Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Image Captioning Generation Using Inception V3 and Attention Mechanism

Yasir Hameed Zaidan [a, *], ***Jumana Waleed*** [b]

[a]*Department of Computer Science, College of Science, University of Diyala, Iraq.Email: scicompms2132@uodiyala.edu.iq*

[b]*Department of Computer Science, College of Science, University of Diyala, Iraq.Email: jumanawaleed@uodiyala.edu.iq*

ARTICLE INFO

Article history:
Received: 15 /02/2023
Rrevised form: 30 /03/2023
Accepted : 10 /04/2023
Available online: 30 /06/2023

Keywords:

Image Captioning Generation

Inception V3

LSTM

Attention Mechanism

ABSTRACT

Captioning an image is the process of using a visual comprehension system with a model of language, by which we can construct sentences that are meaningful and syntactically accurate for an image. The goal is to train a deep learning model to learn the correspondence between an image and its textual description. This is a challenging task due to the inherent complexity and subjectivity of language, as well as the visual variability of images. Computer vision and natural language processing are both used in the difficult task of image captioning. In this paper, an end to end deep learning-based image captioning system using Inception V3 and Long-Short Term Memory (LSTM) with an attention mechanism is implemented. Expansive experimentation has been realized on one of the benchmark datasets named MS COCO, and the experiential results signify that this intended system is capable of surpassing diverse related systems concerning the extensively utilized measures of evaluation, and the accomplished results were 0.543, 0.87, 0.66, 0.51, 0.42 for Meteor and BLEU(B1-B4), respectively.

MSC..

https://doi.org/10.29304/jqcm.2023.15.2.1228

1. Introduction

Image caption generation represents an essential process in computer vision and multimedia domains, pointing to produce depictive representations for the specified image. In particular, image caption generation could provide the capability of computers to comprehend and explain the surroundings [1]. Practically, the approaches utilized for describing images can be categorized into several kinds; template, retrieval, and deep learning based approaches [2]. These approaches should supply informative, precise, and natural expressions (phrases), and accurately specify the image content like scene, relationship, and objects and their properties in the image. But, during the process of image caption generation, an accurate explanation of image content can be a challenging task since it is not potential to exploit all the visual information of the image [3].

∗Corresponding author

Email addresses: scicompms2132@uodiyala.edu.iq

Communicated by ‘sub etitor’

Deep learning approaches like recurrent neural network (RNN), and convolutional neural network (CNN) exhibited efficient explanations to conquer the low accuracy issue [4]. These approaches are beneficial in various scopes like speech recognition [5], medical image processing [6], human activity recognition [7], and so on.

This paper's major goal is to produce captions for photographs that are of the highest quality and contain accurate and useful item information.

To achieve this aim, several objectives will be done:

1- Providing deep networks with more tools, like attention to producing insightful and excellent images captions.

2- Including local, past, and future context for creating semantically rich images captions.

3- Showing how creating captions for both real and created photographs can benefit from using synthetic images.

2. Related Works

The caption generation systems of images work on automatically generating descriptions in natural languages for given images. Recently, those systems took the attention of various researchers and it has been considerable proceed with image captioning systems based on deep learning with attention mechanisms.

Xiao et al. [8], 2019, presented a captioning system that combined two segregated networks of Long-short Term Memory (LSTM) with a developed attention mechanism. In this system, the 1st network with an attention mechanism is capable of adaptably making a compromise between the textual contented and visual semantic area. The 2nd network integrates the representation of the hidden state of the 1st network and the vector of attention state, and it produces the sequence of words. This system has been comprehensively assessed on the MS COCO dataset, and the results depict that provided a considerable performance.

Deng et al. [9], 2020, presented a deep learning-based image description system using an adapted attention technique. This system involves two stages. In the 1st stage, DenseNet was utilized for extracting the image global features. Simultaneously, for all time axes, sentinel gates are established via the adapted attention technique for determining whether to utilize the feature information of the image for creating words. In the 2nd stage, the LSTM was carried out as a model of language creation concerning image description processes for enhancing image captioning quality. Experimentation on the MS COCO dataset signifies that this presented system revealed substantial development in both METEOR and BLEU evaluation measures.

Tian et al. [10], 2021, proposed a multiple levels network of semantic context information using a symmetrical overall structure for exploiting reciprocal connections and extracting context information among the three distinct semantic layers for jointly solving the distinct vision tasks to provide a comprehensive and accurate description of the image scene. The regions of relationship are created via aggregating and linking regions of the object, the ROI pooling technique is utilized to extract the object features, relations, and the proposals of region caption. Then, these features are passed to the relationship context and scene context networks. After that, the updated object's characteristics and extracted relationship context information are utilized to perform object relationship detection. Finally, the model feeds the retrieved relationship and scene context information to an attention mechanism; which generates information and caption characteristics that are then combined with the decoder to construct caption sentences. Experiments using COCO and VRD datasets illustrate that the presented model was capable of leveraging the context information among semantic layers for improving the visual task generation accuracy.

Wang and Gu [11], 2022, implemented the global and local visual cooperation attention approach for exploring the inherent interactions between global and local image features. In particular, a new network was devised which included a visual interaction encoder and fusion modules. The 1st module worked on the implicit encoding of visual relations between global and local image features for getting an improved rich depiction. While the 2nd module worked on fusing the former attained features to obtain additional improved various-level relation property information. Furthermore, another module (LSTM) was introduced to direct the generation of words. The expansive results of the experiment illustrated the superiority of the presented approach implemented on the MS COCO dataset.

3. Materials and Methods

3.1. Inception V3

Inception V3 represents the 3rd version in the series of evolutionary deep learning architectures presented via Google. After the development of Inception version 1 architecture, Inception Version 2 was performed using batch normalization. While Inception Version 3 utilized the concept of factorization for reducing the number of parameters and connections without minimizing the network efficiency. This model includes various asymmetric and symmetric blocks involving layers of convolutions, dropouts, max and average pooling, and fully-connected, as illustrated in Fig. 1. In total, the architecture of Inception version 3 includes forty-two layers, the first layer takes 299×299 input images, and the last layer outputs the classes through the Softmax function [12].

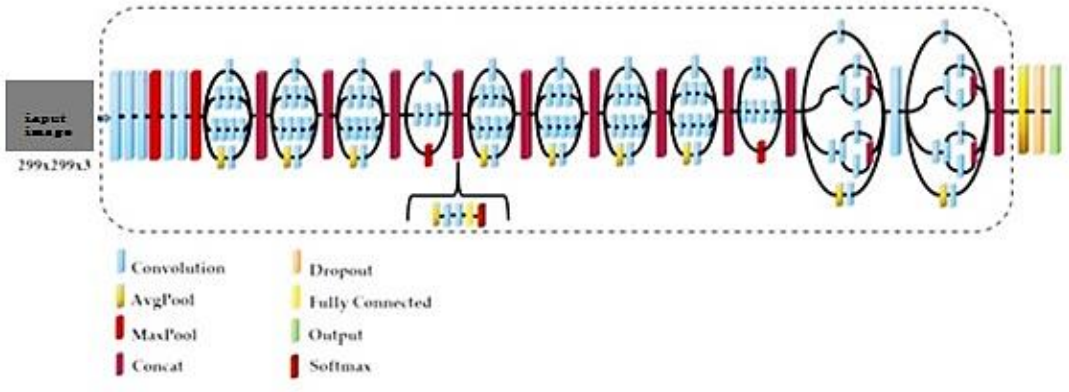

Fig. 1 - Schematic representation of Inception V3.

The main conception behind Inception V3 is to decrease the cost of computation for deep network models without impacting the generality. Therefore, asymmetric and small-size filters (1x5 &1x7) were utilized instead of large filters (7x7 & 5x5). Furthermore, convolution (1x1) was placed before every filter of large size, and this led to making the operation of convolution similar to a cross-channel connection [13].

3.2. RNN

RNNs are a broad and diverse class of computational models that are created through more or less precise analogies with the functional components of biological brain modules. Many abstract neurons (also known as units or processing elements) are connected by analogously abstracted synaptic connections (or links) in an RNN, allowing activations to spread throughout the network. RNNs are characterized by the link topology's cycles, which set them apart from more commonly used feedforward neural networks [14]. In RNNs, the output of the former phase is utilized as input in the subsequent phase. The most popular uses of RNNs are in the classification of images, videos, sequences, and sentiments. RNN is a type of artificial neural network that has connections between nodes that sequentially construct a directed graph [15].

Systems based on RNNs can be illustrated in Fig. 2, differ from feedforward systems in that they have a temporal proportion. It frequently depends on the past, and its decisions are based on what it has discovered before [16].

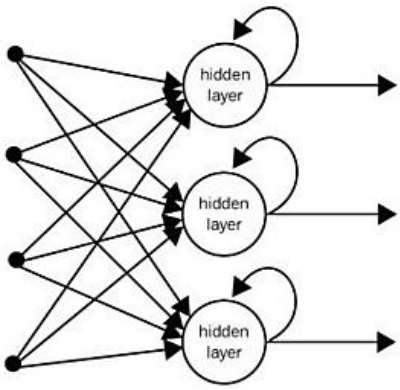

Fig. 2 - The architecture of RNN.

In an RNN, each output signal from the perceptron is transmitted back to the hidden layer of the perceptron. Recurrent signifies that the current output will re-enter as an input for subsequent entries. The system remembers the value of the previous element while simultaneously accounting for fresh input in each subsequent element. In Fig. 3, the RNN's unfolded repeating input is displayed. Through the hidden layer h, the input comes from "x" and the output comes from "y" [16].

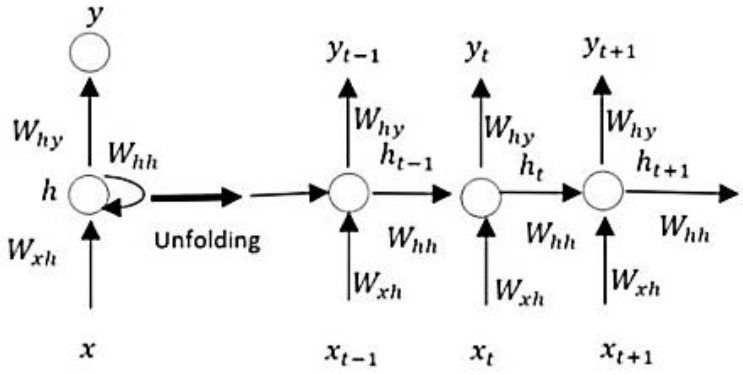

Fig. 3 - RNN representation, both folded and unfolded.

3.3. LSTM

In order to handle the issue of vanishing and exploding gradients, a dissimilar architecture to RNNs named LSTM was presented in which gates were added between every transaction point to protect its hidden activation to be cell state. In other words, The three gates of LSTM (forget, input, and output) are responsible for protecting the cell state. The forget gate represents the 1st gate that affects the cell during the forward pass. It establishes which and with how much the activations of the cell are disregarded. It accomplishes this via multiplying each element of the cell by a vector, ft(0,1)mh. The comparable element in the state of the cell will be erased and when the forget gate gives out a value that is close to 0, then it will be set to 0. But, it will fully maintain its value if it emits a value that is close to one. The input gate represents the following gate that influences the cell. It works on specifying the new information part that will be included in the protected cell state, and this can happen during the computation of a new candidate state. In the input gate, It∈(0,1)mh is multiplied by the candidate cell state and included in the cell state to inhibit needless inclusions to the cell state. The final gate represents the output gate which is significant for back-propagation. This gate works on specifying the cell state portions required to be forward propagated and added to the network output [17], see Fig. 4.

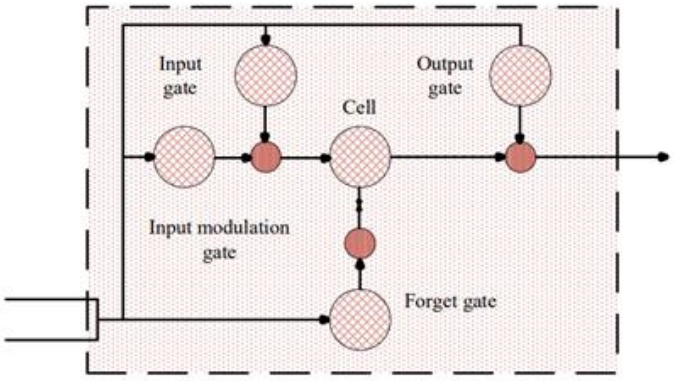

Fig. 4 - LSTM architecture.

4. The Proposed Captioning System

In the proposed system, the Encoder-Decoder model is utilized. The proposed system includes several stages: Firstly, pre-processing the images; Secondly, Encoder (Feature extraction by using Inception V3); Thirdly, Decoder (Sequence generator by using RNN with visual attention); And finally, sentence generation. The general structure illustrates in Fig. 5.

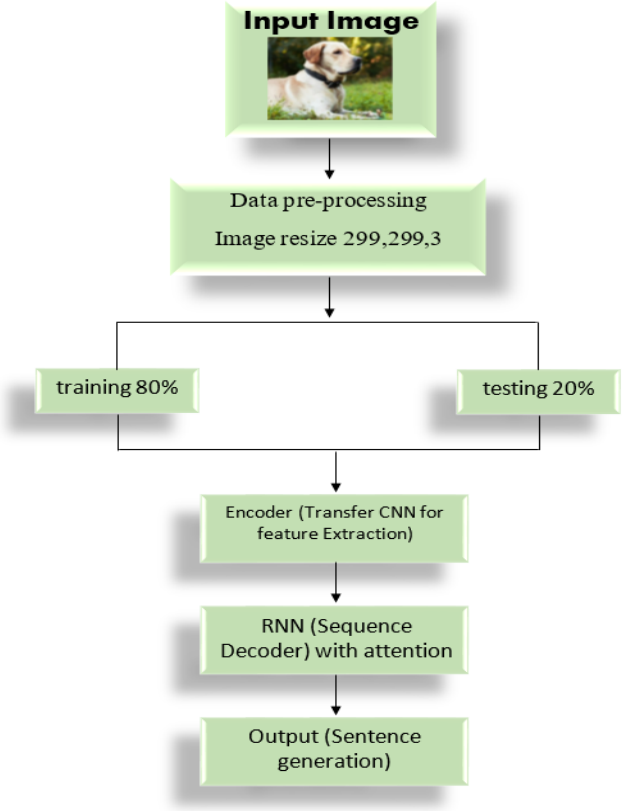

Fig. 5 - General structure of the proposed system.

4.1. Pre-processing the images

Pre-processing the data for the system is the first step. The data for MSCOCO is divided into two folders. One folder with images and one with captions. The first step is to map these to each other. Using the given token file, a dictionary is created with the images as keys and their value is a set of 5 captions. And then, images must be resized to 299*299*3 to fit the inception V3 model input requirements prior to feature extraction.

4.2. Encoder the image (Feature Extraction)

In this system, a pre-trained model named inception v3 is utilized. This model picks the pre-processed image as input and generates the vector of the encoded image that occupies the crucial and necessary image features. Then, these features of the encoded image are input, instead raw image itself for the next step in the image captioning system. The targeted captions that correspond to the encoded image are passed as well. After that, the image features are decoded and learned for predicting explanations (captions) that correspond to the targeted captions. Generally, this step comprises two parts; In the 1st part, inception v3 layers gradually extract the pertinent features from the pre-processed image to generate a compacted representation of the features map. In the 2nd part, a classifier with several linear layers works on taking the features map and predicting a category such as; house, dog, car, and so on for specifying to which the feature is categorized. This part requires only the features map and there is no requirement for the classifier. Table 1 illustrates the layered structure of inception V3.

Table 1 - The layered structure of inception V3.

No.	Layers	Stride	Size of Image
1	Conv 3*3	2	299*299*3
2	Conv 3*3	1	149*149*32
3	Conv 3*3	1	147*147*32
4	Pool 3*3	2	147*147*64
5	Conv 3*3	1	73*73*64
6	Conv 3*3	2	71*71*80
7	Conv 3*3	1	35*35*192
8	inception modules(3 modules)	-	35*35*288
9	inception modules(5 modules)	-	17*17*768
10	inception modules(2 modules)	-	8*8*1280

4.3. Decoder (RNN with Attention Mechanism)

The decoder in the proposed system represents the utilization of the LSTM network with the mechanism of visual attention. As mentioned before, LSTM is an advanced RNN, principally utilized for solving the gradient disappearing and explosion issue through long succession training. LSTM is appropriate to handle and predict significant occurrences with extended intervals and time series delay. In deep learning technology, the mechanism of visual attention is fundamentally close to human discriminating attention. The vision of humans is capable of fast image scanning to specify the targeted regions that require to be concentrated on. Therefore, high attention should be given to those regions for obtaining precise information concerning the targets.

The LSTM inputs are taken from the previous convolution network and the vector of word embedding. In every stage of LSTM, the output represents the distribution of probability created via the model concerning the following word in the sentence.

The model is not capable of importing illustrative sentences directly, therefore, every word in sentences can be pictured as a vector (one-hot) Vi with a dimension equivalent to the words' vocabulary size. For instance, the vocabulary is constructed depending on the superior two thousand words relating to their reoccurrence in the set of training. Thus, the whole sentences are depicted in the sort of $\{V_1, V_2, ..., V_L\}$, and the main intent of the model is to increase the sentences' log-likelihood:

$$\log p\,(V|h, att) = \sum_{i=1}^{L} \log\, p(Vi|V1{:}\,i-1, h, att) \quad (1)$$

where log p (V|h, att) indicates the generating probability of Vi word considering former words V1: i − 1, h indicates low-level feature of image and att indicates high-level feature of image. The Last two stages are illustrated in Fig. 6.

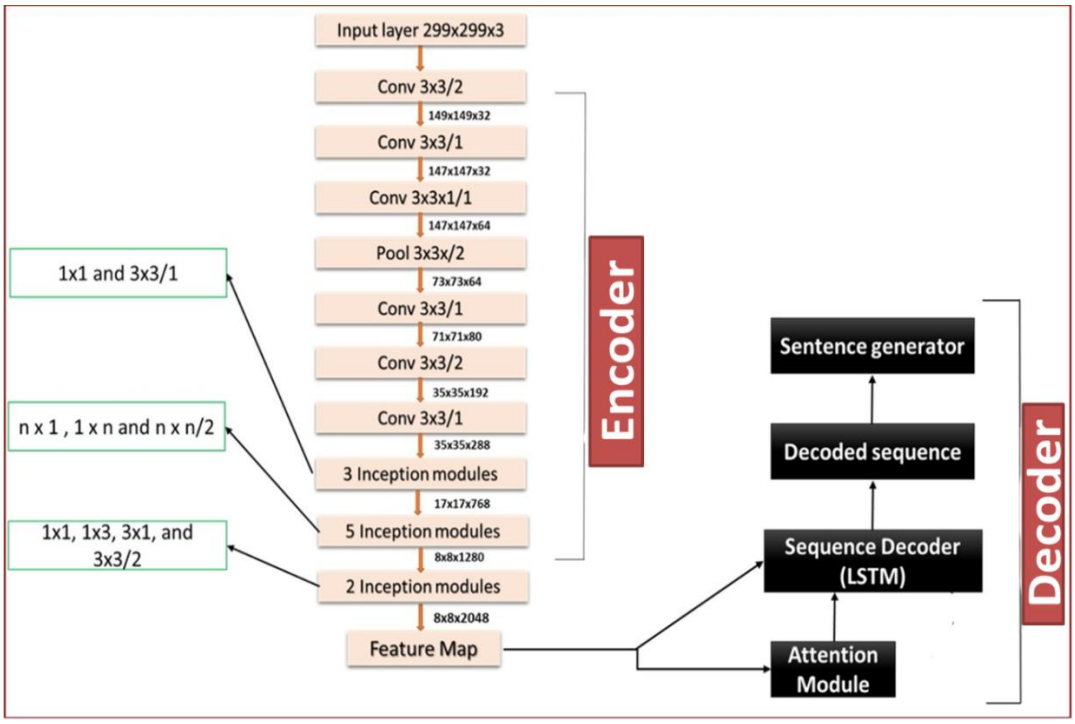

Fig. 6 - System's architecture with inception V3.

5. Experimental Results

The MS COCO dataset was utilized for performance evaluation for the proposed scheme. This dataset involves images of prevalent objects in the real world of human life. MS COCO dataset holds complicated background characteristics, various object instances and kinds, and objects with small sizes, therefore, it is increasingly challenging. This dataset is approximately large (involving 123,287 images each of which holds five ground truth sentences of various lengths). Next, in the splitting process, 113,287, 5000, and 5000 images are exploited for training, testing, and validation, respectively [18]. Fig. 7 shows samples of annotated images in the MS COCO dataset. And image caption process overview is illustrated in Fig. 8.

There are two standard evaluation measures METEOR and BLEU that were widely utilized in the recently presented image captioning schemes. These measures were selected and implemented for obtaining a more objective evaluation of our proposed scheme. The METEOR works on specifying the entire correspondence between sentences by utilizing specific matching criteria, such as paraphrase, synonym, and exact-word matching. While the BLEU has various grams that can be assigned as B1, B2, B3, and B4. Concerning these measures, higher values depict better performance.

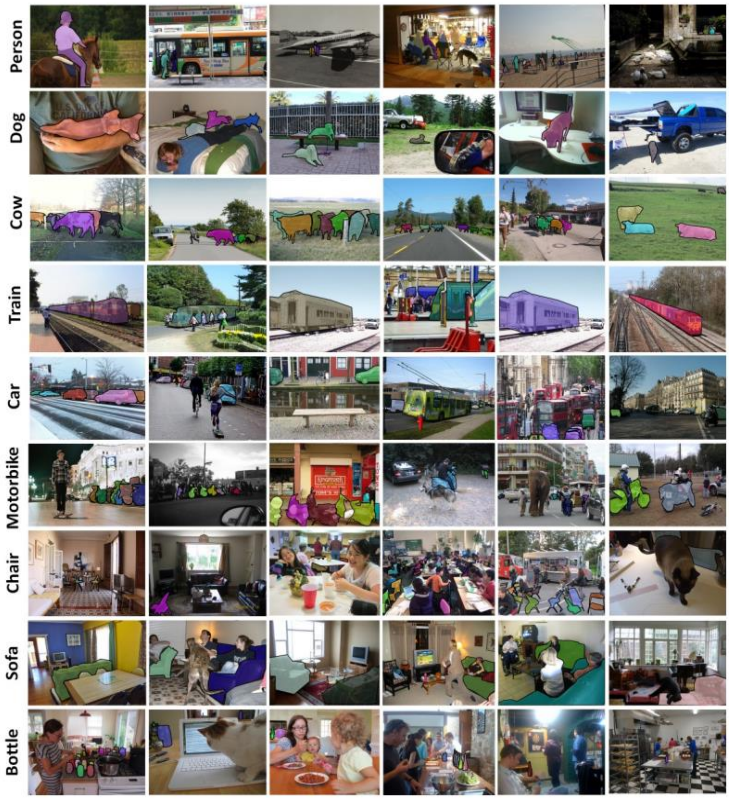

Fig. 7 - Samples of annotated images in the MS COCO dataset.

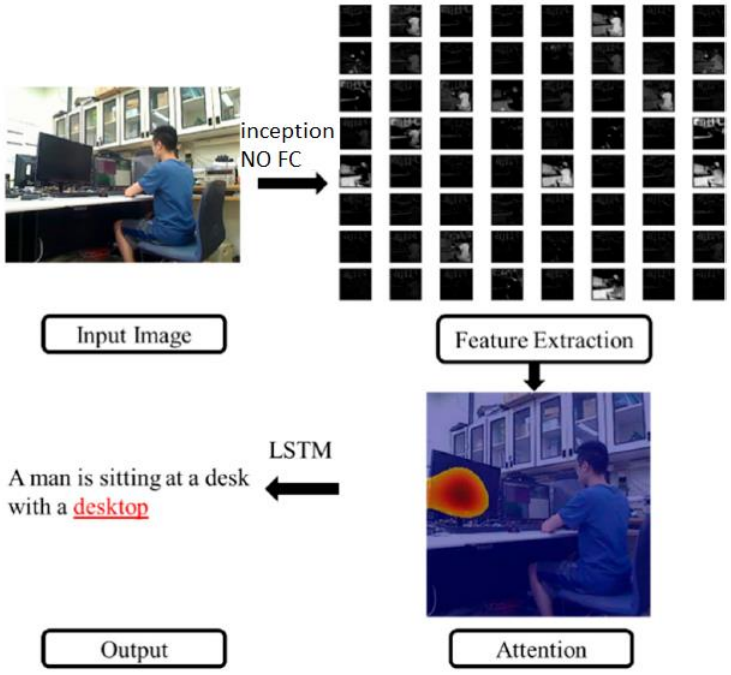

Fig. 8 - Overview of the image captioning system.

Approximately, 20% of samples (from every category in the utilized dataset) are randomly picked for validation, and the training set is 80%. The results of the various measures carried out on the MS COCO dataset under various

percentages of validation and the training sets are illustrated in Table 2. Fig. 9 illustrates examples of the implementation of a deep learning-based image captioning system using Inception V3 and LSTM with an attention mechanism. Table 3 illustrates a comparison stating the superiority of the proposed system over the existing works.

Table 2 - A comparison of the various percentages on the utilized data.

Percentages of Training - Testing	Meteor	BLEU			
		B1	B2	B3	B4
50% _ 50%	0.42	0.47	0.32	0.25	0.20
60% _ 40%	0.22	0.41	0.35	0.25	0.31
70% _ 30%	0.33	0.57	0.37	0.22	0.18
80% _ 20%	**0.543**	**0.87**	**0.66**	**0.51**	**0.42**
90% _ 10%	0.44	0.69	0.55	0.49	0.38

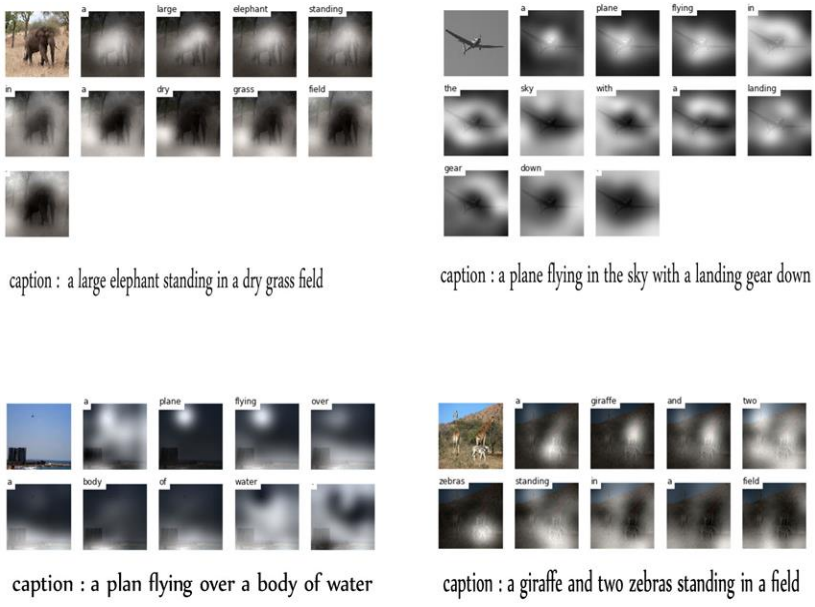

Fig. 9 - Examples of images captioning.

Table 3 - The superiority of the proposed system over the existing works.

Systems	METEOR	BLEU			
		B1	B2	B3	B4
Xiao et al. [8], 2019	0.271	0.758	0.594	0.455	0.346
Deng et al., 2020 [9]	0.270	0.739	0.570	0.422	0.326
Tian et al. [10], 2021	0.295	0.812	0.655	0.502	0.393
Wang and Gu [11], 2022	0.285	0.814	0.651	0.505	0.389
Proposed System	0.543	0.87	0.66	0.51	0.42

6. Conclusion

The proposed system is capable of accurately identifying the contents of an image and generating a caption that describes it. And, the LSTM component allows for the generation of natural language text that is grammatically correct and semantically meaningful. Furthermore, the utilization of the attention mechanism enabled the system to concentrate on the key elements in the image, producing captions that are more precise and insightful. The success of this system highlights the importance of combining different techniques in deep learning to achieve better results.

In the future, it is possible to apply more sophisticated attention mechanisms, such as self-attention or transformer-based attention, to enhance the model's capacity to capture long-term connections between image features and text.

References

[1] Q. Wang, H. Deng, X. Wu, Z. Yang, Y. Liu, Y. Wang, G. Hao, "LCM-Captioner: A lightweight text-based image captioning method with collaborative mechanism between vision and text", *Neural Networks*, vol. 162, (2023), pp. 318-329.

[2] H. Parvin, A. R. Naghsh-Nilchi, H. M. Mohammadi, "Transformer-based local-global guidance for image captioning", *Expert Systems with Applications*, vol. 223, 119774, (2023).

[3] R. Padate, A. Jain, M. Kalla, A. Sharma, "Image caption generation using a dual attention mechanism", *Engineering Applications of Artificial Intelligence*, vol. 123, Part A, 106112, (2023).

[4] S. Mohsen, A. Elkaseer and S. G. Scholz, "Industry 4.0-Oriented Deep Learning Models for Human Activity Recognition," *in IEEE Access*, vol. 9, pp. 150508-150521, (2021).

[5] M. H. Abdul-Hadi, J. Waleed, "Human Speech and Facial Emotion Recognition Technique Using SVM," *2020 International Conference on Computer Science and Software Engineering (CSASE)*, Duhok, Iraq, (2020), pp. 191-196.

[6] Saad Albawi, Muhanad Hameed Arif, Jumana Waleed, "Skin cancer classification dermatologist-level based on deep learning model", *Acta Scientiarum. Technology*, vol. 45, pp. e61531-e61531, 2023. doi: 10.4025/actascitechnol.v45i1.61531.

[7] M. F. Asghar, M. H. Ali, J. Waleed, "Pedestrian Attributes and Activity Recognition Using Deep Learning: A Comprehensive Survey", Al-Iraqia Journal for Scientific Engineering Research, vol. 2, no.1, pp. 40-56, (2022).

[8] F. Xiao, X. Gong, Y. Zhang, Y. Shen, J. Li, X. Gao, "DAA: Dual LSTMs with adaptive attention for image captioning", *Neurocomputing*, vol. 364, pp. 322-329, (2019).

[9] Z. Deng, Z. Jiang, R. Lan, W. Huang, X. Luo, "Image captioning using DenseNet network and adaptive attention", *Signal Processing: Image Communication*, vol. 85, 115836, (2020).

[10] P. Tian, H. Mo, L. Jiang, "Image Caption Generation Using Multi-Level Semantic Context Information", *Symmetry*, vol. 13, no. 7, 1184, (2021).

[11] C. Wang, X. Gu, "Local-global visual interaction attention for image captioning", *Digital Signal Processing*, vol. 130, 103707, (2022).

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, (2016), pp. 2818-2826.

[13] X. Xia, C. Xu and B. Nan, "Inception-v3 for flower classification," *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdu, (2017), pp. 783-787.

[14] M. Lukoševičius, H. Jaeger, "Reservoir computing approaches to recurrent neural network training", *Computer Science Review*, vol. 3, no. 3, pp. 127-149, (2009).

[15] H. S. Gill, O. I. Khalaf, Y. Alotaibi, S. Alghamdi, F. Alassery, "Multi-Model CNN-RNN-LSTM Based Fruit Recognition and Classification", *Intelligent Automation and Soft Computing*, vol. 33, no. 1, pp.637-650, (2022).

[16] A. Jaffar, N. M. Thamrin, M. Syahirul Amin, M. Ali, M. Farid Misnan, A. Ihsan Mohd Yassin, "WATER QUALITY PREDICTION USING LSTM-RNN: A REVIEW", *Penerbit UMT Journal of Sustainability Science and Management*, vol.17, pp. 205-226, (2022).

[17] A. Tsantekidis, N. Passalis, A. Tefas, "Chapter 5 - Recurrent neural networks", *Deep Learning for Robot Perception and Cognition*, Academic Press, pp. 101-115, (2022).

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, "Microsoft COCO: Common Objects in Context", *Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*, vol. 8693. Springer, Cham, (2014).