Proceedings of the 6th International Conference on Communication and Electronics Systems (ICCES-2021)
IEEE Xplore Part Number: CFP21AWO-ART; ISBN: 978-0-7381-1405-7

Intelligence Embedded Image Caption Generator using LSTM based RNN Model

Akash Verma[1], Harshit Saxena[1], Mugdha Jaiswal[1], Dr. Poonam Tanwar[2]
[1]Student, *CSE,FET Manav Rachna International Institute of Research and Sciences*, *Faridabad*
[2] Faculty,*CSE,FET Manav Rachna International Institute of Research and Sciences*, *Faridabad*
akashpool103@gmail.com, harshitsaxenaonnet@gmail.com , mugdha101299@gmail.com, poonamtanar.fet@mriu.edu.in

2021 6th International Conference on Communication and Electronics Systems (ICCES) | 978-1-6654-3587-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCES51350.2021.9489253

***Abstract*—Humans tend to extract information from everything they see be it living or non-living. This whole phenomenon motivated us to move in this direction and explore the field of computer vision and how this can be used with recurrent neural networks to generate captions from any image. By witnessing the recent increase in natural language processing-based applications; various other researchers have also worked on this concept and produced commendable results. Describing an image is not an easy task to implement, the structure and semantics of a sentence hold an important weight age in sentence formation. This paper approaches the problem of caption generation with an LSTM (Long-Short Term Memory) based RNN model and builds architecture based on the same to generate efficient and meaningful captions by training the dataset effectively. Flicker8k dataset is used to train our model and worked well. The accuracy of the model is evaluated based on standard evaluation metrics.**

Keywords:- Computer Vision, Flicker8k Dataset, Long-Short Term Memory(LSTM), Natural Language Processing(NLP), RNN, IOT.

1. Introduction

Primary goal of computer vision is scene understanding, not only extracting the information about the objects that are present in the image but also displaying it in a user readable and understandable natural language and format. This paper briefly discusses how the concept of image generation has changed in recent times and how LSTM can be utilized in the existing model to improve the efficiency and accuracy of this model. Many sophisticated models have been developed to extract visual information from images based on visual categorization of objects in the images [9, 17, 19, 20]. The visual recognition procedures thus pursued in most cases are demanding both in terms of computational complexity and obtaining desired accuracy [10, 25].For the validation of our model, we have used flicker8k dataset comprising of 8000 images, which were preprocessed for accurate readings. The paper is organized as- the first part talks about the work that is currently present and which has been done till date in this area, along with the proposed model. The second part describes the working of the model, its accuracy and result analysis and the last part discusses the conclusion along with what in future can be done in order to make this concept more deployable to the end user, and how this can be used for visually challenged people for their betterment and making their life more safe and secure.

II. Literature review

In current era of Machine Learning (ML) and Artificial Intelligence (AI), many researchers have made use of their work and various studies to complement the areas of image caption generation [2,4]. Not a long ago, people proposed the concept of neural language models for caption generation [1]-[5].The idea of translating an image into a sentence that describes it was introduced. Earlier Neural Networks was used for caption generation, two techniques caption templates which works for object detections and attribute discovery [5][20]. Whereas the 2nd technique was based on identifying the similar captioned images first from a huge database after that modifying them by retrieved captions to fit the given query [21]. Karpathy *et al*[7] brought the concept of using multimodal recurrent neural network model for caption generation. Fang et al. proposed a three-step pipeline for generation by including object detections. The first step in their model training was to learn detectors for several visual concepts and understandings which were based on a multi-instance learning framework. Second step included the application of trained model to detect outputs, lastly rescoring from a joint image text embedding space was implemented [6,15].LSTM was also studied by many researchers within the same timeline, many were successful in implementing the architectures while others contemplated towards a theoretical study more.[7,11]They recently articulated the integration of visual attention phenomenon into a desired LSTM model thereby building a model which fix its focus on different entities or objects during the generation of corresponding words. Great prospects have been observed in generation of human-like image captions.

III Architecture and working

A. Convolution Neural Network

CNN Is one form of Deep Neural networks that can process the data/input in the form of 2D matrix to classify and indentify the images/objects. The features of image can be extracted by reading /scanning the image top to bottom/left to right. The features can be analyzed to extract the

information from the image. Transformed image can also be used [32][33][34].

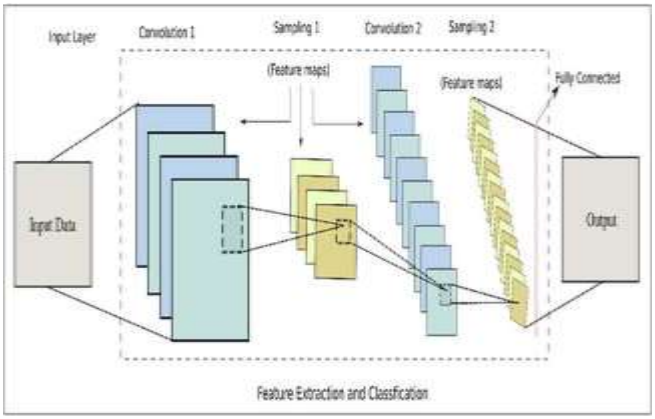

Figure1.Architecture of CNN [23]

Further LSTM which is a RNN can be used for sequence prediction purpose which can predict on the bases of previous input/text and can predict what should be new word[23][26]-[31] .

B. **Working**

This work was brought to life by using python. Many python libraries were used for implementation including Keras which consisted of VCG net responsible for object identification; TensorFlow - developed by Google which is used in the construction of deep learning neural networks by executing various fundamental algorithms

1. Starting with feature extraction using CNN, we applied various layers to our model on the training set for whom we had a corresponding sentence which maps to the activity taking place in the picture.
On applying various layers like conv2d, max pooling, dropout and thus using activation function, we successfully extracted features of every image.
2. Now since we were supposed to extract features from other testing images too so for that we used Google's word2vec model. About word2vec :It is a model by Google which helps us describe our word into digits form for further processing where involvement of words takes place. It is more commonly used in NLP, RNN, and LSTM related works.
3. In word2vec on providing a word, it is converted into a vector of fixed dimensions.
4. Now further using LSTM layerwe maintain the context of words forming sentence because with each word and its position the context of sentence may change.
For example:

1. He is good.
2. He is good for nothing.

So, this "for nothing" changed the context of the complete sentence so this work is handled using LSTM which goes with the weightage of the words also RNN is used as a recurrent model using tan function where the output of one cell is used as the input of next cell.The model worked fine for the training set and also on testing dataset samples also it gave almost 85% good meaningful and accurate captioning.

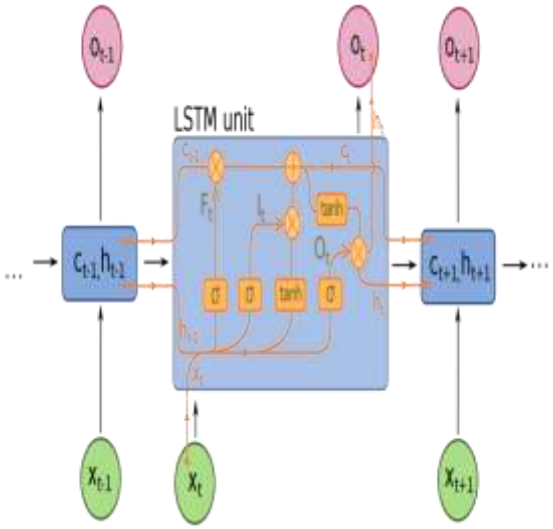

Figure: 2 Architecture of LSTM by François Deloche [24]

IV RESULTS

The dataset that we have used for our research work is available online named as "flicker8k". The dataset was preprocessed and made fit for further evaluation and working. It comprised of 8000 images, out of which we used for training and for testing in a ratio of 70:30. At the time of feature extraction, we had a total of 25,636,712 parameters out of which 25,583,592 were trained successfully and 53,120 were non-trainable parameters. The general confusion matrix was used for analyzing the system performance. This matrix contains result of all models with their predictions. A total of 150 iterations were executed over a standard batch size of 50.

C. Accuracy

The system accuracy was measure using the standard equation given below.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions made}}$$

```
In [8]: trainingmodel(epochs)
Epoch 1/1
2000/2000 [==============================] - 334s 167ms/step - loss: 3.2977
Epoch 1/1
2000/2000 [==============================] - 337s 168ms/step - loss: 3.0494
Epoch 1/1
2000/2000 [==============================] - 333s 166ms/step - loss: 2.8968
Epoch 1/1
2000/2000 [==============================] - 332s 166ms/step - loss: 2.7884
Epoch 1/1
2000/2000 [==============================] - 334s 167ms/step - loss: 2.7053
Epoch 1/1
2000/2000 [==============================] - 333s 167ms/step - loss: 2.6389
Epoch 1/1
2000/2000 [==============================] - 331s 165ms/step - loss: 2.5841
Epoch 1/1
2000/2000 [==============================] - 337s 168ms/step - loss: 2.5392
Epoch 1/1
2000/2000 [==============================] - 337s 168ms/step - loss: 2.5001
Epoch 1/1
2000/2000 [==============================] - 331s 165ms/step - loss: 2.4646
```

Figure.3. Number of epochs executed with loss values

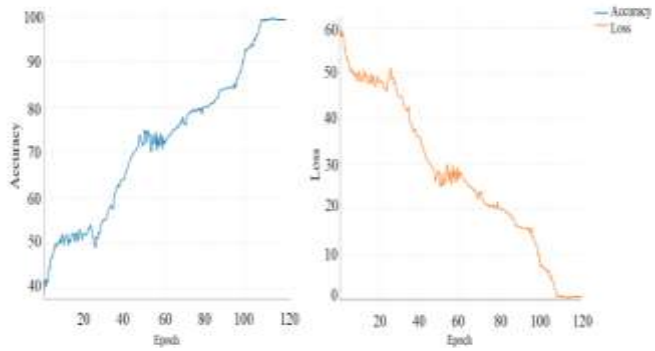

Figure 4. Epoch vs. lossand Epoch vs. accuracy

From the graph we can signify that our model was successful in implementation phase, we look forward to train it on a larger dataset as well.

II. CONCLUSION

Here, we have made the best use of our knowledge and skills in order to design an architecture, overcoming previous drawbacks that were faced in the field of image captioning in constructing a model based on LSTM based RNN capable of scanning and extracting information from any image provided and transforming it to a single line sentence based in a natural language English. Most of the times, it is noted that its quite hard to avoid over fitting of data, but we are glad that we have overcome this difficulty. Main focus was on the algorithmic essence of different attention mechanisms and summarized how the attention mechanism is applied. Hereby we can say that we are successful in building a model that is significantly an improved version of all other previously available image caption generators.

III. FUTURE SCOPE

In the future, we would like to train our model on a larger dataset comprising of a greater number of images thereby resulting in a more accurate and efficient model with wider scope and horizon.Also, we want to deploy our model to a larger audience comprising of blind people mainly.By making use of IoT equipment like Arduino kit, Electrical Equipment, Cameras, and few more things like Bluetooth a product can be manufactured which will help blind people in crossing roads and will ensure that they can travel safely anywhere without depending upon the mercy of others.

As seen in the image, we can embed a camera in the shoe front face so as to get live environment video and we can get a mechanism to connect it wirelessly to Bluetooth in-ear of the blind person. Now using this Arduino equipment, the only difference comes is now the captions will be generated in a dynamic environment and the captions generated will be made to be played in Bluetooth of the blind person via which he can cross with more cautions and this would surely decline the accident and mishappening around with blind people to be specific.

REFERENCES

[1]D.Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In ICLR, 2015.

[2] P. Anderson, X. He, C. Buehler et al., “Bottom-up and top-down attention for image captioning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

[3] K. Cho, B. van Merrienboer, C¸ .Gulc¸ehre, D. Bahdanau, ¨ F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In EMNLP, 2014

[4] I.Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In NIPS, 2014.

[5] Kulkarni, Girish, Premraj, Visruth, Ordonez, Vicente, Dhar, Sagnik, Li, Siming, Choi, Yejin, Berg, Alexander C, and Berg, Tamara L. Babytalk: Understanding and generating simple image descriptions. PAMI, IEEE Transactions on, 35(12):2891– 2903, 2013

[6] Fang, Hao, Gupta, Saurabh, Iandola, Forrest, Srivastava, Rupesh, Deng, Li, Dollar, Piotr, Gao, Jianfeng, He, Xiaodong, Mitchell, ´ Margaret, Platt, John, et al. From

captions to visual concepts and back. arXiv:1411.4952 [cs.CV], November 2014

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015

[8] J. Aneja, A. Deshpande, and S. Alexander, "Convolutional image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

[9] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE Conference on International Conferenceon Computer Vision*, pp. 4904–4912, Las Vegas, NV, USA, June 2016.

[10] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," in *Proceedings of the IEEE Conference on International Conference on Computer Vision*, pp. 1251–1259, Venice, Italy, October 2017.

[12] H. R. Tavakoli, R. Shetty, B. Ali, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in *Proceedings of the IEEE Conference on International Conference on Computer Vision*, pp. 2506–2515, Venice, Italy, October 2017.

[13] A. Mathews, L. Xie, and X. He, "SemStyle: learning to generate stylised image captions using unaligned text," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

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

[15] C. C. Park, B. Kim, and G. Kim, "Towards personalized image captioning via multimodal memory networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 99, p. 1, 2018.

[16] X. Chen, Ma Lin, W. Jiang, J. Yao, and W. Liu, "Regularizing RNNs for caption generation by reconstructing the past with the present," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.

[17] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015.

[18] A. Karpathy, A. Joulin, and F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In NIPS, 2014.

[19] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Multimodal neural language models. In ICML, 2014.

[20] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. TPAMI, 35(12):2891– 2903, 2013.

[21] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In ACL, 2012.

[22] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Generalizing image captions for image-text parallel corpus. In ACL, 2013.

[23] Deep learning in big data Analytics: A comparative study - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/CNN-general-architecture_fig3_321787151 [accessed 5 May, 2021]

[24]Batch 2.0 - generating classical music using recurrent neural networks - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/GRU-unit-courtesy-of-Francois-Deloche-from-Wikipedia_fig3_336547497 [accessed 5 May, 2021]

[25] I. Sutskever, J. Martens, and G. Hinton. Generating text with recurrent neural networks. In ICML, 2011.

[26] Kaustav et.al. "A Facial Expression Recognition System To Predict Emotions", 2020 International Conference on Intelligent Engineering and Management (ICIEM), June, 2020.

[27] Anshu & Poonam," Deep Analysis of Autism Spectrum Disorder Detection Techniques", 2020 International Conference on Intelligent Engineering and Management (ICIEM), June 2020.

[28] Poonam Tanwar, Priyanka," Spam Diffusion in Social Networking Media using Latent Dirichlet Allocation", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-12, October, 2019.

[29] Banita, Poonam Tanwar ,"Evaluation of facial paralysis using Image Computation", International Journal of Engineering and Technology(UAE), International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-3, February 2019

[30] Poonam Tanwar; T.V. Prasad; Kamlesh Dutta," Natural language processing for hybrid knowledge representation", International Journal of Advances Paradigms, Vol 10, issue 3, 2018

[31] Tanwar et.al., A Tour towards Sentiments Analysis using Data Mining, International Journal of Engineering Research & Technology, Volume 05, Issue 12, December 2016.

[32] Suma, V. "A Novel Information retrieval system for distributed cloud using Hybrid Deep Fuzzy Hashing Algorithm." JITDW 2, no. 03 (2020): 151-160.

[33] Manoharan, Samuel. "Supervised Learning for Microclimatic parameter Estimation in a Greenhouse environment for productive Agronomics." Journal of Artificial Intelligence 2, no. 03 (2020): 170-176.

[34] Tanwar Poonam & Rai Priyanka," A proposed system for opinion mining using machine learning, NLP and classifiers", IAES International Journal of Artificial Intelligence (IJ-AI) Vol. 9, No. 4, December 2020, pp. 726~733 ISSN: 2252-8938, DOI: 10.11591/ijai.v9.i4.pp726-733.