International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 4, August 2019, pp. 2932~2940
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i4.pp2932-2940

Natural language description of images using hybrid recurrent neural network

Md. Asifuzzaman Jishan, Khan Raqib Mahmud, Abul Kalam Al Azad
Department of Computer Science and Engineering, University of Liberal Arts Bangladesh, Dhaka-1209, Bangladesh

Article Info

Article history:

Received Sep 26, 2018
Revised Mar 21, 2019
Accepted Apr 4, 2019

Keywords:

Bi-directional recurrent neural network
Long short-term memory
Natural language descriptors
Convolutional neural network
Hybrid recurrent neural network

ABSTRACT

We presented a learning model that generated natural language description of images. The model utilized the connections between natural language and visual data by produced text line based contents from a given image. Our Hybrid Recurrent Neural Network model is based on the intricacies of Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Bi-directional Recurrent Neural Network (BRNN) models. We conducted experiments on three benchmark datasets, e.g., Flickr8K, Flickr30K, and MS COCO. Our hybrid model utilized LSTM model to encode text line or sentences independent of the object location and BRNN for word representation, this reduced the computational complexities without compromising the accuracy of the descriptor. The model produced better accuracy in retrieving natural language based description on the dataset.

Corresponding Author:

Md. Asifuzzaman Jishan,
Department of Computer Science and Engineering,
University of Liberal Arts Bangladesh,
Dhanmondi, Dhaka-1209, Bangladesh.
Email: jishan900@gmail.com

1. INTRODUCTION

A primary motivation of computational visual tasks is to emulate the remarkable human capability to comprehend visual scenes and interpret detailed information from them with astonishing accuracy [1]. For an artificial system to emulate this capability of image description is not merely restricted to recognizing images, rather it is important to understand both syntactic and semantic meaning of the images, that is to say, the task must involve understanding the contents of the image and also the interactions among the contents [2-6]. Image description typically is the generation of natural language based textual description of an image which has been an active area of research [7-12].

Figure 1 depicts an example where the image has been utilized to extract a natural language based single sentence description from the apparent visual information. Here the simple description demonstrates the quite remarkable depth in perception of the image in both syntactical and semantic meaning where apparently the object and spatial contents in the image (e.g., people and street) are connected semantically with the action walking. The content based image interpretation task of this kind is crucial in various practical applications such as automatic image indexing, image-based web-searching, automatic image captioning in news and social media sphere and more importantly in automatic diagnosis of diseases followed by potential automated medical advice generation from biomedical images and so on. To further elucidate the potential applications of automated image description the following motivational examples may be noted: in an image of road with complicated traffic congestion, a focused extraction of visual information might help with simple retrieval features like number of vehicles or type of vehicles or average separation length between vehicles in the image, or in a crowded space if any person poses a threat with aggressive gesture or by

exposing weapons, a simple and quick context retrieval of the image implying immediate potential threat to public would be most desired.

Figure 1. Extraction of a simple natural language description from visual data

Given the scientific and practical importance of the natural language based description of images, it has been a very dynamic research endeavour with tools and techniques of both traditional machine learning and deep machine learning have been brought to bear on achieving expected performance [13-15]. However, restricted scope of the vocabularies for describing visual contents limits the varieties of narratives about a visual space, and the template based image description restricts complex and varied sematic interpretation, though descriptor models can produce grammatically correct texts. Moreover, the growing surge of image and video datasets [16-18] puts up challenging bars against the computational modeling efforts to generate syntactically and semantically viable natural language based description beyond the pre-assumed templates and closed vocabularies.

To seek to circumvent the said limitations in developing a working artificial neural system tool to generate natural language based description of images, a rather complex model is required to yield novel textual description from visual scenes with multimodal complexities. In this study, we align with this approach by implementing a deep learning hybrid image descriptor model concatenating Convolutional Neural Networks (CNN) [19, 20], Long Short Term Memory (LSTM) [21] and Bi-directional Neural Networks (BRNN) [22] models. With this hybrid model our approach is to employ CNN to learn categorical features from images by using softmax classifier followed by the language model LSTM to learn longer patterns typical of natural language based texts, which in turn followed by a BRNN model to learn word representation. This concatenation of image classifier and language models ensures learning of multimodal aspects of image contents along with the related natural language text. Thus, by bi-directional sequencing of images and texts, the deep leaning model along with its recurrent neural networks counterparts learns relation between finer portions of image along with the relevant portion of the sentences. Further, for the learning and execution of the model we have used three benchmark visual datasets for natural language based description, e.g., Flickr8K, Flickr30K, and MSCOCO datasets utilizing the BLEU and METEOR metric [23]. We report achieving significant improvement in the textual retrieval from the datasets in the learning and testing phases by fine-tuning architecture and hyperparameters of the model.

2. METHODOLOGY

2.1. Backend computational model

Neural system is instilled in computational framework to emulate the cognitive functions of human cerebrum in recognizing and processing visual information. It is increasingly a popular computational framework nowadays to extract natural language based description to visual information. There are essentially three imperative parts consisting a Neural System: ANN (Artificial Neural Network), CNN (Convolutional neural system), and RNN (Recurrent Neural Network).

Convolutional Neural Network (CNN) is contained at least one convolutional layers and after that took after by at least one completely associated layers as in a standard multilayer neural system. CNN basically use for image recognition, video analysis system, natural language processing, and many more. In CNN, input layer, convolutional layer, polling layer, fully connected layer, and output layer exist [24] (Figure 2). In input layer there are three measurements and they are width, height and depth. It is a framework of pixel esteem. At that point the convolutional layer existing. A piece of the picture is associated with the following Convolutional layer in light of the fact that if every one of the pixels of the info is associated

with the Convolutional layer. Filter, Kernel, or Feature Detector is a little matrix used for highlights location. After convolutional layer, at that point the pooling layer part exists. Pool Layer plays out a capacity to decrease the spatial measurements of the information, and the computational unpredictability of our model. What's more, it additionally controls overfitting. After pooling layer, fully connected layer part existing and fully connected layers interface each neuron in one layer to each neuron in another layer. The last fully connected layer utilizes a softmax initiation work for characterizing the produced highlights of the information picture into different classes in light of the training dataset and after completing this layer then we get an output [25].

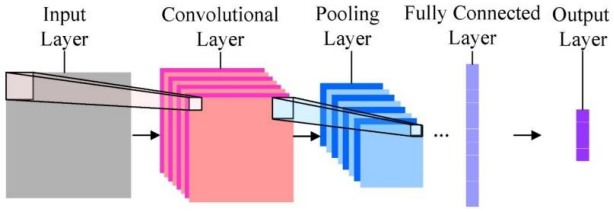


Figure 2. Convolutional neural network (CNN)

Recurrent Neural Network (RNN) is employed to make utilization of consecutive data. Recurrent Neural Network takes the previous output or hidden state as inputs. RNN basically utilized for language demonstrating and creating content, machine translation, speech recognition, generating image description. When it makes a decision, it thinks about the current input state and furthermore what it has gained from the information sources it received already [26]. A recurrent neural network is able to remember processes exactly while a word process running because of it has internal memory. It also predict which character will be come in next and produces output, copies the output and loops it back into the network part. Moreover, in a RNN have two inputs, present and the recent past [27].

Long short-term memory (LSTM) is a special kind of RNN enabled to learn long term dependencies. It is widely used because of its feature of remembering information for long periods of time [28]. This is done by creating special modules that is designed to allow information to be gated-in and gated-out when needed. Unlike traditional RNN, LSTM stores information using a memory cell with linear activation function. The following system of Equation (1) governs the activities of LSTM which includes the design of a memory cell using logistic and linear units with multiple interactions [29].

$$
\begin{aligned}
i_t &= \sigma(w^{(xi)}x_t + w^{(hi)}h_{t-1} + w^{(ci)}c_{t-1} + b^{(i)}) \\
f_t &= \sigma(w^{(xf)}x_t + w^{(hf)}h_{t-1} + w^{(cf)}c_{t-1} + b^{(f)}) \\
c_t &= f_t * c_{t-1} + i_t * tanh\left(w^{(xc)}x_t + w^{(hc)}h_{t-1} + b^{(c)}\right) \\
o_t &= \sigma(w^{(xo)}x_t + w^{(ho)}h_{t-1} + w^{(co)}c_{t-1} + b^{(o)}) \\
h_t &= \sigma_t * tanh(c_t)
\end{aligned}
\qquad (1)
$$

Here *I*, *f*, *o* and *c* are respectively the input, forget, output gate and memory cell activation vectors. Each memory cell c_t has its net input modulated by the activity of an input gate, and has its output modulated by the activity of an output gate. These input and output gates provide a context-sensitive way to update the contents of a memory cell. The forget gate modulates amount of activation of memory cell kept from the previous time step, providing a method to quickly erase the contents of memory cells. Thus the resulting interplay of these gates paves the way to learning of patterns with long dependencies especially featured in the natural languages [30].

2.2. Implementation

2.2.1. Representation

Representing image is most important part for image processing and we get a lot of ideas to review many recent works [9]. We watch that sentence description make visit references to objects and their attributes [31]. The CNN is pre-prepared on ImageNet [16] and finetuned on the 200 classes of the Image Net Detection Challenge [32]. We maintain the technique for Girshick et al. [33] to detect each object in each

image with a Region Convolutional Neural Network (RCNN). The RCNN model has two parts, a region proposal network and another one is binary mask classifier. Following Karpathy et al. [31], we use the primary 19 identified area despite the whole picture. Then compute the representation in light of the pixels I_b inside each bounding box as takes after:

$$v = W_m[CNN\theta_c(I_b)] + b_m \quad (2)$$

The CNN (I_b) changes the pixels inside the bounding box (I_b) to 4096-dimensional enactment of the fully connected layer in a split second before the classifier. The CNN parameters θ_c contain around 60 million parameters. The matrix W_m has measurements h×4096, where h is the extent of the multimodal inserting space. Each image represent as h-dimensional vectors.

Representing sentence is a crucial part of our model. We utilized a Bidirectional Recurrent Neural Network (BRNN) [22] to compute the word representation. Bidirectional Recurrent Neural Network (BRNN) is a part of RNN section and which is use a finite sequence to prediction. In BRNN model, there are label each element of the sequence based on the past and future context element. BRNN conducts this sequencing by close-output of two RNNs and one processing of the sequence is from left to right, the another sequence from right to left. The joined outputs are the prediction of the given target signals. For our model, the BRNN takes a sequence of N words and then it transforms each to h-dimensional vector. Utilizing the list t = 1...N to indicate the situation of a word in a sentence, the exact shape of the BRNN is as per the following:

$$\begin{aligned} x_t &= W_w I_t \\ e_t &= f(W_e x_t + b_e) \\ h_t^f &= f\left(e_t + W_f h_{t-1}^f + b_f\right) \\ h_t^b &= f(e_t + W_b h_{t+1}^b + b_b) \\ s_t &= f(W_d\left(h_t^f + h_t^b\right) + b_d) \end{aligned} \quad (3)$$

The weights W_w determine a word inserting network that we instate with 300-dimensional word2vec [34] weights and keep fixed because of overfitting concerns. I_t is a pointer column vector that has a single one at the record of the t-th word in a word vocabulary. The BRNN comprises of two independent streams of handling, one moving left to right (h_t^f) and the other right to left (h_t^b). We set the activation function f to the rectifier linear unit (ReLU).

2.2.2. Decoding

Decoding considers a picture from the training set and its comparing sentence. We are ultimately interested in producing snippets of content of single words, we might want to align extended, adjacent sequences of words to a single bounding box. We can translate the amount $v^T s_t$ as the unnormalized log likelihood of the t-th word depicting any of the bounding boxes in the image. Note that the naive arrangement that assigns each word freely to the highest scoring locale is lacking in light of the fact that it prompts words getting scattered conflictingly to various regions. We regard the genuine arrangements as inactive factors in a Markov Random Field (MRF) where the binary collaborations between neighboring words urge an arrangement to a similar district. Solidly, given a sentence with N words and a picture with M jumping boxes, we present the inactive arrangement variable $a_j \epsilon 1...M$ for $\boldsymbol{j}$ = 1...N. Here, define a MRF in a chain structure along the sentence as takes after:

$$\begin{aligned} E(a) &= \sum_{(j=1\ldots N)} \psi_j^U\left(a_j\right) + \sum_{(j=1\ldots N-1)} \psi_j^B(a_j, a_{j+1}) \\ \psi_j^U\left(a_j = t\right) &= u_t^T s_t \\ \psi_j^B\left(a_j, a_{j+1}\right) &= \beta I[a_j = a_{j+1}] \end{aligned} \quad (4)$$

Here, β is a hyperparameter that controls the partiality towards longer word phrases. This parameter enables us to introduce between single-word arrangements (β = 0) and adjusting the whole sentence to a solitary, maximally scoring area when β is extensive. The yield of this procedure is a set of image areas explained with fragments of content.

2.2.3. Optimization

We utilize SGD with mini batch of 100 picture sentence sets furthermore, speed of 0.9 to optimization to the alignment model. We cross-approve the learning rate and the weight rot. We likewise utilize dropout regularization in all layers with the exception of in the recurrent layers [35] and clip gradient element wise at 5 (essential). The generative RNN is harder to optimization because of the word frequency difference between uncommon words and common words. We accomplish the best outcomes utilizing RMSprop [36]. That method is a versatile advance size strategy that scales the refresh of each weight by a running normal of its gradient standard.

3. SIMULATION

3.1. Dataset

We utilize the Flickr8K [17], Flickr30K [23] and MSCOCO [18] datasets for our experiment. Flickr8K dataset contain 8,000, Flickr30K dataset contain 31,000 and MSCOCO dataset contain 123,000 images. For Flickr8K and Flickr30K dataset, we utilize 1,000 pictures for validation, 1,000 for testing and the rest pictures for training. For MS COCO, we utilize 5,000 images for validation and testing both parts. We use NVIDIA G1 GAMING GPU for train the dataset.

3.2. Data preprocessing

We preprocess our dataset before training task. We convert all sentences of our dataset to lower case, discard non-alphanumeric characters. We filter words which is occur 5 times in the training set, which result in 2538 words for Flickr8K, 7414 words for Flickr30K, and 8791 words for MSCOCO dataset.

3.3. Image processing

We resized the images of all our datasets to ensure better generality and to avoid any numerical inconsistency during training and testing phases. We use raw image files of each dataset alongside JSON file and VGG CNN features for our three benchmark dataset Flickr8K, Flickr30K, and MSCOCO. The input is a dataset of images and 5 sentence descriptions which were collected with Amazon Mechanical Turk. In particular, this code base is set up for Flickr8K, Flickr30K, and MSCOCO datasets. In the training section, all of images are fed as input to RNN and RNN asked to predict the word of the sentences. For the prediction part, images are passed to RNN and RNN generates the sentence word at a time and we get result of our evaluation with BLEU and METEOR scale.

We use json, datetime, pickle, math, caffe, numpy, scipy, tensorflow, code, socket, argparse, os, and time library for our image to text generation work. We also use vgg_feats.mat which is a .mat file and that stores the CNN features. We use 512 hidden layers and from imagernn.data_provider use getDataProvider for this project. We also involve solver, decode generator, eval_split from the imagernn.data_provider. We also use imread, imresize for image resizing or reshaping. After completing resize of images, then we attempt to train the whole dataset. As regards to the computational duration, Flickr8K takes 1 day, Flickr30K takes 10 days, and MSCOCO takes 24 days to complete the training of whole dataset.

4. RESULTS

We investigate the ability of the working hybrid deep learning model by exploring how well it can generate realistic description of the test images. We trained our model to learn the relation between finer portions of image along with the relevant portion of the sentences. We present the BLEU and METEOR score to assess the performance of our model. These techniques allow us to compute a score the measures how sensible is the image descriptions. The intuition is to measure how close the model generated sentence matches with any of the five reference sentences provided with the dataset. We report these evaluation metrics of our model and present a comparison with other state-of-the-artresults.

We train our model on Flickr8K and Flickr30K datasets and observe the evaluation of full image predictions on 1000 test images. The BLEU-1, 2, 3, 4 evaluation scores and METEOR metric score are assessed and a comparison of the results with other state-of-the-art results is delineated in Table 1 and Table 2. For the experiments, 1,000 images from the datasets are used for testing and validation purpose and the rests for the training purpose. Here in the Tables, (-) indicates an unknown metric of this dataset result.

From the experiment of training our model on MSCOCO dataset, we study the evaluation of full image predictions on 5,000 test images. The BLEU-1, 2, 3, 4 evaluation scores and METEOR metric score are assessed and a comparison of the results with other state of the art results is delineated in Table 3. For this experiment, 5,000 images from the datasets are used for testing and validation purpose and the rests are used for the training purpose. Here in the Table 3, (-) also indicates an unknown metric of this dataset result.

Table 1. BLEU scores and METEOR score for Flickr8K dataset

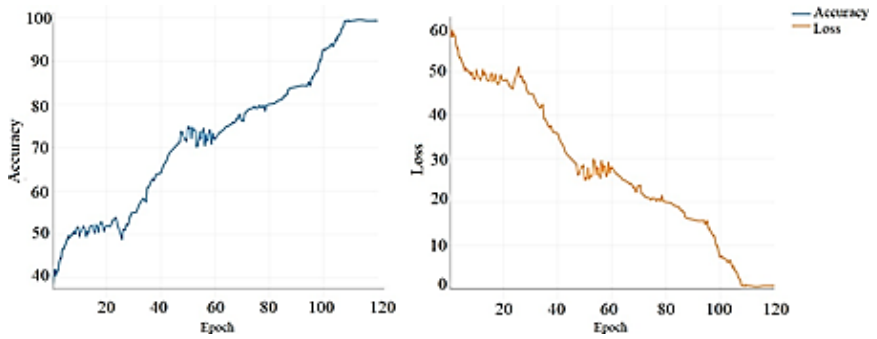
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8K	Mao et al. [37]	58	28	23	-	-
	Google NIC [2]	63	41	27	-	-
	LRCN [38]	-	-	-	-	-
	MS Research [39]	-	-	-	-	-
	Chen and Zitnick [40]	-	-	-	14.1	-
	Hybrid RNN Model	52.6	34.4	21.8	14.1	16.495543

Table 2. BLEU scores and METEOR score for Flickr30K dataset

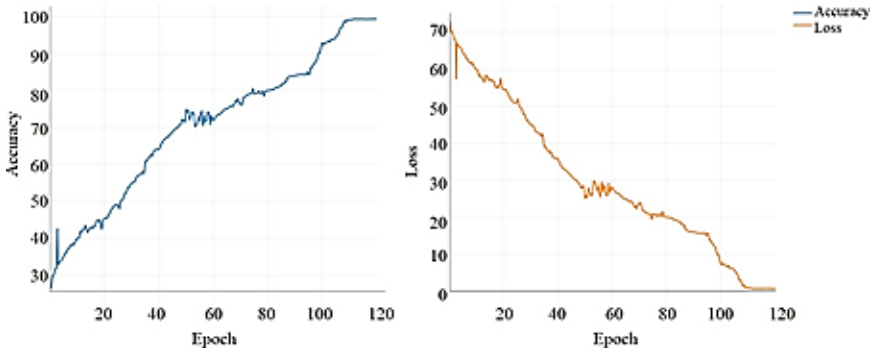
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr30K	Mao et al. [37]	55	24	20	-	-
	Google NIC [2]	66.3	42.3	27.7	18.3	-
	LRCN [38]	58.8	39.1	25.1	16.5	-
	MS Research [39]	-	-	-	-	-
	Chen and Zitnick [40]	-	-	-	12.6	-
	Hybrid RNN Model	56.8	37.3	24.1	15.6	19.441452

Table 3. BLEU scores and METEOR score for MSCOCO dataset

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
MSCOCO	Mao et al. [37]	-	-	-	-	-
	Google NIC [2]	66.6	46.1	32.9	24.6	-
	LRCN [38]	62.8	44.2	30.4	-	-
	MS Research [39]	-	-	-	21.1	20.7
	Chen and Zitnick [40]	-	-	-	19.0	20.4
	Hybrid RNN Model	64.4	45.4	30.9	21.2	19.613227

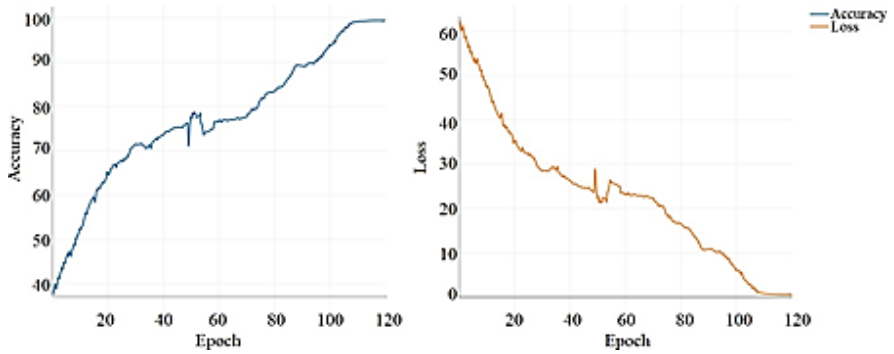


(a)



(b)

Figure 3. Graphical representation of training time using three benchmark dataset (epoch vs. accuracy and epoch vs. loss), (a) Epoch vs. loss, and Epoch vs. accuracy for Flickr8k, (b) Epoch vs. loss, and Epoch vs. accuracy for Flickr30k



(c)

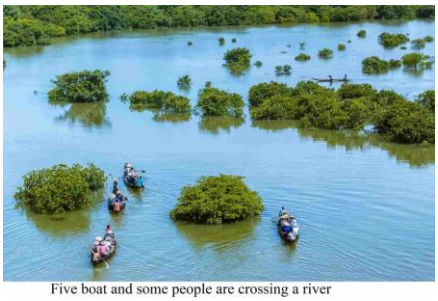
Figure 3. Graphical representation of training time using three benchmark dataset (epoch vs. accuracy and epoch vs. loss), (c) Epoch vs. loss, and Epoch vs. accuracy for MSCOCO

We train our hybrid deep learning model that ensures learning of multimodal aspects of image contents along with the related natural language text. We represent the training result of each dataset using in different graph. In Figure 3, we represent the training accuracy vs. epoch and loss vs. epoch in different graph. This shows that our model does not encounter overfitting problem.

4.1. Discussion

We train our hybrid model to generate sentences on full images where we use CNN to learn categorical features from images and a language model to assist the mapping from image data to sequences of words, followed by a BRNN that learns the word representation. This concatenation of image classifier and language models ensures learning of multimodal aspects of image contents along with the related natural language text. We observe that our hybrid model can generate reasonable descriptions of images as shown in Figure 4(a) even for relatively small or rare objects refer Figure 4(b) which is a significant improvement in the textual retrieval from the images. For the learning and testing phase of our model we have used three benchmark visual datasets for natural language based description, e.g., Flickr8K, Flickr30K and MSCOCO datasets and we have reported the BLEU and METEOR scores for the comparison. Compared to the other state of the art model, our model shows the better performance or comparable to them, as our model fine-tunes the architecture and hyperparameters of the model, results in Table 1-3.

(a) For each test picture, we got the most perfect test sentence



(b) We got the absolute best test sentence for test image

Figure 4. Example of sentence predicted by our model. For every test image, we got the most compatible test sentence

We evaluated the BLEU-1, 2, 3, 4 scores and METEOR scores and compared our results with the benchmark results of Mao et al. [37], Google NIC [2], LRCN [38], MS Research [39], and Chen and Zitnick [40] model. For BLEU-1 score, it is observed that for Flickr30k our model gives better accuracy than the Mao et al. model [37] and for MSCOCO dataset we get better result than the LRCN model [38]. Secondly, in BLEU-2 evaluation, our model gives better result for all three benchmark datasets. For BLEU-3 evaluation, we get better performance for Flickr30k and MSCOCO which is better than the Mao el al. model [37] and LRCN [38] model respectively. For BLEU-4 score, Flickr30k and MSCOCO give better

performance compared with MS Research [39], and Chen and Zitnick [40] model. Finally, we use METEOR evaluation and get 16.495543, 19.441452 and 19.613227 for the benchmark datasets respectively and observe improvements in our results. One limitations of other model is that they are unable to generate different pattern of sentence realizations as the datasets consists of handmade annotations, but our model can generate dynamic output as our model learns to modulate the magnitude of the region and word embedding.

In spite of the fact that our outcomes are encouraging, the model of Multimodal RNN (Recurrent Neural Network) has different type of limitations. First of all, this Multimodal RNN model can only generate a description or sentence of only one input array and that array of pixels at a fixed resolution. Another sensible approach is to use multiple saccades identify the all of entities around the image and their common collaborations and more extensive setting before producing a description. Also, the RNN (Recurrent Neural Network) can receive the information of all images only through additive bias interactions which are less expressive than more complicated multiplicative interactions.

5. CONCLUSION

We study in this paper a complex hybrid neural network model which shows remarkable ability to generate natural language based single sentence description from a given test image. The model identifies the image region and generates natural language description of images. Our approach includes a lowering of resolution images that adjusted parts of visual and language modalities through the interplay of deep convolution learning model with its efficient LSTM and BRNN counterparts. Moreover, we obtain better performance compared to benchmark results by earlier attempts. We report performance results with appropriate representation along with complementary illustrations for better understanding. Our exploration of the model infers that better performance across widening range of datasets may be achieved via model fine-tuning and architectural augmentation.

REFERENCES

[1] L. Fei-Fei, et al., "What do we perceive in a glance of a real-world scene?' *Journal of vision,* vol/issue: 7(1), pp. 10, 2007.
[2] O. Vinyals, et al., "Show and tell: A neural image caption generator," *arXiv: 1411.4555v2*, 2015.
[3] O. Vinyals, et al., "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *arXiv: 1609.06647v1*, 2016.
[4] S. Venugopalan, et al., "Captioning images with diverse objects," *arXiv: 1606.07770v3*, 2017.
[5] L. J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," *ICCV*, 2007.
[6] L. J. Li, et al., "Towards total scene un-derstanding: Classification, annotation and seg- mentation in an automatic framework," *Computer Vision and Pattern Recognition, IEEE Conference on,* pp. 2036-2043, 2009.
[7] S. Fidler, et al., "A sentence is worth a thousand pixels," *CVPR,* 2013.
[8] A. Gupta and P. Mannem, "From image annotation to image description," Neural information processing, Springer, 2012.
[9] G. Kulkarni, et al., "Baby talk: Understanding and generating simple image descriptions," *CVPR*, 2011.
[10] P. Kuznetsova, et al., "Collective generation of natural image descriptions," *ACL*, 2012.
[11] P. Kuznetsova, et al., "Treetalk: Composition and compression of trees for image descriptions," *Transactions of the Association for Computational Linguistics*, vol/issue: 2(10), pp. 351-362, 2014.
[12] A. Farhadi, et al., "Every picture tells a story: Generating sentences from images," *ECCV*, 2010.
[13] S. Bai and S. An, "A Survey on Automatic Image Caption Generation," *Neurocomputing*, 2018.
[14] R. Bernardi, et al., "Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures," *Journal of Artificial Intelligence Research (JAIR)*, vol. 55, pp. 409-442, 2016.
[15] A. Kumar and S. Goel, "A survey of evolution of image captioning techniques," *International Journal of Hybrid Intelligent Systems Preprint*, pp. 1-19, 2017.
[16] J. Deng, et al., "Imagenet: A large-scale hierarchical image database," *CVPR*, 2009.
[17] M. Hodosh, et al., "Framing image description as a ranking task: data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, 2013.
[18] T. Y. Lin, et al., "Microsoft coco: Com- mon objects in context," *arXiv preprint arXiv: 1405.0312*, 2014.
[19] Y. LeCun, et al., "Gradient- based learning applied to document recognition," *Proceed- ings of the IEEE*, vol/issue: 86(11), pp. 2278-2324, 1998.
[20] A. Krizhevsky, et al., "Imagenet classification with deep convolutional neural networks," *NIPS*, 2012.
[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol/issue: 9(8), pp. 1735-1780, 1997.
[22] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *Signal Processing, IEEE Transactions*, 1997.
[23] P. Young, et al., "From image descriptions to visual denotations: New simi- larity metrics for semantic inference over event descriptions," *TACL*, 2014.

[24] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *arXiv: 1408.5882v2,* 2014.
[25] D. Cires Ã., et al., "Multi-column Deep Neural Networks for Image Classification," *arXiv: 1202.2745v1,* 2012.
[26] J. Chun, et al., "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *arXiv: 1412.3555v1,* 2014.
[27] Y. Fan, et al., "TTS Synthesis with Bidi rectional LSTM based Recurrent Neural Networks," *Conference of the International Speech Communication Association*, 2014.
[28] J. Song, et al., "LSTM-in-LSTM for generating long descriptions of images," *Computational Visual Media,* 2016.
[29] Z. C. Lipton, et al., "A Critical Review of Recurrent Neural Networks for Sequence Learning," *arXiv: 1506.00019v4,* 2015.
[30] J. Oh, et al., "Action-Conditional Video Prediction using Deep Networks in Atari Games," *arXiv: 1507.08750v2*, 2015.
[31] A. Karpathy, et al., "Deep fragment embeddings for bidirectional image sentence mapping," *arXiv preprint arXiv: 1406.5679*, 2014.
[32] O. Russakovsky, et al., "Imagenet large scale visual recognition challenge," *arXiv: 1409.0575v3*, 2015.
[33] R. Girshick, et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," *CVPR*, 2014.
[34] T. Mikolov, et al., "Distributed representations of words and phrases and their compositionality," *NIPS*, 2013.
[35] W. Zaremba, et al., "Recurrent neural network regularization," *arXiv preprint arXiv: 1409.2329*, 2014.
[36] T. Tieleman and G. E. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," 2012.
[37] J. Mao, et al., "Explain images with multimodal recurrent neural networks," *arXiv preprint arXiv: 1410.1090*, 2014.
[38] J. Donahue, et al., "Long-term recurrent convolutional networks for visual recognition and description," *arXiv preprint arXiv: 1411.4389,* 2014.
[39] H. Fang, et al., "From captions to visual concepts and back," *arXiv preprint arXiv: 1411.4952*, 2014.
[40] X. Chen and C. L. Zitnick, "Learning a recurrent visual representation for image caption generation," *CoRR, abs/1411.5654*, 2014.

BIOGRAPHIES OF AUTHORS

Md. Asifuzzaman Jishan is studying as a student of Bachelor of Science in Computer Science and Engineering within the Department of Computer Science and Engineering at the University of Liberal Arts Bangladesh (ULAB). He has expertise in C, Java, Python, MATLAB and C++ programming language. He has also working knowledge in different web programming language: HTML, CSS, JavaScript (JS), Laravel framework, and database system. He has been active in the research with research interest in the area of image processing, artificial intelligence, machine learning and neural system.

Khan Raqib Mahmud currently working as a lecturer within the department of Computer Science and Engineering at the University of Liberal Arts Bangladesh (ULAB). He has completed Bachelor of Science (Honors) and Master of Science in Mathematics from Shah Jalal University of Science and Technology, Bangladesh. He received an Erasmus Mundus Scholarship from the Education, Audiovisual and Culture Executive Agency of the European Commission, to pursue a double Masters in Science degree in Computer Simulation for Science and Engineering and Computational Engineering, from Germany and Sweden. He was an MSc thesis student within the Computational Technology Laboratory of the Department of High Performance Computing and Visualization at KTH Royal Institute of Technology, Sweden. His research work concentrated on the study of the sensitivity analysis of Near Wall Turbulence Modeling of Incompressible Flows. His current research interest includes machine learning and pattern recognition, image processing and computer vision and adaptive dynamic system.

Abul Kalam al Azad received his PhD in Applied Mathematics from University of Exeter, United Kingdom, Masters of Science in Theoretical Physics and Bachelor of Science in Physics from University of Dhaka. He is currently an Associate Professor at the Department of Computer Science and Engineering, University of Liberal Arts Bangladesh (ULAB). Previously, he undertook post-doctoral research at Department of Computing and Mathematics, University of Plymouth, United Kingdom, and School of Biological Sciences, University of Bristol, United Kingdom, on a BBSRC fellowship. His research interest includes areas of theoretical and computational neuroscience, con- nectomics, multi-timescale dynamics, self-organized criticality (SOC) and artificial intelligence. He has published a number of papers in peer-reviewed international journals and presented original research articles in numerous international conferences. He received various scholarships, research and travel grants as recognition of his reach work. He was a member of MNN, OCNS and SIAM.