IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 November 2023, accepted 11 December 2023, date of publication 13 December 2023,
date of current version 21 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3342703

RESEARCH ARTICLE

Lighting Search Algorithm With Convolutional Neural Network-Based Image Captioning System for Natural Language Processing

RANA OTHMAN ALNASHWAN[1], SAMIA ALLAOUA CHELLOUG[1], NABIL SHARAF ALMALKI[2], IMÈNE ISSAOUI[3], ABDELWAHED MOTWAKEL[4], AND AHMED SAYED[5]

[1]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[2]Department of Special Education, College of Education, King Saud University, Riyadh 12372, Saudi Arabia
[3]Unit of Scientific Research, Applied College, Qassim University, Buraydah 52571, Saudi Arabia
[4]Department of Information Systems, College of Business Administration in Hawtat Bani Tamim, Prince Sattam bin Abdulaziz University, Al-Kharj 16278, Saudi Arabia
[5]Research Center, Future University in Egypt, New Cairo 11835, Egypt

Corresponding author: Abdelwahed Motwakel (am.ismaeil@psau.edu.sa)

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R408), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Research Supporting Project number (RSPD2024R521), King Saud University, Riyadh, Saudi Arabia. This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2023/R/1444). This study is partially funded by the Future University in Egypt (FUE).

ABSTRACT Recently, deep learning models have become more prominent due to their tremendous performance for real-time tasks like face recognition, object detection, natural language processing (NLP), instance segmentation, image classification, gesture recognition, and video classification. Image captioning is one of the critical tasks in NLP and computer vision (CV). It completes conversion from image to text; specifically, the model produces description text automatically based on the input images. In this aspect, this article develops a Lighting Search Algorithm (LSA) with a Hybrid Convolutional Neural Network Image Captioning System (LSAHCNN-ICS) for NLP. This introduced LSAHCNN-ICS system develops an end-to-end model which employs convolutional neural network (CNN) based ShuffleNet as an encoder and HCNN as a decoder. At the encoding part, the ShuffleNet model derives feature descriptors of the image. Besides, in the decoding part, the description of text can be generated using the proposed hybrid convolutional neural network (HCNN) model. To achieve improved captioning results, the LSA is applied as a hyperparameter tuning strategy, representing the innovation of the study. The simulation analysis of the presented LSAHCNN-ICS technique is performed on a benchmark database, and the obtained results demonstrated the enhanced outcomes of the LSAHCNN-ICS algorithm over other recent methods with maximum Consensus-based Image Description Evaluation (CIDEr Code) of 43.60, 59.54, and 135.14 on Flickr8k, Flickr30k, and MSCOCO datasets correspondingly.

INDEX TERMS Deep convolutional neural network, natural language processing, image captioning, machine learning, hyperparameter tuning.

I. INTRODUCTION

Recently, a massive number of images have been stored digitally and transferred on the Internet as a significant source of information [1], [2]. CV techniques enable computers to define the visual world and, therefore, could bring various promising applications, namely information retrieval, the interaction of human computers, assistance for visually impaired people, and child education [3], [4]. Image captioning is an extensive process in (NLP) and (CV) that could finalize multi-modal transformation from images to texts [5]. For example, an important and challenging domain of artificial intelligence (AI), automatically generating image portrayals is gaining considerable interest [6]. The objective

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

of image captioning can be for producing linguistically plausible sentences, which are semantically correct for the image contents [7], [8]. Thus, the description of an image can have 2 main features: language processing and visual understanding. To guarantee that the created sentence has been grammatically and semantically correct, NLP and CV technologies could be utilized to properly incorporate them and deal with the issue created by the resultant modality [9].

Considerating an image depends essentially on attaining image features [10], [11]. This system utilized for these purposes is widely classified into (1) Deep machine learning (DML) assisted technique and (2) Traditional machine learning (ML) based technique [12]. In the DML technique, features are automatically learned from the training dataset, and they can handle a diverse and large set of videos and images [13], [14]. For instance, (CNN) is extensively employed for learning features, and classifiers like Softmax are utilized for classification [15]. Generally, CNN is followed by Recurrent Neural Network (RNN) to create captions. At the same time, in the traditional ML technique, the handcrafted feature was broadly applied [16]. In this technique, features have been removed from an input dataset. Then, it passed to classifiers, namely Support Vector Machines (SVM), to categorize the object. Since the handcrafted feature is a particular task, removing features from a diverse and large collection of information is not possible [17]. Furthermore, real-time information like video and images have different semantic interpretations and are complex [18].

Although image captioning research has made significant progress, several challenges should be solved. It is significant for the annotation of image captioning models that may struggle with ambiguous images. In addition, data processing is a key procedure for image captioning. More specifically, selecting the optimal hyperparameter and handling imbalanced datasets are two main issues that affect the training process. So far, real-time and multi-modal processing are two main limitations for most of the existing image captioning models. Recent approaches to image captioning systems did not focus on the hyperparameter selection method to affect the effectiveness of the classification model. Mainly, hyperparameters like batch size, epoch count, and learning rate selection could be required to gain improved performance. As the trial and error technique for hyperparameter tuning was a tiresome and erroneous procedure, meta-heuristic algorithms can be implemented. Consequently, in this work, the LSA is used for the parameter selection of the Hybrid Convolutional Neural Network (HCNN) model.

This article develops an LSA with an HCNN-based image Captioning System called (LSAHCNN-ICS) for NLP. The presented LSAHCNN-ICS method develops an end-to-end model which employs CNN-based ShuffleNet as an encoder and HCNN as a decoder. At the encoding part, the ShuffleNet model derives feature descriptors of the image. Besides, in the decoding part, the description of text can be generated using the HCNN model. To achieve improved captioning results, the LSA is applied as a hyperparameter tuning strategy. The investigational validation of this presented LSAHCNN-ICS system is implemented on a benchmark dataset.

The rest of this study is systematized as follows. Section II gives a literature review of image captioning techniques. Then, section III presents the proposed LSAHCNN-ICS method and section IV delivers the experimental validation. Lastly, section V accomplishes the work.

II. RELATED WORKS

Wang and Huang [19] have presented a local representation-improved recurrent convolution network (Lore-RCN). The authors have developed a visual convolution network for obtaining improved local linguistic context that integrates selective local visual data and methods of short-term neighbouring. In addition, they have designed a linguistic convolution network for obtaining improved linguistic representations that techniques long- and short-term connections explicitly for leveraging administrative data in preceding linguistic tokens. He and Lu [20] have suggested an end-to-end method that relies on RNN as a decoder deep and CNN as an encoder. For obtaining superior image captioning extracting, the authors have presented an extremely modularized multi-branch CNN that can improve accuracy while retaining the count of hyperparameters unaltered.

Al-Malla et al. [21] projected an attention-related, Encoding-Decoding deep framework that generates convolution feature extracting in CNNs technique pretraining on ImageNet (Xception), along with object extracting feature in YOLOv4 method, pre-training on MS COCO. Prudviraj et al. [22] have introduced a new multiscale FF network (M-FFN) to ICS tasks for incorporating distinct features and image contextual data of images. Specifically, the author gets benefits of MSFPN for incorporating global contextual data through atrous convolutional at top layers of CNNs. Faiyaz Khan et al. [23] elaborated an end-wise image captioning method employing a multi-modal infrastructure integrating a 1D-CNN for encoding sequence data with pre-training ResNet50 method image encoding to extract region-based visual features.

An effectual structure to caption the remote sensing image (RSI) was presented in [24]. This structure is dependent upon multi-level attention and multilabel feature graph convolutional. More precisely, the presented multi-level attention component is adaptably concentrated on particular spatial features, among them on features of certain scales. In addition, the attribute graph convolution component (GCN) has utilized the attribute graph for learning highly efficient attribute features to the image caption. Dong et al. [25] have investigated a Dual Graph Convolution Network (Dual-GCN) with curriculum and transformer learning to image caption. It is worth mentioning that the authors of [25] did not only utilize an object-level GCN for capturing the object-to-object spatial relationship in a single image.

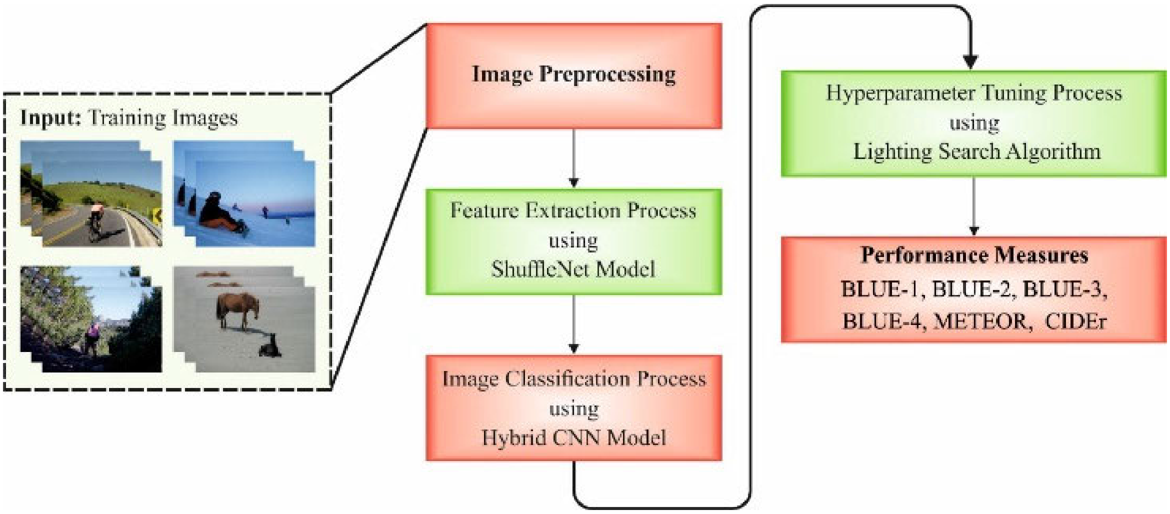
Input: Training Images
Image Preprocessing
Feature Extraction Process using ShuffleNet Model
Image Classification Process using Hybrid CNN Model
Hyperparameter Tuning Process using Lighting Search Algorithm
Performance Measures
BLUE-1, BLUE-2, BLUE-3, BLUE-4, METEOR, CIDEr

FIGURE 1. Working process of the proposed model.

With the well-planned Dual-GCN, the authors of [25] create the linguistic transformer superior to understand the connection betwixt distinct objects from the single image and generate complete utilization of the same images as auxiliary data for generating a reasonable caption explanation to a single image.

Wang and Gu [26] present a novel Joint Relationship Attention Network (JRAN), which newly discovers the connections among the feature from the image. In theory, the JRAN exploits semantic features as supplementary to region features, completely learning 2 kinds of connections, the visual connections among region features and the visual–semantic connections among the region and semantic features. Wang and Gu [27] examine the Double-Level Relationship Networks (DLRN) that newly act as the complementary local and global features from the image and improves the connection among features. The former learn distinct hierarchies of visual relationships by applying graph attention for local-level relationship improvement and pixel-level relationship improvement correspondingly.

In [28], the authors analyse the local visual modelling with grid features for image captioning that can be vital to generating correct and detailed captions. To accomplish this objective, the author presents a Locality-Sensitive Transformer Network (LSTNet) with two novel designs Locality-Sensitive Attention and Locality-Sensitive Fusion (LSF). In [29], a Local Relation Network (LRN) was planned over the objects and image regions that not only determines the connection among the object and image regions among them creates major context-based features equivalent to all the regions from the image. Lastly, a different typical LSTM utilizes an attention process that concentrates on related contextual data, spatial places, and deep visual features.

Different from the above recent methods, we propose to enhance the process of image captioning by incorporating CNN and LSA. The former is important for feature extraction, while the latter is useful for optimizing parameter tuning. We indicate that some of the related works have been applied for Flickr8k, Flickr30k and MSCOCO datasets, while others have been tested for a limited number of datasets. Besides, the related works on image captioning did not consider the number of channels.

III. THE PROPOSED MODEL

In this study, an innovative LSAHCNN-ICS algorithm can be developed for captioning images in the NLP. This introduced LSAHCNN-ICS technique depends upon an end-to-end model comprising two major parts: CNN-based ShuffleNet as an encoder and HCNN as a decoder. Automated ICS employ an encoder and decoder framework to extract features from an image using the encoder, whereas the role of the decoder consists of generating a transcript. In this case, the ShuffleNet model has been exploited to remove features from the image, and the HCNN model acts as a decoder which produces the transcript. Fig. 1 represents the block diagram of the LSAHCNN-ICS model.

A. ENCODER UNIT: SHUFFLENET

In the encoding part, the ShuffleNet model derives feature descriptors of the image. The proposed model uses ShuffleNet as the backbone CNN to remove visual features at the input image. ShuffleNet's compact architecture and efficient computation make it suitable for extracting image features. The encoding unit captures the fine-grained information, encodes it, and generates fixed-size vectors. ShuffleNet has related concepts to ResNet, MobileNet, and Xception. Depthwise separable convolution and channel shuffle are used to enhance the ResNet architecture, which ensures network performance and enhances operational efficacy [30]. Different from the residual structure that straightaway incorporates the deep and non-deep features accomplished by numerous convolutions, the inverted residual model splits the input

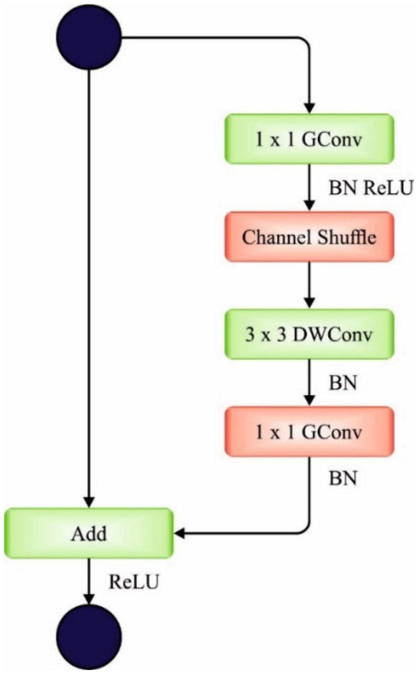

FIGURE 2. Structure of ShuffleNet model.

feature maps into two divisions, X1 and X2, they are merged with deep and non-deep features, and lastly, it uses channel shuffle to fuse deep and non-deep features. Fig. 2 illustrates the framework of the ShuffleNet technique. Assume that the input layer is separated into G groups, and the overall no. of channels is G × n. Selective Kernel Networks (SK) and Squeeze-and-Excitation Networks (SE) are added to the model to accurately make the classification correspondingly. Initially, a feature map U with a size of H × W and the overall amount of channels C are compressed into feature vectors of (1, 1, C) through a global pooled $\mathrm{F_{sq}}$ given in the following [30].

$$Z_c = F_{sq}\left(U_c\right) = \frac{1}{H \times W} \sum\nolimits_{i=1}^{H} \sum\nolimits_{j=1}^{W} U_c\left(I, j\right). \tag{1}$$

Then, add linear mapping and the activation function to the feature vectors for handling non-linear conditions that could best adapt the complex correlations amongst channels. Lastly, the evaluated channel feature was multiplied with the deep feature maps to get the output. The SE method weakens the insignificant feature and strengthens the significant feature by controlling the size of the channel to make the extracted feature more directional. Channel attention is allowable to be inserted among all the feature maps.

B. DECODER UNIT: LSA WITH HCNN MODEL

In this study, in the decoding part, the description of text can be generated using the HCNN model. The outcomes of the image feature and word sequence encoders can be integrated by combination and fed as input into the HCNN model. The HCNN model produces a softmax forecast in all vocabulary words to be the succeeding word in a sequence, and the word at the maximum possibility can be chosen. These procedures are continued till the ending token is produced. The HCNN methodology generates text description, which comprises a sequence connection of CNN and LSTM [31]. The presented method could extract complicated features amongst many sensor parameters gathered for forecasting power demands and save complex irregular trends. Firstly, the upper layer of HCNN comprises CNN. The CNN could obtain different parameters that affect power utilization, namely, sub-metering, voltage, and intensity. Furthermore, household features like time, date, household occupancy, and behaviour of the residents are modelled as Metadata in the CNN layer. CNN comprises of input layer which accepts sensor variables as input, an output unit that extracts features to LSTM, and multiple hidden layers. The convolution layer employs the convolutional process to the incoming multi-variate time sequence and passes the outcomes to the following layer. Every convolutional neuron processes power utilization information for the receptive field. The convolution process could decrease the parameter count and make the HCNN network deeper. Where $x_i^0 = \{x_1, x_2, \ldots, x_n\}$ is the power utilization input vector, and n indicates the number of normalized 60 min units for every window. Eq. (2) denotes the outcome of vector y_{ij}^1 output from the initial convolution layer, y_{ij}^1 is computed using output vector x_{ij}^1 of the preceding layer, m indicates the index value of filter, b_j^1 characterizes the bias for j^{rh} feature maps, w denotes the weight of the kernel, and 0 denotes the activation function as follows [31].

$$y_{ii}^1 = s\left(b_i^1 + \sum_{m=1}^{M} w_{m,j}^1 x_{i+m-1,j}^0\right) \tag{2}$$

$$y_{ij}^1 = s\left(b_i^l + \sum_{\mathrm{m}=1}^{M} w_{\mathrm{m},j}^l x_{i+\mathrm{m}-1,j}^0\right) \tag{3}$$

The pooling layer decreases the space size of the demonstration to decrease the network computation and cost number of parameters. The convolutional layer employs a pooling layer that integrates the output of neuron clusters in a single layer into one neuron in the following layer. Eq. (4) characterizes the max-pooling layer operation. R indicates the pooling size lesser than the input size y, and T denotes the stride that decides how much further to move the region of the input dataset.

$$p_{ij}^l = \max_{r \in R} y_{i \times T + r, j}^{l-1} \tag{4}$$

LSTM is a lower layer of HCNN, which store time dataset regarding significant features. The output value in the preceding CNN layer could be accepted by the gate unit. The latter encompasses forget, input and output gates. The memory cell makes up the LSTM upgrade the state with activation of all the gating units that are controlled to constant values among zero and one as follows [31]:

$$i_t = \sigma\left(W_{pi}p_t + VV_{hi}h_{t-1} + W_{ci}\circ c_{t-1} + b_i\right) \tag{5}$$

$$f_t = \sigma\left(W_{pf}p_t + VV_{hf}h_{t-1} + W_{cf}\circ c_{f-1} + b_f\right) \tag{6}$$

$$\mathrm{O}_t = \sigma\left(W_{po}p_t + W_{ho}h_{t-1} + VV_{co}\circ c_f + b_o\right) \tag{7}$$

The input, forget, and output gates constitute the LSTM, as well as output of all the gates is characterized as i, f, and o. The cell and hidden states c and h are defined by input, forget, and output gates. 0 denotes an activation function, namely tanh. This activation function has nonlinearity and correspondingly squashes the input within [−1, 1]. W indicates the weighted matrixes, and b indicates the bias vector. p_f encompasses the critical feature of power utilization as the output.

$$c_f = f_f \circ c_{f-1} + i_r \circ \sigma\left(W_{pc}p_t + VV_{hc}h_{t-1} + b_c\right) \tag{8}$$

$$h_r = 0_t \circ \sigma\left(c_t\right) \tag{9}$$

The final layer of HCNN is composed of FC layers. They are utilized for generating the text and $h^l = \{h_1, h_2, h_l\}$, whereas l denotes the number of units in LSTM. The output of LSTM was applied as input for the FC layer.

To achieve improved captioning results, the LSA is applied as a hyperparameter tuning strategy. The LSA is a novel meta-heuristic method that depends on the lightning occurrence in nature [32]. The main purpose has been a generalization of the hypotheses through the model of step leader propagation. LSA executes a search procedure through faster particles called projectile that moves in the searching space. The projectile technique is the same as other positions utilized in evolutionary mechanisms, comprising "chromosome", "particle", or "individual". The LSA combines three kinds of projectiles that are determined as follows:

- Transition projectile: projectiles form an early population of step leaders. The projectile was generated using a random value derived at a uniform likelihood distribution, and it can be expressed in the following [32]:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & x < a \text{ or } x > b \end{cases} \tag{10}$$

- Space projectile: the projectiles are upgraded and evolved such that one of them becomes a leader. The upgrading process is formulated as follows:

$$p_j^s = p_i^s \pm exprnd(D) \tag{11}$$

In the expression, p_j^s represents the upgraded space projectile, p_i^s indicates the older space projectile, and *exprnd* generates arbitrary value at the exponential distribution, and it can be given as follows:

$$f(x) = \begin{cases} \dfrac{1}{\mu}e^{\frac{-x}{\mu}} & x > 0 \\ 0 & x \le 0 \end{cases} \tag{12}$$

μ is considered as the distance D within p_i^s and leader projectile p^L is shown below.

$$D = \left|p^L - p_i^s\right| \tag{13}$$

- The leader projectile: characterize the optimal solution. This can be upgraded as [32]:

$$p_{new}^L = p^L + normrnd\,(0, E_k) \tag{14}$$

Assume p_{new}^L as the upgraded leader projectile, $andp^L$ shows the old leader projectiles, $normran(\mu, \sigma)$ arbitrarily generated number with μ mean and σ standard deviation. The parameter σ is employed as E_k kinetic energy, which is reduced exponentially through the evolution of iteration. t defines the existing iteration count, and T refers to the overall iteration count. Eq. (15) shows that the randomly produced lead projectile could be searched in every direction at the current place described by the shape parameter [32].

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{15}$$

$$E_k = 2.05 - 2\exp\left(\frac{-5(T-t)}{T^\gamma}\right) \tag{16}$$

The upgraded space and lead projectile for replace the old projectile and make the channel development as long as the energy (quality) is more proficient than the old one.

Algorithm 1 Pseudocode of LSA

```
Initializing Max iteration, channel time
Initializing lead tips energy
Produce transition projectiles at random using equation (10)
Assess the performance of projectiles
Iteration = 1
While iteration ≤ Max_iteration do
    Upgrading lead tips energy using equation (16)
    Upgrading best and worst leaders using equation (14)
    If Max channel time is obtained, then
        Move step leader from the worst location towards the best
            Rearrange channel time
    End if
    Upgrading kinetic energy and its direction using equation (16)
    Upgrading space and leader projectile
    Assess the performance of the projectile
    If fork then
        Produce two symmetrical channels at the fork point
            Remove the channel that has lower energy
    End if
    Iteration = iteration +1
End while
Return optimum step leader
```

The novel p_j projectiles are produced by the subsequent formula. a and b refer to the upper and lower boundaries. Later, the quality of *the* p_j is estimated. The better of 2 projectiles remain in the population, whereas others are disregarded. This process is utilized in the LSA with the lowest rate.

IV. PERFORMANCE VALIDATION

The proposed model is simulated using the Python 3.8.5 tool on PC i5-8600k, GeForce 1050Ti 4GB, 16GB RAM, 250GB SSD, and 1TB HDD. The parameter settings are given as follows: learning rate: 0.01, dropout: 0.5, batch size: 5, epoch count: 50, and activation: ReLU.

TABLE 1. Dataset details.

Dataset	Number of images	Reference
Flickr8k	8000	[33]
Flickr30k	31000	[34]
MSCOCO	164062	[35]

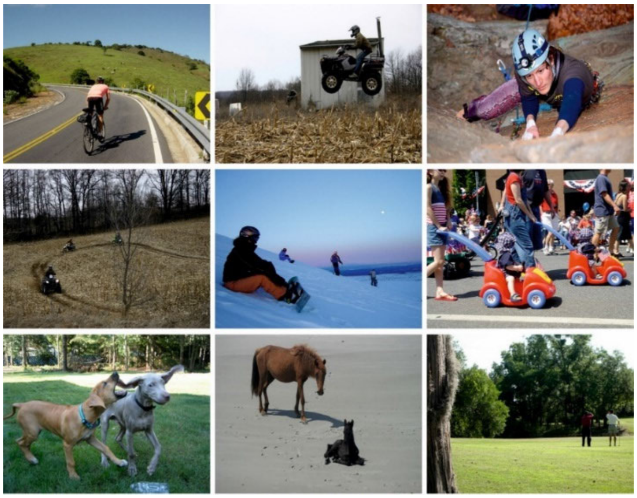

FIGURE 3. Sample images.

TABLE 2. Image captioning result of LSAHCNN-ICS technique with other approaches under Flickr8K dataset [36], [37], [38], [39].

Flickr8K Database						
Methodologies	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	CIDEr
NIC	57.93	43.03	32.13	18.63	14.53	31.56
Soft-Attention	60.52	44.84	34.78	20.65	16.96	33.60
Hard-Attention	62.70	46.81	36.80	22.93	18.56	36.20
SCA-CNN-VGG	65.06	49.73	39.01	24.52	21.46	37.90
CNN Model	67.57	51.80	41.11	26.54	23.77	40.81
LSAHCNN-ICS	70.15	53.61	43.70	29.50	26.66	43.60

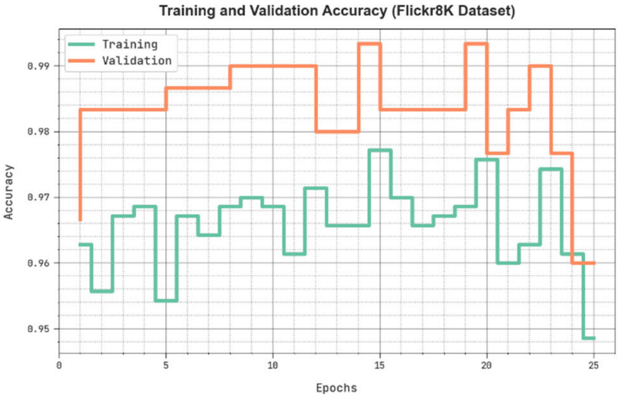

FIGURE 4. TR_{acc} and VL_{acc} analysis of LSAHCNN-ICS model at Flickr8K dataset.

The experimental image captioning results of the LSAHCNN-ICS model are investigated on 3 databases (Flickr8k, Flickr30K, and MSCOCO), as given in Table 1. Flickr8k dataset is a novel standard collection for sentence-based image description and search, containing 8,000 images that have been each paired with 5 diverse captions, which present clear descriptions of the events and salient entities. The Flickr30k database comprises 31,000 images gathered at Flickr, along with 5 reference sentences offered via human annotators. The MS COCO (Microsoft Common Objects in Context) database can be a massive quantity of object detection, captioning dataset, segmentation, and key-point detection. These datasets have been widely used in various computer vision and natural language processing tasks, providing rich and comprehensive resources for image captioning and understanding research.

Fig. 3 illustrates some sample images. A set of brief comparative analyses is made with recent methods, namely hard-attention, Neural Image Caption (NIC) [36], soft-attention [37], hard attention [37], Spatial and Channel-wise Attention with CNN VGG (SCA-CNN-VGG) [38], and CNN [39] methods.

In this study, three parameters were employed for investigational validation such as CIDEr, BLEU, and Meter. Bleu is a commonly used evaluation to predict the quality of the produced text. The values of Blue should be more to increase the efficiency of machine translation. METEOR measure mostly relies upon single precision weighted harmonic mean and word recall rate. It evaluates the reconciliation accuracy as well as recalls among optimal candidate and reference translations. The CIDEr index considered all sentences as a "document" and reported it in the type of a TF-IDF vector. It determines the cosine resemblance between the reference caption and the generated caption utilizing a score value.

Table 2 presents an entire image captioning analysis of the LSAHCNN-ICS algorithm in the Flickr8K database. The experimental value demonstrates that the NIC approach has shown the least performance while the soft-attention and hard-attention models have certainly exhibited increased outcomes.

Along with that, the SCA-CNN-VGG and CNN models have tried to depict closer image captioning performance. However, the LSAHCNN-ICS model has outperformed the existing approaches with increased BLEU-1 of 70.15, BLEU-2 of 53.61, BLEU-3 of 43.70, BLEU-4 of 29.50, METEOR of 26.66, and CIDEr of 43.60.

The training accuracy (TR_{acc}) and validation accuracy (VL_{acc}) accomplished with the LSAHCNN-ICS system with the Flickr8K database are showcased in Fig. 4. The obtained value pointed out the LSAHCNN-ICS system has realized increased values of TR_{acc} and VL_{acc}. Specifically, the VL_{acc} looked better than the TR_{acc}.

Fig. 5 displays the training loss (TR_{loss}) and validation loss (VL_{loss}) accomplished using LSAHCNN-ICS methodology in the Flickr8K database. The simulation outcome stated that the LSAHCNN-ICS algorithm had attained decreased values of TR_{loss} and VL_{loss}. In certain, the VL_{loss} is lesser than TR_{loss}.

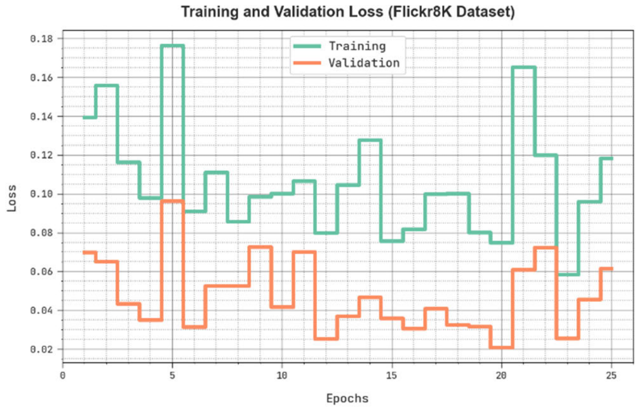

FIGURE 5. TR_{loss} and VL_{loss} analysis of LSAHCNN-ICS methodology under Flickr8K dataset.

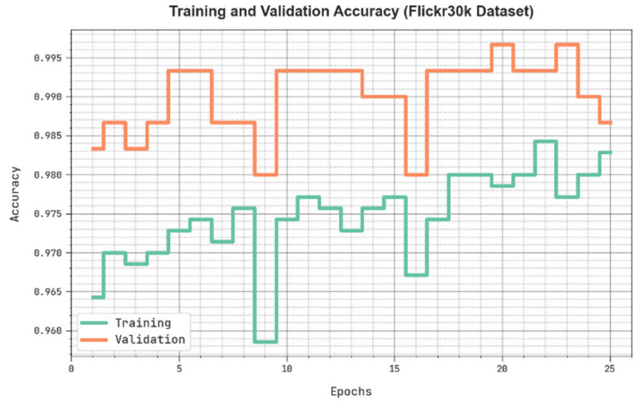

FIGURE 6. TR_{acc} and VL_{acc} analysis of LSAHCNN-ICS technique with Flickr30K database.

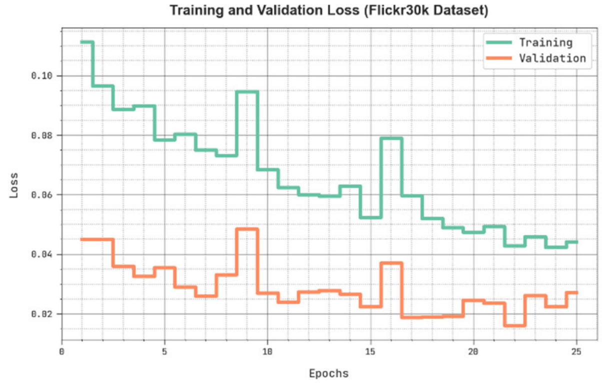

FIGURE 7. TR_{loss} and VL_{loss} outcome of LSAHCNN-ICS technique on Flickr30K database.

A widespread comparative image captioning results of the LSAHCNN-ICS methodology with other systems on the Flickr30K database can be described in Table 3. The obtained result represents that the NIC methodology has reached minimal image captioning outcomes.

The TR_{acc} and VL_{acc} gained by the LSAHCNN-ICS methodology in the Flickr30K database are defined in Fig. 6.

TABLE 3. Image captioning result of LSAHCNN-ICS technique with other approaches under Flickr30K dataset [36], [37], [38], [39].

Flickr30k Dataset						
Methodologies	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	CIDEr
NIC	58.95	48.95	38.45	27.55	21.65	37.99
Soft-Attention	61.23	51.40	41.32	29.06	23.90	39.87
Hard-Attention	62.79	53.33	44.30	31.27	25.96	42.61
SCA- CNN-VGG	64.88	56.00	47.27	34.23	28.50	45.27
CNN Model	67.59	57.64	49.01	36.15	29.01	56.82
LSAHCNN-ICS	69.23	60.39	51.06	38.07	32.68	59.54

TABLE 4. Image captioning result of LSAHCNN-ICS technique with other approaches under the MSCOCO dataset [36], [37], [38], [39].

MSCOCO Database						
Methodologies	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	CIDEr
NIC	62.47	48.47	33.87	22.77	19.57	68.49
Soft-Attention	64.14	50.65	35.41	25.45	21.64	71.15
Hard-Attention	67.01	52.40	36.99	28.12	24.54	89.19
SCA- CNN-VGG	69.15	55.25	39.89	31.01	26.51	105.82
CNN Model	75.54	57.84	42.71	32.51	29.41	117.90
LSAHCNN-ICS	77.34	59.84	45.46	35.44	30.95	135.41

TABLE 5. Image captioning result for the state-of-art methods.

Related works	Dataset	BLEU _1	BLEU _2	BLEU _3	BLEU _4	METEO R	CIDE r
[20]	Flickr8k	69.9	50.6	36.8	26.5	23	-
	Flickr30k	69.1	50.6	36.3	26.0	20.9	55.4
	MSCOCO	76.3	59.8	44.9	34.2	27.5	111.3
[22]	Flickr30k	76.4	62.4	53.1	39.5	29.2	62.1
[23]	BanglaLekhaImageCapt ions	65.1	42.6	27.8	17.5	29.7	50.1
[25]	MSCOCO	82.2	67.6	52.4	39.7	29.7	129.2

The simulation result outperformed the LSAHCNN-ICS method is getting superior values of TR_{acc} and VL_{acc}. Precisely, the VL_{acc} appeared higher than TR_{acc}.

Fig. 7 shows the TR_{loss} and VL_{loss} accomplished by the LSAHCNN-ICS system with the Flickr30K database. The result revealed that the LSAHCNN-ICS approach had reached reduced values of TR_{loss} and VL_{loss}. Explicitly, the VL_{loss} is smaller than TR_{loss}.

Fig. 8 shows the TR_{acc} and VL_{acc} achieved by the LSAHCNN-ICS algorithm at the MSCOCO database. The achieved result states that the LSAHCNN-ICS technique can be realized improved values of TR_{acc} and VL_{acc}. Notably, VL_{acc} appears to exist better than TR_{acc}.

Fig. 9 shows the TR_{loss} and VL_{loss} accomplished by the LSAHCNN-ICS method under the MSCOCO database are exhibited. The simulation result demonstrated that the LSAHCNN-ICS approach is capable of minimal values of TR_{loss} and VL_{loss}. At certain, the VL_{loss} is reduced than TR_{loss}.

V. DISCUSSION

The results assured the enhancements of the LSAHCNN-ICS model on the image captioning process. The better efficiency of this developed approach is because of the integration

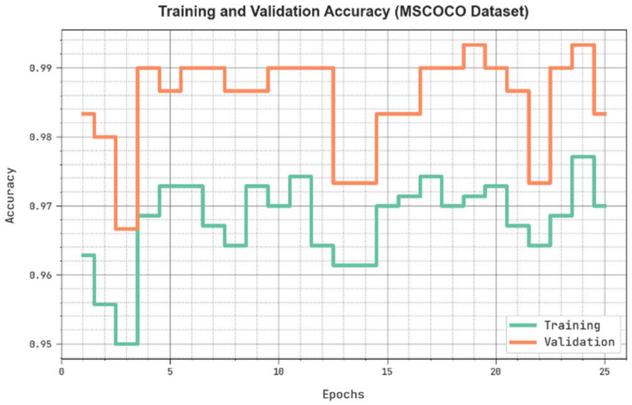

FIGURE 8. ***TR_{acc}* and *VL_{acc}* analysis of LSAHCNN-ICS technique under the MSCOCO dataset.**

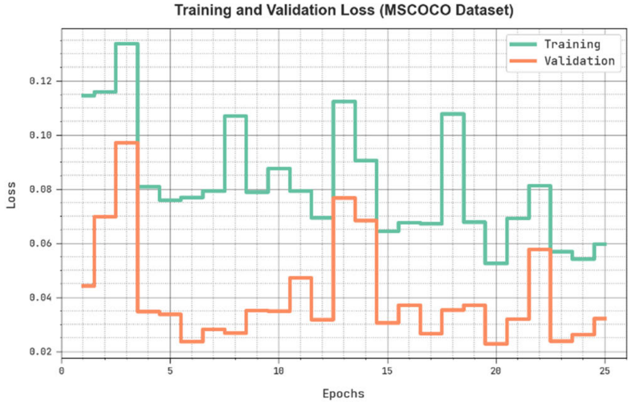

FIGURE 9. ***TR_{loss}* and *VL_{loss}* result of LSAHCNN-ICS methodology under MSCOCO database.**

of LSA-based hyperparameter tuning manner and unique characteristics of the HCNN model. Since the manual hyperparameter tuning process impacts the effectiveness of the DL model, the automated hyperparameter tuning using LSA helps to accomplish improved performance over other DL models. By systematically searching through the space of possible architectures, the LSA determines the optimal combination of hyperparameter values that optimize the model's performance on a specific task, such as image captioning. In addition, the LSA explores different architectural configurations to find the optimal combination that enhances the model's understanding of image content and improves the quality of generated captions.

More precisely, table 5 shows the effectiveness of recent systems that have considered the same datasets tested in our work. The results obtained by our framework are superior to the results obtained in [20] for Flickr8K and Flickr30k datasets. In addition, our results are better than the results obtained in [22] in terms of the METEOR metric. Table 5 also illustrates that the LSAHCNN-ICS approach has demonstrated better results for the MSCOCO dataset compared to the results achieved in [20] and [23]. Moreover, LSAHCNN-ICS has provided better results than [25] in terms of the METEOR metric.

Despite the advantages of our framework, we indicate that our study is limited to three datasets including Flickr8k, Flickr30K, and MSCOCO. In addition, our system allows us to generate only one caption per image, while some applications need to generate multiple captions per image based on a specific purpose and perspective.

VI. CONCLUSION

In this article, a novel LSAHCNN-ICS methodology can be developed for captioning images in the NLP. The presented LSAHCNN-ICS technique developed an end-to-end model comprising two major parts: CNN-based ShuffleNet as an encoder and HCNN as a decoder. At the encoding part, the ShuffleNet model derives feature descriptors of the image. Besides, in the decoding part, the description of text can be generated using the HCNN model. To achieve improved captioning results, the LSA is applied as a hyperparameter tuning strategy. The simulation analysis of the presented LSAHCNN-ICS technique is performed on a benchmark database, and the achieved results reported the superior outcomes of the LSAHCNN-ICS algorithm over existing systems with maximum CIDEr of 43.60, 59.54, and 135.14 on Flickr8k, Flickr30k, and MSCO-CO datasets respectively. The enhanced performance is owing to the addition of LSA assisted hyperparameter tuning process and the unique characteristics of the HCNN model. Therefore, the proposed model can be used to improve assistive technology and aid the visually impaired in comprehending their environment. In the forthcoming, the efficiency of the LSAHCNN-ICS method can be improved with the usage of the weighted voting ensemble DL model. Also, this developed method could be enriched for the development of hybrid meta-heuristic algorithms in hyperparameter tuning. This proposed image captioning model can be computationally expensive, especially when dealing with large images and complex architectures. Real-time applications require efficient and optimized models to generate captions quickly. The time required for generating captions can introduce significant latency in real-time applications. Reducing the inference time is essential to ensure a smooth user experience. In real-time scenarios, the model may encounter objects or scenes it has not seen during training. Robustness to handle such unseen concepts is crucial for accurate and relevant captions.

ACKNOWLEDGMENT

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R408), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Research Supporting Project number (RSPD2024R521), King Saud University, Riyadh, Saudi Arabia. This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2023/R/1444). This study is partially funded by the Future University in Egypt (FUE).

REFERENCES

[1] S. Kalra and A. Leekha, "Survey of convolutional neural networks for image captioning," *J. Inf. Optim. Sci.*, vol. 41, no. 1, pp. 239–260, Jan. 2020.

[2] G. Geetha, T. Kirthigadevi, G. G. Ponsam, T. Karthik, and M. Safa, "Image captioning using deep convolutional neural networks (CNNs)," *J. Phys., Conf. Ser.*, vol. 1712, no. 1, Dec. 2020, Art. no. 012015.

[3] W. Ren, A. H. Bashkandi, J. A. Jahanshahi, A. Q. M. AlHamad, D. Javaheri, and M. Mohammadi, "Brain tumor diagnosis using a step-by-step methodology based on courtship learning-based water strider algorithm," *Biomed. Signal Process. Control*, vol. 83, May 2023, Art. no. 104614.

[4] R. Li, H. Liang, Y. Shi, F. Feng, and X. Wang, "Dual-CNN: A convolutional language decoder for paragraph image captioning," *Neurocomputing*, vol. 396, pp. 92–101, Jul. 2020.

[5] S. Srivastava, H. Sharma, and P. Dixit, "Image captioning based on deep convolutional neural networks and LSTM," in *Proc. 2nd Int. Conf. Power Electron. IoT Appl. Renew. Energy Control (PARC)*, Jan. 2022, pp. 1–4.

[6] J. Wei, Z. Li, J. Zhu, and H. Ma, "Enhance understanding and reasoning ability for image captioning," *Appl. Intell.*, vol. 53, no. 3, pp. 2706–2722, Feb. 2023

[7] G. Hoxha and F. Melgani, "A novel SVM-based decoder for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3105004.

[8] Q.-H. Kha, Q.-T. Ho, and N. Q. K. Le, "Identifying SNARE proteins using an alignment-free method based on multiscan convolutional neural network and PSSM profiles," *J. Chem. Inf. Model.*, vol. 62, no. 19, pp. 4820–4826, Oct. 2022.

[9] Q. Yuan, K. Chen, Y. Yu, N. Q. K. Le, and M. C. H. Chua, "Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding," *Briefings Bioinf.*, vol. 24, no. 1, Jan. 2023, Art. no. bbac630.

[10] J. Qiu, F. P.-W. Lo, X. Gu, M. L. Jobarteh, W. Jia, T. Baranowski, M. Steiner-Asiedu, A. K. Anderson, M. A. Mccrory, E. Sazonov, M. Sun, G. Frost, and B. Lo, "Egocentric image captioning for privacy-preserved passive dietary intake monitoring," *IEEE Trans. Cybern.*, to be published.

[11] C. Bai, A. Zheng, Y. Huang, X. Pan, and N. Chen, "Boosting convolutional image captioning with semantic content and visual relationship," *Displays*, vol. 70, Dec. 2021, Art. no. 102069.

[12] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[13] X. Zhang, Y. Li, X. Wang, F. Liu, Z. Wu, X. Cheng, and L. Jiao, "Multi-source interactive stair attention for remote sensing image captioning," *Remote Sens.*, vol. 15, no. 3, p. 579, 2023.

[14] V. Atliha and D. Šešok, "Comparison of VGG and ResNet used as encoders for image captioning," in *Proc. IEEE Open Conf. Electr., Electron. Inf. Sci. (eStream)*, Apr. 2020, pp. 1–4.

[15] A. M. Rinaldi, C. Russo, and C. Tommasino, "Automatic image captioning combining natural language processing and deep neural networks," *Results Eng.*, vol. 18, Jun. 2023, Art. no. 101107.

[16] S. Sehgal, J. Sharma, and N. Chaudhary, "Generating image captions based on deep learning and natural language processing," in *Proc. 8th Int. Conf. Rel., INFOCOM Technol. Optim. (Trends Future Directions) (ICRITO)*, Jun. 2020, pp. 165–169.

[17] B. Makav and V. Kilic, "Smartphone-based image captioning for visually and hearing impaired," in *Proc. 11th Int. Conf. Electr. Electron. Eng. (ELECO)*, Nov. 2019, pp. 950–953.

[18] N. Gupta and A. S. Jalal, "Integration of textual cues for fine-grained image captioning using deep CNN and LSTM," *Neural Comput. Appl.*, vol. 32, no. 24, pp. 17899–17908, Dec. 2020.

[19] X. Wang and J. Huang, "A local representation-enhanced recurrent convolutional network for image captioning," *Int. J. Multimedia Inf. Retr.*, vol. 11, no. 2, pp. 149–157, Jun. 2022.

[20] S. He and Y. Lu, "A modularized architecture of multi-branch convolutional neural network for image captioning," *Electronics*, vol. 8, no. 12, p. 1417, Nov. 2019.

[21] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," *J. Big Data*, vol. 9, no. 1, pp. 1–16, Dec. 2022.

[22] J. Prudviraj, C. Vishnu, and C. K. Mohan, "M-FFN: Multi-scale feature fusion network for image captioning," *Int. J. Speech Technol.*, vol. 52, no. 13, pp. 14711–14723, Oct. 2022.

[23] M. F. Khan, S. M. Sadiq-Ur-Rahman, and S. Islam, "Improved Bengali image captioning via deep convolutional neural network-based encoder–decoder model," in *Proc. Int. Joint Conf. Adv. Comput. Intell.* Singapore: Springer, 2021, pp. 217–229.

[24] Z. Yuan, X. Li, and Q. Wang, "Exploring multi-level attention and semantic relationship for remote sensing image captioning," *IEEE Access*, vol. 8, pp. 2608–2620, 2020.

[25] X. Dong, C. Long, W. Xu, and C. Xiao, "Dual graph convolutional networks with transformer and curriculum learning for image captioning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2615–2624.

[26] C. Wang and X. Gu, "Learning joint relationship attention network for image captioning," *Expert Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118474.

[27] C. Wang and X. Gu, "Learning double-level relationship networks for image captioning," *Inf. Process. Manage.*, vol. 60, no. 3, May 2023, Art. no. 103288.

[28] Y. Ma, J. Ji, X. Sun, Y. Zhou, and R. Ji, "Towards local visual modeling for image captioning," *Pattern Recognit.*, vol. 138, Jun. 2023, Art. no. 109420.

[29] H. Sharma and S. Srivastava, "Multilevel attention and relation network based image captioning model," *Multimedia Tools Appl.*, vol. 82, no. 7, pp. 10981–11003, Mar. 2023.

[30] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[31] S. Yin, Y. Zhang, and S. Karim, "Region search based on hybrid convolutional neural network in optical remote sensing images," *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 5, May 2019, Art. no. 155014771985203.

[32] H. Shareef, A. A. Ibrahim, and A. H. Mutlag, "Lightning search algorithm," *Appl. Soft Comput.*, vol. 36, pp. 315–333, Nov. 2015.

[33] *Flickr 8K Dataset*. Accessed: Mar. 8, 2023. [Online]. Available: https://www.kaggle.com/datasets/adityajn105/flickr8k

[34] *Flickr Image Dataset*. Accessed: Mar. 2, 2023. [Online]. Available: https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset

[35] T. Y. Lin, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[36] E. 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.

[37] K. Xu, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2015, pp. 2048–2057.

[38] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6298–6306.

[39] M. Al Duhayyim, S. Alazwari, H. A. Mengash, R. Marzouk, J. S. Alzahrani, H. Mahgoub, F. Althukair, and A. S. Salama, "Metaheuristics optimization with deep learning enabled automated image captioning system," *Appl. Sci.*, vol. 12, no. 15, p. 7724, Jul. 2022.