Indonesian Journal of Electrical Engineering and Computer Science
Vol. 21, No. 2, February 2021, pp. 757~767
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v21.i2.pp757-767

Bangla language textual image description by hybrid neural network model

Md. Asifuzzaman Jishan[1], Khan Raqib Mahmud[2], Abul Kalam Al Azad[3], Mohammad Rifat Ahmmad Rashid[4], Bijan Paul[5], Md. Shahabub Alam[6]
[1,2,3,4,5]Department of Computer Science and Engineering, University of Liberal Arts Bangladesh, Dhaka, Bangladesh
[6]Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh

Article Info

Article history:

Received Jul 18, 2020
Revised Sep 20, 2020
Accepted Oct 4, 2020

Keywords:

Bangla natural language descriptors
Convolutional neural network
Hybrid recurrent neural network
Long short-term memory bi-directional recurrent neural network

ABSTRACT

Automatic image captioning task in different language is a challenging task which has not been well investigated yet due to the lack of dataset and effective models. It also requires good understanding of scene and contextual embedding for robust semantic interpretation of images for natural language image descriptor. To generate image descriptor in Bangla, we created a new Bangla dataset of images paired with target language label, named as Bangla natural language image to text (BNLIT) dataset. To deal with the image understanding, we propose a hybrid encoder-decoder model based on encoder-decoder architecture and the model is evaluated on our newly created dataset. This proposed approach achieves significance performance improvement on task of semantic retrieval of images. Our hybrid model uses the convolutional neural network as an encoder whereas the bidirectional long short term memory is used for the sentence representation that decreases the computational complexities without trading off the exactness of the descriptor. The model yielded benchmark accuracy in recovering Bangla natural language and we also conducted a thorough numerical analysis of the model performance on the BNLIT 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 fundamental motivation of computational visual tasks is to imitate the remarkable capability of human to cognize and comprehend visual information with astonishing speed and accuracy. For an artificial framework to emulate this ability of image description is not simply confined to perceiving images, rather it is imperative to comprehend both syntactic and semantic importance of the images, in other words, the undertaking must include understanding the substances of the picture as well as the communications among the substances [1-5]. Image description is essentially the language based textual description of an image, which has been an active field of research in computer vision and natural language processing [6-13]. Image captioning has drawn a lot of interest of the researchers because of its many practical applications, such as text based image search, image curation, assisting of visual impaired individuals to better understand the real world, image understanding in social media, etc.

Where most of the studies of image caption generation are in English language, we focus on generating caption in another language: Bangla (To the 'Bengali' speaking people, the language is mainly known as 'Bangla'). Demographically, Bangla is one of the most widely spoken languages. It is spoken by in excess of 210 million individuals as a first or second language, with somewhere in the range of 100 million Bengali speakers in Bangladesh, around 85 million in India, mainly in the regions of West Bengal, Assam, and Tripura, and sizable migrant networks in the United Kingdom, the United States, and the Middle East. Given the recent advances in natural language processing, this study aims at generating Bangla-textual captions of contextual images to the serve the Bangla-speaking community.

The motivational Figure 1 portrays a case of model generated image-captioning, where the image has been used to extricate a natural language based single sentence depiction from the clear and visual data. Here the straightforward captioning shows the very exceptional profundity in view of the image in both grammatical and semantic significance, where the item and spatial substance in the image (e.g. people and road) are associ- ated semantically, and concurrent with the activity of standing together. The perception of saliency into images could be culture dependent so it is necessary to generate captions in different languages, which is referred to as the cross lingual image captioning [13, 14].

Figure 1. Extraction of a basic common language portrayal from visual information

To create an image captioning model, one of the main challenges is to create a dataset in the target language. So, we first build a new target language dataset, named BNLIT dataset, of a reasonable size by annotating each image with a single annotation and refining these annotations through experts. With the best of our knowledge, the dataset of an image to Bangla caption generation is not available in the public literature. Given the logical and functional significance of the common language based depiction of images, it has been a unique study employing both traditional and deep machine learning methods for accomplishing expected outcome. Furthermore, the ever-growing number of image and video datasets raise testing bars against the computational endeavors to produce linguistically and semantically viable natural language based portrayal, limited by templates and closed vocabularies.

In order to build an image caption generation model, it is imperative to improve the visual relevancy of image descriptor of an image, i.e., how well the model understands the image context and then how efficiently it generates descriptive sentences, which is coherent with the image content. It is also important to consider how contextual semantic embedding can be adapted to different scenarios of an image. In order to circumvent these complexities in captioning task, we propose a hybrid encoder-decoder model, and the challenging part in the encoder-decoder architectures is to design the interface that controls the information flow between applied CNN [14], long short term memory (LSTM) [15] and bi-directional neural networks (BRNN) [16, 17] model constructs.

So the main contribution of the paper are 1) creating a target language dataset; 2) building a Bangla caption generation model based on a hybrid encoder-decoder model, and 3) experimenting successfully with the proposed model on tasks of semantic retrieval of images. The full version of BNLIT dataset has been already uploaded and published in four different dataverse [18].

2. RELATED WORK

In the computer vision, image classification, and image to text generation research dataset play a critical role. The production of ground truth stereo and optical stream datasets [19, 20] animated a surge of enthusiasm for these territories. The early advancement of item acknowledgment datasets [21-23] encouraged the immediate examination of several picture acknowledgment calculations while at the same time pushing the field towards increasingly complex issues. There are different datasets existing e.g. Flickr8K, Flickr30K, MS COCO, ImageNet for the image processing sector. As of late, the ImageNet dataset [24] containing a great many pictures has empowered leaps forward in both item arrangement and recognition investigate utilizing another class of profound learning calculation [24-26].

Image Classification The main focus of image classification is to identify objects from the images. Early datasets of this sort included pictures containing a solitary article with clear foundations, for example, the MNIST written by hand digits [24] or COIL family unit objects. Caltech 101 [21] and Caltech 256 [22] denoted the change to additional reasonable article pictures recovered from the web while additionally expanding the number of item classes to 101 and 256, separately. Most popular and available image datasets in the AI, machine learning, and deep learning people group because of the bigger number of preparing models, CIFAR- 10 and CIFAR-100 offered 10 and 100 classes from a dataset of modest 3232 pictures [26, 27]. As of late, ImageNet [28] made a striking takeoff from the gradual increment in dataset sizes.

Image to Text Generation Generate text from the given input image is the main focus on image to text generate. Since the paper of AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks [28] and ChatPainter: Improving Text to Image Generation using Dialogue [29], we see that they were mainly focused on CNN features of their paper. Meanwhile, in the event that we talk about the paper of Grounded Compositional Semantics for Finding and Describing Images with Sentences, Exploring Models and Data for Image Question Answering [28-30] and DenseCap: Fully Convolutional Localization Networks for Dense Captioning [27], we see that they dependent on DT-RNN demonstrate for producing content from the picture district. They likewise centered on utilizing a semantic implanting framework and demonstrated how a neural system can function and distinguish pictures area. They utilized the dataset of COCO-QA, DAQUAR. They utilized the VGG-16 design for its cutting-edge Performance but, their result of the model was so poor and that was only 0.27. These papers are state-of-the-art for our work [30].

3. DATASET

Image collection is the most important, popular for various significant applications and also challeng- ing. We created a new dataset which name is BNLIT and that dataset contains 8,743 images. We choose Bangladesh perspective images for create a new dataset. Flickr8K, Flickr30K, and MS COCO datasets are containing western cultural images but we use only our countrys cultural image. For create a new dataset, main challenge is collect data or images from various sources. We choose Bangladesh perspective images for creating this new dataset thats why we collect images from the villages, rivers, humans, animals, shops, cow, dog, field, station and many more. We collect those images from different sources like as cell phone gallery, camera, university gallery, marriage function, tour images and also internet sources.

Annotation is another important part for this Bangla dataset. We give one annotation for each image and that caption language is Bangla. Image annotation is that the method by that an automatic data processing adaptive dynamic programming (ADP) system mechanically assigns data within the variety of captioning to a digital image. This application of laptop vision techniques is employed in image retrieval systems to arrange and find pictures of interest from information. These techniques are often considered a sort of multi category image classification with an awfully sizable amount of categories with the vocabulary size. Typically, image analysis within the variety of extracted feature vectors and also the coaching annotation words square measure utilized by machine learning techniques to aim to mechanically apply annotations to new pictures.

BNLIT is containing 8,743 images with different types of image classification. At first, we need to classification whole dataset. For classification, we use 30 types of classes and they are cat, horse, dog, house, cow, window, village, human, town, chair, table, board, spoon, cake, mirror, bottle, pen, pencil, book, car, truck, sky, train, bus, aero plane, bird, tree, fish, water, flower. If dataset size is larger, then machine get better learning and give better accuracy. In our country, it is difficult to collect images for train up the system. We use one sentence for each image. After collect and annotate of each image, need to resize all images of whole dataset. There are a huge number of images in our dataset and they are staying different pixels. So, before training, we resized and set same pixels of whole dataset. We write a script in python which is resizing all images of dataset and save a new directory.

Early stuff datasets centered on texture classification and had straightforward pictures fully coated with one rough patch. Each dataset has a specific number of images and classes. In the Table 1, we compared

our dataset with the other existing dataset with respectively classes and image number. In MSRC dataset, containing 591 images with 21 classes and KITTI dataset containing 203 classes. In another side CamVid and SIFT FLOW containing 700 and 2,688 classes respectively. We use 30 classes for 8,743 images in our BNLIT dataset.

Table 1. Overview of datasets with classes

Dataset	Images	Classes	Year
MSRC [31]	591	21	2006
KITTI [32]	203	14	2012
CamVid [33]	700	32	2008
SIFT Flow [34]	2,688	15	2009
Barcelona [35]	15,150	31	2010
ADE20K [36]	25,210	2,693	2017
BNLIT [18]	8,743	30	2019

4. HYBRID ENCODER - DECODER MODEL

The Neural System for the interpretation and handling of visual data is incorporated into calculative frameworks to copy the subjective elements of human brain. There are basically three basic parts comprising a Neural System: convolutional neural network (CNN), long short-term memory (LSTM), and Bi-directional recurrent neural network (BRNN) models. We illustrated of our implemented model in the Figure 2.

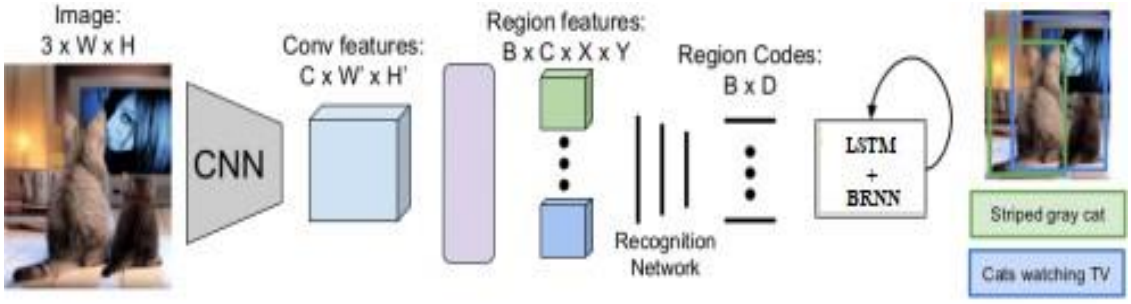

Figure 2. Overview of our proposed model. First of all, an input image processed by CNN. After that, these regions are processed with a fully-connected recognition network and described with a BRNN and LSTM language model. The model is trained end-to-end with stochastic gradient descent

Convolutional Neural Network is an important part of image processing and classification of images using neural networks. In the architecture of a CNN, input layer, convolutional layer, polling layer, fully connected layer and output layer exist [9-12]. In input layer there are three measurements and they are width, height and depth. 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. 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. To extend, 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 [7, 25].

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 [4]. 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 [5, 6]. The LSTM has the capacity to evacuate or add data to the cell state, painstakingly managed by structures called gates. Gates are an approach to alternatively let data through. They are made out of a sigmoid neural net layer and a pointwise multiplication activity [7].

Bi-directional recurrent neural network (BRNN) is a part of RNN and in BRNN demonstrate, there are mark every component of the arrangement in view of the past and future setting component. BRNN conducts this sequencing by close-yield of two RNNs and one handling of the grouping is from left to right, the another arrangement from right to left. It always can avoid gradient vanishing problem which is a common problem for normal RNN model [7, 8].

5. SIMULATION

5.1. Image processing

For the image processing section, at first, we resize the full dataset images to confirm that each images staying in the same pixels. The images of the dataset are without a doubt shading images with pixel esteems running from 0 to 255 with a component of 224 x 224, so before feed the information into the model, it is vital to preprocess it. Firstly, we classify full dataset using CNN and VGG16 features. We do the classification for 30 classes. BRNN mainly use for generating text from the given input images. Finally, we combine the both model of CNN, LSTM, and BRNN features of our dataset and train up full model. Then we take attempt to evaluate our trained model for these datasets to get better result.

5.2. Implementation

Representing image is most vital half for image process and those we get tons of concepts to review several recent works [22]. We have a tendency to watch that sentence description build visit references to things and their attributes [23]. The CNN is pre-prepared on ImageNet [24, 25] and finetuned on the two hundred categories of the ImageNet Detection Challenge [26]. We have a tendency to maintain the technique to discover every object in every image with a part region-based convolutional neural network (RCNN). Following the paper [7], we have a tendency to use the first nineteen known space despite the total images pixel using bounding box as takes after :

$$v = Wm[CNN\theta c(Ib)] + bm \quad (1)$$

The CNN (Ib) changes the pixels inside the bounding box (Ib) to 4096-dimensional establishment of the completely associated layer in a brief moment before the classifier. The CNN parameters θc contain around 60 million parameters. The framework Wm has estimations h 4096, where h is the degree of the multimodal embeddings space. Each picture speak to as h-dimensional vectors.

Representing sentence is also a crucial part of our research. We have a tendency to use a BRNN [6, 7] to cypher the word illustration. BRNN could be a part of RNN section and that is use a finite sequence to prediction. In BRNN model, there are label every component of the sequence supported the past and future context component. For our model, the BRNN takes a sequence of N words and so it transforms every to h-dimensional vector.

5.3. Optimization

We used stochastic gradient descent (SGD) to optimize the CNN part with a mini batch of 16 frame sentence sets. We are using learning rate 0.01, decay rate 1e-6, momentum=0.9, nesterov = True. We cross-approve of the learning rate and the weight of rot. We also use dropout regularization in all layers except for recurrent layers [21]. After that, to measure the losses used, use the categorical cross-entropy loss, and to measure accuracy, use the precision metric. Generative BRNN is more difficult to optimize because of the difference in the frequency of words between uncommon words and common words. For the BRNN and LSTM parts, we use Adam's Bangla Caption Generation Image Optimizer.

6. RESULTS AND DISCUSSION

We implemented a hybrid neural network framework that is capable of generates a Bangla full sentence from the given input image. Firstly, let us look at the viewpoint of the CNN features which is very important for image classification. After that, we give concern about the BRNN and LSTM portion which is capable to generate Bangla text from the given image.

6.1. Encoder model: convolutional neural network

In this part, we mainly discussed about CNN implementation result of BNLIT dataset. We showed that, training time accuracy and validation time accuracy vs. epoch for CNN in Figure 3. We showed that result in graphically for whole dataset. We ran 10 epochs and select batch size 16. From the first epoch of during CNN training time, we got better accuracy for dataset. We showed that accuracy vs. loss and

validation accuracy vs. validation loss in CNN classification training time. After ran 8 epochs, we got 0.794538 training accuracy which is best accuracy for this dataset for CNN result. We got 0.782161 validation accuracy for BNLIT dataset and that is benchmark result for this dataset in CNN part because of it is a self-made new dataset.

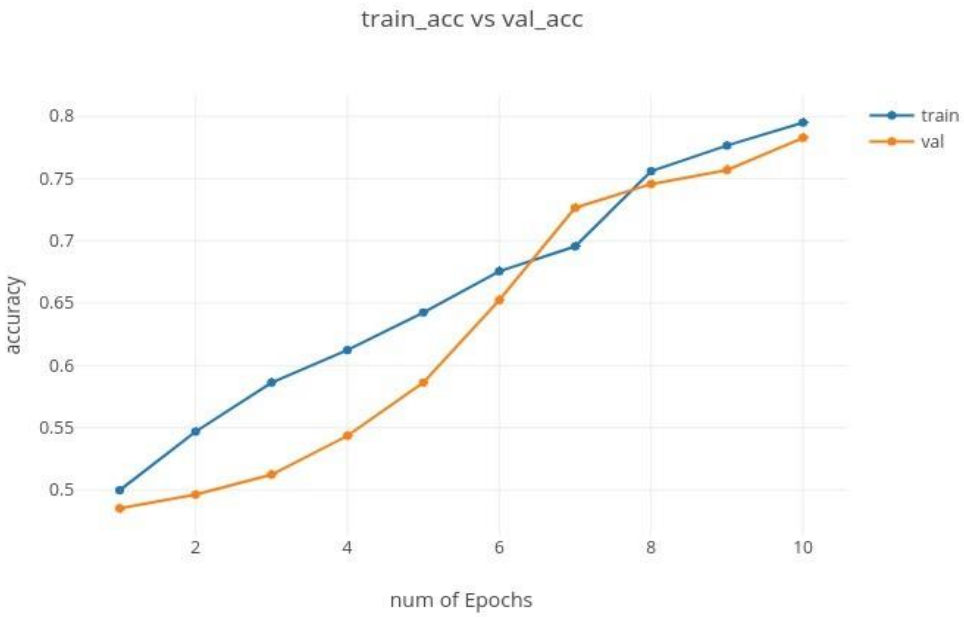

Figure 3. Graphical representation of training time and validation time accuracy for image classification of CNN part

6.2. Decoder model: bidirectional long short term memory

After CNN, we mainly discussed about BRNN and LSTM implementation result of BNLIT dataset. We showed training time accuracy vs. epochs in Figure 4. We also represented that, training time loss vs. epoch for BRNN and LSTM in Figure 5. We showed that result in graphically for whole dataset. We showed that accuracy vs. loss in BRNN and LSTM during training time. After ran 50 epochs, we got 0.8739 accuracy which is best accuracy for this dataset for BRNN and LSTM result and that is benchmark result for BNLIT dataset. We select batch size 128 during BRNN and LSTM train up for BNLIT dataset.

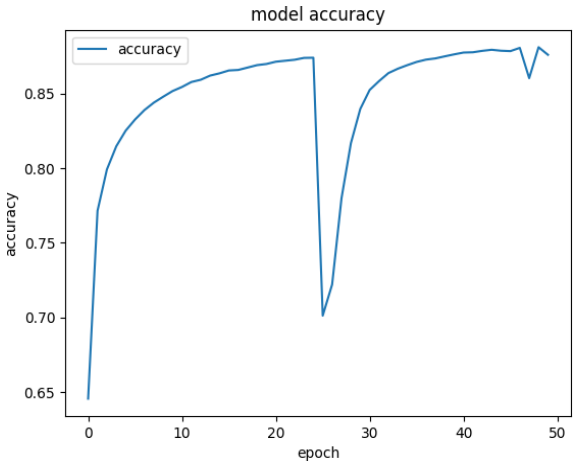

Figure 4. Graphical representation of BRNN and LSTM part of BNLIT dataset result (epoch vs. training time accuracy)

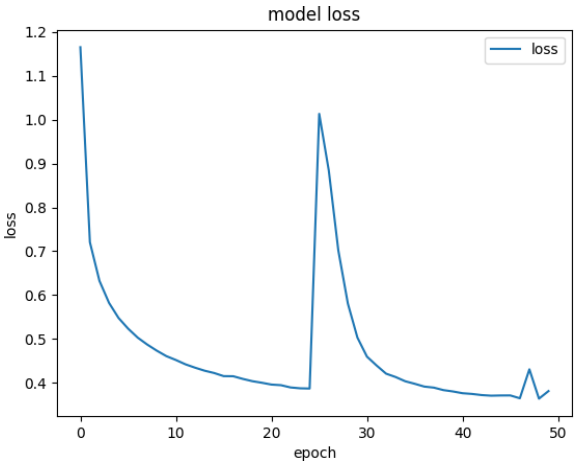

Figure 5. Graphical representation of BRNN and LSTM part of BNLIT dataset result (epoch vs. training time loss)

6.3. Hybrid model to generate text

We generated a pickle file from the whole dataset which is containing 8,743 images. We ran 25 epochs for final training. Each epoch took approximately 1 hour 20 minutes and our accuracy reached 0.942546 for training and 0.758651 for validation. We got approximately 0.197432 losses in training period and 1.615326 losses in validation purpose and that is benchmark result. After complete all epochs, generated training accu- racy vs. validation accuracy graph and training loss vs. validation graph. We illustrated graphically training and validation accuracy in Figure 6 and showed training and validation loss graphically in Figure 7.

To reduce the loss value of the model, the model was trained 25 epochs. From the second epoch, accuracy got improvement comparison with first epoch and generated model and save in a specific directory. The initial accuracy value was therefore 0.8128 in first epochs for training period. But, from the second epoch with the accuracy value coming down to 0.8296.

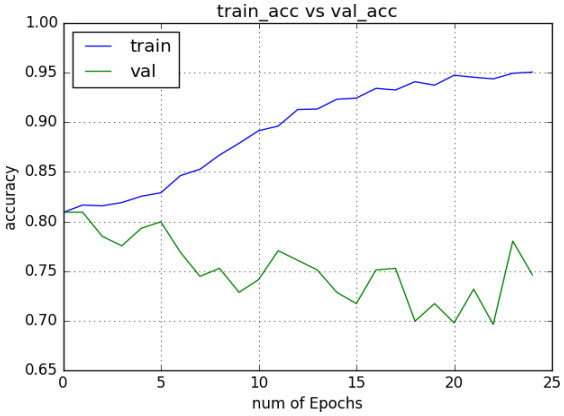

Figure 6. Graphical representation of during final train up for training and validation accuracy

Figure 7. Graphical representation of during final train up for training and validation loss

6.4. Model evaluation

We researched the capacity of the working cross breed profound learning model by investigating how well it can create a reasonable depiction of the test images. We prepared our model to become familiar. With the connection between better parts of the images alongside the applicable bit of the sentences. We represent the BLEU and METEOR scores to evaluate the presentation of our model. These methods permit us to process a score the measures how reasonable is the picture portrayals. The instinct is to quantify how close the model created sentence coordinates the reference sentences gave the dataset. We report these assessment measurements of our model and illustrated them in Table 2.

We prepared our model on BNLIT dataset and watched the assessment of full image expectations on 1000 test pictures. The BLEU-1, 2, 3, 4 assessment scores and METEOR metric scores are surveyed outlined in Table 2. We actualized the concealed layers size of 64, 128, 256, and 512 separately.

Table 2. BLEU scores and METEOR score for BNLIT dataset

Hidden Layer Size	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
64	64.5	45.6	31.8	22.1	19.613227
128	63.8	42.3	30.4	19.6	18.625489
256	64.8	46.5	32.3	22.9	19.683625
512	64.9	46.8	33.1	23.3	19.968532

6.5. Discussion

We implemented our hybrid model using BNLIT dataset. We observed that our model gives better accuracy using our self-made Bangla dataset. During the classification of using CNN, we see how this new dataset can capture to learning from dataset and image classification using VGG16. We get better accuracy which is 0.794538 training time accuracy and 0.782161 which is validation accuracy for BNLIT dataset for CNN result. Furthermore, we got 0.8739 training time accuracy during in the BRNN and LSTM period. Then combined both model and train up full dataset again and finally our accuracy reached 0.942546 for training

time and 0.758651 for validation. To extend, we showed in Figure 8 and Figure 9 that how to generate text from given input image. Finally, we represented our evaluation results in the Table 2.

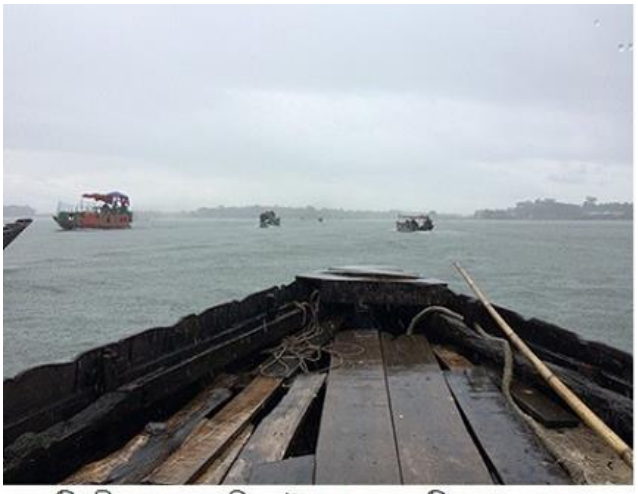

Figure 8. Case of sentence anticipated by our model. We showed that how much the perfect Bangla text our model can generate

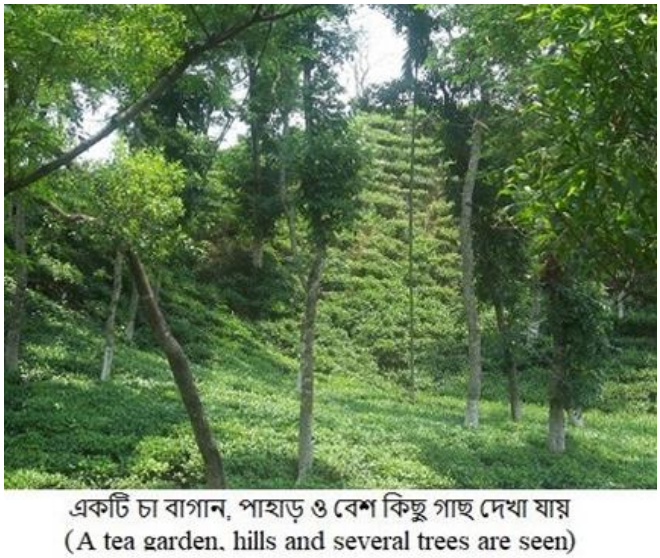

Figure 9. Case of sentence anticipated by our model. For each test picture, we got the most perfect test sentence

7. CONCLUSION

In this study, a complex hybrid neural network model is proposed, which demonstrates exceptional capacity to create Bangla natural language based single sentence depiction from a given test image. The model is capable of detecting images with embedded multimodal and semantic complexities, and is able to generate natural language description based on the context of images. Our methodology incorporates modification to the model to capture visual and language modalities by employing effective LSTM and BRNN counterparts. Moreover, we report acceptable performance and accuracy as the necessary for our self-made dataset. Our experiments with the model shows that better execution across wider scope of datasets may be accomplished by means of model fine-tuning and architectural augmentation.

REFERENCES

[1] 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", In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2321-2334, 2017.

[2] L. Chen, H. Zhang, J. Xiao, L. Nie,J. Shao, W. Liu, T. Chua, "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning", In *Computer Vision and Pattern Recognition (CVPR)*, pp. 6298-6306, 2017.

[3] P. Anderson, B. Fernando, M. John-son, S. Gould, "SPICE: semantic propositional image caption evalu- ation", In *European Conference on Computer Vision (ECCV)*, pp. 382-398, 2016.

[4] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, "Image captioning with semantic attention", In *Computer Vision and Pattern Recognition (CVPR)*, pp. 4651-4659, 2016.

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator", *arXiv:1411.4555v2*, 2015.

[6] Wang, H., Zhang, Y., Yu, X., "An Overview of Image Caption Generation Methods", *Computational intelligence and neuroscience*, pp. 1-13, 2020.

[7] Md. Asifuzzaman Jishan, K. R. Mahmud, A. K. Al Azad, "Natural language description of images using hybrid recurrent neural network", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 2932-2940, 2019.

[8] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, M. Sun, "Show, adapt and tell: adversarial training of cross-domain image captioner", in *Proceedings of the IEEE Conference on International Con- ference on Computer Vision and Pattern Recognition*, pp. 521-530, Honolulu, HI, USA, July 2017.

[9] J. Aneja, A. Deshpande, A. G. Schwing, "Convolutional image captioning", In *Computer Vision and Pattern Recognition (CVPR)*, pp. 5561-5570, 2018.

[10] F. Fang, H. Wang, Y. Chen, P. Tang, "Looking deeper and transferring attention for image captioning", *Multimedia Tools and Application*, vol. 77, no. 23, pp. 31159-31175, 2018.

[11] T. Yao, Y. Pan, Y. Li, T. Mei, "Exploring visual relationship for image captioning", In *European Confer- ence on Computer Vision (ECCV)*, pp. 711-727, 2018.

[12] Q. Wang and A. B. Chan, "CNN+CNN: convolutional decoders for image captioning", *arXiv:1805.09019v1 [cs.CV]*, 2018.

[13] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, "Bottom-up and top-down at- tention for image captioning and visualquestion answering", In *Computer Vision and Pattern Recognition (CVPR)*, pp. 6077-6086, 2018.

[14] Takashi Miyazaki, Nobuyuki Shimizu, "Cross-Lingual Image Caption Generation", *Association for Computational Linguistics (ACL)*, pp. 1780-1790, 2016.

[15] Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks", *Neural Information Processing Systems (NIPS)*, vol. 1, pp. 1097-1105, 2012.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[17] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks", *Signal Processing, IEEE Trans- actions*, vol. 45, no. 11, pp. 2673-2681, 1997.

[18] Md. Asifuzzaman Jishan, Khan Raqib Mahmud, and Abul Kalam Al Azad, "Bangla Natural Language Image to Text (BNLIT)", 2019. [Online]. Available: https://www.kaggle.com/jishan900/bangla- natural-language-image-to-text-bnlit, https://doi.org/10.7910/DVN/DZZ1ZB (Har- vard Dataverse), http://dx.doi.org/10.17632/ws3r82gnm8.4 (Mendeley-ELSEVIER), http://doi.org/10.5281/zenodo.3372752 (Zenodo).

[19] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski, "A database and evaluation method- ology for optical flow", *International Journal of Computer Vision (IJCV)*, vol. 92, no. 1, pp. 1-31, 2011.

[20] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories", *Computer Vision and Pattern Recogni-tion (CVPR)*, Workshop of Generative Model Based Vision (WGMBV), 2004.

[21] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset", *California Institute of Tech- nology, Tech. Rep. 7694*, 2007.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886-893, 2005.

[23] Y. Lecun and C. Cortes, "The MNIST database of handwritten digits", 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[24] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (coil-20)", *Columbia Universty, Tech. Rep.*, 1996.

[25] Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images", *Computer Science Department, University of Toronto*, Tech. Rep, 2009.

[26] Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition", *The Pattern Analysis and Machine Intelligence (PAMI)*, vol. 30, no. 11, pp. 1958-1970, 2008.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database", *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248-255, 2009.

[28] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He, "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks", *arXiv:1711.10485v1 [cs.CV]*, 2017.

[29] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio, ”ChatPainter: Improving Text to Image Generation using Dialogue”, *arXiv:1802.08216v1 [cs.CV]*, 2018.
[30] Richard Socher, Andrej Karpathy, Quoc V. Le*, Christopher D. Manning, and Andrew Y. Ng, ”Grounded Compositional Semantics for Finding and Describing Images with Sentences”, *Tennessee Association of Community Leadership (TACL)*, vol. 2, no. 1, pp. 207-218, 2014.
[31] J. Shotton, J. Winn, C. Rother, and A. Criminisi, Texton-Boost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation, In *European Conference on Computer Vision (ECCV)*, pp. 1-15, 2006.
[32] Geiger, P. Lenz, and R. Urtasun, Are we ready for autonomous driving? the KITTI vision benchmark suite, *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354-3361, 2012.
[33] G. J. Brostow, J. Fauqueur, and R. Cipolla, Semantic object classes in video: A high-definition ground truth database, *Patt. Rec. Letters*, vol. 30, no. 2, p. 8897, 2009.
[34] Liu, J. Yuen, and A. Torralba, “Nonparametric scene parsing via label transfer”, *IEEE Trans, on The Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, no. 12, pp. 2368-2382, 2011.
[35] J. Tighe and S. Lazebnik, Superparsing: Scalable nonparametric image parsing with superpixels, In *Eu- ropean Conference on Computer Vision (ECCV)*, pp. 352-365, 2010.
[36] Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, Scene parsing through ADE20K dataset, In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 352-365, 2017.

BIOGRAPHIES OF AUTHORS

Md. Asifuzzaman Jishan has completed Bachelor of Science in Computer Science and Engineer- ing 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 lan- guage. He has also working knowledge in different web programming language: HTML, CSS, JavaScript (JS), Laravel framework and database system. His research has resulted into a research article which has been published in an international journal, another one research published in an in- ternational conference, and one full image dataset published in four different Dataverse. 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, Au- diovisual 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 Engineer- ing, 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 In- stitute of Technology, Sweden. 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 Uni- versity 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 pub- lished a number of papers in peer-reviewed international journals and presented original research articles in numerous international conferences.

Mohammad Rifat Ahmmad Rashid is serving as an Assistant Professor in the Department of Com- puter Science and Engineering of ULAB. Before joining ULAB, he worked as a researcher in the Pervasive Technologies Research Area within the IoT Service Management Unit in LINKS founda- tion, Italy. He received his Ph.D. degree from Polytechnic University of Turin, Italy in 2018 with a focus on empirical software engineering. His research interests include energy consumption analysis, model-based process optimization and data quality analysis.

Bijan Paul is currently working as a senior lecturer at the Department of Computer Science and Engineering at the University of Liberal Arts Bangladesh. He has completed his B.Sc and M.Sc in Computer Science and Engineering department from Shahjalal University of Science and Tech- nology. He has overall eight years of teaching and research experience within both academia and industry. Moreover, he has served as a consultant in multinational companies. He was awarded for his work on Mongol Dip (A Bilingual Screen Reading Software for Visually Impaired People) from the Honorable Prime Minister of the Govt. of the Peoples Republic of Bangladesh for the tremendous effort and involvement to building a system for the visually impaired people. His research interest includes Software Engineering, Machine Learning, Human-Computer Interaction, Internet of Things, Vehicular Adhoc Network and Wireless Sensor Network.

Md. Shahabub Alam is a full-stack Software Engineer with a demonstrated development history using the front-end and back-end technologies. He has expertise in Python, ASP.NET Web API, NET Framework, and skilled in implementing web applications using JavaScript, Angular JS, etc. Besides, he is also keen on the intersection of applied Data Science, Machine learning, and Software Engineering. He is a strong engineering professional with a B.Sc. in Computer Science and Engi- neering from Ahsanullah University of Science and Technology. He loves working in a diverse team, collaborating and sharing ideas, achieving a common goal and driving an organization forward.