Journal of Applied Science and Engineering, Vol. 22, No. 4, pp. 691–698 (2019) DOI: 10.6180/jase.201912_22(4).0011

Experimental Assessment of Beam Search Algorithm for Improvement in Image Caption Generation

Chiranji Lal Chowdhary[1]*, Aman Goyal[2] and Bhavesh Kumar Vasnani[2]

[1]*School of Information Technology and Engineering, VIT Vellore, India*
[2]*School of Computer Science and Engineering, VIT Vellore, India*

Abstract

One of the biggest problems faced regarding images in various fields is the generation of meaningful description for the image i.e. caption for the image. The images are being used for numerous purposes with major on the web however they had to spend a great share of their time to generate a proper and accurate description for the image. This makes it very complex as the machine has to learn from the datasets and then describe the objects, activities and the places. The fact that humans can do it quite easily for small sets but fail when the number of images is more. This make it a rather interesting challenge for deep learning algorithms. The applied approach for the image caption generation would be based on long-short-term memory networks (LSTM) and recurrent neural networks (RNN). Such network model allows to select the next word of the sequence in a better manner. In this paper, Python is used to form this caption generating platform with the help of TensorFlow library which can easily generate the LSTM model for a given images. In this research work, machines are trained by deep learning approach. To improve the efficiency of the caption generation, the training has to be quite deep with more sample images. Additionally, detailed analysis is done on the improvement which can be brought to implement by including Beam Search in it.

Key Words: LSTM, Recurrent Neural Networks, TensorFlow, Deep Learning

1. Introduction

With exponential development in the multimedia industry, the increase in the image data, on and off the web, has seen rapid growth. Thus, the work on image related areas is currently one of the major fields of work. Image caption generation and retrieval are still emerging research areas and development scope for these fields is quite wide. The image caption is a popular problem where the aim is to come up with the best description for all the features and actions that are there in the image. There have certain developments and research in this field using deep neural networks (DNN) but most of them do not accommodate hidden layers and so lack efficiency and accuracy. All these researches accommodate the use of an artificial neural network (ANN) but have not achieved better results.

The image captions were used in a diversity of applications, such as, (i) Image captions were in practice to define imageries to blind persons or for those with little vision and depends on sounds /texts to designate a sight. (ii) Another good example is in web development. It is a worthy exercise to deliver an explanation aimed at some image which looks scheduled the page so as to an image could read or perceived as contrasting to objective seen. This is leading to web content accessibility.

Since the time machine learning has come to play it is highly used in image caption generation field. But simple machine learning models without additional techniques are simply not sufficient for the requirements out

*Corresponding author. E-mail: chiranji.lal@vit.ac.in there for image caption generation. Long short-term memory (LSTM) approach can prove to improve the accuracy to if implemented along with recurrent neural network (RNN) which makes it capable of identifying and learning long-term dependencies present in the model. This was the modification which has improved the ability of the machine learning program to predict better results through LSTM units which are composed of smaller cells, and input, forget and output gate. LSTMs are well suited for various important functionalities including processing predictions and classification. Also, LSTM overshadows other methods like RNN and hidden Markov models and others in image caption generating application due to relative insensitivity to gaps. Beam Search is also a modification currently not focused much but has a significant improvement in the image caption generation application. Therefore, authors focused on depicting the benefit that it can provide to the application just by adding this to predict even deeper results.

This paper is aimed at a beam search algorithm for improvement in image caption generation. In section 2, the related work in image caption generation is addressed. In section 3, the theory and proposed work are discussed which is followed by the implementation of proposed model in section 4. Section 5 is result discussions of the proposed models and methods of image caption generations. To conclude, the conclusion is presented in section 6 which is followed by future work in section 7.

2. Related Work

In the modern-day, with developments being made in the field of artificial intelligence, authors are seeing an increase in the number of problems from real-world scenarios that can be tackled with the help of deep learning. Automatic caption generation of images is one such application that is slowly being popularized in today’s world. The main challenge behind this task is to efficiently combine artificial intelligence techniques involving both computer vision as well as parsing of natural languages. One such method that researchers wish to implement to efficiently combine these two fields is the implementation of a deep recurrent architecture that has been successfully utilized in existing image caption generators [1]. A deep recurrent neural network involves the units of the network being connected to each other in the form of a directed cycle, thus exhibiting temporal behavior of a dynamic nature wherein internal memory can be efficiently utilized to process any random input sequence, thus making this kind of network suitable for the real-world applications like voice recognition and image captioning [2]. To delve into the appropriate structuring of such a recurring network, the researchers have researched the 2-directional mapping that is bound to be involved between the images themselves and the sentence-based descriptions authors wish to generate in with the help of this network [3].

One of the most relevant architectures currently being used for automatic image captioning is the CNN-LSTM (long short-term memory) architectures that have been frequently used on popular datasets such as Microsoft Common Objects in Context (MCSCOCO) to generate captions across visual spaces [4]. Another field of the survey to tackle the task of the generation of annotations for images would be to classify the currently existing methodologies based on their point of views used in conceptualizing the problem, such as models that focus on image description as a retrieval problem versus the models that view image description as a generation problem across the visual representations or multimodal space [5]. Su Mei Xi [6] propose a framework which produces sentential comments for general pictures. Under the state of the new unlabeled picture districts, authors ascertains the restrictive likelihood of each semantic catchphrase and comment on the new pictures with maximal contingent likelihood. Investigations on the Corel picture set demonstrate the viability of the new algorithm [6].

Additionally, cascade recurrent neural network (CRNN) for picture subtitle generation is proposed. Unique in relation to the traditional multimodal recurrent neural network, which just uses a solitary system for removing unidirectional syntactic highlights, CRNN embraces a cascade network for taking in visual-dialect associations from forward and in reverse headings, which can abuse the profound semantic settings contained in the picture. In the proposed system, two implanting layers for thick word articulation are developed [7]. To enhance the execution of image caption generation Dong-Jin Kim [8] proposed a technique in which transfer learning CNN is used in sentences and portrayal image features with deep Fisher Kernel. With this, the generation of sentences is done using gLSTM and with good efficiency in performance. Chunting Zhou [9] consolidate the qualities of the two structures and proposed a new model called C-LSTM for sentence formation and content classification. C-LSTM can catch both local highlights of expressions as well as additional worldwide and transient sentence semantics. The results demonstrate that the C-LSTM beats both CNN and LSTM and can achieve astounding execution on these tasks [9].

Likewise, it is also proposed the utilization of LSTM to join diverse features. Also, it made utilization of the rule that LSTM hidden layer hubs can remember, and gather the features removed from each model as contributions of a movement of LSTM [10]. Martin Sundermeyer et al. [11] connected the LSTM neural system engineering to two languages modeling tasks. This system compose is particularly appropriate to language displaying as in principle it permits the correct displaying of the likelihood of a word succession. Yansong Feng et al. [12] is focused on the task of naturally producing caption for pictures, which is vital for some image-related applications. Reference [13] gives a general review of the light beam search (LBS) philosophy and applications. LBS empowers an intelligent investigation of different target choice issues because of the introduction of tests of a substantial arrangement of non-dominated focuses, to the decision-maker (DM) in every iteration. Reference Yongfei Shen et al. [14] utilize CNN and LSTM models to create picture description. Be that as it may, it is proposed that an enhanced LSTM demonstrate, a bidirectional LSTM. Aghasi [15] discussed most of the current works which utilize long short-term memory (LSTM) as a repetitive neural system cell to tackle this undertaking. He presented a model that can naturally create a pictorial depiction and depends on an intermittent neural system with changed LSTM cell with an extra door in charge of picture features.

There are two distinctive approaches to play out the undertaking of picture captioning. These two types are fundamentally recovery-based strategy and generative technique. A large portion of work is done because of recovery-based strategy [16]. Language models have customarily been generally assessed of relative frequencies, utilizing tally insights that can be extricated from enormous measures of content information [17]. Su Mei Xi et al. [18] proposed a novel framework which produces sentential explanations for general pictures. First of all, a weighted component grouping algorithm is utilized on the semantic idea bunches of the picture region. For a given bunch/cluster, authors decide applicable relevant factual conveyance and relegate more noteworthy weights to significant highlights as contrasted with less applicable highlights [18]. Image-based web index is the procedure of seeking data by utilizing related pictures [19]. Additionally, these units can be distinctive sorts of RNNs, for example, a basic RNN and an LSTM. Via preparing regularly utilizing NeuralTalk1 stage on Flickr8k dataset, without extra preparing information, also it shows signs of improvement results than that of ruling structure and especially, the proposed show outperforms Google-NIC in picture subtitle age [20].

3. Theory and Proposed Work

The major objective is to generate real-time caption for any input image in a single pass and ensuring the accuracy of the result by training the machine with good datasets. The researchers would make the machine to generate a suitable caption from the images based on the learning that authors feed into it. Deep learning concept with the use of recurrent neural networks would be quite useful and a big dataset would be required to provide learning to the machine. The process contains a series of sub-processes that would be learning through the datasets, providing the input and then generating the caption from it. Ultimately, authors are going to compare the accuracy in the results that are predicted from the two approaches one with beam search and another without beam search.

In this proposed work, authors are first taking the dataset from *Flickr8k* datasets. These datasets are then first passed through a feature capturing model which captures the feature of each image and start generating a model for producing a caption for the image. To generate the model the program reads each image and its feature in details and try to compare with other images if possible. These images have a predefined caption with are given in the datasets and the model tries to read and learn from these captions and features of the images. After training the dataset it sees the accuracy of the model by applying it to any random trained images. And then the model is ready for testing. This model is the given different varieties of images for testing whether the model is giving the proper output/caption for images or not. The broad architecture of the project can be divided into two separate parts one of which is the LSTM part and the other is the Beam Search Algorithm. The detailed description for each of the module is as follows.

3.1 LSTM Architecture

In the above diagram, the lowest layer consists of the words of the caption that are to be predicted and the layer are the word embedding vectors for every word that are present in the caption. The topmost layer is the outputs from the LSTM model which are probability distributions that are there for the next word present in the caption so that the best caption is outputted as the result. The model is trained to minimize the negative sum of the log probabilities of each word. In the second approach, authors added a beam search algorithm to increase the accuracy and compare the results. This is shown in Figure 1.

3.2 Beam Search

Beam search is a heuristic finding algorithm that investigates a chart by growing the most encouraging hub in a restricted set. Beam search is an advancement of the best-first hunt that diminishes its memory requirements. A beam search (Figure 2) is regularly used to keep up tractability in huge frameworks with the lacking measure of memory to store the whole inquiry tree. For case, it is utilized as a part of numerous machine interpretation frameworks.

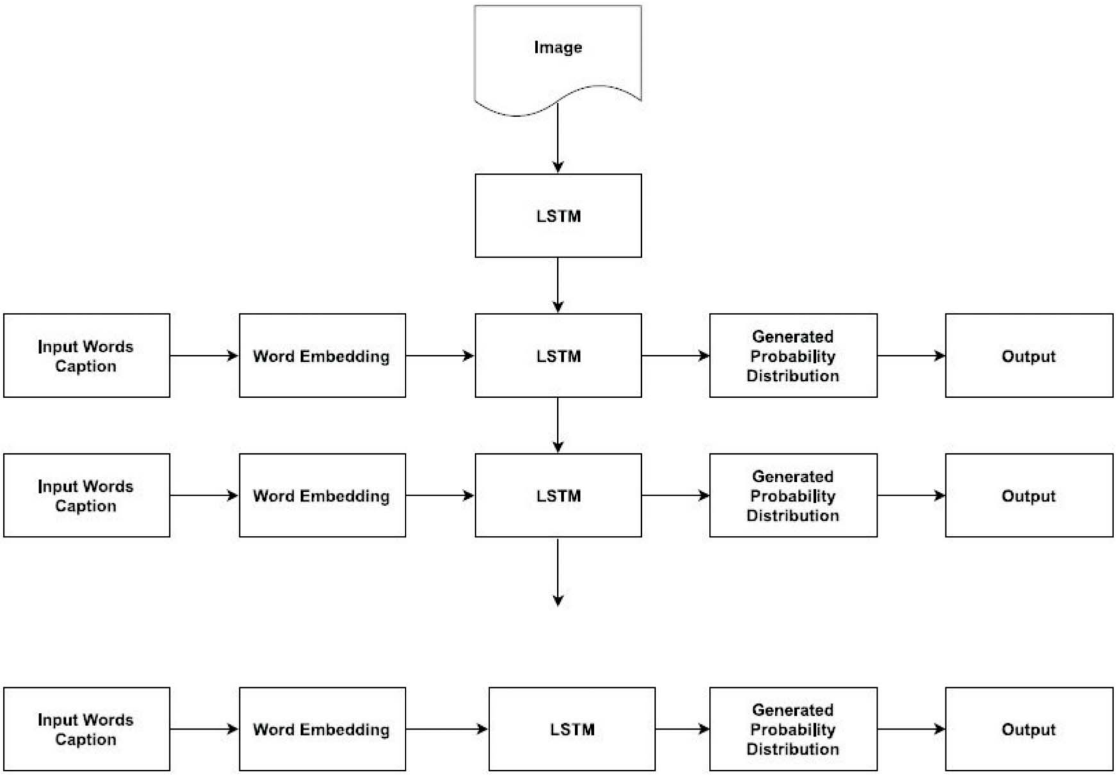

Figure 1. Caption generator.

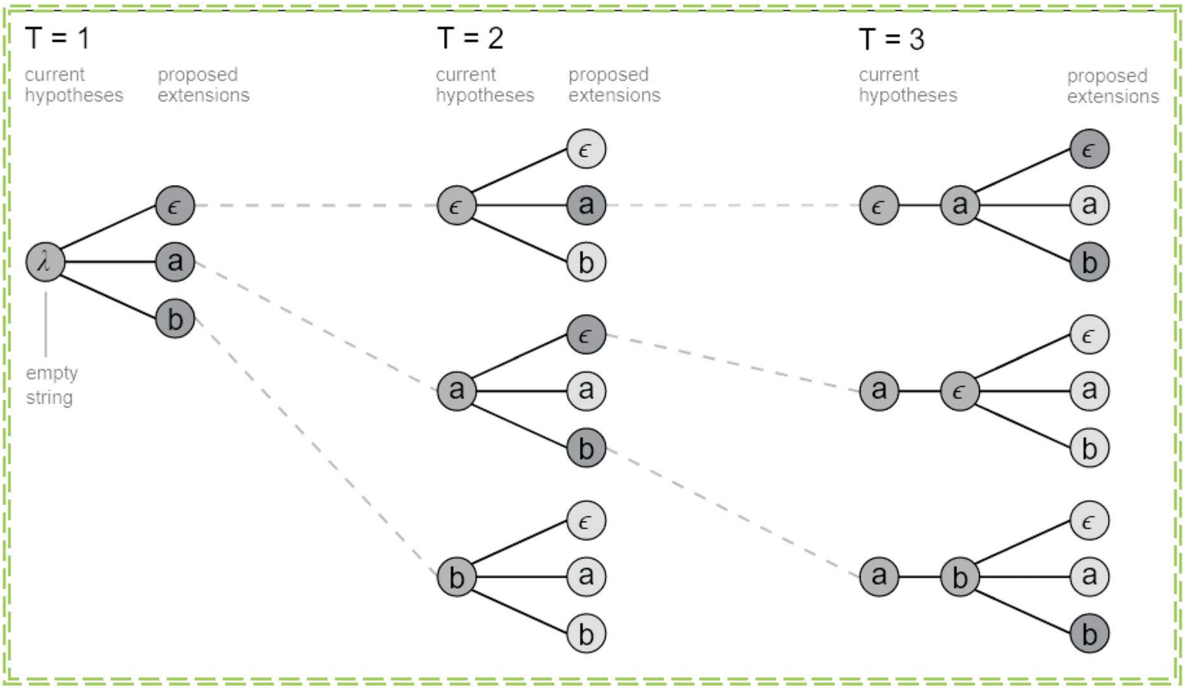

Figure 2. Beam search steps.

Pseudocode for Beam Search:

```
1. Start
2. Define sequence of the words in caption
3. Walk over each step-in sequence
   3.1 For every row in the data:
      3.1.1 For every candidate update possible candidate sequence and score
      3.1.2 Order all the candidates by score
      3.1.3 Select the best k predictions
4. Return top k predictions
```

4. Implementation of Proposed Model

The implementation is done using Python environment along with the help of TensorFlow which is a machine learning library useful for machine learning programs. The model generates the basic caption through the aid of the LSTM and RNN implementation with InceptionV3 which is a pre-trained model for Keras and eases the computation as implementation is performed on a CPU machine to extract features from training which is time-consuming. Also, InceptionV3 based image encoding which is the previously generated captions is further improved by Beam Search which checks the previous layers for the best captions. InceptionV3 is the best implementation as compared to others like VGG16 which is the most basic and the slowest. In this work, researchers have tested the image for different index values of the Beam Search algorithm. In the implementation, authors focus on getting the top 5 predictions based on the score for every image given as input. The researchers are also analyzing the change in the caption for different values of k as well starting with the value k = 3 to k = 7. The following are the images along with the captions generated using normal implementation and through-beam search based implementation.

5. Result Discussions

The researchers validated the improvements and working through accuracy and loss scores obtained from the implementation. The machine was trained for 50 EPOCHS and the results shown below are taken from this trained machine. Both accuracy and loss are good metrics that can tell us the performance of the trained machine. Greater the accuracy better are the results as compared to the original ones. Similarly, the loss is the loss of the required data that should be there in the results but is missing. Both these metrics can be improved by training the machine for a greater number of EPOCHS.

The results (Table 1) obtained by both the approaches were quite different. The results varied quite a lot due to the difference in accuracy and loss of the trained model in both the approaches. The accuracy is more in the implementation of the beam search algorithm and the loss was quite less even in 1 EPOCH while it was more in the normal implementation for 4 EPOCHS. The reason for this is the Beam Search which is quite helpful for the application of image generation as the image generation model are usually Recurrent Neural Networks and generating the best captions require the selections of best nodes by comparing the previous layers. Beam Search checks the previous layers of the neural network and then select the best result from the previous layers which would best suit the image.

The analysis shows us that Beam Search has significantly improved the results of the implementation. These experimental observations prove the importance of Beam Search in the image caption generation application. Also, with beam search authors can even decide the number of hidden layers it should traverse back so that even for small computing devices, this application can be applied with almost the same accuracy. Overall, the prediction through-beam search is a better caption generated keeping in mind the natural language processing aspects in any caption which is the major requirement for human-readable captions and which is not focused by other machine learning and neural network models.

Deep learning-based image captioning methods are limited to categorize on learning methods of supervised, reinforcement, and unsupervised learning. The method is congregated by reinforcement learning and unsupervised learning. Typically captions are engendered aimed at an entire prospect in the image. Though, captions can also be produced for dissimilar regions of dense caption imagery. Image caption methods are limited to either unassuming encode-decode building. In the maximum of the image-caption methods, LSTM is used as a language prototypical.

Table 1. Accuracy analysis

Analysis	Without Beam Search	With Beam Search
Accuracy	0.0803	0.4539
Loss	6.6720	3.1809

6. Conclusion

In this work, authors conclude that image caption generation task can be simplified with the help of deep learning concepts of deep neural networks. The Show and tell the algorithm that incorporates the concept of long short-term models prove to be beneficial in simplifying the task to generate captions. Also, the conclusion can be drawn that the accuracy and the loss are directly and indirectly proportional respectively to the extent of training. Thus, it can be concluded that Beam Search can highly improve the efficiency of image caption generation as is very helpful for this application and does not increase the complexity of the project as well. As a whole Beam Search with recurrent neural networks and long short-term model is one of the best implementations of image caption generation available till date. The basis for this conclusion is the gradual change in accuracy for the same number of EPOCHS for both the approaches. For human preferable captions, it is necessary to have a broad idea about the long-term dependencies that can play a vital in modulating the ultimate caption that is generated. And hence, it can be easily concluded with proof that LSTM, Beam Search and RNN is very suitable for any real-time applications and can provide users with the results that they can directly accommodate into their daily applications. This application can be applied in various other applications as a subpart which can aid the host application by improving its feasibility, functionality, and usability by any general user. Some of the future applications that can accommodate caption generation are photo management through captions, real-time news caption, a caption for lifelogging/vlogging, and aiding blind people medical image caption for diagnosis.

References

[1] Vinyals, O., A. Toshev, S. Bengio, and D. Erhan (2014)

Show and tell: a neural image caption generator, Proc. of 2015 CVPR 2015 Computer Vision Foundation, Boston, England, 3156–3164.

[2] Sak, H., A. Senior, and F. Beaufays (2014) Long short-term memory recurrent neural network architectures for large scale acoustic modeling, Proc. of Fifteenth Annual Conference of the International Speech Communication Association, Singapore.

[3] Chen, X., and C. L. Zitnick (2015) Mind's eye: a recurrent visual representation for image caption generation, Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, England. doi: 10.1109/CVPR.2015.7298856

[4] Soh, M. (2016) *Learning CNN-LSTM Architectures for Image Caption Generation*, https://cs224d.stanf, ord.edu/reports/msoh.pdf.

[5] Bernardi, R., R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank (2016) *Automatic Description Generation from Images: a Survey of Models, Datasets, and Evaluation Measures*, https://www.jair.org/media/4900/live-4900-9139-jair.pdf. doi: 10.1613/jair.4900

[6] Xi, S. M., and Y. I. Cho (2014) Image caption automatic generation method based on weighted feature, Proc. of 2013 13th International Conference on Control, Automation and Systems (ICCAS 2013), Gwangju, South Korea. doi: 10.1109/ICCAS.2013.6703998

[7] Wu, J., and H. Hu (2017) Cascade recurrent neural network for image caption generation, *Electronics Letters* 53(25), 1642–1643. doi: 10.1049/el.2017.3159

[8] Kim, D., D. Yoo, B. Sim, and I. S. Kweon (2016) Sentence learning on deep convolutional networks for image Caption Generation, Proc. of 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), 246–247. doi: 10.1109/URAI.2016.7625747

[9] Chowdhary, C. L., and D. P. Acharjya (2017) Clustering algorithm in possibilistic exponential fuzzy c-mean segmenting medical images, *Journal of Biomimetics, Biomaterials and Biomedical Engineering* 30, 12–23. doi: 10.4028/www.scientific.net/JBBBE.30.12

[10] Chen, J. L., Y. L. Wang, Y. J. Wu, and C. Q. Cai (2017) An ensemble of convolutional neural networks for image classification based on LSTM, 2017 International Conference on Green Informatics (ICGI), Fuzhou, China, 217–222. doi: 10.1109/ICGI.2017.36

[11] Sundermeyer, M., R. Schluter, and H. Ney (2012) LSTM *Neural Networks for Language Modeling*, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.248.4448&rep=rep1&type=pdf.

[12] Feng, Y., and M. Lapata (2013) Automatic caption generation for news images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(4), 797–812. doi: 10.1109/TPAMI.2012.118

[13] Chowdhary, C. L., and D. P. Acharjya (2018) Segmentation of mammograms using a novel intuitionistic possibilistic fuzzy c-mean clustering algorithm, *Nature Inspired Computing* 75–82.

[14] Li, J., and Y. Shen (2017) Image describing based on bidirectional LSTM and improved sequence sampling, Proc. of 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), 735–739. doi: 10.1109/ICBDA.2017.8078733

[15] Poghosyan, A., and H. Sarukhanyan (2017) Short-term memory with read-only unit in neural image caption generator, 2017 Computer Science and Information Technologies (CSIT), Yerevan, Armenia. doi: 10.1109/CSITechnol.2017.8312163

[16] Shah, P., V. Bakrola, and S. Pati (2017) Image captioning using deep neural architectures, Pro. of 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India. doi: 10.1109/ICIIECS.2017.8276124

[17] Sundermeyer, M., H. Ney, and U. Schluter (2015) From feedforward to recurrent LSTM neural networks for language modeling, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23(3), 517–529. doi: 10.1109/TASLP.2015.2400218

[18] Chowdhary, C. L., and D. P. Acharjya (2016) A hybrid scheme for breast cancer detection using intuitionistic fuzzy rough set technique, *Biometrics: Concepts, Methodologies, Tools, and Applications* 1195–1219.

[19] Vijay, K., and D. Ramya (2015) Generation of caption selection for news images using stemming algorithm, Pro of 2015 International Conference on Computation of Power, Energy, Information and Communication

(ICCPEIC), Chennai, India, 536–540. doi: 10.1109/ICCPEIC.2015.7259513

[20] Wang, M., L. Song, X. Yang, and C. Luo (2016) A parallel-fusion RNN-LSTM architecture for image caption generation, Proc of 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA. doi: 10.1109/ICIP.2016.7533201

Manuscript Received: Oct. 29, 2018
Accepted: Jul. 22, 2019