Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.024553
Article

Tech Science Press

Oppositional Harris Hawks Optimization with Deep Learning-Based Image Captioning

V. R. Kavitha[1], K. Nimala[2], A. Beno[3], K. C. Ramya[4], Seifedine Kadry[5], Byeong-Gwon Kang[6] and Yunyoung Nam[7,*]

[1]Department of Computer Science and Engineering, Prathyusha Engineering College, Thiruvallur, 602025, India
[2]Department of Networking and Communications, SRM Institute of Science and Technology, Chennai, India
[3]Department of Electronics and Communication Engineering, Dr. Sivanthi Aditanar College of Engineering, Tiruchendur, 628215, India
[4]Department of Electrical and Electronics Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, 641008, India
[5]Deparmtent of Applied Data Science, Noroff University College, Kristiansand, Norway
[6]Department of Information and Communication Engineering, Soonchunhyang University, Asan, Korea
[7]Department of Computer Science and Engineering, Soonchunhyang University, Asan, Korea
*Corresponding Author: Yunyoung Nam. Email: ynam@sch.ac.kr

Received: 21 October 2021; Accepted: 29 December 2021

Abstract: Image Captioning is an emergent topic of research in the domain of artificial intelligence (AI). It utilizes an integration of Computer Vision (CV) and Natural Language Processing (NLP) for generating the image descriptions. It finds use in several application areas namely recommendation in editing applications, utilization in virtual assistance, etc. The development of NLP and deep learning (DL) models find useful to derive a bridge among the visual details and textual semantics. In this view, this paper introduces an Oppositional Harris Hawks Optimization with Deep Learning based Image Captioning (OHHO-DLIC) technique. The OHHO-DLIC technique involves the design of distinct levels of pre-processing. Moreover, the feature extraction of the images is carried out by the use of EfficientNet model. Furthermore, the image captioning is performed by bidirectional long short term memory (BiLSTM) model, comprising encoder as well as decoder. At last, the oppositional Harris Hawks optimization (OHHO) based hyperparameter tuning process is performed for effectively adjusting the hyperparameter of the EfficientNet and BiLSTM models. The experimental analysis of the OHHO-DLIC technique is carried out on the Flickr 8k Dataset and a comprehensive comparative analysis highlighted the better performance over the recent approaches.

Keywords: Image captioning; natural language processing; artificial intelligence; machine learning; deep learning

1 Introduction

The image captioning method is a cross-disciplinary research problem including machine learning (ML), computer vision (CV), and natural language processing (NLP). The input of an image captioning method is an image and the output is a text representing the image. These tasks require the method to understand the relationship between objects, identify the objects in an image, and expressed in terms of natural language sentence [1]. In fact, various real time application needs the image captioning technique. E.g., afterward taking a picture, user could utilize this technique for matching the proper text that could replace the users' automatic filling using text. As well, it could assist blind people to understand the image contents. Related task includes video caption, wherein the input is a video and the output is its descriptions [2]. These tasks require the approach to be capable of understanding the relationship among the objects to take the semantic data of an image and make human-readable sentences, but still, they are challenging for the machine. The primary objective of image caption is to generate automatically human-like and content-rich descriptions of provided images. It has gained significant attention in industry and academic sectors since it could be extensively employed in video retrieval, earlier childhood education [3]. Unlike another computer vision task (object detection, image classification), training an efficient image caption method is very difficult by the fact that it needs a complete understanding of the fundamental entities in an image, along with its relations. In general, an encoder–encoder architecture is utilized as an essential to produce image description.

A large number of research efforts have been dedicated to automated image caption, also it is classified as retrieval-based image caption, new image captioning generation, and template-based image caption. First, Retrieval-based approach initially discovers the visually related images with its caption from the trained datasets, and next the image caption can be elected from related images with captions. Template-based image caption detects the actions or objects or attributes and next filling the blank slots in a fixed template [4]. In various way, the new image captioning generation method is to investigate the visual content of an image and next create image captioning from the visual contents through language models.

In comparison with the initial 2 classes, new caption generation could produce novel caption for a provided image which is semantically more specific when compared to prior methods [5]. The majority of the researches in these classes are based on deep learning (DL) and ML technique that is also the method adopted in this work. One usual architecture utilized in this class is the encoder-decoder architecture for image caption. Current studies have employed the deep recurrent neural network (DRNN) as the decoder and the deep convolutional neural network (DCNN) as the encoder that is shown to be very effective [6]. But still, it continues to be a challenge for identifying an appropriate DCNN and DRNN method for the image caption.

In Wei et al. [7], a multi-Attention method was initially presented by using the local and non-local evidence for efficient feature reasoning and in representation image caption. The presented generator is developed for generating more precise sentences, whereas the presented discriminator was applied for determining either the generated sentence is machine generated/human described. Wang et al. [8] develop a multi-layer dense attention mechanism for image captioning. A Fast region based convolutional neural network (RCNN) method is used for extracting image features as the coding layer, the long short term memory (LSTM) is employed for decoding the multi-layer dense attention mechanism, and generated the descriptive text. The model parameter is enhanced by strategy gradient optimization in reinforcement learning (RL) method.

Wang et al. [9] present a cascade semantic fusion architecture (CSF) for mining the representation feature for encoding image content by using attention model. Especially, the CSF advantages from 3 kinds of visual attention semantics, involving image-level, spatial attention, and object-level features, in a new 3-phase cascade manner. Initially, object-level attention feature is extracted for capturing the detailed content of an object depending on the pre-trained detector. Next, a fusion model to fuse object-level attention feature using spatial feature, in that way induce image level attention feature for enriching the context data around the object. Lastly, spatial attention feature is learned to expose the salient region representations as a complement for 2 formerly learned attention features.

Zhao et al. [10] propose a multi-modal fusion mechanism to generate descriptions for explaining the content of an image. The method contains 4 sub-networks: a convolutional neural network (CNN) for image feature extraction, image features extraction, a language CNN method for sentence modelling, and a recurrent network for word predictions. In Wang et al. [11], the intermediary associate connections are estimated to the topological inner structure of recurrent neural network (RNN) cell, i.e., additionally developed as an evolutionary method on the proxy of image caption tasks. On Microsoft Common Objects in Context (MSCOCO) datasets, the presented method starts from scratch, discover over 100 RNN variations with the performance of 100 on CIDEr: Consensus-Based Image Description Evaluation (CIDEr) and 31 on bilingual evaluation understudy (BLEU)-4, and the topmost performances achieve 101.4 and 32.6 respectively. Shen et al. [12] develop a Variational Autoencoder (VAE) and Reinforcement Learning based Two-stage Multi-task Learning Model (VRTMM) for the remote sensing image caption tasks. Initially, fine-tune the CNN together with the VAE. Next, the Transformer generates the text descriptions with spatial and semantic features. Then, RL algorithm is employed for enhancing the quality of the generated sentence. Deng et al. [13], proposed an adoptive attention mechanism using a visual sentinel. In the encoding stage, the method presents DenseNet for extracting the global feature of an image. Simultaneously, on every time axis, the sentinel gate is set by the adoptive attention model to utilize the image feature data for generating words. Chu et al. [14] designed an automatic image captioning based on resnet50 and LSTM (AICRL) model that is capable of conducting the automated image captions. AICRL contains one encoder as well as decoder.

This paper introduces an Oppositional Harris Hawks Optimization with Deep Learning based Image Captioning (OHHO-DLIC) technique. The OHHO-DLIC technique involves the design of distinct levels of pre-processing. Moreover, the feature extraction of the images is carried out by the use of EfficientNet model. Furthermore, the image captioning is performed by bidirectional long short term memory (BiLSTM) model, comprising encoder as well as decoder. At last, the oppositional Harris Hawks optimization (OHHO) based hyperparameter tuning process is performed for effectively adjusting the hyperparameter of the EfficientNet and BiLSTM models. The experimental analysis of the OHHO-DLIC technique is carried out on the Flickr 8k Dataset and a comprehensive comparative analysis highlighted the better performance over the recent approaches.

2 The Proposed Image Captioning Technique

This paper has developed an efficient OHHO-DLIC technique to generate captions for the images automatically. The OHHO-DLIC technique involves distinct levels of pre-processing, EfficientNet based feature extraction, BiLSTM based image captioning, and OHHO based hyperparameter tuning. The OHHO based hyperparameter tuning process is performed for effectively adjusting the hyperparameter of the EfficientNet and BiLSTM models. Fig. 1 illustrates the overall working process of proposed OHHO-DLIC model.

2.1 Stage I: Pre-Processing

Primarily, data pre-processing take place in different levels as listed below.

- Lower case conversion: The dataset text contains words with differing letter cases that cause problems to the module similar to the words with differing capitalizations will be considered as different thus increases the problem vocabulary and then leads to difficulty. Therefore, it is necessary for converting the whole text into lower case to prevent this issue.
- Punctuation removal: the study's aim is to generate descriptive sentence without punctuation for images so the existence of punctuation adds difficulty to the problems i.e., beyond the scope of this research.

- Number removal: Numerical information existing in the text possess a problem to the module since it rises vocabulary which must be eliminated.
- Indicates starting and ending sequence: Word tokens '<start>' and '<end>' are added at starting and ending of all the sentences to represent the starting and ending of the predictive sequence to the module.
- Tokenization: The clean text is divided as to constituent words and a dictionary comprising the whole vocabulary for word-to-index and index-to-word matching has been made.
- Vectorization: The word in the text data have been encoded with single numerical representation in the words to index dictionary hence it converts the cleaned sentence description into numerical sequence beforehand the word representation vector was learned in this sequence. In order to resolve these differing sentence lengths, the short sentence is padded to the length of the long sentence sequences.

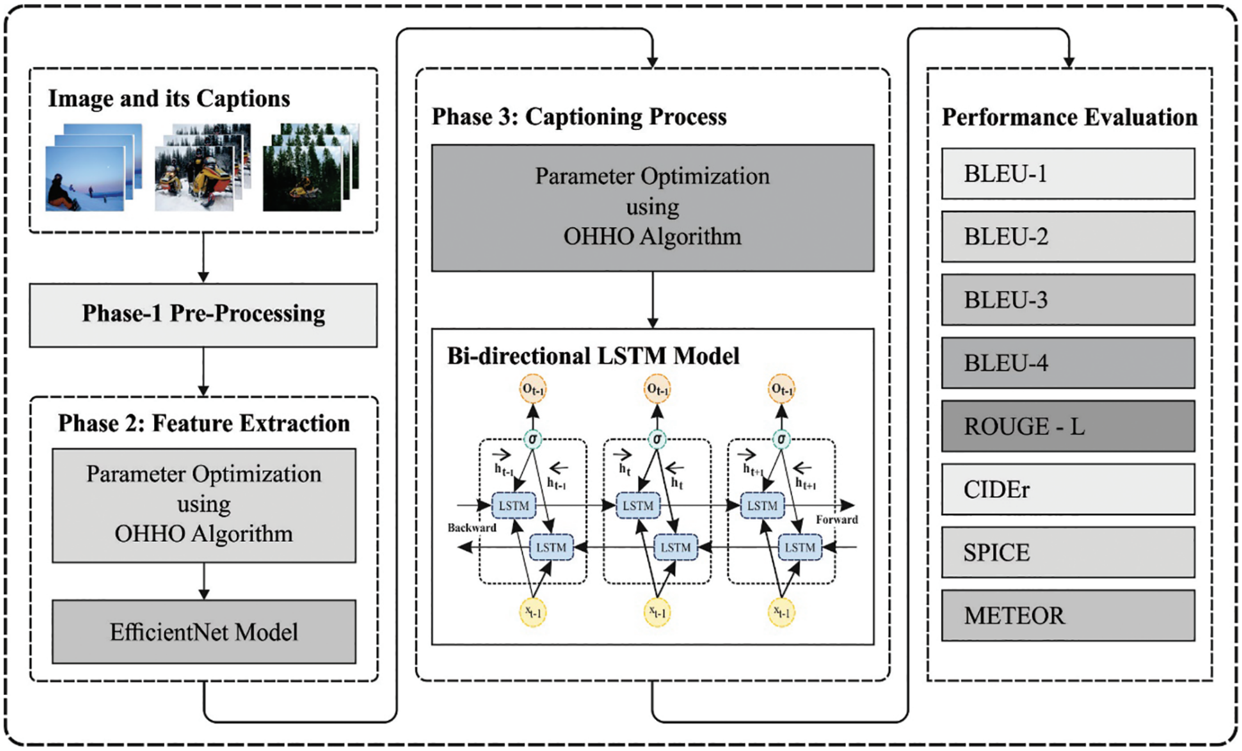
Image and its Captions
Phase-1 Pre-Processing
Phase 2: Feature Extraction
Parameter Optimization using OHHO Algorithm
EfficientNet Model
Phase 3: Captioning Process
Parameter Optimization using OHHO Algorithm
Bi-directional LSTM Model
O_{t-1}
σ
$\overrightarrow{h}_{t-1}$
$\overleftarrow{h}_{t-1}$
$\overrightarrow{h}_{t}$
$\overleftarrow{h}_{t}$
$\overrightarrow{h}_{t+1}$
$\overleftarrow{h}_{t+1}$
LSTM
Backward
Forward
x_{t-1}
Performance Evaluation
BLEU-1
BLEU-2
BLEU-3
BLEU-4
ROUGE - L
CIDEr
SPICE
METEOR

Figure 1: Overall process of OHHO-DLIC model

2.2 *Stage II: EfficientNet Based Feature Extraction*

During feature extraction, the images are fed into the EfficientNet model to generate feature vectors. The method is applied as a feature extraction model for generating appropriate sets of feature vectors of the input satellite image. The DL method resolves complicated problems mostly well and rapidly with minimum error rate and better classifier accuracy. The DL method is composed of several models (activation function, pooling, convolution, and FC layers). The DL framework's aim is to attain optimal performance and accuracy with less important models. Unlike DL methods, the EfficientNet technique is compound scaling model that employs the compound coefficients for uniformly scale network depth, resolution, and width [15]. The EfficientNet contains eight different models in B0 to B7. EfficientNet uses inverted bottleneck convolution i.e., initially developed in the MobileNetV2 technique that has a layer that initially expand the network and later compresses the channel. This architecture decrease evaluation by the factor of 2 as associated with standard convolution, wherein f means the filter size. The scientists outperformed that EfficientNetB0 is the most convenient of every eight methods and employs lesser parameters. Thus, it is directly used EfficientNetB0 to evaluate efficacy.

2.3 Stage III: BiLSTM Based Image Captioning

At the time of image captioning, the textual and image feature vectors are passed into the encoder of the BiLSTM model to generate image captions. Because of the long-term connection of speech signals, the RNN is suitable to complete the speech detection modeling tasks than the conventional deep neural network (DNN). In other words, part of input of the RNN hidden layer at present moment is the hidden layer outcome of earlier moment that permits the RNN for obtaining data of every prior moment with the recurrent feedback association, and it provides RNN memory. If the outcome is around the appropriate input data, a normal RNN is competent. If this time interval has been extended, while theoretically, RNN manages this long-term dependency issue, it could not be effective in practice. It can be established that the basic purpose to make RNN trained complexity has been vanishing gradient and gradient explosion issues.

The LSTM preserves the error at further constants level with particular gating unit, permitting the RNN for learning several time steps, so avoided the vanishing gradients. The LSTM saves data from gated units outside the ordinary flow of RNN [16]. These units save, write, or read data, same as information under the computer memory. The unit defines that data has been saved with the switch of door and if it can be permitted for reading, writing, or clearing the data. This gate switch is on the fundamental of received signal. Related to the nodes of neural network (NN), it can be utilized its individual group of weights for filtering the data and choose whether to permit the data for passing dependent upon its strength and the content of import.

The LSTM block has of group of memory cells *c* and 3 gates: input *i*, forget *f*, and output gates *o* that are utilized for controlling the flow of data.

The subsequent is a detailed explanation of many gating units:

- Memory unit *c*: it saves the network time status;
- Input gate *i*: it can be chosen that for passing input data as to cell *c*;
- Output gate *o*: it resolves that to output the cell data; and,
- Forget gate *f*: This gate appropriately resets the cell states.

During the time step *t*, the equation of LSTM block is explained as:

$$\begin{aligned}
i_t^1 &= \sigma(W_i^1 x_t^1 + V_i^1 h_{t-1}^1 + U_i^1 c_{t-1}^1 + b_i^1)\\
f_t^1 &= \sigma(W_f^1 x_t^1 + V_f^1 h_{t-1}^1 + U_f^1 c_{t-1}^1 + b_f^1)\\
\hat{c}_t^1 &= \tanh(W_c^1 x_t^1 + V_c^1 h_{t-1}^1 + b_c^1)\\
c_t^1 &= f_t^1 \circ c_{t-1}^1 + i_t^1 \circ \hat{c}_t^1\\
o_t^1 &= \sigma(W_o^1 x_t^1 + V_o^1 h_{t-1}^1 + U_o^1 c_t^1 + b_o^1)\\
h_t^1 &= W_{proj}^{'1}(o_t^1 \circ \tanh(c_t^1))
\end{aligned} \tag{1}$$

c_t^1 and h_t^1 are corresponding, the cell state and cell output of 1st layer at time *t*; Particularly, the cell output h_t^0 of layer 0 at time *t* represents the input feature vector at time *t*. *W*, *V*, *U*, and *b*, correspondingly imply the weight matrix as well as bias vector linking distinct gates. As only one cell has been fixed to all LSTM blocks, thus *U* refers the diagonal matrix at this time, and *b* refers the bias vectors. σ denotes the sigmoid activation functions. ° stands for the element-wise product. $W_{proj}^{'1}$ demonstrates the projection matrix.

The Bi-LSTM NN has been collected of LSTM units which function in combined ways to incorporate past and future context data. Bi-LSTM is study of long-term dependency without recollecting duplicate context data. So, it can be exhibited excellent efficiency to sequential modeling issues and has been extremely utilized in text classification. Different from the LSTM networks, the Bi-LSTM networks have 2 parallel layers which propagate from 2 manners with forward and reverse passes for capturing dependency from 2 contexts.

2.4 Stage IV: OHHO Based Hyperparameter Tuning

At last, the OHHO algorithm is utilized to determine the hyperparameters involved in the EfficientNet and BiLSTM models. Heidari et al. [17] presented a nature simulated computation intelligence model adapting the harris hawk's behavior of prey pursuit. Harris Hawks has a single cooperative pursuit approach that depends on the condition of dynamic nature and escape strategy of prey. The hawk shows advanced team spirit for chasing strength with respect to encircling, hunting, and getting out of the hunt. The exploitation and exploration phases of the HHO model is given in the following:

Exploration Phase:

During the exploration, the harris hawk uses their powerful eye for finding the prey. Harris Hawks is arbitrarily perched from various positions, and they examine the probability of hunting on 2 occasions on the basis of q value.

$$X(t+1) = \begin{cases} X_{rand}(t) - r_1 |X_{rand}(t) - 2r_2 X(t)| & q \geq 0.5 \\ (X_{rabbit}(t) - X_m(t)) - r_3(LB + r_4(UB - LB)) & q < 0.5 \end{cases} \tag{2}$$

In which $X(t+1)$ represent the following t iteration of the hawk's vector, $X_{rabbit}(t)$ demonstrates the present location of the rabbit, $X(t)$ indicates the present location of the hawkers, r_1, r_2, r_3, r_4 and q have arbitrary number in the range of zero and one, LB and UB are the upper as well as lower limits of the parameter., $X_{rand}(t)$ denotes an arbitrarily elected hawk from the existing population, and $X_m(t)$ signifies the average location of Hawke's present location.

Exploitation phase:

In this stage, there is a possibility to attack a previously recognized prey. Fig. 2 depicts the flowchart of HHO technique [18].

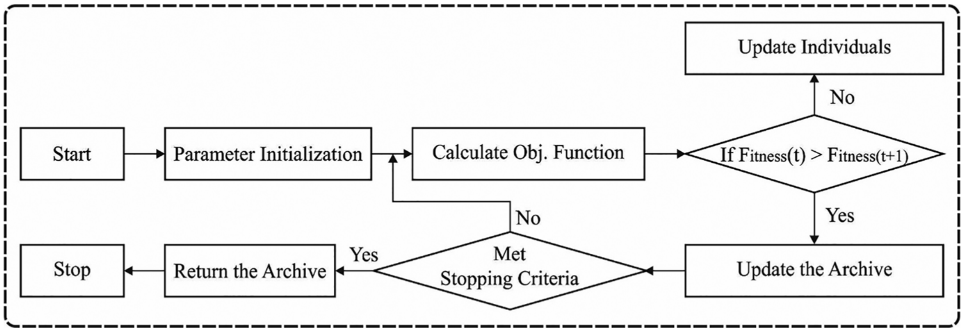

Figure 2: Flowchart of HHO technique

This technique of opposition based learning (OBL) has been established by Rahnamayan, Tizhoosh, and Salama. An essential knowledge behind OBL was concurrent concern of evaluating and their equivalent opposite evaluation for achieving an optimum estimate to the present candidate solution. To some solution $x(x_1, x_2, \ldots, x_D) \in \mathbb{R}^D$, their opposite solution $y(y_1, y_2, \ldots, y_D) \in \mathbb{R}^D$ is evaluate as:

$$y_j = lb_j + ub_j - x_j \tag{3}$$

where lb_j and ub_j represents the lower as well as upper bounds to j^{th} module of vector x. For finding unknown optimum solutions, the exploring performance in an arbitrary manner as well as their opposite way concurrently offers a greater chance for seeing the promising region. Therefore, it can be logical that when the presently evaluated point is distant in the unknown optimum solutions, afterward, the calculation of their opposite evaluated place refers the opposite way which is near the unknown optimum solutions.

In all the iterations of presented technique, it can be creating the group of opposite solutions with use of top fixed candidate solution of swarm. These amount of opposite solutions have been reduced for allowing the exploration at primary iteration and for transiting in exploration to exploitation on the iterations. Thus, during this approach, the opposite candidate solutions have been created with use of Eq. (3) that improved not only the exploration of given search space but also uses at condition if the unknown optimum is proper under the opposite way of any optimal solutions of the swarm.

During this presented technique, the amount of opposite solutions which are created from a specific iteration r has been computed as:

$$N_{op} = round\left(N - r \times \frac{N-1}{T}\right) \tag{4}$$

where r implies the present iteration, N refers the size of swarms or amount of hawks from the swarms and T signifies the maximal amount of iterations that have been existing to a technique [19]. During the m-HHO, this N_{op} opposite solution has been created by top N_{op} search agent of swarm. Briefly, the presented m-HHO has been established with relating the 3 approaches such as OBL, recently established arbitrary dives and altered energy parameter manner, and one more approach named greedy selection. Then, do the subsequent to all hawks, upgrade the energy parameter E.

Afterward select for updating the state of hawk dependent upon the value of E and r particularly are chosen if $|E| \geq 0.5$, $r < 0.5$ and $|E| < 0.5$, $r < 0.5$, correspondingly. Then, the presented greedy selection was implemented, if the swarm upgrade technique has been ended. The next step is that OBL approach was applied dependent upon Eqs. (3) and (4). Lastly, to prove if the maximal amount of iterations T was gained. Return the optimum X_{rabbit} when it can be attained, else, remain the iteration model.

3 Experimental Validation

This section investigates the image captioning efficiency of the OHHO-DLIC technique on Flickr 8k dataset [20]. The results are inspected under varying aspects. Fig. 3 illustrates a few sample images.

Fig. 4 illustrates the sample image captioning results obtained by the OHHO-DLIC technique. The figure has shown that the OHHO-DLIC technique has effectively derived the image captions for the applied input image.

Fig. 5 demonstrates the accuracy analysis of the OHHO-DLIC technique on the test dataset applied. The figure has shown that the training as well as testing accuracies get increased with an increase in epochs. At the same time, it is notified that the validation accuracy seems to be higher than training accuracy.

Fig. 6 depicts the loss analysis of OHHO-DLIC approach on the test dataset applied. The figure outperformed that the training as well as testing losses obtains minimum with a higher in epochs. Likewise, it can be stated that the validation loss seems that lesser than training loss.

Tab. 1 and Fig. 7 provide a detailed results analysis of the OHHO-DLIC technique interms of different measures. A detailed comparative analysis is made with Attribute LSTM, LSTM with attention layer (LSTM-A3), Global–Local Attention (GLA), CNN, Cascade Semantic Fusion (CSF), Adaptive, Action context (AC), Self-critical Sequence Training (SCST), UpDown (UD), and Multi-Attention Generative Adversarial Image Captioning Network (MAGAN) techniques.

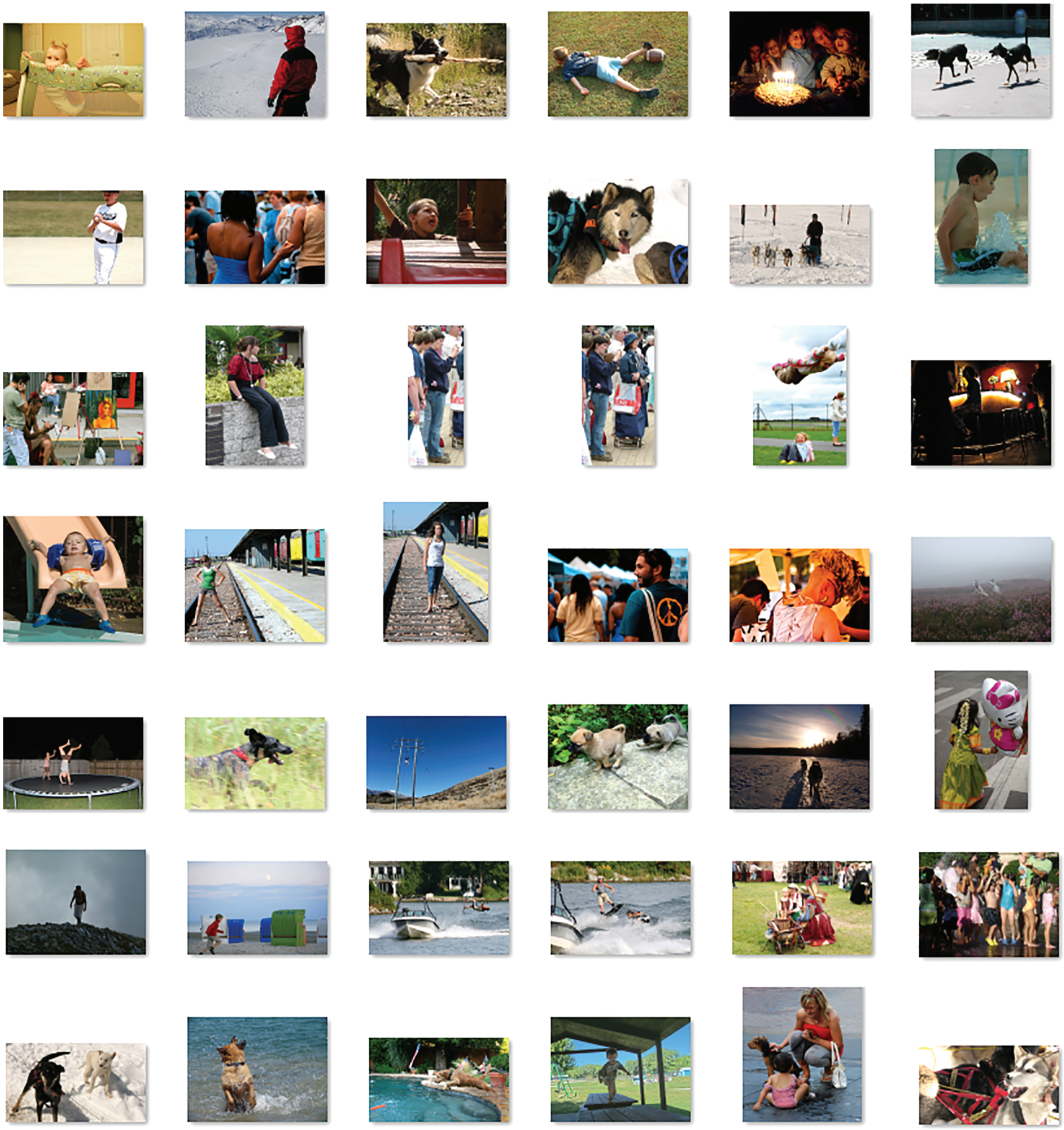

Figure 3: Sample images

The results demonstrated that the A-FCN, GLA, CNN, LSTM-A3, Attribute LSTM, and Adaptive models have obtained poor performance with the lower BLEU values. At the same time, the CSF, AC, SCST, UD, and MAGAN approaches have resulted in moderate BLEU values. Moreover, the proposed OHHO-DLIC technique has accomplished improved outcomes with the maximum BLEU-1, BLEU-2, BLEU-3, and BLEU-4 of 82.97%, 67.21%, 51.64%, and 41.86%.

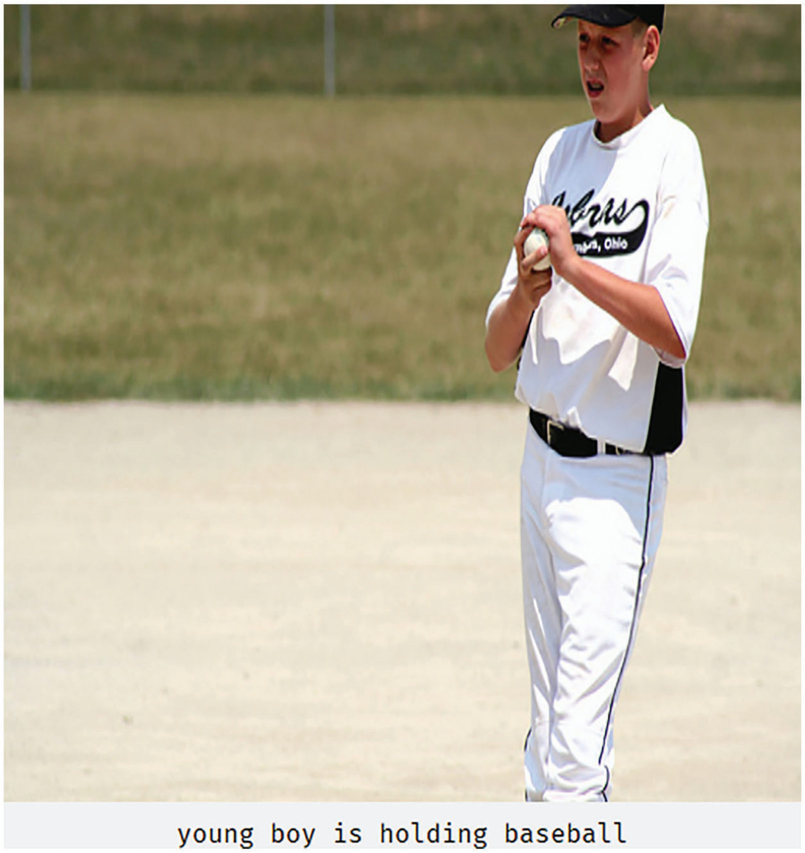

Figure 4: Sample result image of OHHO-DLIC model

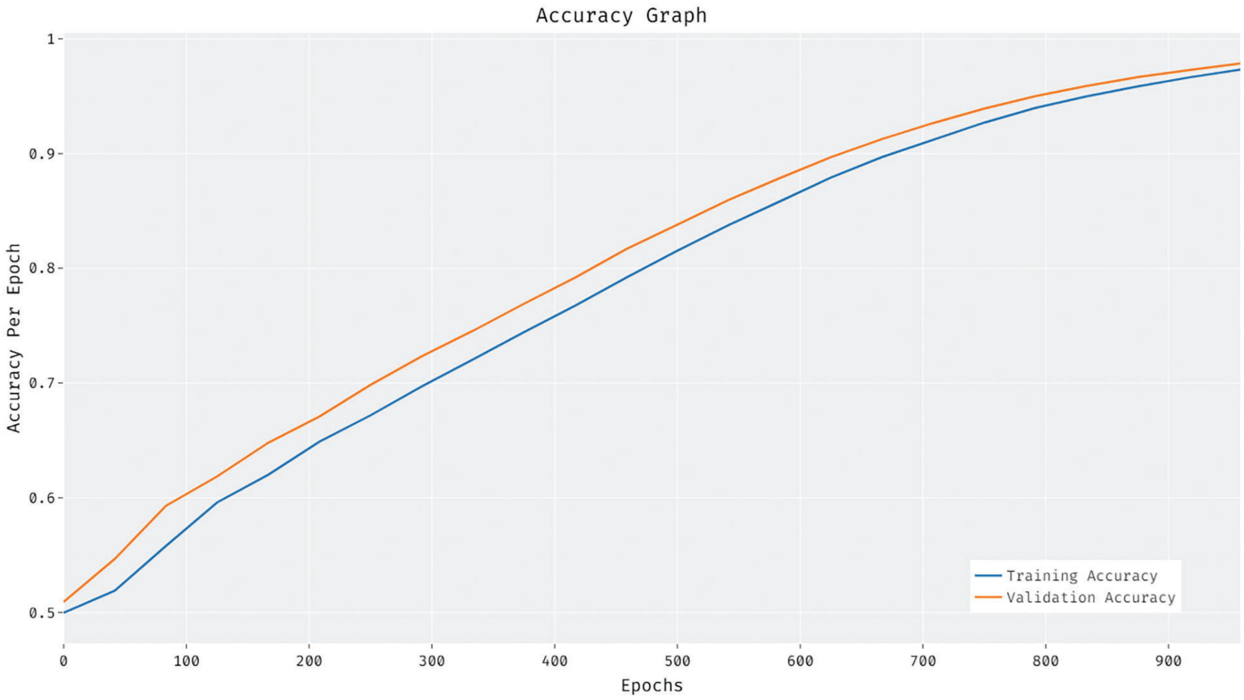

Figure 5: Accuracy analysis of OHHO-DLIC model

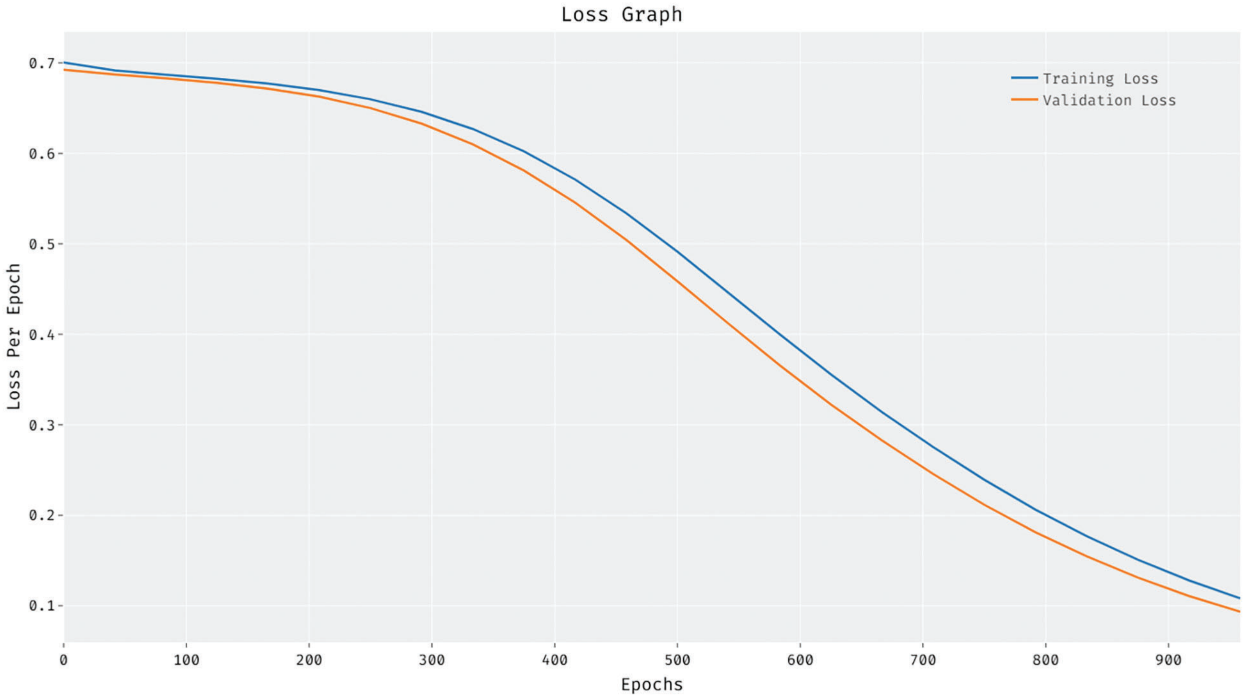

Figure 6: Loss analysis of OHHO-DLIC model

Table 1: Result analysis of OHHO-DLIC model with different measures

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Attribute LSTM	74.73	57.21	42.44	32.48
LSTM-A3	74.19	57.94	43.43	33.88
A-FCN	71.54	53.94	41.67	30.84
GLA model	73.71	56.83	42.34	32.42
CNN	73.97	55.92	42.30	30.50
CSF	77.47	61.08	47.43	35.81
Adaptive	74.80	58.40	44.40	33.60
AC technique	77.80	61.20	45.90	33.70
SCST technique	78.10	61.90	47.00	35.20
UD technique	80.20	64.10	49.10	36.90
MAGAN	80.60	64.70	49.90	37.30
OHHO-DLIC	82.97	67.21	51.64	41.86

Tab. 2 offers a brief outcomes analysis of the OHHO-DLIC manner with existing techniques with respect to metric for machine translation evaluation (METEOR) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-L.

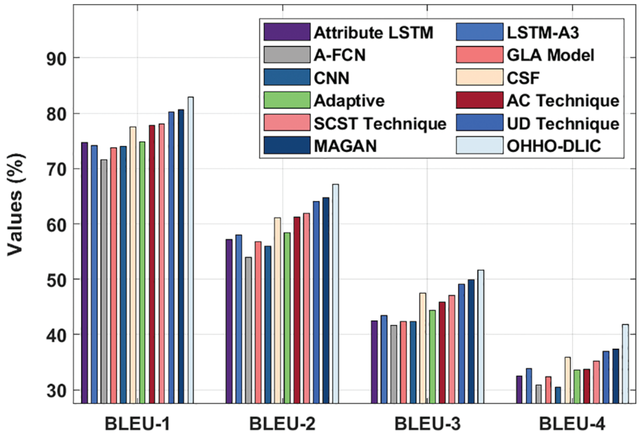

Figure 7: Result analysis of OHHO-DLIC model with varying measures

Table 2: Comparative analysis of OHHO-DLIC model with existing approaches

Methods	METEOR	ROUGE-L
Attribute LSTM	27.12	–
LSTM-A3	26.58	54.59
A-FCN	25.01	–
GLA model	26.13	53.56
CNN	25.61	–
CSF	28.13	57.30
Adaptive	26.40	55.00
AC technique	26.40	55.40
SCST technique	27.00	56.30
UD technique	27.60	57.10
MAGAN	27.60	57.70
OHHO-DLIC	31.39	59.81

Fig. 8 portrays the METEOR analysis of the OHHO-DLIC technique with recent approaches. The figure reported that the A-FCN, CNN, GLA, Adaptive, AC, and LSTM-A3 techniques have obtained lower METEOR values of 25.01, 25.61, 26.13, 26.4, 26.4, and 26.58 respectively. Along with that, the SCST, attribute LSTM, UD, and MAGAN techniques have attained slightly enhanced METEOR values of 27, 27.12, 27.6, and 27.6 respectively. Though the CSF technique has achieved a near optimal METEOR value of 28.13, the proposed OHHO-DLIC technique has accomplished higher METEOR value of 31.39.

Fig. 9 showcases the ROUGE-L analysis of the OHHO-DLIC system with present manners. The figure demonstrated that the GLA, LSTM-A3, and Adaptive algorithms have reached minimum ROUGE-L values of 53.56, 54.59, and 55 correspondingly. In line with, the AC, SCST, UD, and CSF schemes have gained slightly increased ROUGE-L values of 55.40, 56.30, 57.10, and 57.30 correspondingly. But the MAGAN algorithm has reached a near optimal ROUGE-L value of 57.70, the presented OHHO-DLIC methodology has accomplished superior ROUGE-L value of 59.81.

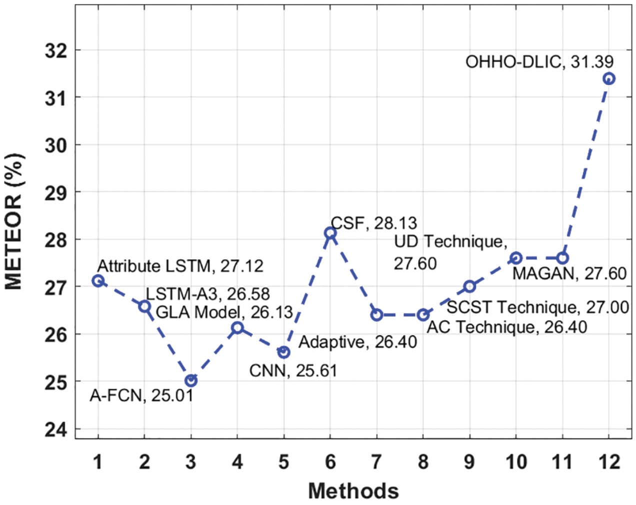

Figure 8: METEOR analysis of OHHO-DLIC system with recent manners

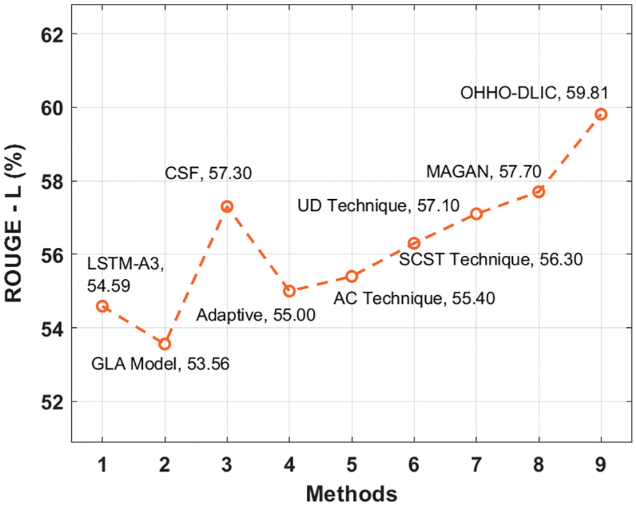

Figure 9: ROUGE-L analysis of OHHO-DLIC system with recent manners

Tab. 3 and Fig. 10 outperform the CIDEr analysis of the OHHO-DLIC technique with recent approaches. The figure demonstrated that the Attribute LSTM, GLA, CNN, LSTM-A3, and Adaptive techniques have obtained minimum CIDEr values of 94.71, 96.98, 97.11, 101.25, and 104.20 correspondingly. Besides, the AC, CSF, SCST, and UD techniques have attained somewhat increased CIDEr values of 110.20, 110.89, 114.70, and 117.90 correspondingly. Finally, the MAGAN technique has achieved a near optimal CIDEr value of 118.50, the proposed OHHO-DLIC technique has accomplished maximum CIDEr value of 122.35.

Tab. 4 and Fig. 11 exhibits the SPICE analysis of the OHHO-DLIC algorithm with recent techniques. The figure clear that the CNN, Adaptive, and LSTM-A3 techniques have obtained lower SPICE values of 18.61, 19.70, and 19.99 correspondingly. Similarly, the Adaptive, UD, and SCST manners have attained slightly enhanced SPICE values of 19.70, 20.30, and 20.70 respectively. Eventually, the CSF technique has achieved a near optimal SPICE value of 21.85, the projected OHHO-DLIC methodology has accomplished superior SPICE value of 23.76.

Table 3: CIDEr analysis of OHHO-DLIC model with existing approaches

Methods	CIDEr
Attribute LSTM	94.71
LSTM-A3	101.25
GLA model	96.98
CNN	97.11
CSF	110.89
Adaptive	104.20
AC technique	110.20
SCST technique	114.70
UD technique	117.90
MAGAN	118.50
OHHO-DLIC	122.35

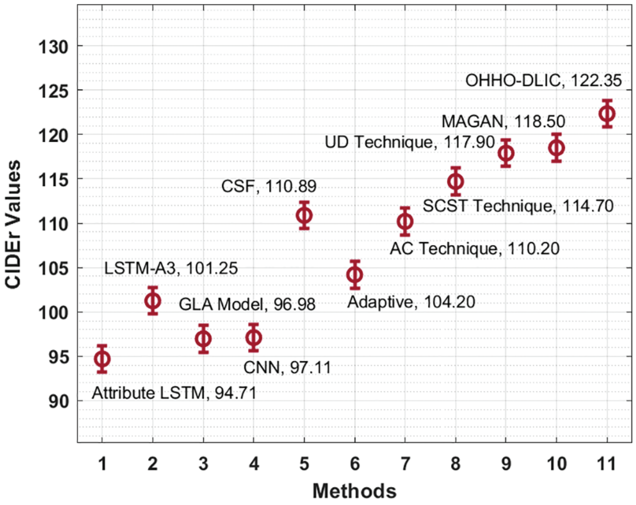

Figure 10: CIDEr analysis of OHHO-DLIC system with recent manners

Table 4: SPICE analysis of OHHO-DLIC model with recent approaches

Methods	SPICE
LSTM-A3	19.99
CNN	18.61
CSF	21.85
Adaptive	19.70
AC technique	20.30
SCST technique	20.70
UD technique	21.50
OHHO-DLIC	23.76

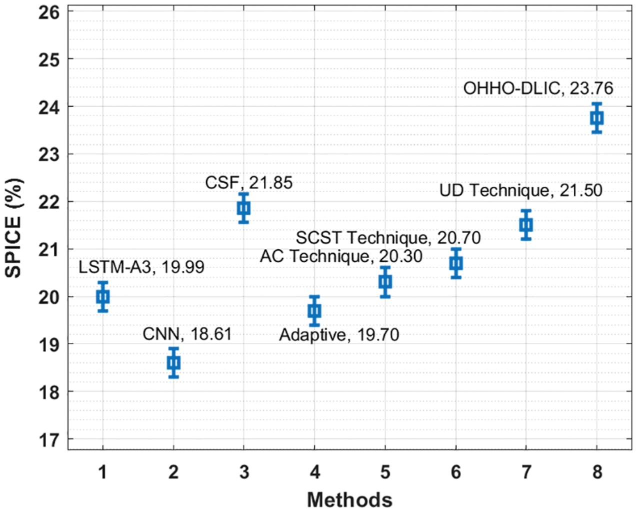

Figure 11: SPICE analysis of OHHO-DLIC system with recent manners

By looking into the above tables and figures, it is apparent that the OHHO-DLIC technique has resulted in effective image captioning outcomes over the other techniques.

4 Conclusion

This paper has developed an efficient OHHO-DLIC technique to generate captions for the images automatically. The OHHO-DLIC technique involves distinct levels of pre-processing, EfficientNet based feature extraction, BiLSTM based image captioning, and OHHO based hyperparameter tuning. The OHHO based hyperparameter tuning process is performed for effectively adjusting the hyperparameter of the EfficientNet and BiLSTM models. The experimental analysis of the OHHO-DLIC technique is carried out on the Flickr 8k Dataset and a comprehensive comparative analysis highlighted the better performance over the recent approaches. Therefore, the OHHO-DLIC technique is found to be a proficient tool for image captioning. In future, multihead attention mechanism can be included to improve the image captioning performance.

Funding Statement: This work was supported by the Soonchunhyang University Research Fund and University Innovation Support Project.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, “Show and tell: A neural image caption generator,” in *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 3156–3164, 2015.

[2] S. K. Lakshmanaprabu, S. N. Mohanty, K. Shankar, N. Arunkumar and G. Ramirez, “Optimal deep learning model for classification of lung cancer on CT images,” *Future Generation Computer Systems*, vol. 92, pp. 374–382, 2019.

[3] M. D. Z. Hossain, F. Sohel, M. F. Shiratuddin and H. Laga, “A comprehensive survey of deep learning for image captioning,” *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–36, 2019.

[4] K. Shankar, A. R. W. Sait, D. Gupta, S. K. Lakshmanaprabu, A. Khanna *et al.,* “Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model,” *Pattern Recognition Letters*, vol. 133, pp. 210–216, 2020.

[5] A. Karpathy and L. F. Fei, “Deep visual-semantic alignments for generating image descriptions,” in *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 3128–3137, 2015.

[6] N. Krishnaraj, M. Elhoseny, M. Thenmozhi, M. M. Selim and K. Shankar, “Deep learning model for real-time image compression in Internet of Underwater Things (IoUT),” *Journal of Real-Time Image Processing*, vol. 17, no. 6, pp. 2097–2111, 2020.

[7] Y. Wei, L. Wang, H. Cao, M. Shao and C. Wu, “Multi-attention generative adversarial network for image captioning,” *Neurocomputing*, vol. 387, pp. 91–99, 2020.

[8] K. Wang, X. Zhang, F. Wang, T. Y. Wu and C. M. Chen, “Multilayer dense attention model for image caption,” *IEEE Access*, vol. 7, pp. 66358–66368, 2019.

[9] S. Wang, L. Lan, X. Zhang, G. Dong and Z. Luo, “Cascade semantic fusion for image captioning,” *IEEE Access*, vol. 7, pp. 66680–66688, 2019.

[10] D. Zhao, Z. Chang and S. Guo, “A multimodal fusion approach for image captioning,” *Neurocomputing*, vol. 329, pp. 476–485, 2019.

[11] H. Wang, H. Wang and K. Xu, “Evolutionary recurrent neural network for image captioning,” *Neurocomputing*, vol. 401, pp. 249–256, 2020.

[12] X. Shen, B. Liu, Y. Zhou, J. Zhao and M. Liu, “Remote sensing image captioning via variational autoencoder and reinforcement learning,” *Knowledge-Based Systems*, vol. 203, pp. 105920, 2020.

[13] Z. Deng, Z. Jiang, R. Lan, W. Huang and X. Luo, “Image captioning using DenseNet network and adaptive attention,” *Signal Processing: Image Communication*, vol. 85, pp. 115836, 2020.

[14] Y. Chu, X. Yue, L. Yu, M. Sergei and Z. Wang, “Automatic image captioning based on resnet50 and lstm with soft attention,” *Wireless Communications and Mobile Computing*, vol. 2020, pp. 1–7, 2020.

[15] S. Zhang, Y. Bu, B. Chen and X. Lu, “Transfer learning for encrypted malicious traffic detection based on efficientnet,” in *2021 3rd Int. Conf. on Advances in Computer Technology, Information Science and Communication (CTISC)*, Shanghai, China, pp. 72–76, 2021.

[16] H. Fei and F. Tan, “Bidirectional grid long short-term memory (bigridlstm): A method to address context-sensitivity and vanishing gradient,” *Algorithms*, vol. 11, no. 11, pp. 172, 2018.

[17] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja *et al.,* “Harris hawks optimization: Algorithm and applications,” *Future Generation Computer Systems*, vol. 97, pp. 849–872, 2019.

[18] R. Sharma and S. Prakash, “HHO-LPWSN: Harris hawks optimization algorithm for sensor nodes localization problem in wireless sensor networks,” *ICST Transactions on Scalable Information Systems*, vol. 8, no. 31, pp. 1–10, 2018.

[19] S. Gupta, K. Deep, A. A. Heidari, H. Moayedi and M. Wang, “Opposition-based learning harris hawks optimization with advanced transition rules: Principles and analysis,” *Expert Systems with Applications*, vol. 158, pp. 113510, 2020.

[20] Dataset, 2021. https://www.kaggle.com/adityajn105/flickr8k/activity.