International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6965
***Article Received**: 26 March 2023 **Revised**: 15 May 2023 **Accepted**: 04 June 2023*

A CNN and LSTM-based Model for Creating Captions for Photos

Bala Murali Krishna Thati[1], Swathi Voddi[2], Srikanth Busa[3], Surendra[4],J N V R Swarup Kumar[5], M.V.L.N. Raja Rao[6]
[1]Professor, Department of CSE, Dhanekula Institute of Engineering & Technology. Ganguru, Vijayawada.
Email: balu.thati9@gmail.com
[2]Assistant Professor, Department of Computer science and Engineering, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada,
Email: Potluri.swathi@gmail.com
[3]Professor, Department of CSE, Kallam Haranadhareddy Institute of Technology, Chowdavaram, Guntur, AP.
Email: Srikanth.busa@gmail.com
[4]Assistant Professor, Department of CSE, K L Deemed to be University, Green Fields, Vaddeswaram, AP.
Email: guntisurendra@gmail.com
[5]Assistant Professor, Department of CSE, GITAM School of Technology, GITAM (Deemed to be University), Visakhapatnam. Email: sjavvadi2@gitam.edu
[6]Professor, Department of Information Technology, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, AP.
Email: rajarao.mamidanna@gmail.com

Abstract—Can a machine interpret an image's meaning with the same speed as the human brain when it is seen? This problem was heavily researched by computer vision specialists, who believed it to be unsolvable until recently. It is now possible to develop models that can generate captions for pictures because of advancements in deep learning techniques, accessibility to large datasets, and processing power. This will be accomplished by the Python-based implementation of the article's deep learning convolutional neural network technique and a particular kind of recurrent neural network. Here the proposed model uses CNN and LSTM methods to achieve desired task

Keywords- Feature Extraction; Image Analysis; Neural Network; Deep Learning; Text Analysis.

I. INTRODUCTION

An image caption generator employs computer vision and natural language processing methods to comprehend the context of a picture and describe it in a language like English [1]. The goal of this research is to introduce readers to the concepts behind a CNN and LSTM model and show them how to utilize them to build an image caption generator [2]. CNN, a Deep Learning approach, uses learnable weights and biases to priorities different objects and elements in an input image to help with image recognition [3]. Image classification is one of this architecture's most well-liked uses [4]. The neural network combines nonlinear, pooling, and several convolutional layers [5]. The output of the first convolution layer serves as the input for the second layer after the image has gone through one convolution layer. The identical procedure is then done for each subsequent layer [6].

After a sequence of convolutional, nonlinear, and pooling layers, it is crucial to include a completely linked layer [7]. In this layer, the convolutional network's output data is utilised [8]. A completely linked layer is added to the end of the network to create an N-dimensional vector, where N is the number of classes from which the model selects the desired class [9]. Recurrent neural networks (RNNs) have a subclass known as networks that may learn order dependency in sequence prediction problems [10]. The hardest problems, such speech recognition, machine translation, and many others, are where technology is most frequently applied [11]. This issue was found during the training of traditional RNNs because as neural networks get more complex, little to no training may occur if the gradients are very tiny or zero, which would result in poor predicting performance [12]. Since there may be delays of various lengths between important events in a time series, LSTM networks are ideal for categorizing, analyzing, and generating predictions based on time series data [13]. The image characteristics will be given to the LSTM model, which will provide the image descriptions, by the CNN model Exception, which was trained on the ImageNet dataset.

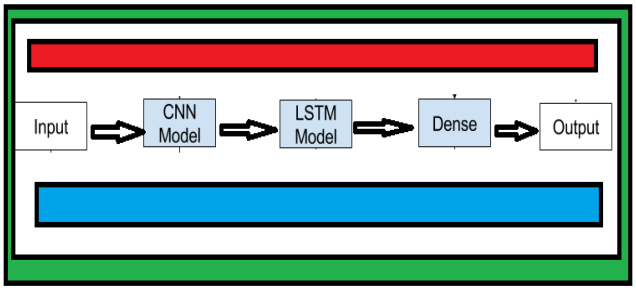


Figure-1 Overall representation

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6965
***Article Received**: 26 March 2023 **Revised**: 15 May 2023 **Accepted**: 04 June 2023*

Due to its ability to overcome the short-term memory limitations of the RNN, the LSTM is noticeably more efficient and superior than the RNN in general. The LSTM may process inputs while processing important data and avoiding irrelevant data. The positioning of the images in a movie or the words in a text document are two instances of inputs with both spatial and temporal structure. CNN-LSTMs are frequently employed when the output requires temporal structure, such as the words in a textual description. The 2D structure of pixels in an image or the 1D structure of words in a phrase, paragraph, or page are two examples of inputs with spatial and temporal structure. In order to recognize the context of images and explain them in a natural language like English, the model of the image caption generator that applies the CNN and LSTM principles uses computer vision and natural language processing.

II. RELATED WORK

[14] The model makes use of textual attention to increase the data's correctness.

[15] suggested a brand-new component termed a review network for the encoder-decoder structure. Along with CNN and RNN encoders, RNN decoders are taken into account in this study. Because it is ubiquitous, the review network can improve any present encoder-decoder paradigm. The review network creates a thought vector each time a review step is finished, and this thought vector serves as the input for the decoder's attention mechanism. The review network performs a series of review steps on the encoder concealed states using an attention mechanism. It has demonstrated that traditional encoder-decoders cannot use our technique. We demonstrate empirically that our approach outperforms cutting-edge encoder-decoder systems when it comes to source code and image captioning.

[16] A key component of scene understanding, which integrates the expertise of computer vision with natural language processing, is image caption, which automatically generates natural language descriptions in accordance with the content detected in an image. In the implementation of human-computer interaction, for example, image captions are often utilized and essential. The attention mechanism, a crucial component of computer vision that has lately received extensive application to tasks involving the creation of image captions, is the subject of the summary of related techniques in this study. Along with the benefits and drawbacks of these methods, the most popular datasets and evaluation criteria in this area are also covered. In the study's conclusion, a number of unresolved issues with the image caption task are listed.

III. DATASET

Flickr8k_Dataset: Contains 8092 JPEG-formatted photos. There are several files with various sources for the photo descriptions in the Flickr8k text collection. The primary file of our dataset, Flickr8k.token, which includes image names and their corresponding captions separated by newlines, can be found in the Flickr 8k text folder ("n"). Six thousand photos are utilized in the image dataset for training, one thousand for validation, and one thousand for testing.

IV PROPOSED METHOD

Our brains have the ability to categorise or annotate each image that is shown to us. What about computers, though? How is it possible for a computer to analyses a picture and provide an extraordinarily accurate and pertinent caption? Making a useful caption generator for an image was once thought to be extremely difficult, but thanks to improvements in computer vision and deep learning techniques, the availability of pertinent datasets, and AI models, it is now much simpler. Several data annotation enterprises are making billions of dollars thanks to the international growth of caption production. The goal of this project is to create an annotation tool that can generate descriptions for photographs that are exceptionally relevant using databases. Two deep learning techniques, LSTM (a kind of recurrent neural network) and Convolutional Neural Networks, must be understood in its fundamentals in order to do the same (CNN). Although we will just briefly discuss these strategies here, if you are interested in learning more, please click here. Deep learning and computer vision are used to create image caption generators, which recognise the context of a picture and annotate it with pertinent captions. It entails labelling an image with English keywords using datasets provided during model training. The CNN model known as Exception is used to train the ImageNet dataset and is in charge of extracting picture features. The LSTM model, which creates the picture description, will be given these extracted characteristics.

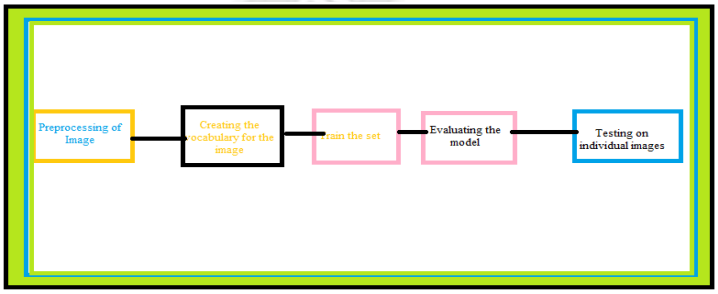


Figure-2 Steps to follow to achieve proposed model

A convolutional neural network (ConvNet/CNN) may be used to analyses a photo, give different visual elements and objects weights and biases, and distinguish between them. ConvNet are simpler to construct than other classification methods. Convolutional neural networks are deep neural networks with the capacity to handle data represented as a 2D matrix. Since photos can be easily represented as a 2D matrix, CNN is a suitable method for working with photographs. Prior to

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6965
***Article Received**: 26 March 2023 **Revised**: 15 May 2023 **Accepted**: 04 June 2023*

combining the images to identify them, it analyses the photographs from top to bottom and left to right to gather important information. When using perspective, you may use images that have been cropped, rotated, or scaled.

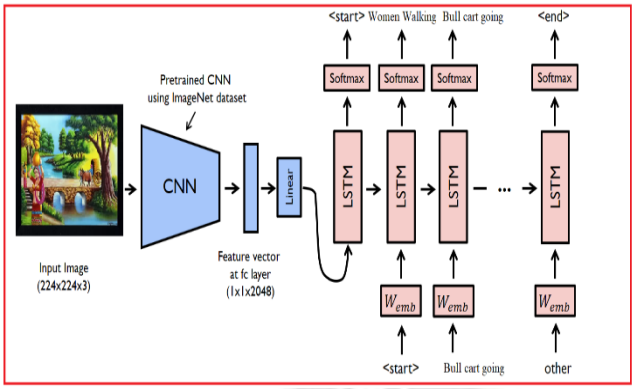
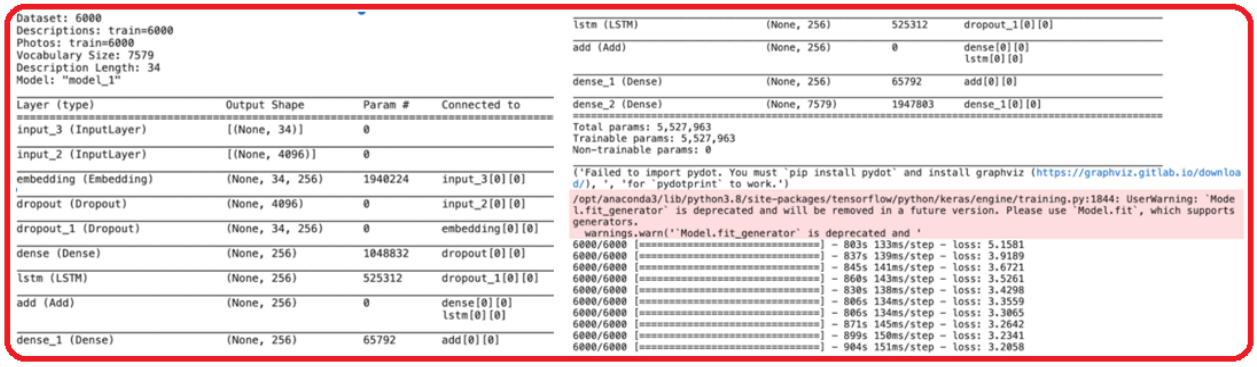


Figure-3 Proposed Model Architectural View

LSTM is best for puzzle-based problem solving. Based on the sentence before, we can infer what the subsequent phrase will be. By solving the drawbacks of RNNs with short time period memory, it has proven to be more effective than conventional RNNs. While processing inputs, LSTM may perform the necessary statistics. Via an overlook gate, it may also discard the unnecessary statistics.

V. EXPERIMENTAL RESULTS

The subsequent sections of the study will clarify our ideas and conclusions as we use a CNN-LSTM model for picture captioning. Please be aware that this work does not serve as a tutorial on how to construct photo captioning; rather, it examines the CNN-LSTM architecture and some of its real-world applications. The program's Python 3 code was created and run in Keras. The following is a list of the prerequisites and requirements you'll need to understand the implementation completely. Image positioning and object detection are the areas of computer vision that have gotten the greatest attention. Social networking site users are now able to upload photographs of any size or complexity and search for descriptions on Google. All of the following are lacking: upgradeability, performance, flexibility, and scalability. High-quality pictures must be used as the input. difficult to notice details in photographs with poor quality. Complex scene analysis could be tough. Utilizing a proxy is intended to speed up the photo search procedure. If the input image is complex, it won't be possible to submit it and read out the caption because processing will take some time. Deep Neural Networks are capable of producing precise, expressive, and flowing subtitles that address the issues present in both versions. speed up the production of subtitles. Thanks to the materials we provide, users of social media won't need to spend hours searching Google for subtitles. Our service makes it simple for users to submit certain photos to social networks. Users are not forced to manually add captions when uploading images. The picture search problem can be solved using the suggested design. You can upload either color or black and white images, and you can read the English caption out loud. Neural networks may solve any issue and provide precise, well-written, and fluid subtitles by utilizing tensor flows and algorithms.



```
Dataset: 6000
Descriptions: train=6000
Photos: train=6000
Vocabulary Size: 7579
Description Length: 34
Model: "model_1"
```

Layer (type)	Output Shape	Param #	Connected to
input_3 (InputLayer)	[(None, 34)]	0	
input_2 (InputLayer)	[(None, 4096)]	0	
embedding (Embedding)	(None, 34, 256)	1940224	input_3[0][0]
dropout (Dropout)	(None, 4096)	0	input_2[0][0]
dropout_1 (Dropout)	(None, 34, 256)	0	embedding[0][0]
dense (Dense)	(None, 256)	1048832	dropout[0][0]
lstm (LSTM)	(None, 256)	525312	dropout_1[0][0]
add (Add)	(None, 256)	0	dense[0][0] lstm[0][0]
dense_1 (Dense)	(None, 256)	65792	add[0][0]
dense_2 (Dense)	(None, 7579)	1947803	dense_1[0][0]

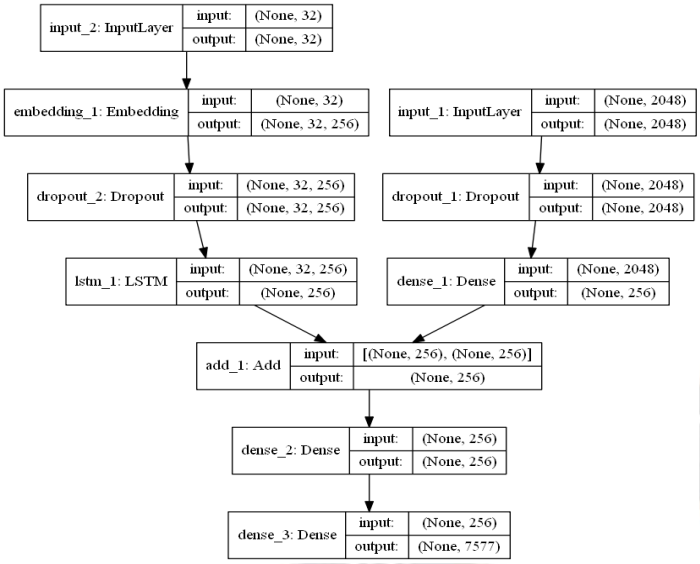
```
Total params: 5,527,963
Trainable params: 5,527,963
Non-trainable params: 0

('Failed to import pydot. You must `pip install pydot` and install graphviz (https://graphviz.gitlab.io/downloa
d/), ', 'for `pydotprint` to work.')
/opt/anaconda3/lib/python3.8/site-packages/tensorflow/python/keras/engine/training.py:1844: UserWarning: `Mode
l.fit_generator` is deprecated and will be removed in a future version. Please use `Model.fit`, which supports
generators.
  warnings.warn('`Model.fit_generator` is deprecated and '
6000/6000 [==============================] - 803s 133ms/step - loss: 5.1581
6000/6000 [==============================] - 837s 139ms/step - loss: 3.9189
6000/6000 [==============================] - 845s 141ms/step - loss: 3.6721
6000/6000 [==============================] - 860s 143ms/step - loss: 3.5261
6000/6000 [==============================] - 830s 138ms/step - loss: 3.4298
6000/6000 [==============================] - 806s 134ms/step - loss: 3.3559
6000/6000 [==============================] - 806s 134ms/step - loss: 3.3065
6000/6000 [==============================] - 871s 145ms/step - loss: 3.2642
6000/6000 [==============================] - 899s 150ms/step - loss: 3.2341
6000/6000 [==============================] - 904s 151ms/step - loss: 3.2058
```

"Describing" An image is a representation of a thing or a scene that has been converted into a written description that can be read by humans. It combines machine learning with computer vision. Neural network captioning models are made out of these two essential elements: Removal of Features.

Specifically designed for issues involving the sequence prediction of spatial inputs like images or videos, the Language Model. Convolutional neural network (CNN) layers are used in this design to extract features from the input data, and LSTMs are used to forecast sequences based on the feature vectors. CNN LSTMs are essentially a subclass of deep models that combine aspects of natural language processing and computer vision. They are deep both geographically and chronologically. These candidates show great promise and are being

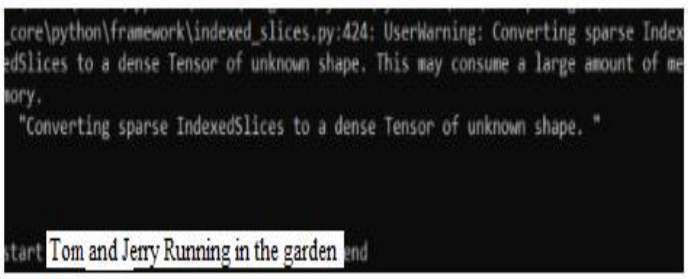
International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6965
***Article Received**: 26 March 2023 **Revised**: 15 May 2023 **Accepted**: 04 June 2023*



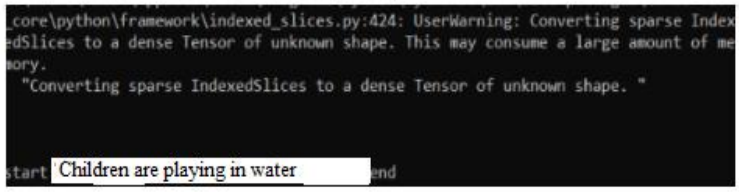
Encoder-decoder architecture serves as the foundation for models for image caption generators, which use effort paths to create precise and relevant captions. This paradigm bridges the gap between natural language processing and computer vision. It entails recognizing and interpreting the image's context before summarizing everything in a language like English.

The two primary models utilized in the creation of our model were RNN-LSTM and CNN (Convolutional Neural Network) (Recurrent Neural Networks- Long Short-Term Memory).

In the derived application, RNN-LSTM acts as the decoder, organizing the words and generating captions, while CNN acts as the encoder, extracting the characteristics from the image or photograph. Some of the program's most important applications include self-driving cars.



The majority of calculations happen in the convolutional layer, which is the main part of a CNN. There could be more convolutional layers after the first. During convolution, a kernel or filter inside of this layer examines the image's receptive fields to see if a feature is present. The entire picture is iterated through by the kernel. At the conclusion of each cycle, a dot product between the input pixels and the filter is computed. The ultimate result of a certain method of connecting the dots is a feature map or a feature that has been convolved. The image is finally converted to numerical values in this layer so that the CNN can recognise them and extract useful patterns from them.



The important features of a picture are extracted by the feature extraction model using a neural network, often as a fixed-length vector. The feature extraction sub model is a deep convolutional neural network, or CNN. It is possible to immediately train this network using the images in your dataset. Another option is a convolutional model that has already been trained. Information on the datasets and the data preparation for the model is available at the same site above.

Here, we'll just focus on the crucial sections of the programme that was utilized to create and run the model. You should get your dataset ready and use a different dataset.

VI. CONCLUSION AND FUTURE WORK

The CNN-LSTM architecture, which controls computer vision and natural language processing, has several uses. Modern neural networks can be used to do NLP tasks such as the transformer for sequential image and video data. For sequential data, like as natural language, extremely powerful CNN networks can be used simultaneously. Therefore, it enables us to apply the advantages of robust models to recently unknown activities. Introduce hybrid neural models and urge readers to employ various CNN-LSTM model designs more frequently were the only objectives of this essay.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 6s
DOI: https://doi.org/10.17762/ijritcc.v11i6s.6965
***Article Received**: 26 March 2023 **Revised**: 15 May 2023 **Accepted**: 04 June 2023*

REFERENCES

[1] Gupta, N., & Jalal, A. S. (2020). Integration of textual cues for fine-grained image captioning using deep CNN and LSTM. *Neural Computing and Applications*, *32*(24), 17899-17908.

[2] Khamparia, A., Pandey, B., Tiwari, S., Gupta, D., Khanna, A., & Rodrigues, J. J. (2020). An integrated hybrid CNN–RNN model for visual description and generation of captions. *Circuits, Systems, and Signal Processing*, *39*(2), 776-788.

[3] Ms. Madhuri Zambre. (2012). Performance Analysis of Positive Lift LUO Converter . International Journal of New Practices in Management and Engineering, 1(01), 09 - 14. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/3

[4] Soh, M. (2016). Learning CNN-LSTM architectures for image caption generation. *Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep*, *1*.

[5] Alzubi, J. A., Jain, R., Nagrath, P., Satapathy, S., Taneja, S., & Gupta, P. (2021). Deep image captioning using an ensemble of CNN and LSTM based deep neural networks. *Journal of Intelligent & Fuzzy Systems*, *40*(4), 5761-5769.

[6] Mondal , D. (2021). Green Channel Roi Estimation in The Ovarian Diseases Classification with The Machine Learning Model . Machine Learning Applications in Engineering Education and Management, 1(1), 07–12.

[7] Al-Muzaini, H. A., Al-Yahya, T. N., & Benhidour, H. (2018). Automatic Arabic image captioning using RNN-LSTM-based language model and CNN. *International Journal of Advanced Computer Science and Applications*, *9*(6).

[8] Sharma, H., Agrahari, M., Singh, S. K., Firoj, M., & Mishra, R. K. (2020, February). Image captioning: a comprehensive survey. In *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)* (pp. 325-328). IEEE.

[9] Khatri, K. ., & Sharma, D. A. . (2020). ECG Signal Analysis for Heart Disease Detection Based on Sensor Data Analysis with Signal Processing by Deep Learning Architectures. Research Journal of Computer Systems and Engineering, 1(1), 06–10. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/11

[10] Chen, M., Ding, G., Zhao, S., Chen, H., Liu, Q., & Han, J. (2017, February). Reference based LSTM for image captioning. In *Thirty-first AAAI conference on artificial intelligence*.

[11] Wang, M., Song, L., Yang, X., & Luo, C. (2016, September). A parallel-fusion RNN-LSTM architecture for image caption generation. In *2016 IEEE international conference on image processing (ICIP)* (pp. 4448-4452). IEEE.

[12] Johnson, M., Williams, P., González, M., Hernandez, M., & Muñoz, S. Applying Machine Learning in Engineering Management: Challenges and Opportunities. Kuwait Journal of Machine Learning, 1(1). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/90

[13] Loganathan, K., Kumar, R. S., Nagaraj, V., & John, T. J. (2020). Cnn & lstm using python for automatic image captioning. *Materials Today: Proceedings*.

[14] Tan, Y. H., & Chan, C. S. (2017). phi-LSTM: a phrase-based hierarchical LSTM model for image captioning. In *Asian conference on computer vision* (pp. 101-117). Springer, Cham.

[15] Al Fatta, H., & Fajar, U. (2019, December). Captioning image using convolutional neural network (CNN) and long-short term memory (LSTM). In *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 263-268). IEEE.

[16] Pa, W. P., & Nwe, T. L. (2020, May). Automatic Myanmar image captioning using CNN and LSTM-based language model. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)* (pp. 139-143).

[17] Ana Silva, Deep Learning Approaches for Computer Vision in Autonomous Vehicles , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[18] Xu, K., Wang, H., & Tang, P. (2017, July). Image captioning with deep LSTM based on sequential residual. In *2017 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 361-366). IEEE.

[19] Liu, M., Li, L., Hu, H., Guan, W., & Tian, J. (2020). Image caption generation with dual attention mechanism. *Information Processing & Management*, *57*(2), 102178.

[20] Yang, Z., Yuan, Y., Wu, Y., Cohen, W. W., & Salakhutdinov, R. R. (2016). Review networks for caption generation. *Advances in neural information processing systems*, *29*.

[21] Wang, H., Zhang, Y., & Yu, X. (2020). An overview of image caption generation methods. *Computational intelligence and neuroscience*, *2020*.