I.J. Image, Graphics and Signal Processing, **2019, 6, 1-8**
Published Online June 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2019.06.01

Attention-Based Deep Learning Model for Image Captioning: A Comparative Study

Phyu Phyu Khaing
University of Computer Studies, Mandalay, Myanmar
Email: phyuphyukhaing@ucsm.edu.mm

May The` Yu
University of Computer Studies, Mandalay, Myanmar
Email: maytheyu@ucsm.edu.mm

Received: 10 March 2019; Accepted: 22 May 2019; Published: 08 June 2019

Abstract—Image captioning is the description generated from images. Generating the caption of an image is one part of computer vision or image processing from artificial intelligence (AI). Image captioning is also the bridge between the vision process and natural language process. In image captioning, there are two parts: sentence based generation and single word generation. Deep Learning has become the main driver of many new applications and is also much more accessible in terms of the learning curve. Image captioning by applying deep learning model can enhance the description accuracy. Attention mechanisms are the upward trend in the model of deep learning for image caption generation. This paper proposes the comparative study for attention-based deep learning model for image captioning. This presents the basic analyzing techniques for performance, advantages, and weakness. This also discusses the datasets for image captioning and the evaluation metrics to test the accuracy.

Index Terms—Attention Mechanism, Deep Learning Model, Image Captioning

I. INTRODUCTION

Image captioning, describing natural language description of images, is still challenges in computer vision. Image captioning has two kinds of approaches: top-down and bottom-up to success the machine translation. The top-down approaches apply the encoder-decoder network architecture (Convolutional Neural Network as an encoder and LSTM as a decoder). It initially takes the image into the encoder to get the feature and the features were fed into the decoder to generate the image description. The bottom-up approaches include several separated tasks, such as identifying objects or attributes, arranging words and sentences, describing sentences using a language model to extract the image caption [1].

Deep learning is also a learning technique for data to encourage the implementation of machine learning that is the function and structure of the brain known as an artificial neural network. Deep learning is also called hierarchical learning or deep structured learning. Neural network architectures of deep learning differ from the original neural network because of more hidden layers, and they can be trained in a supervised and unsupervised method for both supervised and unsupervised learning task. Neural network architecture was typically applying for deep learning. The term "deep" point the number of network layers. Although the neural networks traditionally contain only two or three layers, deep neural networks can contain hundreds. So, more layers are the deeper networks [2, 3, 4, 5, 6, 7, 8, 9].

In recent years, attention is a popular and useful tool for deep learning. Neural Network using Attention Mechanisms are based on the mechanism of visual attention that exists in humans. A standard technique for machine translator and image captioning is an attention mechanism to capture distant dependencies. Attention in the context of deep learning could be embedded in the network. Attention is used for speech recognition, machine translation, image captioning, object identification, reasoning, and summarization [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. Many types of image captioning research have been finished but it has still many biggest challenges.

This research paper is constructed with the following sections. Section 2 presents a critical review that related to this research. Section 3 points the comparison of the previous captioning models. Section 4 shows the steps for image captioning process and section 5 discusses different image captioning models that commonly used by different researchers. Section 6 describes the most famous datasets which have been applying for captioning of image. Different evaluation metrics are examined in section 7. Section 8 summarizes the captioning approaches.

II. RELATED WORK

Cho et al. (2015) describe many contents that used encoder-decoder networks based on attention. This learned based on a convolutional neural network that works on trained attention mechanisms, and gated recurrent neural networks. Bidirectional recurrent neural network (BiRNN) was implemented for an encoder and recurrent neural network learning model (RNN-LM) based on attention was used for a decoder in the neural translation of machine [10]. Xu et al. (2015) proposed two types of image caption generators based on attention: soft deterministic attention and hard stochastic attention. And it attended by visualizing focus on “where” and “what” and quantitatively validated the effectiveness of attention for generating image caption. This operated on Convolutional Neural Network (CNN) as the vectors that extract features from the image as input and Long Short-Term Memory (LSTM) for generating word at every step on context vector [11].

Li et al. (2017) introduced the attention model of deep neural network (DNN) for applying for the recognition of scene text without requiring segmentation of input image. This is the integration to extract feature using CNN, to attend feature and to recognize sequence using LSTM network in a cooperatively trainable network. Furthermore, this proposed the attention model, applied the IIIT5K, SVT, ICDA2003, and ICDAR2013 datasets by comparing previous approaches, and worked with 6.5 million parameters. Results are better than the previous methods published in state-of-the-art research. However, this has some incorrect recognition of text. The future works should be considered to get a more practical representation in scene text image by using a deeper CNN [12].

Fu et al. (2017) described an automatic image captioning system by transforming images into accurate and meaningful sentences. Before generating the words, giving the other ones as the input and then it was arranged to the visual perception experience. An image was encoded with higher-level semantic information by introducing scene-specific contexts. Some benchmark datasets including Flickr8K, Flickr30K, and MSCOCO were used to produce the results by applying both human evaluation and automatic evaluation metrics. The performance of the work was improved either scene-specific context or region-based attention. The combination of two modeling ingredients suggests attaining the achievement of the state-of-the-art as the future scope [13].

Gao et al. (2017) submitted an attention-based LSTM (aLSTMs) for video captioning by applying semantic consistency. This used local two-dimensional convolutional neural network for dynamic weighted sum, LSTM encoder for a visual feature and word-embedding feature extraction, and multimodal embedding for mapping the visual and sentence feature [14]. Bin et al. (2017) developed an adaptive attention strategy for visual captioning with attending salient visual content or linguistic knowledge. In this work, there are two processes: image captioning and video captioning. This applied recurrent neural network as an encoder and LSTM as a decoder, and also applied adaptive attention by designing visual captioning model, linguistic knowledge embedding and learning attributes for visual. This constructs the context for visual attention by using a pre-trained multi-label classifier to control the visual captioning model by applying a visual gate and embed linguistic knowledge from all previous hidden state by using a latent representation. The experimental results show the supercilious of adaptive attention strategy [15].

Qu et al. (2017) propounded a visual attention mechanism applied to long-short term memory for staring the salient object in image for captioning. In this work, CNN is used for extracting features such as colors, size, and location; LSTM is used to generate a sentence, and; attention mechanism is used to describe the important objects in the image. CNN extract features from the image with VggNet. LSTM is work with four gates (input gate, forget gate, attend gate and output gate) and a memory cell. Attention has two aspects: color stimulus-driven and dimension stimulus-driven. The proposed model was validated on three popular benchmark datasets: Flick8k, Flick30k, and MSCOCO and the performance shown by using standard evaluation metrics: BLEU. The proposed model can generate more interpretability sentence and get more accuracy in object recognition. The future work should use unsupervised data to understand comprehensively and precisely about a whole picture [16].

Li et al. (2018) proposed the global-local attention (GLA) method for describing the caption of the image. Features based on object-level integrated with image-level by applying attention mechanism. This used VGG16 for image feature extractor, Faster R-CNN for object detector, attention mechanism for integration of global feature and local feature, and stacked two-layer LSTM for the model of language. The proposed GLA method implemented on Microsoft COCO caption data set by checking with many favored evaluation metrics such as BLEU, ROUGE-L, METEOR, and CIDEr. This can create more appropriate and reasonable sentences that related the image context but cannot jointly the language model and train CNN part. So, the integration of image feature extractor and object detector is still as the future study to train and test of the end-to-end model [17].

Ye et al. (2018) initiated the attentive linear transformation (ALT) for automatically generating image caption as a novel attention framework. That model used Convolutional Neural Network (CNN) for encoding input image to extract features, transformation matrix with high-dimension for converting from image features to context vector and Recurrent Neural Network (RNN) for decoding from the context vector to a sentence that related with the image. This experiments on the benchmark dataset such as MS COCO and Flickr30K by measuring evaluation metrics. ALT’s advantage is that the weight for linear transformation can show information unless a concrete form uses as a featured channel or spatial region. ALT can nicely describe than existing attention models but cannot correctly recognize words on the sign, cannot distinguish some-part-redundant object, cannot correctly count the quality of the object, and mistakes the gender. This paper suggested using text detector to recognize the words and objecting detector to count the quantity of the objects as the future works [18].

Cornia et al. (2018) exploited the salient and contextual region on the results of the saliency prediction model for image captioning by exploiting. This too used a Fully Convolutional Network as an encoder on high-level features, LSTM layer on features and the saliency map, attention mechanism that selects an image region from previous LSTM and supplies it to next LSTM [19]. Wang et al. (2018) proposed Affective Guiding and Selective Attention Mechanism (AG-SAM) as a novel image captioning model. This used the encoder-decoder network: Convolutional Neural Network (CNN) like encoder for extracting features of the image and then long-short-term-memory (LSTM) like decoder for selective attention-based and emotional awareness. This also introduced a gate to generate the context vector with the attention mechanism [20].

Zhu et al. (2018) developed a Captioning Transformer (CT) model by applying stacked attention modules without the time dependencies to address the issues of long-short-term-memory (LSTM) structure and also proposed multi-level supervision training. The encoder of this model is Convolutional Neural Network (CNN) that used ResNet and ResNext as image classification models to extract image features and transformer model with stacked attention mechanism as the decoder to decode from image features to the sentence. There are three methods for integrating image features to transformer model: 1) image spatial feature map, 2) spatial image feature map that merge the image feature and each word, and 3) spatial image feature map that used image feature in front of text embedding. This used MSCOCO dataset and standard evaluation metrics for evaluating the performance by juxtaposition with several start-of-the-art methods. The accuracy of the study is better than the original models. This pointed out to study the method in the digital virtual asset security field for future scope [21].

Wang et al. (2016) initiated deep bidirectional LSTM model that designed for image caption generation. This model is based on a deep CNN and two separate LSTM network for learning long-term interaction between image and text. This caption generation model is evaluated with Flickr8K, Flickr30K, and MSCOCO benchmark datasets. Bidirectional LSTM model achieves high performance on both generation and retrieval tasks. As the future scope, more sophisticated language representation, multitask learning and attention mechanism can extend in the model [29].

Wang et al. (2016) demonstrated the architecture of RNN-LSTM applied parallel-fusion for image captioning by combining the advantages of simple RNN and LSTM. This approach improves the performance and the efficiency by evaluating with BLEU and METEOR on Flickr8K dataset. To focus the higher performance, future work needs to examine the limitation of parallel threads by using more complex image features [30].

Lu et al. (2017) introduced an encoder-decoder framework with adaptive attention for image caption generation. Adaptive attention learns when to attend and where to attend on the image. This framework implements on Flickr30K and 2015 MSCOCO image captioning datasets to analyze the adaptive attention. The framework is efficiently evaluated on the caption that generated from the image, and it can be applied in other application domains [31]. Chen et al. (2017) developed Spatial and Channel-wise Attention in Convolutional Neural Network (SCA-CNN) for image caption generation. In multi-layer feature maps, SCA-CNN concise for the sentence generation by encoding what and where the visual attention is. This is manipulated on three benchmark datasets: MSCOCO, Flickr8K, and Flickr30K. Future work intends to work temporal attention in SCA-CNN, by attending video frames features for video captioning and to increase the attentive layers without overfitting [32].

Gan et al. (2017) initiated a Semantic Compositional Network (SCN) for image caption generation and video clip captioning. SCN detect semantic concepts from the image and use the probability of each task for parameter composition in LSTM. This is quantitatively evaluated and qualitatively analyzed on COCO, Flickr30K, and Youtube2Text datasets; and the performance significantly outperforms with multiple evaluation metrics [33]. Liu et al. (2017) found a quantitative evaluation metric by focusing on evaluating and improving the correctness of attention in neural image captioning. The metric evaluates between human annotations and the generated attention maps by using Flickr30K and COCO datasets. This can close the gap between human perception and machine attention and can experiment in related fields [34].

Gu et al. (2017) exploited the CNN language model for image caption generation. MSCOCO and Flickr30K datasets have been using to conduct the experiments for analysis. The model can generate a sentence that is relevant with image but model is wrong when visual attributes are predicted. It can integrate extra attributes that learning for image captioning as future scope [35]. Wu et al. (2018) proposed a method that integrates the high-level concepts to the CNN-RNN architecture, approved the improvements of the method with image captioning and visual question answering, and also used to incorporate external knowledge in answering for high-level visual questions. Image captioning results reported by testing on the popular Flickr8k, Flickr30k, and Microsoft COCO dataset; and visual question answering tested on the DAQURA-ALL, DAQURA-REDUCED, Toronto COCO-QA, and VQA datasets. Further work should extract more specifically related information by creating queries based on the knowledge that reflect the content of the question and image [36].

Aneja et al. (2018) explored a convolutional image captioning technique, demonstrated its efficacy on the MSCOCO dataset and the performance with baseline. The model with attention can improve performance [37]. Wang et al. (2018) discovered a framework that only employs convolutional neural networks (CNNs) to generate captions. They conduct extensive experiments on MSCOCO and investigate the influence of the model width and depth. Compared with LSTM-based models that apply similar attention mechanisms, our proposed models achieves comparable scores of BLEU-1,2,3,4, METEOR, and higher scores of CIDEr [38].

III. Comparison of Models

Comparison of most prominent papers is shown below in Table 1 based on the comparative study of many research papers.

Table 1. Comparison of major previous research work

Reference	Findings and Limitations
Ref [13], 2017	1. Either scene-specific context or region-based attention can raise the performance of the image captioning method. 2. Scene-specific context can generate new caption from distorted scene image. 3. Region-based attention does not require proposing and selecting visual regions. 4. Region-based attention model needs more data to train.
Ref [12], 2017	1. Deep Neural Network based on attention (DANN) is more powerful for the image that contains most scene text (non-horizontal character, unusual font, or texture background). 2. Image with a complex background and text has a large deformation, are not correctly identified. 3. If ‘o’ and ‘n’ are adjacent, read ‘n’ as ‘m’.
Ref [17], 2018	1. Global-Local Attention (GLA) can a generate sentence description that is more relevant to the input image. 2. GLA can obtain global information as well as local object information. 3. The model cannot connectively train the language model and feature extractor. 4. This cannot fine-tune the Faster R-CNN on Flickr dataset because Flickr datasets lack object information.
Ref [18], 2018	1. Attentive Linear Transformation (ALT) can attend to subtler and more abstract visual concepts. 2. The linear transformation weights can apprehend more information if a concrete form does not like spatial region or feature channel. 3. The words on the sign cannot correctly recognize. 4. Objects are difficult to distinguish and the quantity of object cannot correctly count. 5. The prediction for gender is a mistake.
Ref [19], (2018)	1. Using saliency and context attention can avoid the repetition of words, and the failure to describe the context. 2. Performance loss when the salient regions of the image are not described in the corresponding ground-truth caption. 3. Some problems arise in the presence of complex scenes.

IV. Steps for Image Captioning

Image captioning using deep learning based on attention mechanism contains some important stages such as data preparation, feature extraction, encoder network, attention mechanism, decoder network, and evaluation stages. Data preparation takes images to ready for feature extraction. Feature extraction, encoder-decoder network, and attention mechanism are differently used in previous research work. The most common implemented models, datasets and evaluation metrics discussed in the next sections. The sequence of stages for image captioning is shown in Fig.1.

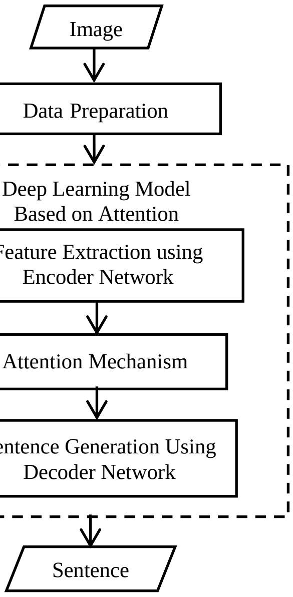
Image
Data Preparation
Deep Learning Model Based on Attention
Feature Extraction using Encoder Network
Attention Mechanism
Sentence Generation Using Decoder Network
Sentence

Fig.1. Stages for Image Captioning

V. Captioning Models

There are many methodologies to utilize for image captioning or description generation. The researchers' groups have commonly implemented little famous architecture. CNN, RNN, DNN, and LSTM are famous architectures and are commonly used for image caption generation processes. A brief overview of the deep learning-based approaches for image captioning is shown in Table 2. The following subsection will introduce about of these techniques.

A. CNN

Convolutional Neural Network (CNN or ConvNet), that has applied in numerous fields of artificial intelligence (AI) just as speech and image recognition, pattern recognition, video analysis, and natural language processing. CNN is a feed-forward neural network in machine learning and one category of deep neural network in deep learning [2-9] and also used as the encoder network for extracting the features of the image in most deep learning model based on attention mechanism [10-21].

B. RNN

Recurrent Neural Network (RNN) is also called the looped based neural network. RNNs can apply their memory (internal state) to perform input sequences, unlike feed-forward neural networks. In image captioning process, RNN is implemented to guess the next word from the current word by learning and thus RNN is also known as the language model as a decoder network [3, 7, 9, 18]. But RNN is implemented for both encoder and decoder network in some image captioning model [10].

Table 2. Overview of deep learning-based approaches for image captioning

Reference	Model for Image	Model for Language
Ref [11], 2015	VGGNet	LSTM
Ref [29], 2016	VGGNet, AlexNet	LSTM
Ref [31], 2017	ResNet	LSTM
Ref [32], 2017	VGGNet, ResNet	LSTM
Ref [33], 2017	ResNet	LSTM
Ref [34], 2017	VGGNet	LSTM
Ref [35], 2017	VGGNet	Language CNN, LSTM
Ref [13], 2017	VGGNet, AlexNet	LSTM
Ref [18], 2018	VGGNet	LSTM
Ref [21], 2018	ResNet and ResNext	LSTM

C. LSTM

Long-and-Short-Term Memory (LSTM) is a component of a recurrent neural network (RNN). An RNN, which developed with LSTM units, is also called an LSTM network. LSTM is mostly composed with four gates: input gate, forget gate, cell, and output gate. LSTM is implemented for sentence representation in image caption generation [4, 6, 8, 9, 11, 12, 13, 15, 16, 19, 20, 21]. LSTM is also applied for feature extraction for image and word [17].

D. Encoder-Decoder Framework

There have been many types of research that represent the encoder-decoder framework. Among them, one is an image encoder by using CNN and a text decoder by applying RNN. In that framework, CNN extracts various visual features from the image and RNN generates the caption of the image. Fig.2 shows the encoder-decoder framework for image captioning [39].

E. Attention Mechanism

Many methods implement a pre-trained CNN model in the encoding of image and use the fixed content of the image in the decoding process to generate the natural language sentence. It, however, have difficulties to convert all important information into one single vector.

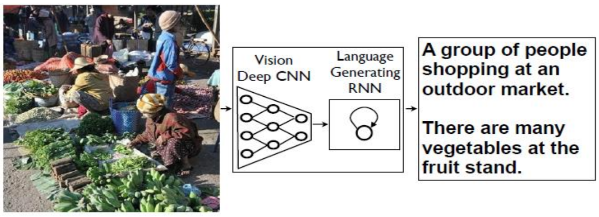
Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

Fig.2. Encoder-decoder framework for image captioning. Figure extracted from [39]

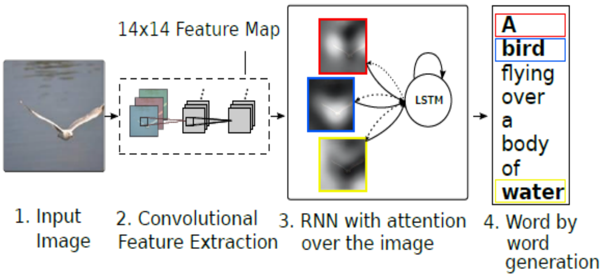
14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Fig.3. Attention Mechanism. Figure adapted from [11]

Attention will be helpful for generating the image caption by extracting the important image regions in accordance with the context of the image. In image captioning, so, the attention mechanism has been widely applied. Attention mechanism studies what and where the decoder should attend to. Fig.3 shows the attention mechanism for generation image caption.

VI. Datasets

There are various kinds of datasets that applied for detection, classification, recognition, and caption generation. The most famous standard benchmark datasets, used for image captioning based on the sentence, are MSCOCO [22], FLICKR8K [23], and FLICKR30K [24], etc. A summary of datasets that most mostly used in image captioning is listed in Table 3.

A. MSCOCO

MSCOCO [22] caption dataset is implemented for image captioning method. MSCOCO [30] caption dataset is comprised of three parts: the training, the testing and the validation. Each image of training and validation is described with five captions but there do not have annotated sentences for the testing set. The dataset focused on the individual instance for object segmentation. The dataset, which released in 2014, have 82,783t images for training 40,775 images for testing and 40,504 images for validation. In 2015, this cumulatively released 165,482 train images, 81,434 testing images and 81,208 validation images. The 2017 release, that is the last, contain 118,287 training images, 40,670 testing images, and 5,000 validation images.

B. Flickr8K

The images for Flickr8K [23] dataset collected from the Flickr.com website and consists of 8,092 images that attend on the performing action of animals or people. Each image in the dataset contains five sentences that characterized with entities (animals, people, and objects), situation, scenes, and events. The images in the dataset are annotated with the test-passed worker for grammar and spelling checking from United State. There have been using 6,000 images to train, 1,000 images to test and 1,000 images to validate.

C. Flickr30K

Flickr30K [24] is a large image description dataset and accommodates with region-phrase correspondence for ground-truth comprehension. This dataset emerges by combining the embedding of image and text, common object detectors, color classifier and bias that select larger objects. There has classified 513,644 mentions for scene and entity and there has been working with five sentences per image. There are 28,000 images for training, 1,000 images for testing and 1,000 images for validation.

Table 3. Summarization for datasets

Reference	**Datasets**
Ref [29], 2016	Flickr8K, Flickr30K, MSCOCO
Ref [30], 2016	Flickr8K
Ref [16], 2017	Flickr8K, Flickr30K, MSCOCO
Ref [31], 2017	Flickr30K, MSCOCO
Ref [32], 2017	Flickr8K, Flickr30K, MSCOCO
Ref [33], 2017	Flickr30K, MSCOCO
Ref [34], 2017	Flickr30K, MSCOCO
Ref [35], 2017	Flickr30K, MSCOCO
Ref [17], 2018	MSCOCO
Ref [18], 2018	Flickr30K, MSCOCO
Ref [21], 2018	MSCOCO
Ref [36], 2018	Flickr8K, Flickr30K, MSCOCO
Ref [37], 2018	MSCOCO
Ref [38], 2018	MSCOCO

VII. Evaluation

The evaluation metrics are commonly applied to automatically manipulate the accuracy and effectiveness of caption generation. The commonly used evaluation metrics are Bilingual Evaluation Understudy (BLEU) [25], Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [26], Metric for Evaluation based Image Description Evaluation (METEOR) [27], and Consensus-based Image Description Evaluation (CIDEr) [28]. A similarity-based measure between ground truth sentence and machines generated sentence is calculated by all of these methods. Each of these evaluation methods is introduced with the following sections. Table 4 shows an overview of the evaluation metrics used in image captioning. And, Table 5 shows the performance comparison of the previous research on MSCOCO dataset.

A. BLEU

BLEU [25] is extensively used for machine translation and it is an automatic human-like evaluation. It is language-independence, speedy, and cheaply evaluation method. The semantic similarity between the human description of the image and machine-generated caption can be determined by applying the BLEU score. It measures n-grams' fraction that is in common between a reference and a hypothesis. The strength of BLEU [33] evaluation metrics highly correlates with the judgments of human by an average of judgment errors of the individual sentence. The judgment over a test corpus is divine rather than the judgment of human for every sentence.

B. ROUGE

ROUGE [26] is an automatic evaluation package for the comparison of the quality of a summary and human-created summaries. It is very effective for automatic evaluation of machine translation. Four different ROUGE measures are ROUGE-L, ROUGE-N, ROUGE-S, and ROUGE-W. ROUGE-L identifies the longest common subsequence (LCS) and it has sentence-level LCS and summary-level LCS. ROUGE-N is a recall-related n-gram measure between a set of reference summaries and a candidate summary. ROUGE-S named skip-bigram co-occurrence statistics and measure the skip-bigram overlapping between a set of reference translations and a candidate translation. ROUGE-W calls the weighted longest common subsequence (WLCS) and uses the polynomial function to calculate.

Table 4. Overview of evaluation metrics

Reference	**Evaluation Metrics**
Ref [29], 2016	BLEU
Ref [30], 2016	BLEU, METEOR
Ref [16], 2017	BLEU
Ref [31], 2017	BLEU, METEOR, CIDEr
Ref [32], 2017	BLEU, METEOR
Ref [33], 2017	BLEU, METEOR, CIDEr
Ref [34], 2017	BLEU, METEOR
Ref [35], 2017	BLEU, METEOR, CIDEr, SPICE
Ref [17], 2018	BLEU, METEOR, CIDEr, ROUGE_L
Ref [18], 2018	BLEU, METEOR, CIDEr, ROUGE_L, SPICE
Ref [21], 2018	BLEU, METEOR, CIDEr, ROUGE_L
Ref [36], 2018	BLEU, METEOR, CIDEr
Ref [37], 2018	BLEU, METEOR, CIDEr, ROUGE_L, SPICE
Ref [38], 2018	BLEU, METEOR, CIDEr, ROUGE_L

C. METEOR

METEOR [27] score has been highly applied in comparison with other metrics because of highly correlated with human subjects' annotations. It can evaluate on any target language to construct the system of statistical translation by applying the same resources. It is freely available as open source software.

D. CIDEr

The goal of CIDEr [28] is to automatically evaluate the image caption. This evaluation metrics show how many matching the consensus of image description sets with a candidate sentence. This is more suitable for the evaluation of image description generation for consensus measuring.

Table 5. Performance comparison of previous research

References	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE_L	SPICE
Ref [29], 2016	67.2	49.2	35.2	24.4	-	-	-	-
Ref [16], 2017	72.3	52.2	37.1	25.2	-	-	-	-
Ref [31], 2017	74.2	58.0	43.9	33.2	26.6	108.5	-	-
Ref [32], 2017	71.9	54.8	41.1	31.1	25.0	-	-	-
Ref [33], 2017	74.1	57.8	44.4	34.1	26.1	104.1	-	-
Ref [34], 2017	-	-	37.2	27.6	24.78	-	-	-
Ref [35], 2017	72.6	55.4	41.1	30.3	24.6	96.1	-	-
Ref [17], 2018	72.5	55.6	41.7	31.2	24.9	96.4	53.3	-
Ref [18], 2018	75.1	59.0	45.7	35.5	27.4	110.7	55.9	20.3
Ref [21], 2018	73.0	56.9	43.6	33.3	-	108.1	54.8	-
Ref [36], 2018	80.0	64.0	50.0	40.0	28.0	107.0	-	-
Ref [37], 2018	71.1	53.8	39.4	28.7	24.4	91.2	52.2	17.5
Ref [38], 2018	68.8	51.3	37.0	26.5	23.4	83.9	50.7	-

VIII. CONCLUSION

This paper presents the comparative study of an attention-based deep learning model for image caption generation. For image captioning, encoder-decoder network based on attention mechanism is very useful to generate sentence description. In the encoder-decoder network, CNN is mostly used for feature extraction and RNN is mostly used for sentence generation. Attention mechanism attends a more salient part of the image in the output of the encoder network and converts feature maps to feature vector for the input of the decoder network. The paper also suggests important steps for image caption generation. And also, this presents datasets that mostly used, and evaluation metrics to calculate the performance.

REFERENCES

[1] You, Quanzeng, et al., In Proceedings of the IEEE conference on computer vision and pattern recognition, "Image captioning with semantic attention", 2016.

[2] D.J. Kim, D. Yoo, B. Sim, and I.S. Kweon, "Sentence learning on deep convolutional networks for image Caption Generation", *Ubiquitous Robots and Ambient Intelligence (URAI)*, 2016 13th International Conference on IEEE, pp. 246-247.

[3] L. Yang and H. Hu, "TVPRNN for image caption generation", *Electronics Letters*, 53(22):1471-1473, 2017.

[4] A. Karpathy, and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 39(4):664-676, 2017.

[5] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 39(11):2298-2304, 2017.

[6] Park, Cesc Chunseong, Youngjin Kim, and Gunhee Kim. "Retrieval of sentence sequences for an image stream via coherence recurrent convolutional networks", *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2018): 945-957.

[7] A. Wu, C. Shen, P. Wang, A. Dick, and A.v.d. Hengel, "Image Captioning and Visual Question Answering Based on Attributes and External Knowledge", *IEEE Trans. on Pattern Analysis And Machine Intelligence*, 40(6):1367-1381, 2018.

[8] A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk, "Article Annotation by Image and Text Processing", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(5):1072-1085, 2018.

[9] S. Shabir, and S.Y. Arafat, "An image conveys a message: A brief survey on image description generation", *2018 1st International Conference on Power, Energy and Smart Grid (ICPESG), IEEE*, pp. 1-6, April 2018.

[10] K. Cho, A. Courville, and Y. Bengio, "Describing Multimedia Content Using Attention-Based Encoder-Decoder Networks", *IEEE Trans. on Multimedia*, 17(11):1875-1886, 2015.

[11] K. Xu, J.L. Ba, R. Kiros, K Cho, A. Courville, R. Salakhudinov, R. S.Zemel and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", *International conference on machine learning*, pp. 2048-2057, 2015.

[12] S. Li, M. Tang, A. Guo, J. Lei, and J. Zhang, "Deep Neural Network with Attention Model for Scene Text Recognition", *IET journals of the institution of Engineering and Technology*, 11(7):605-612, 2017.

[13] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning Where to see and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 39(12):2321-2334, 2017.

[14] L. Gao, Z. Guo, H. Zhang, X. Xu, and H.T. Shen, "Video Captioning With Attention-Based LSTM and Semantic Consistency", *IEEE Trans. on Multimedia*, 19(9):2045-2055, 2017

[15] Y. Bin, Y. Yang, J. Zhou, Z. Huang, and H.T. Shen, "Adaptively Attending to Visual Attributes and Linguistic Knowledge for Captioning", *In Proceedings of the 2017 ACM on Multimedia Conference*, pp. 1345-1353, 2017.

[16] S. Qu, Y. Xi, and S. Ding, "Visual Attention Based on Long-Short Term Memory Model for Image Caption Generation", *Control and Decision Conference (CCDC), 2017 29th Chinese, IEEE*, pp. 4789-4794, May 2017.

[17] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian, "GLA: Global-Local Attention for Image Description", *IEEE Trans. on Multimedia*, 20(3):726-737, 2018.

[18] S. Ye, J. Han, and N. Liu, "Attentive Linear Transformation for Image Captioning", *IEEE Trans. on Image Processing*, 27(11):5514-5524, 2018.

[19] Cornia, Marcella, et al. "Paying more attention to saliency: Image captioning with saliency and context attention", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.2 (2018): 48.

[20] A. Wang, H. Hu and L. Yang, “Image Captioning with Affective Guiding and Selective Attention”, *ACM Trans. Multimedia Comput. Commun.* Appl., 14(3):73, 2018.

[21] X. Zhu, L. Li, J. Liu, H. Peng, and X. Niu, “Captioning Transformer with Stacked Attention Model”, *Applied Sciences*, 8(5):739, 2018.

[22] T.Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, and P. Dollar, “Microsoft COCO: Common Objects in Context”, *European Conference on Computer Vision*, pp. 740-755, 2014.

[23] Hodosh, Micah, Peter Young, and Julia Hockenmaier. “Framing image description as a ranking task: Data, models and evaluation metrics”, *Journal of Artificial Intelligence Research* 47 (2013): 853-899.

[24] Plummer, Bryan A., et al., in Proceedings of the IEEE international conference on computer vision, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models”, 2015.

[25] Papineni, Kishore, et al., in Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, “BLEU: a method for automatic evaluation of machine translation”, 2002.

[26] Lin, Chin-Yew. “Rouge: A package for automatic evaluation of summaries”, *Text Summarization Branches Out* (2004).

[27] M. Denkowshi, and A. Lavie, in Proceedings of the Ninth Workshop on Statistical Machine Translation, “Meteor Universal: Language Specific Translation Evaluation for Any Target Language”, 2014, pp. 376-380.

[28] R. Vedantam, C.L. Zitnick, and D. Parikh, in Proceedings of the IEEE conference on computer vision and pattern recognition, “CIDEr: Consensus-based Image Description Evaluation”, 2015, pp. 4566-4575.

[29] C. Wang, H. Yang, C. Bartz, and C. Meinel, In Proceedings of the 2016 ACM on Multimedia Conference, “Image captioning with deep bidirectional LSTMs”, 2016, pp. 988–997.

[30] M. Wang, L. Song, X. Yang, and C. Luo, “A parallel-fusion RNN-LSTM architecture for image caption generation”, *In 2016 IEEE International Conference on Image Processing (ICIP’16)*, pp. 4448–4452, 2016.

[31] J Lu, Jiasen, et al., in Proceedings of the IEEE conference on computer vision and pattern recognition. “Knowing when to look: Adaptive attention via a visual sentinel for image captioning”, 2017.

[32] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, and T.S. Chua, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’17), “SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning, 2017”, pp. 6298–6306.

[33] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, “Semantic compositional networks for visual captioning”, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’17)*. pp. 1141–1150, 2017.

[34] C. Liu, J. Mao, F. Sha, and A.L. Yuille, “Attention Correctness In Neural Image Captioning”, *In AAAI*, pp. 4176–4182, 2017.

[35] J. Gu, G. Wang, J. Cai, and T. Chen, “An Empirical Study Of Language CNN For Image Captioning”, *In Proceedings of the International Conference on Computer Vision (ICCV’17)*, pp. 1231–1240, 2017.

[36] Q. Wu, C. Shen, P. Wang, A. Dick, and A.v.d. Hengel, “Image Captioning And Visual Question Answering Based On Attributes And External Knowledge”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1367–1381, 2018.

[37] J. Aneja, A. Deshpande, and A.G. Schwing, “Convolutional Image Captioning”, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5561–5570, 2018

[38] A. Wang and A.B. Chan, “CNN+ CNN: Convolutional Decoders For Image Captioning”, *arXiv preprint arXiv*:1805.09019, 2018.

[39] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017.

Authors’ Profiles

Phyu Phyu Khaing is an Assistant Lecturer in the Myanmar Institute of Information Technology, Mandalay, Myanmar. She received her B.C.Sc. degree in 2007, B.C.Sc. (Hons.) in 2008 and M.C.Sc. degree in 2010 from Computer University (Monywa), Myanmar. She is currently working toward the Ph.D. degree in University of Computer Studies, Mandalay, Myanmar. Her research interests include computer vision, digital image processing, and Deep Learning.

May The` Yu received the Ph.D in Information Technology from the University of Computer Studies, Yangon at 2014. She is presently serving as an associate professor, Faculty of Information Science, University of Computer Studies, Mandalay. Her research interest is Image Processing.

How to cite this paper: Phyu Phyu Khaing, May The` Yu, "Attention-Based Deep Learning Model for Image Captioning: A Comparative Study", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.11, No.6, pp. 1-8, 2019.DOI: 10.5815/ijigsp.2019.06.01