Received: June 28, 2022. Revised: September 24, 2022. 

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

# EfficientNetB7 and Bi-LSTM with GloVe Vector Based Myanmar Image Captioning

**San Pa Pa Aung[1]*** **Win Pa Pa[1]**

[1]*Natural Language Processing Lab., University of Computer Studies, Yangon, Myanmar*
* Corresponding author's Email: sanpapaaung@ucsy.edu.mm

**Abstract:** Image Captioning (IC) is one of the most widely discussed topic in Artificial Intelligence. In this paper, Myanmar image caption is generated using EfficientNetB7 and Bidirectional Long Short-Term Memory (Bi-LSTM) with GloVe embedding and described about the comparative analysis results. For the purpose of achieving better performance, Myanmar image caption corpus is created and annotated over 50k sentences for 10k images, which are based on Flickr8k dataset and 2k images are selected from Flickr30k dataset. Two different types of segmentations such as word and syllable segmentation level are studied in text pre-processing step and then constructed our own GloVe vectors for both segmentations. As far as being aware and up to our knowledge, this is the first attempt of applying syllable and word vector features in neural network-based Myanmar IC system and then compared with one-hot encoding vectors on various different models. According to the evaluations results, EfficientNetB7 with Bi-LSTM using word and syllable GloVe embedding outperforms than EfficientNetB7 and Bi-LSTM with one-hot encoding, other neural networks such as Gated Recurrent Unit (GRU), Bidirectional Gated Recurrent Unit (Bi-GRU), and Long Short-Term Memory (LSTM), VGG16 with Bi-LSTM, NASNetLarge with Bi-LSTM models as well as baseline models. EffecientNetB7 with Bi-LSTM using GloVe vectors achieved the highest BLEU-4 score of 35.09%, 49.52% of ROUGE-L, 54.34% of ROUGE-SU4 and 21.3% of METEOR score on word vectors, and the highest BLEU-4 score of 46.2%, 65.62% of ROUGE-L, 68.43% of ROUGE-SU4 and 27.07% of METEOR score on syllable vectors.

**Keywords:** EfficientNetB7, NASnetLarge, Gated recurrent unit, Long short-term memory, Bidirectional long short-term memory, Visual geometry group.

## 1. Introduction

Caption generation from an image is one of the challenging tasks in the field of Computer Vision and natural language processing, two of the major fields in Artificial Intelligence. However, most research in this area generated image captions with English while there are a lot of different languages exist in the world. With their distinctive languages, there is a necessity of particular research to generate captions in those isolated language. At part of this work, manually annotated image captions corpus for Myanmar language was developed and proposed to support the evaluation of image caption generation. At present, our image caption corpus contains approximately 50460 sentences for 10092 images, which are based on Flickr8k dataset and 2k images are selected from Flickr30k dataset.

Image captioning for Myanmar language, only two works such as VGG16 with LSTM based language model (around 15k sentences) [1] and improving Myanmar image caption generation using NASNetLarge with Bi-LSTM model (over 40k sentences) [2], are found publicly. As far as being aware and up to our knowledge, none of the previously Myanmar IC has been applied word and syllable embedding vectors in language modelling. Therefore, in this work, we manually created Myanmar image captions corpus, and building our own word and syllable embedding vectors for Myanmar language. We investigated with different dimensions to find which is the best for both word and syllable segmentations and modelling of word

vectors for Myanmar language. More than that, different deep learning models such as EfficientNetB7 with GRU, EfficientNetB7 with Bi-GRU, EfficientNetB7 with LSTM, EfficientNetB7 with Bi-LSTM using GloVe embedding model, EfficientNetB7 with Bi-LSTM using one-hot embedding model, VGG16 with Bi-LSTM models and NASNetLarge with Bi-LSTM models are compared with baseline models [1, 2] and state-of-the-art models [3, 4, 5]. The experimental results showed that the EfficientNetB7 with Bi-LSTM using GloVe embedding vectors can give significantly better performance for both segmentations in Myanmar image captioning compared with other different models.

The main contributions of this paper are threefold:

-EfficientNetB7 feature extraction model and Bidirectional LSTM language generation model with GloVe vectors are applied for Myanmar image caption generation, that can accurately identify the objects in the images and also generate grammatically correct sentences with their relevant images.

-We evaluated the effectiveness of proposed model on Myanmar image captions corpus which contains 50460 sentences for 10092 images. Our evaluation results shown that the proposed model obtained the significantly better performance on caption description.

-We built our own GloVe embedding vectors for both segmentations such as word and syllable segmentation and compared with different models.

The rest of the paper is organized as follows. Section 2 shows the existing work flow of image captioning. In section 3, the proposed encoder-decoder architecture is introduced and the process flow of Myanmar IC is described in section 4. Several groups of experiments are illustrated in section 5. Section 6 summarizes the presented work.

## 2. Related work

Nowadays, Image Captioning (IC) is one of the most widely discussed topic which is the combination of computer vision and natural language processing.

InceptionV3 pre-trained feature extraction model is used for understanding the contents of images as an encoder and Bidirectional Gated Recurrent Unit (Bi-GRU) is applied for image annotation as decoder. Experiments have been done on BNATURE dataset of Bengali language which contains the total 8000 images with five different captions in each image [6]. The authors only achieved the BLEU-4 score 16.41% because InceptinV3 is less powerful in feature extraction which has 94% top-5 classification accuracy on ImageNet dataset. According to our experimental results, Bi-LSTM achieved better results than Bi-GRU although the Bi-LSTM model takes more training time. It still has a gap to generate captions that has little errors and few of them are irrelevant with their corresponding test images.

The pre-trained VGG16 and Alexnet feature extraction models of Convolutional Neural Network (CNN) are applied 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 [3]. In [4], Deep convolutional neural network is used to learn the image contents and two separate LSTM network is applied to learn long-term visual-language interactions and make prediction by the use of history and future context information at high-level semantic space. Then, the deep multimodal bidirectional models also investigated, in which the depth of nonlinearity transition is increased in various approaches to recognize hierarchical visual-language embeddings. The accuracy of proposed models is measured on four benchmark datasets: Flickr8K, Flickr30K, MSCOCO, and Pascal1K datasets. The highest BLEU-N (N=1,2,3,4) scores 66.7%, 48.3%, 33.7% and 23% respectively, and 19.1% of METEOR score are achieved using VGG16 with Bi-LSTM model on Flickr8k dataset. The authors revealed that the model performance on small-scale dataset Flickr8K is not good as large dataset Flickr30K and MSCOCO. It failed to identify the objects in complex background images because feature extraction model is not state-of-the-art model that has only 16 layers deep. The deeper networks yield the better understanding the contents of the images. Moreover, their approach does not consider word embedding in language modelling that make predicting image captions task better.

ResNet101 is used as feature extraction model and Standard LSTM with one cell is utilized as decoder. The pretrained vector representations as Word2Vec and GloVe embedding are compared on the MSCOCO dataset [5]. The model performance with GloVe vectors achieved better results than the model with Word2Vec because image captioning is more suitable with co-occurrence of word pairs in the entire corpus. Moreover, it does not state the generated captions results, and they used only the pre-trained word vectors with English language.

In this work, we manually created the Myanmar image captions corpus (around 50460 sentences for

10k images) and built our own word and syllable GloVe embedding vectors to compare the results of other neural network models as well as baseline models. EfficientNetB7 is the new model and more powerful in feature extraction as an encoder which has the 97% top-5 classification accuracy on ImageNet dataset [7]. Bi-LSTM is used as a decoder to overcome the problems of vanishing gradients which are present in Recurrent Neural Network (RNN). Unlike other image captioning models, in our proposed architecture model, we added our own word and syllable embedding vectors to the Bi-LSTM model for efficiency and better results that is proved to be very effective in generation with Myanmar language as we demonstrate in Fig 4.

## 3. Methodology

Especially, Myanmar image caption generation can be divided into two parts: 1) image features extraction acts as encoder and 2) generating a caption with Myanmar language as decoder. In the image features extraction part, we compared three popular convolution networks architectures- Visual Geometry Group (VGG) OxfordNet 16-layer [8], NASNetLarge [9] and EfficientNetB7 [7] as encoders for Myanmar image captioning in order to find out which is the best at feature extraction to apply for caption generation. According to the experiments, we found that EfficientNetB7 is significantly better performance than for both VGG16 and NASNetLarge models without changing the decoder model, so EfficientNetB7 is used as the encoder of the proposed model. In caption generation part, four different language generation models such as Long Short-Term Memory (LSTM) [10], Gated Recurrent Unit (GRU) [11], Bidirectional Gated Recurrent Unit (Bi-GRU) [12] and Bidirectional Long Short-Term Memory (Bi-LSTM) are investigated to apply which language modelling is the best at image captioning. The best result is obtained from a combination of EfficientNetB7 as an encoder and Bi-LSTM as a decoder. The following subsections are explained in details.

### 3.1 Feature extraction model

To develop the encoder-decoder neural network model, encoder is the vital initial step of image captioning model that extracted all of the features in images and the extracted features are used as input to the decoder. The feature extraction models of CNN have different feature vector size and different capability based on the use of model.

**EfficientNetB7**: MobileNets and ResNet are scaling up to improve the effectiveness of EfficientNetB7 model [7]. The features vectors of input images are defined to be 2560 elements and the default input image size of EfficientNetB7 is 600x600 and then processed by a Dense layer to produce a 256 elements representation of the image. The last layer of EfficientNetB7 model is removed because we need to take feature vectors instead of classification of the images. The output of second fully connected layer is taken as the initial state of Bi-LSTM in the decoder after it is downsized by the dense map layer.

### 3.2 Word embeddings

Word embedding is basically a form of word representation that transforms human understanding language to form vectors of each word. Word2Vec and GloVe are the most common use techniques to learn word vectors. In this work, GloVe is used in word and syllable embedding phase based on Bi-LSTM neural network after pre-processing step.

**GloVe:** GloVe (Global Vectors for Word Representation) is an approach to achieve vector representations utilizing unsupervised learning methods as stated by matrix factorization techniques on the word-context matrix [13]. Word and syllable vectors are created for monolingual Myanmar corpus using the GloVe v.1.2.

### 3.3 Bidirectional long short-term memory (Bi-LSTM)

Long Short-Term Memory (LSTM) is an extension of Recurrent Neural Network (RNN) that is designed to handle long-term dependencies and it is more accurately than conventional RNNs. LSTM is the centre of Bi-LSTM model which involves an input layer, two hidden layers and an output layer.

**Input layer:** During the training phrase, the input layer takes the pre-segmented words in our corpus and their corresponding 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.

**Hidden layer:** The hidden layer consists of two different LSTM networks - forward and backward, connecting to the same output layer. During training, both the 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 forward and backward hidden layer are calculated as the following Eqs. (1) and (2). The concatenation of forward and backward layer constructed the final encoded hidden vector, $h_t = [\vec{h}_t, \overleftarrow{h}_t]$ as in Eq. (3):

$$\vec{h}_t = \sigma\left(W_{\vec{h}}\left[\vec{h}_{t-1}, w_t\right] + b_{\vec{h}}\right) \quad (1)$$

$$\overleftarrow{h}_t = \sigma\left(W_{\overleftarrow{h}}\left[\overleftarrow{h}_{t-1}, w_t\right] + b_{\overleftarrow{h}}\right) \quad (2)$$

$$h_t = W_{\vec{h}}\vec{h}_t + W_{\overleftarrow{h}}\overleftarrow{h}_t + b_{\hat{y}} \quad (3)$$

Where, $\vec{h}_{t-1}$ and $\overleftarrow{h}_{t-1}$ are the previous forward hidden state and backward hidden state, $w_t$ denotes input word embedding, W is the weight matric, b is the bias vector and σ is the sigmoid activation function.

**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 [14]. Fig. 1 shows the architecture of Bi-LSTM model.

## 4. Myanmar image captioning

The basic framework of Myanmar image captioning is depicted in Fig. 3. Data pre-processing plays the vital role in every deep learning algorithm. In training module of Myanmar IC, two different types of data pre-processing are required such as image pre-processing and text pre-processing. In image pre-processing step, the input images must be resized to the expected format, i.e. (331,331) for NASNetLarge, (224,224) for VGG16 and (600,600) for EfficientNetB7 to get the better quality and to

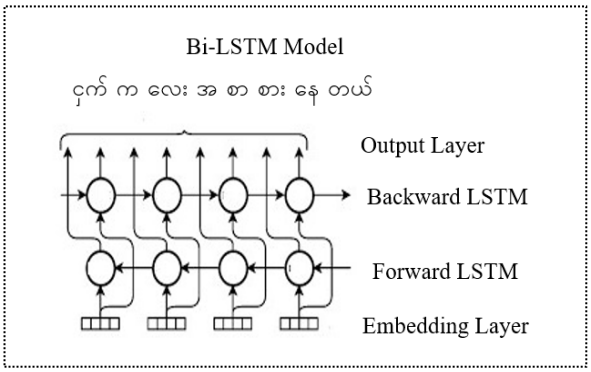


Figure. 1 Architecture of Bi-LSTM model

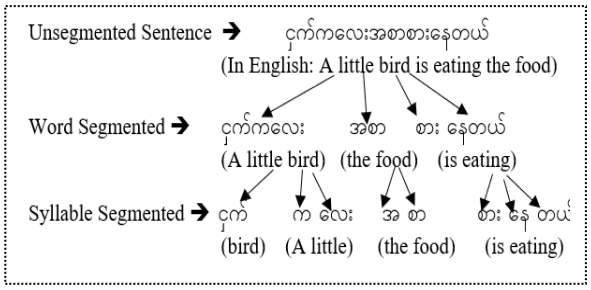


Figure. 2 Myanmar word segmented structure

avoid any numerical inconsistency during training and testing phases. The image pre-processing 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. After the image pre-processing is done, the pre-processed images are provided as input to the EfficientNetB7 features extraction model which extracted the features of the images and feed forward to the Bi-LSTM model.

In text pre-processing step, two different kinds of segmentation such as word segmentation [15] and syllable segmentation [16] are used in training to compare which segmentation level affects in Myanmar image captioning. The process of segmentation for a Myanmar image caption sentence in our corpus is presented as following in Fig. 2:

Text pre-processing is very important role in language modelling, and syllable segmentation is significantly better than word segmentation for Myanmar IC. After text pre-processing step is done, we got the clean Myanmar image captions corpus. GloVe vectors are created for the segmented corpus and feed forward to Bi-LSTM model. We applied Bi-LSTM model for training to get the best learned model using the image feature vectors and GloVe vectors, and then generates the caption with Myanmar language for given test image.

## 5. Experiments

In this section, different groups of experiments are designed to fulfill the following ambitions:

-The benefits and performance of proposed EfficientNetB7 with bidirectional LSTM model is measured on GloVe embedding and one-hot encoding with different ways.

-The influences of text pre-processing are examined based on two different segmentations process such as word and syllable segmentation on Myanmar image captions corpus.

-Our approach is compared with state-of-the-art methods in terms of captions generation on our Myanmar image captions corpus.

-GloVe embedding features are constructed in two different ways- word GloVe vectors and syllable GloVe vectors, and know how language modelling learned to predict a sentence conditioned by visual context information over time.

### 5.1 Dataset preparation for Myanmar language

We have chosen Flickr8k dataset [17] and 2k images from Flickr30k dataset [18], a total 10k images for our experiment which are commonly used dataset for caption annotation in English language. This dataset includes complex everyday activities with common objects in naturally occurring contexts and can be downloaded easily. Therefore, it covered large possible categories of images. As no Myanmar image caption dataset is available in the literature, we have manually annotated the captions of this dataset. It contains five annotated captions per image to generate image Captioning dataset for Myanmar Language. In this work, we have taken all of the English captions from the Flickr10k dataset (i.e., Flickr8k and 2k from Flickr30 dataset) to build Myanmar image caption corpus. Firstly, English to Myanmar machine translation [19] is applied to translate English captions to Myanmar captions.

Then, the translated Myanmar sentences are manually checked and corrected by matching each image and creating sentence descriptions relevant to the picture. Hence, the total images captions for 10092 images with five annotated Myanmar captions are 50460 sentences and the vocabulary size is 3350 words. The maximum sentence length is 24 for word level and 32 for syllable level segmentation.

Validation set contained 650 images to monitor the accuracy of trained model. The model performance improved and stabilized at the end of 15 epoch and then saved that model to get the best-learned model on the training dataset. Test set contained 650 images to measure the performance of the learned model and its prediction on a test set. The rest of the 8792 images are used as training.

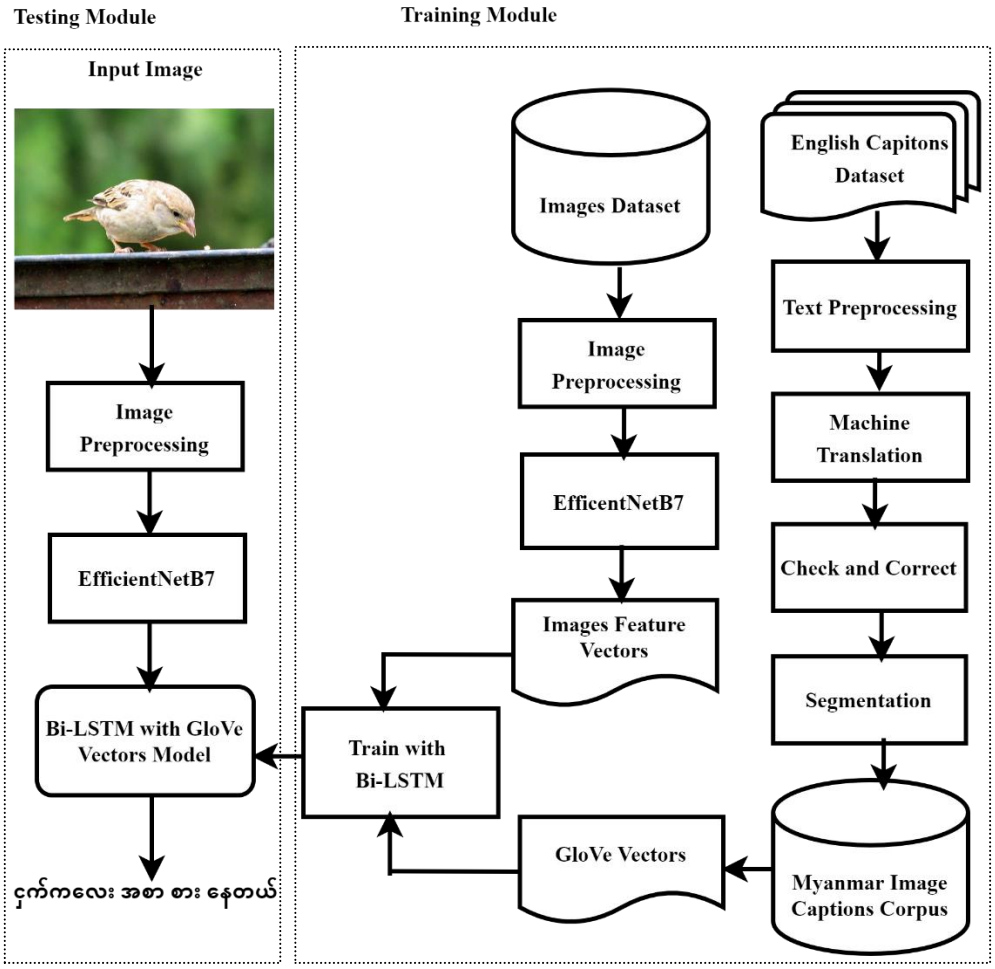


Figure. 3 System flow diagram of Myanmar image captioning

Table 1. Performance comparison of BLEU-N, ROUGE-L, ROUGE-SU4 and METEOR scores on word segmentation. The superscript 'E' means 'EfficientNetB7', 'G' is 'VGG16', 'A' is 'AlexNet', 'R' is 'ResNet101', 'N' is 'NASNetLarge' feature extraction models, '+M' is using multi-task learning and '+W' is using GloVe vector. '-' indicates unused. The superscripts are also applicable in Table 2 and other sections in this manuscript

| **Models** | **Word Segmentation (%)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **B-1** | **B-2** | **B-3** | **B-4** | **ROUGE-L** | **ROUGE-SU4** | **METEOR** |
| $LSTM^{G}$ [1] | 64.14 | 48.58 | 39.86 | 24.38 | - | - | - |
| $Bi\text{-}LSTM^{N}$ [2] | 67.24 | 51.29 | 41.75 | 27.55 | - | - | - |
| $Bi\text{-}LSTM^{G, A}$ [3] | 65.5 | 46.8 | 32 | 21.5 | - | - | 19.4 |
| $Bi\text{-}LSTM^{G, +M}$ [4] | 66.7 | 48.3 | 33.7 | 23 | - | - | 19.1 |
| $LSTM^{R,+W}$ [5] | 69 | 51.6 | 37.6 | 26.9 | 50.3 | - | 22.4 |
| $GRU^{E}$ | 68.1 | 53.62 | 45.6 | 32.22 | 47.55 | 52.39 | 20.72 |
| $LSTM^{E}$ | 69.65 | 54.98 | 47.38 | 33.57 | 47.17 | 51.45 | 21.06 |
| $Bi\text{-}GRU^{E}$ | 68.84 | 54.28 | 46.25 | 32.99 | 47.52 | 54.05 | 21.02 |
| $Bi\text{-}LSTM^{G}$ | 67.07 | 51.37 | 41.82 | 28.03 | 42.96 | 49.45 | 18.74 |
| $Bi\text{-}LSTM^{N}$ | 67.63 | 51.82 | 42.4 | 28.82 | 43.79 | 51.79 | 19.69 |
| $Bi\text{-}LSTM^{E}$ | 70.12 | 55.07 | 47.85 | 34.91 | 46.18 | 52.79 | 21.25 |
| **$Bi\text{-}LSTM^{E,+W}$** | **71.42** | **56.73** | **48.45** | **35.09** | **49.52** | **54.34** | **21.3** |

Table 2. Performance comparison of BLEU-N (N=1,2,3,4), ROUGE-L, ROUGE-SU4 and METEOR scores on syllable segmentation

| **Models** | **Syllable Segmentation (%)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **B-1** | **B-2** | **B-3** | **B-4** | **ROUGE-L** | **ROUGE-SU4** | **METEOR** |
| $LSTM^{G}$ [1] | 64.14 | 48.58 | 39.86 | 24.38 | - | - | - |
| $Bi\text{-}LSTM^{N}$ [2] | 70.74 | 58.74 | 52.44 | 40.05 | - | - | - |
| $Bi\text{-}LSTM^{G, A}$ [3] | 65.5 | 46.8 | 32 | 21.5 | - | - | 19.4 |
| $Bi\text{-}LSTM^{G, +M}$ [4] | 66.7 | 48.3 | 33.7 | 23 | - | - | 19.1 |
| $LSTM^{R,+W}$ [5] | 69 | 51.6 | 37.6 | 26.9 | 50.3 | - | 22.4 |
| $GRU^{E}$ | 72.02 | 61.65 | 56.13 | 44.46 | 62.46 | 65.28 | 26.76 |
| $LSTM^{E}$ | 73.06 | 62.02 | 57.02 | 44.82 | 64.16 | 65.69 | 26.67 |
| $Bi\text{-}GRU^{E}$ | 72.46 | 61.74 | 55.92 | 43.97 | 61.17 | 65.84 | 26.88 |
| $Bi\text{-}LSTM^{G}$ | 69.76 | 58.08 | 51.86 | 39.47 | 58.14 | 63.11 | 24.16 |
| $Bi\text{-}LSTM^{N}$ | 72.19 | 60.71 | 54.44 | 42.11 | 59.41 | 64.34 | 25.17 |
| $Bi\text{-}LSTM^{E}$ | 73.66 | 63.02 | 57.2 | 45.22 | 65.14 | 67.13 | 26.9 |
| **$Bi\text{-}LSTM^{E,+W}$** | **73.9** | **63.45** | **57.8** | **46.2** | **65.62** | **68.43** | **27.07** |

### 5.2 Building of word and syllable GloVe vectors

Recently, word embedding model have been applied in text to speech [20], text summarization [21] with their own corpus for Myanmar language. In [22, 23], only two set of pre-trained word vectors can be accessed publicly for Myanmar language. The pre-trained word vectors cannot be used directly because the words are not relevant with our IC. For the reason that our own word and syllable vectors are constructed with standard Unicode encoding for more coverage and much better performance of Myanmar image captioning. While constructing the

syllable and word GloVe vectors for our own image captions corpus, there are some issues in counting vocabulary and find out unknow terms for each word in the image captions corpus due to insufficient data in building GloVe embedding model. Therefore, the text data is collected to build a huge monolingual Myanmar corpus for the intention of construction better quality word embedding model with broad coverage. Myanmar news corpus [21] (around 10k sentences) is used by collecting various Myanmar News websites which contains different types of news such as World news, business, health, politics, Entertainment, education and sport.

In monolingual Myanmar corpus, the sentences from our image captions corpus (50460 sentences) are also added. Finally, it contains the total 60460 sentences. After collecting the data, the next step is building GloVe embedding model for both word and syllable Myanmar image captions corpus that are mapped with vector value. Each word and syllable are stated as real-valued vectors with different dimensionality (50, 100, 200, 300). Dimension 300 is selected to use in our experiments for better performance. The training iteration is repeated 15 times with negative sampling. The length of training context is set to 8 for all models. Embedding layer consists of the number of vocabularies, dimension of each word vector and maximum length of input vector.

### 5.3 Experiment setting

All of our investigations were conducted on NVIDIA GeForce MX250, RAM 16GB, Ubuntu Linux machine and implemented with Python by using Keras library, which is run on TensorFlow as backend. The system performance stabilized at the end of 15 epochs and saved the best learned model on the training dataset. Sparse softmax cross entropy is used to evaluate the loss which measures the probability error in discrete classification tasks. The Adaptive moment estimation (Adam) optimizer is used for better performance instead of RMSprop optimizer. A dropout of 50 % was set, which is the efficient regularization technique to mitigate the excessive during the training time. The best hyperparameters tuning list of Myanmar image captioning models' architecture is shown in Table 3.

Loss function for our experiments is evaluated as,

$$L\ (I, S) = -\sum_{t=1}^{N} log\ p_t(S_t) \quad (4)$$

Table 3. Hyperparameters setting 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 seeds | 1035 |

Where I is input image and S is generated sentence, N is the length of generated caption. $p_t$ and $S_t$ are probability and predict word at time t respectively. We have tried to reduce the loss values during the training process.

### 5.4 Evaluation metrics

To compare the achievement of models we utilized BLEU-N(N=1,2,3,4) [24], ROUGE-L [25], ROUGE-SU4 [25] and METEOR [26] metrics which are mostly used to evaluate the quality of image description generation.

Bilingual Evaluation Understudy (BLEU) is widely used to measure the performance of image caption generation as well as machine translation. It is evaluated to measure how many words are shared by the generated captions and reference captions. BLEU scores range from 0 to 1, if the value is close to 1, the best score that is approximating similar with human translation and 0 is no match at all [24]. Eqs. (5) and (6) are used to calculate the BLEU scores of n-gram metric (n=1,2,3 and 4).

$$BLEU = min\ (1, \frac{output\_length}{reference\_length})(\prod_{i=1}^{4} p_i\ )^{1/4} \quad (5)$$

Where, p is the modified n-gram precision, output_length is the generated captions length and reference_length is the ground truth captions length.

$$p = \frac{\sum_{n-gram \in c} Count_{clip}(n-gram)}{\sum_{n-gram \in c} Count(n-gram)} \quad (6)$$

Where, c is the generated output caption and $Count_{clip}$(n-gram) is the number of n-grams occur in the reference captions according to the output n-gram, Count(n-gram) is the number of n-grams occur in the generated output captions.

Recall-Oriented Understudy of Gisting Evaluation (ROUGE) [25] was intended for evaluating automatic summarization, machine translation and image captioning. In this work, ROUGE-L (Longest Common Subsequence) and ROUGE-SU4 (Skip-bigram plus unigram-based co-occurrence statistics) are used for comparison with other models. ROUGE scores are evaluated by using the following equations:

$$R = \frac{number\ of\ overlaping\ words}{Total\ words\ in\ reference\ captions} \quad (7)$$

$$P = \frac{number\ of\ overlaping\ words}{Total\ words\ in\ candidate\ captions} \quad (8)$$

$$F\text{-}measure = \frac{(1+\beta^2)RP}{R+\beta^2 P} \quad (9)$$

Where, the value of $\beta$ was set to be 1, R is recall and P is precision value.

Metric for Evaluation of Translation with Explicit Ordering (METEOR) [26] evaluated the mean value of precision and recall scores based on the unigram. METEOR can solve the limitation of strict matching by utilizing the word and synonyms based on unigram. In our experiment, METEOR scores are calculated as in the following equations:

$$P = \frac{m}{w_t} \quad (10)$$

$$R = \frac{m}{w_r} \quad (11)$$

$$F_{mean} = \frac{PR}{\alpha P + (1-\alpha)R} \quad (12)$$

Where, m is the number of unigrams in the candidate captions also found in reference, $w_t$ is the number of unigrams in candidate captions, $w_r$ is the number of unigrams in reference captions, the value of α was set to be 0.9, R is recall value and P is precision value.

### 5.5 Effectiveness of feature extraction models

In this paper, we used three visual models for encoding images: VGG16, NASNetLarge and EfficientNetB7, to investigate the effects of applied encoding approaches. The evaluation results are reported in Table 1 and Table 2, it is clear to see that without utilizing EfficientNetB7 feature extraction model and keep other configurations unchanged (shown as Bi-LSTM$^{G}$ and Bi-LSTM$^{N}$), the performance of the models drops significantly on all evaluation metrics for both segmentations. The experiments result also stated how utilizing encoder effects on image captioning performance. According to the results presented in Table 1 and Table 2, we can be stated that encoder plays a very important role in image captioning and can be significantly improved model performance without changing a decoder architecture. To this end, EfficientNetB7 is used as encoder for our proposed model, because it has higher resolutions, such as 600x600, are also widely used in object detection CNN [7]. We believe that feature extraction on insufficient data is more challenging and assistance to evaluate the benefits brought by encoder. Replacing NASNetLarge with EfficientNetB7 brings significantly better performance on all evaluation metrics.

### 5.6 Effectiveness of GloVe vectors

In addition to using EfficientNetB7 to increase the accuracy and next effective approach is GloVe vectors for Myanmar IC. We collected the text data to build our own GloVe vectors for both segmentations in order to improve the quality of the system. Next, the Bi-LSTM$^{E,+W}$ model using GloVe embedding vectors is trained and evaluation is done on each validation set to investigate the achievement and generality of the system. The model needs more training time than one-hot encoding vector and it will take to 5400 seconds per epoch. The best performing baseline model (Bi-LSTM$^{N}$) [2] without utilizing GloVe vectors is selected to compare with our experiments results. The comparison with baseline models in terms of BLEU and other metrics score results are showed in Table 1 and 2.

The experiments results stated that the GloVe embedding model (shown as Bi-LSTM$^{E,+W}$) significantly improve the BLEU score 4.18%, 5.44%, 6.7% and 7.54% for BLEU-1, BLEU-2, BLEU-3, and BLEU-4 respectively using word GloVe vectors, and 3.16%, 4.71%, 5.36%, 6.15% for BLEU-1, BLEU-2, BLEU-3, and BLEU-4 respectively using syllable GloVe vectors rather than baseline model (Bi-LSTM$^{N}$) [2]. The Bi-LSTM$^{E,+W}$ model also achieved much better performance than other neural network models like GRU$^{E}$, LSTM$^{E}$, Bi-GRU$^{E}$, Bi-LSTM$^{G}$, Bi-LSTM$^{N}$, Bi-LSTM$^{E}$ using one-hot encoding models as well as state-of-the-art models. Furthermore, Bi-LSTM$^{E,+W}$ using syllable GloVe embedding model is better performance than Bi-LSTM$^{E,+W}$ using word GloVe embedding model because a word is composed of one or more syllables (i.e., a word ‘Girl’ in Myanmar ‘မိန်းကလေး’ consists of three syllables like ‘မိန်း’, ‘က’ and ‘လေး’). BLEU score evaluated to measure

how many words or syllables are similar between the machine generated captions and reference captions, that is why, the syllable segmentation results are much better than word segmentation results.

Regarding ROUGE-L, ROUGE-SU4 and METEOR performance, we investigated how to affect the quality of the model using the GloVe embeddings vectors for both segmentations during the overall model training on a specific dataset. As the GloVe embeddings performed the best during all of our experiments, it reaches 49.52%, 54.34% and 21.3% for ROUGE-L, ROUGE-SU4 and METEOR scores respectively on word segmentation. In syllable segmentation, our proposed model (shown as Bi-LSTM$^{E,+W}$) obtains the best results 65.62%, 68.43% and 27.07% for ROUGE-L, ROUGE-SU4 and METEOR scores respectively whereas model without GloVe embedding (shown as Bi-LSTM$^{E}$) achieves 65.14% on ROUGE-L, 67.13% on ROUGE-SU4 and 26.9 on METEOR score.

Based on the experiment results, we can summarize that the effect of GloVe vectors can be seen clearly in Myanmar IC for both tasks although the size of GloVe vectors is small. Even though we believe enlarging the size of GloVe vectors into our approach can get further improvements, note that our proposed model obtained much better results on all evaluation metrics.

### 5.7 Comparison with state-of-the-art methods

In this section, the proposed Bi-LSTM$^{E,+W}$ model is compared with state-of-the-art methods. The comparison results are summarized in Table 1 and Table 2. Our approach achieved the best performance on all evaluation metrics for both segmentations. Bi-LSTM$^{E,+W}$ using GloVe vectors mostly obtained better performance compare to Bi-LSTM$^{E}$ without using GloVe vector as well as other different models. We should be aware that a recent interesting work (Bi-LSTM$^{G,\ A}$) [3] is significantly inferior to 5.92%, 9.93%, 16.45%, and 13.59% for BLEU-1, BLEU-2, BLEU-3, BLEU-4 and 1.9% of METEOR respectively compare to Bi-LSTM$^{E,+W}$ with word level, and 8.4% of BLEU-1, 16.65% of BLEU-2, 25.8% of BLEU-3, 24.7% of BLEU-4 and 7.67% of METEOR score compare to Bi-LSTM$^{E,+W}$ with syllable level. In addition, the previous model (Bi-LSTM$^{G,+M}$) [4] is significantly decrease to 12.09% of BLEU-4 and 2.2% of METEOR score compare with Bi-LSTM$^{E,+W}$ on word level, 23.2% of BLEU-4 score and 7.97% of METEOR score compare to Bi-LSTM$^{E,+W}$ with syllable level on updated corpus. Furthermore, our best results achieved 35.09% of BLEU-4, 49.52% of ROUGE-L and 21.3% of METEOR score (compare to 26.9% of BLEU-4, 50.3% of ROUGE-L and 22.4% of METEOR score in LSTM$^{R,+W}$ [5]) on word segmentation and 46.2% of BLEU-4, 65.62% of ROUGE-L and 27.07% of METERO score (compare to 26.9% of BLEU-4, 50.3% of ROUGE-L and 22.4% of METEOR score in LSTM$^{R,+W}$ [5]) on syllable segmentation.

Moreover, in recent interesting work [3, 4], the authors found that small dataset like Flickr8K which has difficulty to train the deep models with insufficient data. Nonetheless, the proposed Bi-LSTM$^{E,+W}$ model substantially outperforms on all metrics for both word and syllable segmentation tasks although the size of the corpus is small (around 50460 sentences for 10K images), compare with other different models namely GRU$^{E}$, Bi-GRU$^{E}$, LSTM$^{E}$, Bi-LSTM$^{G}$, Bi-LSTM$^{N}$, Bi-LSTM$^{E}$, the baseline models [1, 2] as well as the state-of-the-art models [3, 4, 5].

### 5.8 Experiment results and analysis

According to the generated results, we found that syllable segmentation results are more specific than word segmentation using one-hot encoding model [2]. Nonetheless, EfficientNetB7 with Bi-LSTM using word vector features also achieved the specific generated captions as well as syllable vector features that are not different significantly. In this section, we especially compared on the predicted captions generated by EfficientNetB7 with Bi-LSTM using GloVe vectors features (shown as Bi-LSTM$^{E,+W}$) and EfficientNetB7 with Bi-LSTM without using GloVe vectors model (shown as Bi-LSTM$^{E}$) for both tasks. If so, superscript W is used as the generated captions of Bi-LSTM$^{E,+W}$ model and superscript H is denoted as the generated captions of Bi-LSTM$^{E}$ model. As we noted that in Fig. 4, the generated captions $1^{W}$ and $3^{H}$ are word segmentation results, and generated caption $2^{W}$ and $4^{H}$ are syllable segmentation results. In Fig. 4(a), generated captions $1^{W}$, $2^{W}$ and $4^{H}$ are much more similar to one of the ground-truth captions, but in caption $3^{H}$, it fails to identify the objects correctly like ‘ငါးမျှား’ (‘fishing’) and ‘ရေကန်’ (‘lake’). As can be seen in Fig. 4(b), the generated captions $1^{W}$ and $2^{W}$, the model can capture the objects in details and also cover the different semantic information; for example, generated caption $1^{W}$ captures ‘ယာခင်း’ (‘farm’) while the generated caption $2^{W}$ describes ‘လယ် ကွင်း’ (‘field’). Nonetheless, in Fig. 4(b),

| Input Image | Generated Captions with and without GloVe Vectors |
|---|---|
|  | (a) 1[W]. လူ တစ်ယောက် သစ်ပင် အောက် မှာ ငါးမျှား နေတယ်<br>(In English: A person is fishing under the tree)<br>2[W]. လူ တစ် ယောက် က ရေ ကန် ဘေး မှာ ရပ် နေ တယ်<br>(In English: A person is standing beside the lake)<br>3[H]. လူ တစ်ယောက် က တော ထဲမှာ လမ်းလျှောက် နေတယ်<br>(In English: A person is walking in the forest)<br>4[H]. လူ တစ် ယောက် က သစ် ပင် အောက် မှာ ငါး မျှား နေ တယ်<br>(In English: A person is fishing under the tree) |
|  | (b)1[W]. အမျိုးသား က ယာခင်း ထဲမှာ နွား နှစ်ကောင် နဲ့ ယာထွန် နေတယ်<br>(In English: A man is plowing with two oxen in the farm<br>2[W]. လူ တစ် ယောက် က လယ် ကွင်း ထဲ မှာ နွား နှစ် ကောင် နဲ့ ယာ ထွန် နေ တယ်<br>(In English: A person is plowing with two oxen in the field<br>3[H]. လူ တစ်ယောက် က နွား နှစ်ကောင် ကို ကိုင် နေတယ်<br>(In English: A person is holding two oxen)<br>4[H]. လူ တစ် ယောက် က နွား နှစ် ကောင် ကို ထိန်း ဆွဲ နေ တယ်<br>(In English: A person is pulling two oxen) |
|  | (c )1[W]. လူ တစ်ယောက် က ကြိုးတံတား ပေါ်မှာ လမ်းလျှောက် နေတယ်<br>(In English: A person is walking on the rope bridge)<br>2[W]. အ မျိုး သ မီး က ကြိုး တံ တား ပေါ် မှာ လမ်း လျှောက် နေ တယ်<br>(In English: A woman is walking on the rope bridge)<br>3[H]. လူ များ က ကြိုးတံတား ပေါ်မှာ လမ်းလျှောက် နေ ကြ တယ်<br>(In English: People are walking on the rope bridge)<br>4[H]. အ မျိုး သ မီး က ကြိုး တံ တား ပေါ် မှာ လမ်း လျှောက် နေ တယ်<br>(In English: A woman is walking on the rope bridge) |
|  | (d) 1[W]. လူ တစ်ယောက် က ပင်လယ်ကမ်းခြေ မှာ ထီး ဆောင်း ပြီး ထိုင် နေတယ်<br>(In English: A man is sitting on the beach by holding the umbrella)<br>2[W]. လူ တစ် ယောက် က ထီး ဆောင်း ပြီး ထိုင် နေ တယ်<br>(In English: A man is sitting by holding the umbrella)<br>3[H]. ကလေး နှစ်ယောက် က ကမ်းခြေ မှာ ထိုင် နေတယ်<br>(In English: Two children are sitting on the beach)<br>4[H]. ကောင် မ လေး က ကမ်း ခြေ မှာ ထိုင် နေ တယ်<br>(In English: A girl is sitting on the beach) |
|  | ( e ) 1[W]. လူ တစ်ယောက် က ဝက် ကို ကြိုး ချည် ထား တယ်<br>(In English: A person ties a pig with the rope)<br>2[W]. လူ တစ် ယောက် က ဝက် ကို ကြိုး နဲ့ ဆွဲ ပြီး လမ်း လျှောက် နေ တယ်<br>(In English: A man is walking by pulling a pig which is tied with rope)<br>3[H]. လူ တစ်ယောက် က ခွေး ကို ကြိုး နဲ့ ဆွဲ နေတယ်<br>(In English: A person is pulling a dog by the leash)<br>4[H]. လူ တစ် ယောက် က မြင်း ကို ကိုင် ပြီး လမ်း လျှောက် နေ တယ်<br>(In English: A person is walking by holding a horse) |

Figure. 4 Example of generated captions with Myanmar language for a given image. Superscript W is used as the generated captions of the proposed Bi-LSTM$^{E,+W}$ with GloVe embedding vectors and superscript H is used as the generated captions of Bi-LSTM$^{E}$ without GloVe embedding vectors. In four generated captions for each image, $1^{W}$ and $3^{H}$ are word segmentation results and, $2^{W}$ and $4^{H}$ are syllable segmentation results

generated caption $3^{H}$ and $4^{H}$, Bi-LSTM$^{E}$ using one-hot encoding model cannot capture some objects correctly with their relevant image. It fails to identify the objects like 'plowing' and 'field'. In Fig. 4(c), both models can predict the sentence accurately for both segmentations and also cover the different semantic information. Moreover, as can be observed in Fig. 4(d), the proposed model effectively predicts the activities and information of the main objects although one-hot encoding model fails to identify the main object like 'ထီးဆောင်း' ('holding the umbrella') and 'လိင်' ('gender'). Furthermore, in Fig. 4(e), generated caption $1^{W}$ and $2^{W}$, most of the objects are predicted correctively and also generated grammatically correct sentence although the contents of the image are difficult to identify accurately. Generated captions $3^{H}$ and $4^{H}$ misidentify the objects like 'ခွေး'('dog') and 'မြင်း'('horse') instead of 'ဝက်'('pig').

To conclude the experiment results, EfficientNetB7 with Bi-LSTM using GloVe vectors features for both tasks word and syllable vectors can give highly performance results than EfficientNetB7 with Bi-LSTM without using GloVe vectors as well as the other different models for Myanmar IC even with the open test images. It illustrates that our proposed Bi-LSTM$^{E,+W}$ model has a powerful ability to learn visual-language correlation and predicts grammatically correct captions with Myanmar language for both tasks word and syllable segmentations. Fig. 4(a) to (e) are automatically generated captions with Myanmar language without any human intervention.

## 6. Conclusion and future work

In this work, we examined the effectiveness of word representation on EfficientNetB7 with Bi-LSTM based Myanmar image captioning for both word and syllable segmentation tasks. Myanmar image captions corpus (around 50460 sentences for 10k images) is created based on the Flickr8k and 2k images are selected from Flickr30k dataset. Moreover, words and syllable GloVe vectors were also constructed for Myanmar IC by utilizing the gathered monolingual Myanmar corpus for much better performance. The comparisons are done on various neural network models, namely GRU$^{E}$, Bi-GRU$^{E}$, LSTM$^{E}$, Bi-LSTM$^{G}$, Bi-LSTM$^{N}$, Bi-LSTM$^{E}$, Bi-LSTM$^{E,+W}$, baseline models and state-of-the-art models. According to the experiments results, EfficientNetB7 with Bi-LSTM using GloVe embedding model (Bi-LSTM$^{E,+W}$) achieved significantly better performance than EfficientNetB7 with Bi-LSTM without using GloVe embedding model (Bi- LSTM$^{E}$) as well as other neural network models. Although the size of GloVe vectors is small, word and syllable vectors features can give the effectiveness of Myanmar IC performance. Anyway, this exploration of using word and syllable vectors features for Myanmar IC is the first work to apply Bi-LSTM network in Myanmar language. Incorporating an attention mechanism into our approach will be investigated in the future and we will keep exploring other new feature extraction models.

## Conflicts of Interest

The authors declare no conflict of interest.

## Author Contributions

The paper background work, conceptualization, methodology, dataset collection, implementation, result analysis and comparison, preparing and editing draft, visualization have been done by first author. The supervision, review of work and project administration, have been done by second author.

## References

[1] S. P. P. Aung, W. P. Pa, and T. L. Nwe, "Automatic Myanmar Image Captioning using CNN and LSTM-Based Language Model", In: *Proc. of the 1st Joint SLTU and CCURL Workshop*, pp. 139-143, 2020.

[2] S. P. P. Aung, W. P. Pa, and T. L. Nwe, "Improving Myanmar Image Caption Generation Using NASNetLarge and Bi-directional LSTM", *19th International Conf. on Computer Applications (ICCA)*, 2021.

[3] C. Wang, H. Yang, and C. Meinel, "Image Captioning with Deep Bidirectional LSTMs", *arXiv:1604.00790v3 [cs.CV]*, 2016.

[4] C. Wang, H. Yang, and C. Meinel, "Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning", *ACM Transaction on Multimedia Computing Communications, and Application*, Vol. 14, No. 2s, Article 40, 2018.

[5] V. Atliha and D. Sesok, "Pretrained Word Embeddings for Image Captioning", *IEEE Open Conf. of Electrical, Electronic and Information Sciences (eStream)*, 2021.

[6] M. Faruk, H. A. Faraby, M. M. Azad, M. R. Fedous, and M. K. Morol, "Image to Bengali Caption Generation Using Deep CNN and

Bidirectional Gated Recurrent Unit", In: *Proc. of 23rd International Conf. on Computer and Information Technology (ICCIT), arXiv:2012.12139v1 [cs.CV]*, 2020.

[7] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", *arXiv*:*1905.11946v5 [cs.LG]*, 2020.

[8] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks For Large-Scale Image Recognition", *arXiv:1409.1556v6 [cs.CV]*, 2015.

[9] B. Zoph and Q. V. Le, "Neural Architecture Search with Reinforcement Learning", *International Conf. on Learning Representations*, 2017.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Computation* 9.8, pp. 1735-1780, 1997.

[11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling", *arXiv:1412.3555v1[cs.NE]*, 2014.

[12] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks", *IEEE Transactions on Signal Processing*, 45(11), 2673-2681, 1997.

[13] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation", In: *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532-1543, 2014.

[14] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks", *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673-2681, 1997.

[15] W. P. Pa and N. L. Thein, "Myanmar Word Segmentation using Hybrid Approach", In: *Proc. of 6th International Conf. on Computer Applications*, pp. 166-170, 2008.

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

[17] 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, 2013.

[18] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions", *Transactions of the Association for Computational Linguistics (TACL)*, pp. 67-78, 2014.

[19] Y. M. S. Sin, W. P. Pa, and K. M. Soe, "UCSYNLP-Lab Machine Translation Systems for WAT 2019", In: *Proc. of the 6th Workshop on Asian Translation*, pp. 195-199, 2019.

[20] A. M. Hlaing and W. P. Pa, "Word Representations for Neural Network Based Myanmar Text-to-Speech System", *International Journal of Intelligent Engineering and Systems*, Vol.13, No.2, pp. 239-249, 2020.

[21] Y. M. Thu and W. P. Pa, "Myanmar News Headline Generation with Sequence-to-Sequence model", In: *Proc. of the 23rd Conf. of the Oriental COCOSDA*, pp. 117-122, 2020.

[22] R. A. Rfou, B. Perozzi, and S. Skiena, "Polyglot: Distributed word representations for multilingual NLP", *arXiv Preprint arXiv:1307.1662*, 2013.

[23] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages", *arXiv Preprint arXiv:1802.06893*, 2018.

[24] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation", In: *Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311-318, 2002.

[25] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", In: *Proc. of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004.

[26] M. Denkowski and A. Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language", In: *Proc. of the Ninth Workshop on Statistical Machine Translation*, pp. 376-380, 2014.