generate the description. In the top-down approach, there is an end to end formulation from an image to sentence using the recurrent neural network, and

all the parameters of the recurrent neural network are learned from training data. Limitations of the top-down approach are that sometimes fine details cannot be obtained which may be important to generate the description of an image. Bottom-up approaches do not have this problems they can be operated on any image resolution, but there is a lack of end to end formulation which is going from image to sentence.

Bottom-up approach start with visual concepts, objects, attributes, words and phrases which are combined using the language models. Farhadi et al.[10] defined a method that can compute a score linking to an image. This score is used to attach a description to an image.

Top-down approaches are the recent ones; in these approaches, image captioning problem is formulated as machine translation problem. In machine translation, one language is translated to another one, but in case of image captioning, the visual representation is translated to language. Bahdanau et al.[11] developed the encoder-decoder architecture for machine translation by allowing a model to automatically search the part of the source sentence that is relevant in predicting the target word.

Cheng et al. [20] applied VGG16 and Alexnet as an encoder and Bi-LSTM model as a decoder on three benchmark datasets: Flickr8k, Flickr30k and MSCOCO datasets. Alexnet visual model is less powerful than VGG16. The authors achieved the highest BLEU-N (N=1, 2, 3, 4) scores 65.5%, 46.8%, 32.0% and 21.5% respectively using VGG16 with Bi-LSTM model on Flickr8k dataset.

In this work, top-down approaches is used to solve the image captioning problem. For encoding part, NASNetLarge feature extraction model is applied to understand the contents of the images. GRU, LSTM and Bi-LSTM models are applied to generate the description with Myanmar language for decoding part. Furthermore, various experiments are investigated to know which approaches can give better performance for our system compared with the previous work.

III. METHODOLOGY

In this system, the pre-trained NASNetLarge model is used as encoder to compare the results obtained from each of them. The vector containing the output of the fully connected layer in each model is passed to language generation models as decoder to generate automatic Myanmar image captions for any given image.

A. Feature Extraction Model

The Neural Architecture Search (NAS) framework is proposed by [3] used to find good convolutional architectures on a dataset of interest. To the best of our knowledge, this is the first pre-trained feature extraction model used for Myanmar image descriptions.

NASNetLarge: NASNetLarge is a pre-trained feature extraction model of Convolutional Neural Network (CNN). The model is trained on more than a million images from the ImageNet database. The network can classify images into 1000 object categories. As a result, the network has learned rich features representations for a wide range of images. The features vectors of input images are defined to be 4032 elements and the default input image size is 331x331 for NASNetLarge and then processed by a Dense layer to produce a 256 elements representation of the photo [2]. The output of last layer before the softmax function was removed to avoid classification of the image. This pre-trained model is used to extract the features vector of images within the dataset and saving the transfer-values in a pickle file so that we can be reloaded quickly for further computation [4].

B. Gated Recurrent Unit (GRU)

GRU was introduced by Kyunghyun Cho et al. [12] and applied successfully for machine translation and sequence generation. GRU is an improved version of the Recurrent Neural Network that is used to handle the vanishing gradient problem of standard RNN, it consists of update gate and reset gate. The behavior of the cell is controlled by these two gates.

The core of the GRU model is memory cell which stores knowledge of every time step. Basically, the update gate and reset gate are the vectors responsible for deciding which information should be passed to the output. These two gates are trained to store the information from the previous time steps without wasting those as well as removing the pieces of information which are irrelevant for the prediction [13][17] [18].

C. Long Short-Term Memory (LSTM)

Firstly, LSTM model is briefly introduced which is the center of Bi-LSTM model. The main motivation of applied LSTM network is that it designed to handle long-term dependencies and avoid quick exploding and vanishing gradient problems that suffers from traditional RNN during back propagation optimization. LSTM network consists of three gates mechanisms: input gate, forget gate and output gate. LSTM network takes the inputs from various sources: current input x_t, the previous hidden state of all LSTM units h_{t-1} as well as previous memory cell state c_{t-1} at given time step t. At time step t, the updating of those gates for given inputs x_t, h_{t-1} and c_{t-1} as follows:

$$\text{Input gates: } i_t = \sigma(W_{xi}\, x_t + W_{hi}\, h_{t-1} + b_i) \quad (1)$$

$$\text{Forget gates: } f_t = \sigma(W_{xf}\, x_t + W_{hf}\, h_{t-1} + b_f) \quad (2)$$

$$\text{Output gates: } o_t = \sigma(W_{xo}\, x_t + W_{ho}\, h_{t-1} + b_o) \quad (3)$$

$$g_t = \emptyset(W_{xc}\, x_t + W_{hc}\, h_{t-1} + b_c) \quad (4)$$

$$\text{Cell states: } c_t = f_t \,\Theta\, c_{t-1} + i_t \,\Theta\, g_t \quad (5)$$

$$\text{Cell output: } h_t = o_t \,\Theta\, \emptyset(c_t) \quad (6)$$

Where W is the weight matrices learned from the network and b are bias vectors. σ is the sigmoid activation function, $\emptyset$ presents hyperbolic tangent and Θ denotes the products with gate values [16][17][19].

D. Bidirectional Long Short-Term Memory (Bi-LSTM)

Bi-LSTM network consists of an input layer, two hidden layers and an output layer.

1) Input Layer: During the training phrase, the input layer takes the pre-segmented words in our corpus and their conrresponding image features from the previous feature extraction model. Word embedding layer transformed each word in the image captions sentences into one-hot encoded format. After that, the word embedding vector is the input parameters for the Bi-LSTM neural network model.

2) Hidden Layer: The hidden layer consists of two separate LSTM networks - forward and backward, connecting to the same output layer. During training, both the

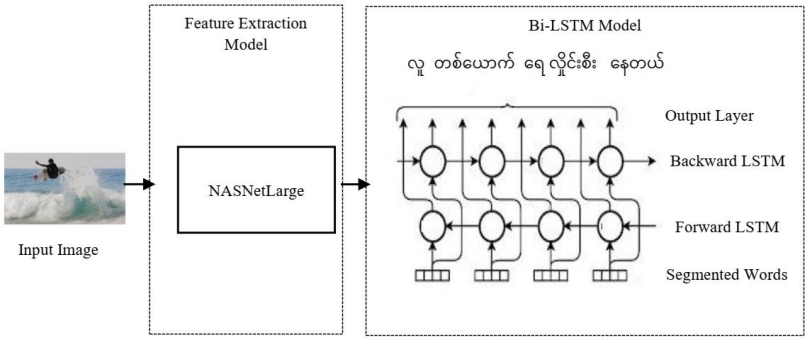

Fig. 1. An Illustration of Bi-LSTM Image Captioning Model

forward hidden sequences $\vec{h}_t = (\vec{h}_1, \vec{h}_2, \ldots..\vec{h}_k)$ and backward hidden sequences $\overleftarrow{h}_t = (\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots..\overleftarrow{h}_k)$ use the same sequences of word vectors coming from the input layer to set the parameters of the system to accurately predict captions. The concatenation of forward and backward layer constructed the final encoded hidden vector .

3) Output Layer: The output layer or dense layer picks the appropriate words based on the sequences of data from both hidden layers using a softmax activation function, which is effective in dealing with multiclassification and probability distribution problems. The output of this function is in the form of one-hot encoded word which is then converted back to word form in a high-level representation for image captions [16][17][20].

E. Beam Search

Beam search is the best-first search algorithm which presents the final step to generate a caption with the highest likelihood of occurrence given the input image. This algorithm considered iteratively the set k of best sentences up to time t as candidates to generate sentences of size t+1, and then keep only the resulting best k of them [19]. In our experiment, the beam search size has been chosen as 3.

IV. MYANMAR IMAGE CAPTION GENERATION PROCESS

Encoder-Decoder framework of Myanmar image caption generation process is described in Fig. 1. Data preprocessing is a vital role in every machine learning algorithm [1]. The image captioning system required two different types of data preprocessing: one for image understanding part using NASNetLarge feature extraction model as encoder and one for text understanding part using Bi-direction LSTM model as decoder.

NASNetLarge pre-trained feature extraction model is a type of deep convolutional neural network as encoder, the input images must be resized to the expected format, i.e. (331,331) for NASNetLarge and (224,224) for VGG16. The image preprocessing module can be offered by Tensorflow that can access easier for them to be read into memory, decoded as jpg, jpeg and resized using pre-trained model.

Data preparation for the deep Bi-LSTM network decoder required the preprocessing of the textual data i.e., Myanmar descriptions. In this work, we compared the results for both word segmentation and syllable segmentation in language understanding part. The preprocessing steps for Myanmar images captions in our corpus are performed with two different ways as follows:

Word Segmentation: The sentences in Myanmar image captions corpus are not segmented correctly and some do not have almost no segmentation is essential for the quality improvement of Myanmar image captioning system. Therefore, UCSYNLP word segmenter [8] is used for Myanmar word segmentation and Myanmar syllable segmenter[1] for syllable segmentation. After Myanmar image captions sentences are segmented by using UCSYNLP word segmenter, the "|" symbol from the segmented results is replaced with space. The process of word segmentation for a Myanmar image caption sentence in our corpus is described and the meaning is "A man is riding the wave":

Unsegmented sentence: လူတစ်ယောက်ကရေလှိုင်းစီးနေတယ်

Segmented sentence: လူ တစ်ယောက် က ရေ လှိုင်းစီး နေတယ်

Syllable Segmentation: Syllable segmentation is a vital preprocessing step for Natural Language Processing. In Myanmar language, words are combined with most of the syllables, and syllables are comprised more than one character. Regular Expression (RE) based Myanmar syllable segmentation tool is used in this work. For Myanmar syllable-based neural image captioning model, "sylbreak" is used to segment the Myanmar image captions sentences into syllable level. After segmenting the Myanmar image captions sentences into syllable segmentation, the "|" symbol from the results is removed and replaced with white space and leading the trim process. The process of syllable segmentation for a Myanmar image caption sentence in our corpus is showed as follows and the meaning is "A man is riding the wave":

Unsegmented sentence: လူတစ်ယောက်ကရေလှိုင်းစီးနေတယ်

Segmented sentence: လူ တစ် ယောက် က ရေ လှိုင်း စီး နေ တယ်

The pre-segmented sentences in our corpus feed forward to the word embedding layer of Bi-LSTM model for handling the text input which provides a dense representation of the words and their relative meanings. Word embedding vector is used to fit a neural network on the text data and given as input to the Bi-LSTM model. This model makes predictions

[1] https://github.com/ye-kyaw-thu/sylbreak

TABLE I. PERFORMANCE COMPARISON ON BLEU-N (HIGHER IS BETTER). THE SUPERSCRIPT '*N*' MEANS THE *NASNetLarge* FEATURE EXTRACTION MODEL AND '*V*' IS *VGG16* MODEL, '-' INDICATES UNUSED

Models	Word Segmentation (%)				Syllable Segmentation (%)			
	BLEU-1	***BLEU-2***	***BLEU-3***	***BLEU-4***	***BLEU-1***	***BLEU-2***	***BLEU-3***	***BLEU-4***
Baseline [5]	64.14	48.58	39.86	24.38	-	-	-	-
Cheng et al. [20]	65.5	46.8	32	21.5	-	-	-	-
GRU^N	66.45	50.36	40.65	26.4	69.35	57.34	50.97	38.2
$LSTM^N$	66.76	50.06	40.31	26.5	69.73	58.17	51.24	39.04
$Bi\text{-}LSTM^V$	65.37	49	39.13	25	69.81	57.69	50.87	38.06
$Bi\text{-}LSTM^N$	**67.24**	**51.29**	**41.75**	**27.55**	**70.74**	**58.74**	**52.44**	**40.05**

over the entire vocabulary for each word in Myanmar image caption corpus based on the previous images feature extraction vectors and word embedding vectors until the end of the sentence.

V. DATASET AND EXPERIMENTS

A. Image Captions Corpus Creation for Myanmar Language

We have used Flickr8k dataset [14] for our experiments which contains 8092 images and each of the image is paired with five different English captions which provide a clear description of the salient entities and events. Myanmar image captions corpus is constructed by using two different approaches.

In First approach, English captions are performed preprocessing to clean the description texts. All words in the sentences convert to lowercase and remove all punctuations to ease the calculation of number of unique words in the dataset. And then, all words that are one character (e.g., 'a', 's') and number in the sentences are removed. After preprocessing English captions are done, Attention based Neural Machine Translation model [10] is applied to convert English captions to Myanmar captions. This model is trained on UCSY corpus that has 220k English Myanmar paralleled sentences. In machine translation, due to the web news and conversations influenced on the domain of the training data, the translated results are not good enough to use directly image captioning model. Although the translated results are not good, the translated Myanmar sentences help to reduce manual captioning time for building Myanmar image captions corpus.

In second approach, we tried to fix the translated Myanmar sentences to get the clean corpus. So, we manually checked and corrected the translated sentences one by one to match the captions with their relevant images. The total Myanmar captions sentences for 8092 images are 40460 sentences in our corpus. The vocabulary size is 3350 words and the maximum sentence length is 24 for word level, 32 for syllable level segmentation. The dataset is divided into three parts, 7092 images for training, 500 images for validation and the rest of 500 images are used for testing.

B. Experimental Framework

All experiments are run on NVIDIA GeForce MX250, RAM 16GB laptop and implemented with Python by using Keras library, which is run on Tensorflow as backend. The system performance stabilized at the end of 12 epochs and saved the best learned model on the training dataset. Loss is evaluated using sparse softmax cross entropy which measures the probability error in discrete classification tasks. The optimizer used is adaptive moment estimation (Adam) instead of RMSprop optimizer for better results [6]. Regularization of 50 % dropout is used to reduce over fitting during the training time. The hyperparameters tuning list of Bi-LSTM Myanmar image captioning models' architecture is shown in Table II:

TABLE II. HYPERPARAMETERS TUNING LIST OF OUR MODELS

Parameters	Values
Embedding size	300
Hidden layer size	256
Max-sequence length	32
Dense layer size	256
Batch size	32
Number of epochs	15
Beam Search (k)	3
Random Seed	1035

C. Evaluation Metrics

Bilingual Evaluation Understudy (BLEU) is mostly used algorithm that checks the quality of generated text output by the image caption generation model. It evaluates the numerical translation similarity between the generated sentence and comparing the text to one or more the ground truth sentences. BLEU scores can measure the fraction of N-grams (N=1, 2, 3, 4) in common and it is focus on the precision. The output of BLEU score range is always between 0 and 1, value close to 1 show that the generated sentence is more analogous to the human-generated sentence. In our experiments, BLEU scores are evaluated as

$$\text{BLEU}=\min\left(1,\frac{hypothesis\_length}{reference\_length}\right)\left(\prod_{i=1}^{4} precision_i\right)^{1/4} \quad (7)$$

Where hypothesis_length is the generated caption length and reference_length is the ground truth caption length.

$$\text{Precision}=\frac{Overlapped\ n-grams}{Total\ no\ of\ n-grams\ in\ reference} \quad (8)$$

a). Generated Caption 1: ခွေး က တန်း ကို ခုန် ကျော် နေတယ်
(In English: The dog is jumping over hurdle)
Generated Caption 2: ခွေး က သံ တန်း ကို ခုန် ကျော် နေ တယ်
(In English: The dog is jumping over the iron bar)

b). Generated Caption 1: လူ တစ်ယောက် က စက်ဘီး စီး နေတယ်
(In English: A man is riding a bicycle)
Generated Caption 2: လူ တစ် ယောက် က မော် တော် ဆိုင် ကယ် စီး နေ တယ်
(In English: A man is riding a motorcycle)

Fig. 2. Automatically Generated Caption by NASNetLarge with Bi-LSTM using Word and Syllable Level Segmentation. In two generated captions of each image, generated caption 1 is word segmentation result and generated caption 2 is syllable segmentation result

D. Performance Discussion

In Table I, we presented the results on Myanmar image captions corpus and provided a comparison with baseline model [5] and state-of-the-art model [20]. The BLEU scores results are achieved using four different models, namely NASNetLarge with GRU, NASNetLarge with LSTM, NASNetLarge with Bi-LSTM, and VGG16 with Bi-LSTM using two different segmentation process such as word level segmentation and syllable level segmentation. As can be seen in Table I, NASNetLarge with Bi-LSTM model using word segmentation achieved the best BLEU scores on 67.24%, 51.29%, 41.75%, 27.55 for BLEU-1, BLEU-2, BLEU-3, and BLEU-4 respectively, compared with other models. Moreover, we found that NASNetLarge with Bi-LSTM model using syllable segmentation perform significantly better than recent method on Myanmar image captioning task. The highest BLEU scores of this model are 70.74%, 58.74%, 52.44%, and 40.05% for BLEU-1, BLEU-2, BLEU-3, and BLEU-4 respectively, compared with other models as well as baseline model.

The baseline model VGG16 with LSTM performed 64.14%, 48.58%, 39.86% and 24.38% of BLEU-1, BLEU-2, BLEU-3 and BLEU-4 respectively, in the previous work [5] on previous Myanmar image caption corpus (15k sentences). Nonetheless, the previous model is significantly decrease to 3.17% of BLEU-4 score using Bi-LSTM^{N} with word level and15.67% of BLEU-4 score using Bi-LSTM^{N} with syllable level on updated corpus. The Bi-LSTM^{N} model is also compared with other neural network models such as GRU^{N}, LSTM^{N} and Bi-LSTM^{V}. These models are better performance than baseline model, but unfortunately, performance is not improved as well as Bi-LSTM^{N}.

In recent interesting work [20], authors achieved the best BLEU scores results on 65.5% of BLEU-1, 46.8% of BLEU-2, 32% of BLEU-3 and 21.5% of BLEU-4 respectively by combining VGG16 with Bi-LSTM model on Flickr8k dataset. By replacing VGG16 with NASNetLarge brings significant better performance on all evaluation BLEU scores in this work compare to previous work [5] [20]. Different feature extraction models are employed in different work, to make a fair and comprehensive comparison, we selected commonly used VGG16 to match NASNetLarge in this work. Furthermore, we found that NASNetLarge feature extraction model outperformed the accuracy than VGG16 model as showed in Table I.

VI. Experiments Result and Analysis

In this section, we especially focused on the predicted captions generated by NASNetLarge model as encoder and Bi-LSTM network as decoder to compare the word segmentation results as well as syllable segmentation results. As we noted that in Fig.2. (a)(b) and Fig.3, the generated caption 1 is word segmentation result and generated caption 2 is syllable segmentation result. The generated captions of other models are not different significantly.

The generated results from NASNetLarge with Bi-LSTM using word segmentation and syllable segmentation are discussed in this section. In Fig. 2 (a) generated caption 1, the model effectively predicts the activities and information of the main objects, but generated caption 2 is more accurately identify like “သံ တန်း” (“iron bar”) with syllable segmentation result although it fails to identify with word segmentation result in generated caption 1. Furthermore, in Fig. 2. (b) generated caption 1, it is noticeable that the prediction is not satisfied and cannot correctly identify the object like “မော်တော်ဆိုင်ကယ်” (“Motorcycle”) because the contents of the image are difficult to identify accurately. Nonetheless, in Fig.2. (b) generated caption 2, we noticed that the model is correctly identified the object “မော်တော်ဆိုင်ကယ်” (“Motorcycle) although the generated caption 1 cannot correctly identify this object.

In Fig.3 generated caption 1, where most of the objects are predicted correctively and also generated grammatically correct sentence but it fails to identify the number of objects like “ကလေး နှစ်ယောက်” (“Two Children”). The model can capture only a boy instead of two children. On the other hand, in generated caption 2, we can be seen that the model accurately generated the major features, number of objects and relationship between these features within images and also generated grammatically correct sentence.

To conclude the experiments results, NASNetLarge with Bi-LSTM model for both segmentations can give the acceptable results for Myanmar image caption generation even with the open test images. NASNetLarge with Bi-LSTM

Generated Caption 1: ကောင်လေး က ကွင်း ထဲမှာ ပြေး နေတယ်
(In English: The boy is running in the field)

Generated Caption 2: က လေး နှစ် ယောက် မြက် ခင်း စိမ်း ပေါ် မှာ က စား နေ တယ်
(In English: Two children are playing on the green grass)

Fig. 3. Automatically Generated Caption by NASNetLarge with Bi-LSTM using Word and Syllable Level Segmentation

using syllable segmentation is more identified the objects accurately according to the descriptions generated sentences of the images rather than word segmentation for Myanmar caption generation system. All of the Fig. 2 and 3, NASNetLarge with Bi-LSTM model automatically generated captions with Myanmar language without any human intervention.

VII. Conclusion and Future Work

In this paper, Myanmar image captions corpus (over 40k sentences) is constructed based on the Flickr8k dataset and compared with baseline model as well as other various neural network models, namely NASNetLarge with GRU, NASNetLarge with LSTM, VGG16 with Bi-LSTM and NASNetLarge with Bi-LSTM models by analyzing BLEU scores. NASNetLarge feature extraction model performed better than VGG16 model as an encoder for the automatic Myanmar image caption generation task. According to the experiment results, we can be stated that encoder plays a very important role in image captioning system and can be significantly improved model without changing a decoder architecture. On the other hand, we can be seen that data preprocessing step plays a vital role in language modeling. In Myanmar image captioning system, syllable segmentation is significantly better than word segmentation. Based on the results showed above, NASNetLarge with Bi-LSTM using syllable segmentation gained better performance than other different models as well as baseline model on the updated dataset, especially for Myanmar language.

In future work, new feature extraction model EfficientNet will be investigated as encoder and also plan to apply attention mechanism into our model to improve the system performance.

References

[1] A. Puscasiu, A. Fanca, D. I. Gota and H. Valean, "Automated image captioning", 2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), 2020.

[2] B. Zoph, V. Vasudevan, J. Shlens and Q. V. Le, “Learning Transferable Architectures for Scalable Image Recognition”, arXiv:1707.07012v4 [cs.CV] 11 Apr 2018.

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

[4] 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.

[5] 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, Marseille, France, 2020.

[6] V. Kesavan, V. Muley and M. Kolhekar, "Deep Learning based Automatic Image Caption Generation", 2019 Global Conference for Advancement in Technology (GCAT), 2019.

[7] 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.

[8] W. P. Pa and N. L. Thein, "Myanmar Word Segmentation using Hybrid Approach", Proceedings of 6th International Conference on Computer Applications, 2008, Yangon, pp-166-170.

[9] Y. M. S. Sin, W. P. Pa and K. M. Soe, “UCSYNLP-Lab Machine Translation Systems for WAT 2019”, Proceedings of the 6th Workshop on Asian Translation, pp.195-199, 2019.

[10] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, “Every Picture Tells a Story: Generating Sentences from Images”, European conference on computer vision, Springer, pp. 15–29, 2010.

[11] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate”, arXiv preprint arXiv:1409.0473, 2014.

[12] 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.

[13] R. Dhir, S. K. Mishra, S. Saha and P. Bhattacharyya, “A Deep Attention based Framework for Image Caption Generation in Hindi Language”, Computacion y Sistemas, Vol.23, No.3, pp. 693-701, 2019.

[14] H. Micah, Y. Peter, and H. Julia, “Framing image description as a ranking task: Data, models and evaluation metrics”, Journal of Artificial Intelligence Research, Vol. 47, pp. 853-899, May, 2013.

[15] I. Sutskever, O. Vinyals and Q. V. Le, “Sequence to Sequence Learning with Neural Networks”, Advances in neural information processing systems, pp. 3104–3112, 2014.

[16] https://medium.com/voice-tech-podcast/an-overview-of-rnn-lstm-gru-79ed642751c6

[17] https://medium.com/@madhuramiah/bi-directional-rnn-basics-of-lstm-and-gru-e114aa4779bb

[18] A. A. Nugraha, A. Arifianto and Suyanto, “Generating Image Description on Indonesian Language using Convolutional Neural Network and Gated Recurrent Unit”, 7th International Conference on Information and Communication Technology (ICoICT), 2019.

[19] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, “Show and Tell: A Neural Image Caption Generator”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156-3164, 2015.

[20] C. Wang, H. Yang, C. Bartz and C. Meinel, “Image Captioning with Deep Bidirectional STMs”, ACM, DOI: http://dx.doi.org/10.1145/2964284.2964299,Amsterdam, Netherlands, 2016.