JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2260 - 2267
https://publishoa.com
ISSN: 1309-3452

Image Caption Generation using Deep Learning Technique

[1]Shailendra S. Aote, [2]Sana Syed

[1,2]Shri Ramdeobaba College of Engineering and Management, Nagpur, India

ABSTRACT
The contents of a picture are auto assembled in Artificial Intelligence (AI), which incorporates computer vision and natural language processing (NLP). The regenerative neuronal model is constructed for which computer vision and machine restatements are needed. This approach is used to create natural-sounding statements that describe the image. In this model, convolutional neural networks (CNN) and intermittent neural networks are used in this model. CNN is used to extract features from images, while the RNN is used to generate rulings. The model has been trained in such a manner that when an input image is handed to it, it creates captions that almost precisely describe the image. On colorful datasets, the model's delicacy, smoothness, and command of language learned from visual descriptions are investigated.

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

1. INTRODUCTION

Language, whether written or spoken, is used to communicate. They frequently use this phrase to describe their surroundings. For physically handicapped people, images and signs are another means of communicating and comprehending. Although automatically generating descriptions from photographs in suitable sentences is a complex and demanding process, it can assist and have a significant influence on visually impaired people's understanding of image descriptions in the web. A popular word used to describe a vision is "visualizing a picture in your head." The development of a picture in one's imagination can help with sentence production. Humans may also describe a picture after taking a cursory look at it. After reviewing current natural visual descriptions, progress in accomplishing challenging goals of human identification will be made. It's significantly more difficult to create captions and characterize photos automatically than it is to classify and identify photographs. The description of a picture must include not just the items in the image, but also the relationships between objects and their properties, as well as the activities depicted in the image. The majority of prior work in visual recognition has focused on labeling pictures using pre-determined groups or categories, resulting in significant advances in this discipline. Finally, closed visual idea vocabularies produce a sufficient and straightforward model for the assumption. These are extremely useful in picture identification, image categorization, image captioning, and a variety of other AI applications. Captioning a picture generates explanations of what is happening in the image. This model takes the image as input and gives a caption for it. With the advancement of technology, the efficiency of image caption generation is also increasing. This image captioning is extremely beneficial for a variety of applications, including the increasingly popular self-driving automobiles. Machine learning tasks for recommendation systems can benefit from image captioning. For picture, captioning several models have been proposed, including the object detection model, visual attention-based image captioning, and Image Captioning using Deep Learning. When compared to the enormous amount of mental capacity that humans possess, these conceptions appear to be severely constrained. However, natural languages such as English should be employed to represent information beyond semantics, i.e., a language model is required for visual comprehension. Most earlier attempts at generating a description from a picture advised combining all of the available answers to the problem. Alternatively, we may train a single model that takes an image as input and generates a series of words, each of which belongs to a dictionary. That adequately characterizes The text summarizing problem in natural language processing is linked to the relationship between visual significance and descriptions (NLP).

2. RELATED WORK

The topic of creating natural language descriptions from visual input has long been researched in computer vision [1] – [3]. The literature on the creation of photo captions lists three different kinds. The first category includes approaches that use templates. This method prioritizes the detection of objects, activities, situations, and attributes [4]–[7]. The second group includes the transfer-based caption producing systems. This method involves retrieving images. This technique finds visually similar pictures, and then creates the query picture from the captions of those pictures [8]. The majority of the researchers suggested that neural networks are beneficial in machine translation, as well as caption synthesis using neural language models [10]As a result, the system has become more complicated. They are composed of visual radical recognizers that were first converted using rule-based systems after being written in a formal language, such as and-or graphs or logic systems. The multimodal recurrent neural network model proposed by Mao et al. [11]

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2260 - 2267
https://publishoa.com
ISSN: 1309-3452

and Karpathy et al. is employed for a visual description generation [12]. Vinyals, Oriol, and colleagues employed the NIC model (Neural Image Caption) [1].
They utilized an advanced form of RNN called LSTM [1]. Xu et al. recently proposed that visual attention be summarized in the LSTM model for focusing its gaze on diverse objects during the creation of associated phrases [13]. To generate human-like image captions, neural language models are useful. Except for the most current approaches, the majority of them use a similar encoding decoding architecture, which combines caption production with visual attention [13]. This research focused on the third category of caption creation 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 employs this last hidden layer as input to create the phrase after pre-training for the picture classification job. CNN is the encoder used in the NIC model. For picture categorization, the pre-trained 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.

3 PROPOSED CNN-LSTM MODEL

This technique is used to construct natural language phrases that express the image. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are both used in this model (RNN). For picture, feature extraction, CNN is employed, and for phrase production, RNN is used.

3.1 Deep Learning

Deep Learning (DL) is an innovation in machine learning (ML) that is based on Artificial Neural Networks (ANN) and incorporates several hidden layers. It's a technique for teaching computers to complete jobs that can only be done by people. Deep Learning algorithms can provide outcomes that are superior to human performance. Several methods, In applications including speech recognition, video recognition, natural language processing, convolutional neural networks, and recurrent neural networks are employed.

3.2 Neural Network

Human beings are blessed with the capability to think, imagine and memorize. Artificial intelligence tries to mimic this behavior and this is what forms the basis for neural networks. They may be seen as a collection of algorithms that are attempting to imitate the workings of the human mind. Neuronal networks serve as the building blocks of the human brain and convey impulses to the brain. Similar to this, neural networks have numerous layers. They attempt to identify the underlying patterns in the supplied data. It has various layers. A perceptron, which is a model of a single neuron and a precursor to larger neural networks, is the name given to each node in a neural network. It feeds a nonlinear or irregular activation function with the signal produced by several linear regressions.

3.3 Convolution Neural Networking

Convolutional neural networks, often known as CNN or ConvNet, are a type of artificial neural network commonly used for image processing. It may also be utilized to solve classification and data analysis issues. CNN is an Artificial Neural Network that detects patterns using classification and attempts to deduce meaning from them.
In more technical terms, CNN is a deep learning technique that allows a user to enter a picture and the algorithm to assign learnable biases and weights to distinct objects or characteristics in the image, allowing it to distinguish one thing from another.

3.3.1 WORKFLOW OF CNN:

All of the necessary ideas for a convolutional neural network have been covered. The input layer is used to provide input to our network, which should be a three-dimensional picture. It might be colorful or black and white. After that, the picture is moved to the convolutional layer. Filters are applied to the input picture in this layer.
Convolution is the term for this procedure. After that, the activation function is used. Here, the RELU activation function is used. Then our output becomes an input to the pooling layer, which helps to reduce the image's resolution. The convolutional layer, on the other hand, goes through the same procedure. It's feasible to have more than a single convolutional layer.
As we progress through the convolutional layers, the patterns get more complex, such as eyes, faces, and birds. The matrix is then flattened before the completely linked layer is added. This layer aids in classifying the items in the supplied picture. The probability of classes is calculated using the Softmax formula. Finally, the output layer is created, which contains the items that were present in the picture.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2260 - 2267
https://publishoa.com
ISSN: 1309-3452

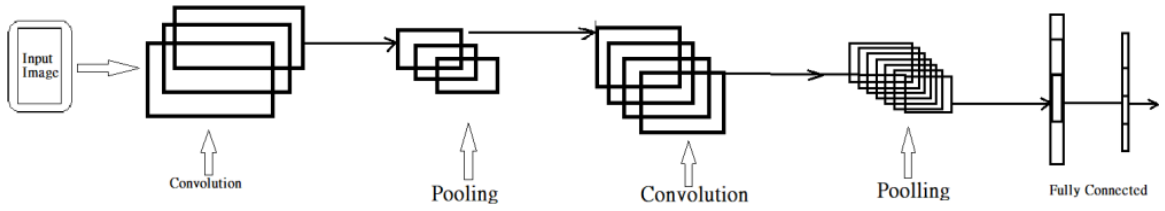

Fig 1. Workflow of CNN

3.4 Recurrent Neural Network

Humans are capable of developing a worldview based on their memories and previous experiences. Thoughts have some permanence, allowing people to employ them as memories for later reference. Traditional neural networks are incapable of storing prior data. This is one of the important reasons hence Recurrent Neural Networks (RNN) are employed.

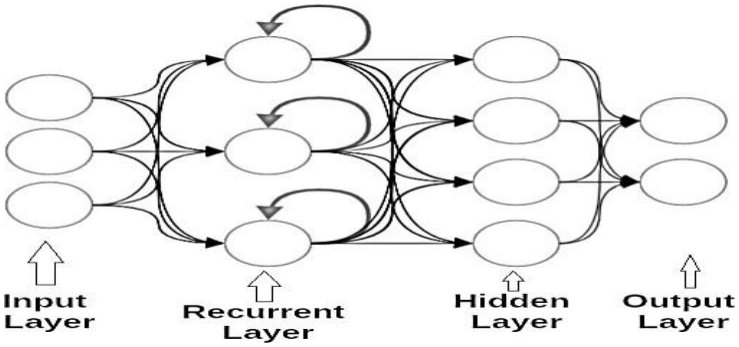

Fig .2 RECURRENT NEURAL NETWORK

Loops are present in RNNs. A loop aids in the transfer of data from one model phase to the next. RNN is designed using no of layers, each of which sends information to the next and allows information to endure.

3.4.1 Long Short Term Memory (LSTM):

LSTM is a type of Recurrent Neural Network, which is used in cases where memory persistence for a longer period is required. It is often considered the more enhanced version of RNN. It can perform almost every task that an RNN network can perform. They can understand long–term dependencies. LSTMs are designed to tackle the issue of long-term dependency.

- **Workflow of LSTM:**

The LSTM is a sort of recurrent neural network that is utilized in situations where memory persistence over time is essential. It is frequently seen as a more advanced form of RNN. It can practically complete every task that an RNN network can. They are aware of long-term dependence. LSTMs were created to address the problem of long-term reliance.

The persistent information can be utilized in concert with the input data for that layer in the case of RNN, and output is anticipated. The information from timestamp 't-1' is given along with the input data at timestamp 't' to forecast the output at timestamp t, as shown in the diagram. The technique used to train the model is called back propagation over time (BTT). As a result, RNN functionality is necessary if we want to use persistence to implement the model.

- **Understanding of LSTMs:**

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2260 - 2267
https://publishoa.com
ISSN: 1309-3452

The "cell state," which resembles a conveyor belt, is one of the major components of LSTM. It's a simple horizontal line that runs the length of the structural chain and has few interactions. "Gates" are used to remove undesired or unneeded data from a cell's state or to introduce new data. The sigmoid layer generates either a 0 or a 1 as its output, Where 1 means "Allows all information to flow through" and 0 means "Dumps all information from the cell state." The cell state is protected and controlled by three gates. The "forget gate layer" is the first layer whereby the input data passes through LSTM. The sigmoid layer makes the judgment on which data to reject in this layer.
The "input gate layer" is the second layer through which the input data passes in LSTM. We determine which new information should be added to the cell state in this layer. This layer consists of two components. The sigmoid function is the first, and the tan (h) the layer is the second.

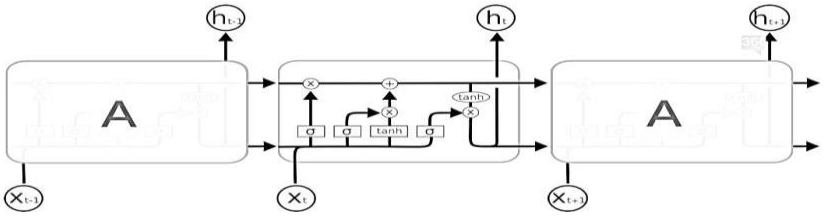

Fig .4 WORKFLOW OF LSTM

In this paper, a neural framework for the caption generation from images is provided, which is based on probability theory. Using a complex mathematical model that maximizes the likelihood of accurate translation for both inference and training, it is possible to get superior results.

3.5 Steps involved in CNN-LSTM Model:

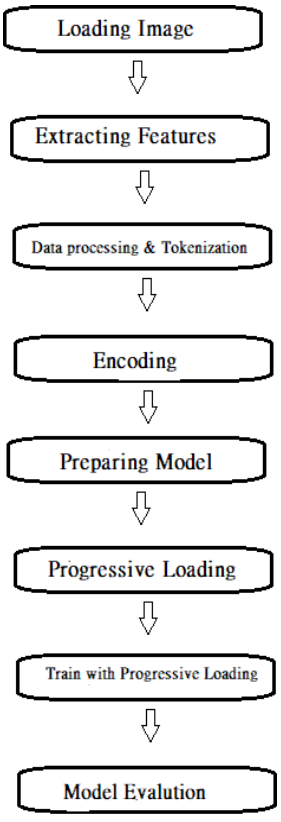

Fig .5 Steps in CNN-LSTM Model

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2260 - 2267
https://publishoa.com
ISSN: 1309-3452

3.5.1 Image Loading:

The image is loaded with the help of keras. It has a function that loads the pictures as a matrix, which is then translated into a NumPy array using keras, which then loads the photos in specified size as specified in a VGG model.

3.5.2 Extracting Features:

The next step is to extract features from the photos that have been loaded. To interpret the important aspects of the pictures, a pre-trained model is employed. The needed weights are obtained using the VGG model, and feature extraction is performed.

3.5.3 Data Processing and Tokenization:

Tokenizing and handling descriptions were necessary. Converting to lowercase, deleting punctuation, removing extraneous words like 'a,' and removing digits are all examples of cleaning. If the vocabulary is limited, the model will train quickly but will not be expressive. A new file is created with the picture IDs and descriptions.

3.5.4 Encoding:

The words must be encoded for processing, and because we will be using probability, Keras will be utilised to build the descriptions. We begin by mapping the picture IDs to the current descriptions.

3.5.5 Preparing Model:

The sentence is generated using a basic model that creates the sentence word by word. The picture is the input, and the recently predicted word is the output, and model is termed recursively because it utilises previously predicted words to produce new words, and it employs input and output pairs, and new words are predicted by probability.

3.5.6 Progressive Loading:

If the processing capacity of the computer facility is insufficient, the photographs and descriptions can be loaded in stages. We may load a function in a progressive manner using Keras. It's used to generate batches of samples for model training, and the generator returns an array of input and output values for the model. The input consists of pictures and encoded-word sequences in the form of an array. The output is the heated encoded words.

3.5.7 Train Progressive Loading:

The model is trained using a data generator and a function on the model. After each epoch (one complete representation of the data set), the model is saved, models are built, and the model with the lowest loss is chosen.

3.5.8 Model Evaluation:

We examine the model after it has been developed. The model is evaluated using the BLEU (Bilingual Evaluation Understudy Score). It provides information about the text's quality. The created sentence is compared to the reference sentence. The model's Bilingual Evaluation Understudy Score is calculated, and the model produces a higher score.

4. RESULTS & ANALYSIS:

As per Hybrid CNN-LSTM Based Approach for Anomaly Detection Systems in SDNs1 [21] experimental results are evaluated below on different models.

Models	**Precision (%)**		**Recall (%)**		**F1-Score (%)**	
	Normal	**Attack**	**Normal**	**Attack**	**Normal**	**Attack**
CNN Standard	76.69	98.86	97.47	88.11	85.84	93.18
LSTM	84.53	98.31	96.02	92.95	89.91	95.55

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2260 - 2267
https://publishoa.com
ISSN: 1309-3452

CNN (L2Reg.L2Reg.)	84.24	**98.56**	**96.62**	92.75	90.00	95.56
CNN-LSTM	**93.18**	97.60	94.04	**97.24**	**93.61**	**97.42**

Table 1: Precision, Recall, F1-score of the different methods.

As can be observed, the suggested CNN-LSTM model outperforms the other techniques. Table 1 shows that when compared to alternative DL models; the typical CNN algorithm performs poorly. The average accuracy of a normal CNN is 90.79 percent; however, when regularization methods are used, the accuracy of CNN is greatly improved, with a score of 93.83 percent. Furthermore, the performance of the LSTM is somewhat better than that of the ordinary CNN, but significantly lower than that of CNN with regularization. Furthermore, combining CNN and LSTM beats all other algorithms, with 96.32 percent accuracy, demonstrating the efficiency of the proposed hybrid CNN-LSTM model in intrusion detection. [21] In properly recognizing assault events, the CNN-LSTM model has a greater degree of accuracy (0.97).

The CNN - LSTM model was designed to determine sequence prediction issues along with spatial inputs like images or videos. In this approach, CNN layers for extracting characteristics on input data are combined with LSTM layers for time series prediction on the feature vectors. In a word, CNN LSTMs are a sort of deep model is disclosed at the junction of computer vision and natural language processing that is both spatially and temporally deep. These models have great potential, and they are increasingly being employed for more difficult jobs like text distribution and video conversion.

A. Datasets

Pictures and descriptions of pictures in the form of text in natural languages, such as English, are included in these databases. Table I showing the descriptive statistics of the datasets. Observers characterize each picture in these datasets with five distinct statements that seem to be generally apparent and neutral.

B. Results

After we have defined and fitted the model for 50 epochs, we trained this model. The correctness of captions produced during the early epochs of training is relatively poor, and they are not closely connected to the test images. We've noticed that the captions generated are somewhat similar to provided test photos after training the model for at least 20 epochs. When the model has been trained for 50 epochs, we see that the model's accuracy improves and the captions it generates become more closely connected to the test pictures, as seen in the following figures. For 50 epochs, the model was trained. Because there are more epochs used, the loss is reduced. If we consider the big dataset then we should use more epochs for more accurate output.

TABLE II : DATASET STATISTICS

Dataset Name	**Size**		
	Train	***Valid***	***Test***
Flickr8k [1]	6000	1200	1200
Flickr30k [1]	28000	1200	1200
MSCOCO [1]	82783	40510	40775

Considering the large dataset then we should use more epochs for accurate results.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2260 - 2267
https://publishoa.com
ISSN: 1309-3452

Fig. 6 Selection of Evaluation Results

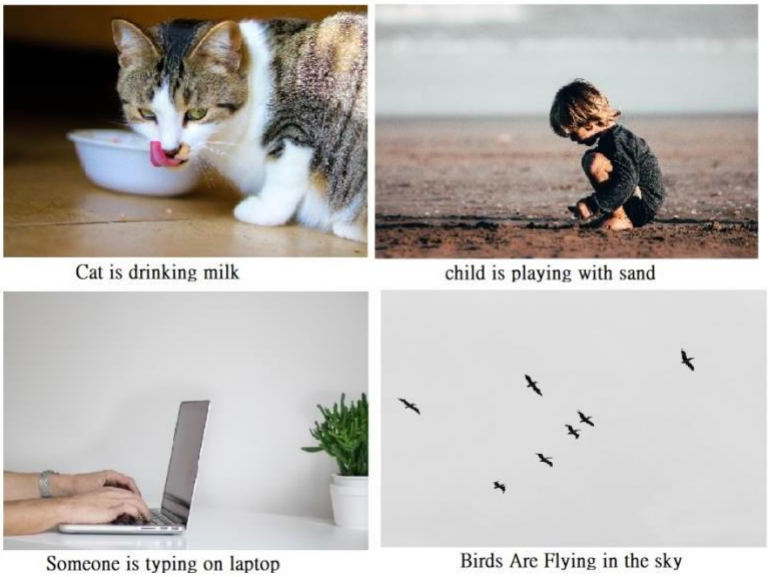

Figure 6 depicts some of the data obtained. BLEU (**Bilingual Evaluation Understudy Score**) = 0.53356 was obtained using the Flickr8k dataset as the training model and a test on 1000 test photos included in the dataset. Conducting a test on the same amount of test photos available in the Flickr30k dataset yields BLEU = 0.61433, whereas running a test on images yields BLEU = 0.67257 for the MS COCO dataset.

5. CONCLUSION & FUTURE WORK:

As a result, the Image Caption Generator is a useful tool that may be used for a variety of applications. It contributes to making life easier to automating the chore of creating labels for photographs in an orderly manner. It might be used to communicate visuals to blind or low-vision persons who rely on noises and text to describe a scene. It's standard practice in web development to offer a description for each picture that appears on the website, so that the image may be read or heard rather than just viewed. This Image Captioning deep learning model is highly effective for analyzing vast volumes of unstructured and unlabeled data to uncover patterns in such photos to lead self-driving cars and construct applications to assist blind people.

We discussed how to generate captions for photographs in our paper. Even if deep learning is evolved, perfect caption production is still not achievable owing to a variety of factors such as hardware requirements, a lack of correct programming logic or model to create exact captions, and the fact that computers cannot think or make judgments as precisely as humans. As hardware and deep learning models progress, we aim to be able to create captions with greater accuracy in the future. It's also possible to enhance this model and create a full Picture-Voice conversion by translating image captions to speech. This is beneficial to blind individuals.

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 theConference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[6] Kulkarni, Girish, et al. "Baby talk: 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] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2018. A Comprehensive Survey of Deep Learning for Image Captioning. ACM Comput. Surv. 0, 0, Article 0 (October 2018), 36 pages. Computing methodologies→Machine learning; Neural networks.

[12] "A gentle introduction to deep learning Caption Generation Models", by Jason Brownlee, November 22, 2017, For deep learning Natural Language Processing.

[13] O.Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge", IEEE transactions on Pattern Analysis and Machine Intelligence, 2016.

[14] Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. CoRR, abs/1506.07285, 2015.

[15] Quanzeng You, HailinJin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semanticattention. CoRR, abs/1603.03925, 2016.

[16] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10, pages 15–29, Berlin, Heidelberg, 2010. SpringerVerlag
.
[17] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, in Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014, pp. 2042–2050

[18] M. Malinowski, M. Rohrbach, M. Fritz, Ask your neurons: a neural-based approach to answering questions about images, in International Conference on Computer Vision, 2015

[19] X. Chen, C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation", CVPR, 2015

[20] Understanding LSTM Networks by Colah, online blog, 2015

[21] Mahmound Abdullah, Nhien An Le Khac, Hamed Jahromi, Anca Delia Jurcut, "A Hybrid CNN-LSTM Based Approach for Anomaly Detection Systems in SDNs" The 16th International Conference on Availability, Reliability and Security, Vienna, Austria, August 2021