International Journal of Innovative Research in Engineering & Management (IJIREM)
ISSN: 2350-0557, Volume-9, Issue-1, February 2022
https://doi.org/10.55524/ijirem.2022.9.1.1
Article ID IJIR-1154, Pages 1-5
www.ijirem.org

Caption Generation of Images Using CNN and LSTM

Ummar Yousuf[1], Ravinder Pal Singh[2], and Monika Mehra[3]

[1]M.Tech Student, Department Electronics and Communication Engineering, RIMT University Mandi Gobindgarh, Punjab India
[2]Technical Head, Department of Research, Innovation and Incubation, RIMT University, Mandi Gobindgarh, Punjab India
[3]Head of Department, Department of Electronics and Communication Engineering, RIMT University, Punjab India

Correspondence should be addressed to Ummar Yousuf; Erummar40@gmail.com

ABSTRACT- The contents of a picture are automatically created in Artificial Intelligence (AI), which combines computer vision and natural language processing (NLP) (Natural Language Processing). It is developed a regenerative neuronal model. Computer vision and machine translation are required. This model is used to produce natural-sounding phrases that describe the picture. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are used in this model (RNN). The CNN is used to extract features from images, while the RNN is used to generate sentences. The model has been trained in such a manner that when an input image is provided to it, it creates captions that almost accurately describe the image. On various datasets, the model's accuracy, smoothness, and command of language learned from picture descriptions are assessed. These tests reveal that the model typically provides correct descriptions of the input image.

Keywords: Long Short Term Memory (LSTM), Deep Learning, Neural Network, Image, Caption, Description

I. INTRODUCTION

Language, whether written or spoken, is used by people to communicate. This language is frequently used by them to describe the visual world around them. For physically handicapped persons, images and signs are another means of communicating and comprehending. Although automatically generating descriptions from photos in correct sentences is a tough and demanding process [1], it can assist and have a significant influence on visually impaired people's understanding of image descriptions on the web. 'Visualizing a picture in the mind' is a common phrase used to describe an image. The ability to conjure up an image in one's imagination can help with sentence production. Humans may also describe a picture after taking a brief look at it. After analyzing current natural picture descriptions, progress in reaching complicated goals of human identification will be made. Automatically creating captions and characterizing images is far more difficult than picture classification and object identification. The description of a picture must
include not just the items in the image, but also the relationships between the objects and their qualities, as well as the activities depicted in the image [20]. The majority of past work in visual recognition has focused on labelling images using pre-determined classes or categories, resulting in significant advances in this discipline. Finally, closed visual concept languages produce an appropriate and straightforward model for assumption.
When compared to the enormous amount of cognitive capacity that humans possess, these notions appear to be severely constrained. However, natural languages such as English should be utilized to communicate information beyond semantics, i.e.,a language model is required for comprehension.

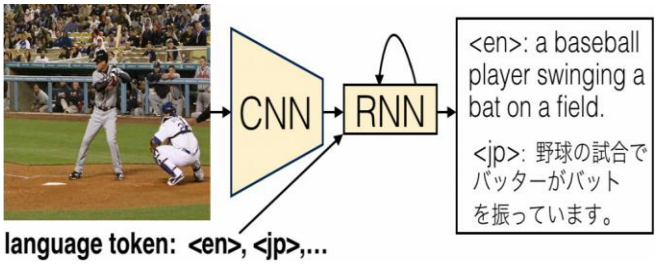

Figure 1: Model based on Neural Networks

Most prior attempts have recommended combining all of the current answers to the above challenge in order to produce description from an image. Alternatively, we will create a single model that accepts an image as input and is trained to produce a series of words, each of which belongs to a dictionary that adequately characterizes the picture, as illustrated in Fig. 1.
The text summarizing problem in natural language processing (NLP) is linked to the relationship between visual significance and descriptions [16][18]. The selection or generation of an abstract for a document is an essential objective of text summarization. In the problem of image

captioning, we want to produce a caption for each image that describes many aspects of that image [21][22].

The approach proposed in this study can generate unique descriptions from pictures. For this job, we used the Flickr 8k dataset, which has 8000 pictures with five descriptions each. Figure 2 shows the dataset structure, which shows that each image has five natural language captions. We use both CNN and RNN in this research. For the picture classification job, a pre-trained Convolutional Neural Network (CNN) is employed. This network performs the function of an image encoder. The last hidden layer is sent into the Recurrent Neural Network as an input (RNN). This network is a sentence-generating decoder. Occasionally, the produced sentence appears to lose track of the original visual information or predicts the incorrect text. This phrase is derived from a description that appears often in the dataset and is only tangentially connected to the input image.

Figure 2: Caption 1: A group of people are seated around a snowy crevasse, Caption 2:

A group of people are seated atop a snowy mountain Caption 3: A group of people sits in the snow, looking out over a mountain landscape. Caption 4: Five youngsters prepare to sled, Caption 5: Five individuals sit in the snow together.

II. RELATED WORK

The challenge of producing natural language descriptions from visual input has long been explored in computer vision [1][2][3]. Three types may be found in the literature on picture caption generation. Template-based techniques [4][5][7] fall under the first group. Detecting objects, actions, scenes, and characteristics takes precedence in this method. The transfer-based caption generating systems [6][8] fall under the second group. Image retrieval is carried out in this method. This method retrieves visually comparable photos, and then uses the captions of these images to generate the query image. The majority of the researchers stated that neural networks are effective in machine translation [10], as well as caption creation using neural language models. Rather of translating a text from a source language into a necessary format, the objective is to transform a picture into a sentence that describes it. As a result, the system has become more complicated. They are made up of visual radical recognizers that are programmed in a formal language, such as And-Or Graphs or logic systems, and then converted utilizing rule-based systems. The multimodal recurrent neural network model proposed by Mao et al. and Karpathy et al. is utilized for picture description generation. Vinyals, Oriol, and colleagues utilized the NIC model (Neural Image Caption). CNN is the encoder used in the NIC model. For picture categorization, the pertained CNN is employed, and the last layer of the network is used as input to the RNN decoder. This RNN decoder is also capable of generating sentences. They employed an advanced form of RNN called LSTM [1]. Recently proposed that visual attention be summarized in the LSTM model for focusing its gaze on distinct objects during the production of associated phrases[13]. In order to generate human-like image captions, neural language models are important. With the exception of the most current techniques, they all use a similar encoding decoding architecture [13][12],which combines caption creation and visual attention. This research focused on the third category of caption generating methods. In this method, a neural model is created that provides natural language descriptions for images. As an image encoder, CNN is employed. The RNN decoder utilizes this last hidden layer as input to construct the phrase after pre-training for the picture classification job.

III. APPROACH

In this paper, a neural framework for generating captions from photos is given, which is based on probability theory. It is feasible to get better outcomes by employing a strong mathematical model that optimizes the probability of proper translation for both inference and training.

A. Convolutional Neural Network (CNN)

Visual recognition is presently using convolutional networks. CNN has a number of convolutional layers. Following these convolutional layers, like in a multilayer neural network [14], the next levels are fully linked layers. The CNN is built in such a way that it can take use of the input image's 2D structure. The amount of local connections and linked weights, as well as different pooling strategies, are used to achieve this goal, resulting in translation invariant characteristics. The primary benefits of CNN are its simplicity of training and the fact that it has fewer parameters than other networks with the same number of hidden states.

We used the Visual Group Geometry (VGG) network for this study, which is a Deep CNN for large-scale picture identification [15]. It is available in both 16 and 19 layer configurations. The classification error values for both 16 and 19 layers are nearly identical in the validation and test sets, at about 7.4% and 7.3 percent respectively. This model describes the characteristics of pictures that are utilized in the caption creation process.

B. Long Short-Term Memory (LSTM)

A recurrent neural network [17][19] is used to simulate the transient dynamics of a group of objects. Ordinary RNNs have a tough time acquiring long-term dynamics because to vanishing and exploding weights or gradients [9]. The memory cell is the LSTM's basic building unit. It keeps track of the current value over a lengthy period of time. Gates are used to regulate the update time of a cell's state. Variants are represented by the number of connections between memory cells and gates.

As illustrated in Fig. 3, our model is based on the LSTM block, which is dependent on the LSTM with no peephole design. The following are the relationships between LSTM memory cells and gates:

il = σ(W ixxl + W imml-1) (1)

fl = σ(W fxxl + W fmml-1) (2)

ol = σ(W oxxl + W ommll-1) (3)

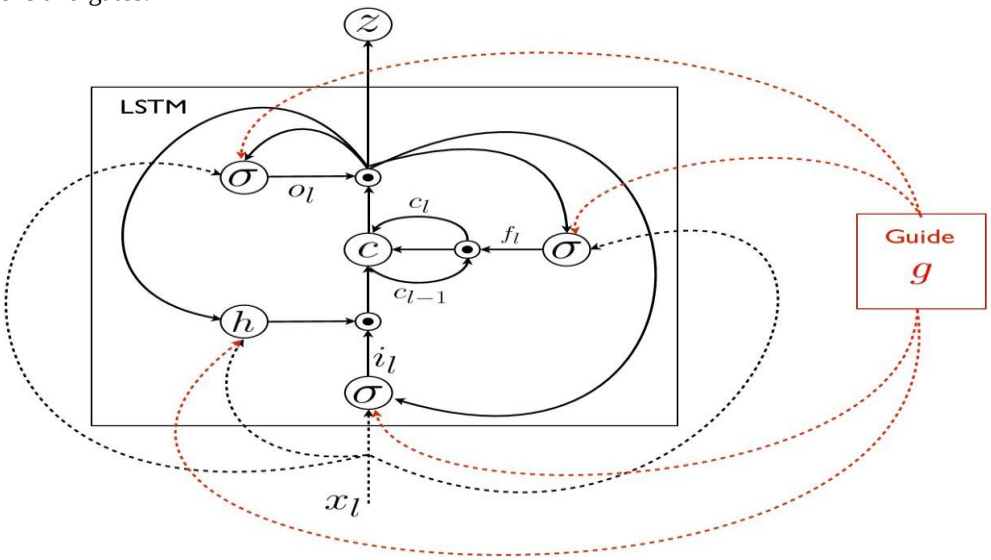

Figure 3: Connection diagram of LSTM [9]

$c_l = f_l \odot c_{l-1} + i_l \odot \hbar(W_{cx} x_l + W_{cm} m_{l-1})$ (4)

$m_l = o_l \odot c_l$ (5)

$$L(I,S) = -\sum_{t=1}^{n} \log(p_t(S_t)) \quad (6)$$

where σ(.) indicates the sigmoid function and $\hbar$ denotes the hyperbolic tangent function, and $\odot$ denotes elementwise multiplication. The variables i1 denotes the input gate, f1 denotes the forget gate, o1 denotes the output gate in the LSTM cell, c1 denotes the state of the memory cell unit, and ml denotes the hidden state, which is the output of the block generated after processing in the LSTM, x1 denotes a sequence parameter at time step l, and variable W[.][.] denotes model parameters. The loss function is represented by Eq. 6, with St denoting the produced sentence at time t. With regard to all LSTM and word embedding settings, this loss is always reduced.

IV. GENERATION OF SENTENCE WITH LSTM

The idea of encoder decoder in network and machine translation modelling [1][11][13][18] is used to represent phrase production in neural networks. The encoder is used to convert a variable sequence of words in natural language to a distributed vector in this paradigm. Then, based on the mapped vectors, a new sequence of words is created using a decoder in natural target language. The goal of the training procedure is to increase the likelihood of flawless translation such that the phrase is in the native source language. When this approach is used to caption generation, the goal is to maximize the quantity of image captions created for a given image, namely:

$\arg\theta \sum i \log(p(s1{:}Li \,|xi,\theta))$ (7)

where xi represents an image, s1:Li signifies collection of words in correctly made sentence of length Li and θ signifies model constraints. For ease of execution, in the following step we disregard the superscript i whenever it is not important or cleared from the situation .As an arrangement of words generate each sentence, the Bayes chain rule is used to split sentence which comprises of words as its elementary element.

$\log(p(s1{:}L|x,\theta)) = \log(p(s1|x,\theta)) + \sum_{i=2} \log(p(s1|x,s1{:}l\text{-}1,\theta))$ (8)

where s1:L signifies the block from sentence produced up to the l-th word. In entire exercise process, to maximize the drive in Eq. 7, we have demarcated the log-likelihood $\log(p(s1{:}Li \,|xi, \theta))$, it can be used with the concealed state in RNN. At time step l+1 the possibility circulation of word for the entire terminology can be calculated with the assistance of softmax function z(.) which is founded on output ml of the retention cell, pl+1 = z (ml) alike to [1] .

Before being used as inputs to LSTM, images and phrases are encoded as fixed-length vectors. First, CNN features are calculated for each picture, and then they are transferred to the embedding matrix. Concatenating a series of words with an image in a phrase yields a new sequence. The picture is considered as the beginning symbol of the new sequence, while the sequence of words is treated as the remaining part of the new sequence. This new sequence is used as an input to the LSTM network for training purpose by iterating the recurrence connection for l from 1 to Li. The neural model's parameters include the linear transfer matrix for picture features, the word embedding matrix, and various LSTM arguments.

The image caption model is divided into three parts. The first is the picture model, which repeats the image feature vector 28 times with a dimension of 28 x 4096, where 28 is the maximum number of words in a caption. The second model is a language model that consists of a single LSTM unit that produces a matrix with dimensions of 28 x 256, where 256 is the LSTM unit's output size, and the final model merges these two vectors and passes them to another LSTM unit with output dimensions of 28 × 915. We send the same encoded text vector as the target vector for training, but for testing, we simply encode "sol" to feature vector along with the test image feature vector, resulting in a matrix of size 28 x 915, which we decode into sequence words.

V. RESULTS

A. Datasets

Images and picture descriptions in the form of sentences in natural language, such as English, are included in these databases. Table I shows the statistics of the datasets. Observers characterize each image in these datasets with five distinct phrases that are generally apparent and neutral.

B. Results

For 5 epochs, the model has been trained. Because there are more epochs used, the loss is reduced to 3.74. If we cogitate the huge dataset then we should practice additional epochs for precise outcomes.

Table 1: Dataset

Dataset Name	Size		
	Train	Valid	Test
Flickr8k [1]	6000	1000	1000

Figure 4: Selection of evaluation of results

Figure 4 depicts some of the data obtained. BLEU = 0.5335 was obtained by training the model on the Flickr8k dataset and testing it on the 1000 test pictures provided in the dataset.

VI. CONCLUSION

The model presented in this paper is a neural network that can automatically view an image and produce relevant captions in natural language such as English. From a given image, the model is trained to generate a phrase or description. The model's descriptions or captions are divided into three categories:

- Descriptions free of mistakes
- Minor mistakes in the description
- A description that is connected to the image
- A description that has nothing to do with the image

The categories in the results are due to the proximity of some words, e.g., for the term automobile, nearby words

such as vehicle, van, cab, and so on are produced, which may be erroneous. After a slew of tests, it's clear that using larger datasets improves the model's performance. The additional dataset will boost accuracy while lowering losses. It will also be fascinating to see how unsupervised data for both photos and text can be used to improve image caption creation methods.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ACKNOWLEDGMENT

This research is not funded by any university or organization.

REFERENCES

[1] Vinyals, Oriol, et al. ”Show and tell: A neural image caption generator.” Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, 2015.

[2] Gerber, Ralf, and N-H. Nagel. ”Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences.” Image Processing, 1996. Proceedings., International Conference on. Vol. 2. IEEE, 1996.

[3] Yao, Benjamin Z., et al. ”I2t: Image parsing to text description.” Proceedings of the IEEE 98.8 (2010): 1485-1508.

[4] Farhadi, Ali, et al. ”Every picture tells a story: Generating sentences from images.” Euro-pean conference on computer vision. Springer, Berlin, Heidelberg, 2010.

[5] Yang, Yezhou, et al. ”Corpus-guided sentence generation of natural images.” Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[6] Kulkarni, Girish, et al. "Babytalk: Understanding and generating simple image descriptions." IEEE Transactions on Pattern Analysis and Machine Intelligence 35.12 (2013): 2891-2903.

[7] Mitchell, Margaret, et al. "Midge: Generating image descriptions from computer vision detections." Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012.

[8] Kuznetsova, Polina, et al. "Collective generation of natural image descriptions." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012.

[9] Jia, Xu, et al. ”Guiding long-short term memory for image caption generation.” arXiv pre-print arXiv:1509.04942 (2015).

[10] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. ”Neural machine translation by jointly learning to align and translate.” arXiv preprint arXiv:1409.0473 (2014).

[11] Mao, Junhua, et al. ”Deep captioning with multimodal recurrent neural networks (m-rnn).” arXiv preprint arXiv:1412.6632 (2014).

[12] Karpathy, Andrej, and Li Fei-Fei. ”Deep visual-semantic alignments for generating image descriptions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[13] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International Conference on Machine Learning. 2015.

[14] El Housseini, Ali, Abdelmalek Toumi, and Ali Khenchaf. ”Deep Learning for target recognition from SAR images.” Detection Systems Architectures and Technologies (DAT), Seminar on. IEEE, 2017.

[15] Simonyan, Karen, and Andrew Zisserman. ”Very deep convolutional networks for large-scale image recognition.” arXiv preprint arXiv:1409.1556 (2014).

[16] Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[17] Lu, Jiasen, et al. ”Knowing when to look: Adaptive attention via a visual sentinel for im-age captioning.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 6. 2017.

[18] Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. ”Im2text: Describing images us-ing 1 million captioned photographs.” Advances in neural information processing systems. 2011.

[19] Chen, Xinlei, and C. Lawrence Zitnick. ”Mind’s eye: A recurrent visual representation for image caption generation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[20] Feng, Yansong, and Mirella Lapata. ”How many words is a picture worth? automatic caption generation for news images.” Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

[21] Rashtchian, Cyrus, et al. ”Collecting image annotations using Amazon’s Mechanical Turk.” Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk.
Association for Computational Linguistics, 2010.