IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 August 2023, accepted 16 September 2023, date of publication 20 September 2023,
date of current version 28 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317276

RESEARCH ARTICLE

Automated Image Captioning Using Sparrow Search Algorithm With Improved Deep Learning Model

MUNYA A. ARASI[1], HAYA MESFER ALSHAHRANI[2], NUHA ALRUWAIS[3], ABDELWAHED MOTWAKEL[4], NOURA ABDELAZIZ AHMED[5], AND ABDULLAH MOHAMED[6]

[1]Department of Computer Science, College of Science and Arts in Rijal Almaa, King Khalid University, Abha 62529, Saudi Arabia
[2]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[3]Department of Computer Science and Engineering, College of Applied Studies and Community Services, King Saud University, P.O. Box 22459, Riyadh 11495, Saudi Arabia
[4]Department of Management Information Systems, College of Business Administration Hawtat Bani Tamim, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[5]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[6]Research Centre, Future University in Egypt, New Cairo 11845, Egypt

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

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups Project under grant number (RGP2/61/44). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R237), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Research Supporting Project number (RSPD2023R608), 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 Image captioning is a deep learning technique that intends to create and generate textual descriptions or captions for images. It integrates computer vision and natural language processing (NLP) to comprehend the visual content of an image and generate human-like descriptions. Deep learning (DL) based image captioning models can be trained on large-scale datasets, allowing them to generalize various types of images and generate captions that apply to a wide range of visual scenarios. By combining computer vision and natural language processing, DL-enabled image captioning models can understand both visual and textual information, which enables them to generate captions that not only describe the visual content but also incorporate contextual and semantic information. This study develops an Automated Image Captioning using Sparrow Search Algorithm with Improved Deep Learning (AIC-SSAIDL) technique. The major intention of the AIC-SSAIDL technique lies in the automated generation of textual captions for the input images. To accomplish this, the AIC-SSAIDL technique utilizes the MobileNetv2 model to generate feature descriptors of the input images and its hyperparameter tuning process takes place using SSA. For the image captioning process, the AIC-SSAIDL technique utilizes an attention mechanism with long short-term memory (AM-LSTM) network. Finally, the hyperparameter selection of the AM-LSTM model is performed by the fruit fly optimization (FFO) algorithm. A wide range of experiments has been conducted on benchmark data to depict the better performance of the AIC-SSAIDL method. The comprehensive result analysis highlighted the enhanced captioning results of the AIC-SSAIDL method with maximum CIDEr of 46.12, 61.89, and 137.45 on Flickr8k, Flickr30k, and MSCOCO datasets, respectively.

INDEX TERMS Image captioning, deep learning, natural language processing, sparrow search algorithm, computer vision.

I. INTRODUCTION

Image captioning is a wide-ranging task in natural language processing (NLP) and computer vision (CV) that enables to completion of the multi-modal transformation from image to text. As a main source of data, multiple images are stored and transferred digitally on the Internet. Meanwhile, social interaction depends mainly on natural language that enables the computer to define the visual world might bring a large number of applications, namely assistance

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

for visually impaired people, child education, information retrieval, and natural human-computer interaction [1]. Based on the input image, this model automatically generates the text description [2]. As a meaningful and challenging field of artificial intelligence (AI), automatically generated image description has attracted considerable attention. The objective is to linguistically generate a possible sentence that is semantically correct to the image content [3]. Thus, language processing and visual understanding of image description are the two major aspects of image captioning [4]. The CV and NLP approaches must be used to properly integrate them and to handle the problem created by the corresponding modalities to guarantee that the generated sentence is semantically and grammatically true. With the rapid development of computing capability and data scale, machine learning (ML) based on hardware and data shows exclusive benefits that directly stimulate the prosperity of AI in different applications [5].

Various research was dedicated to automatic image captions, as well as it is categorized into various approaches [6]. The image caption technique is designated from similar images through captioning, and the retrieval-based method identifies visually similar images through the captioning from the training data [7], [8]. Several types of research are based on machine learning (ML) and deep learning (DL) methods. A deep neural network (DNN) technique was exploited for the image caption technique due to efficient approximation capabilities. The image caption method has tremendously grown due to the considerable expansion of the DNN technique [9]. In recent years, Convolutional Neural Network (CNN) has gained significant attention in CV tasks namely image classification, and object detection. Furthermore, Recurrent Neural Network (RNN) plays a crucial role in NLP [10]. Even though various types of research were conducted, it is still necessary to establish efficient image caption methods for better performance.

The contribution of the paper is summarized as follows. This study develops an Automated Image Captioning using Sparrow Search Algorithm with Improved Deep Learning (AIC-SSAIDL) technique. The AIC-SSAIDL technique utilizes the MobileNetv2 model to generate feature descriptors of the input images and its hyperparameter tuning process takes place using SSA. For the image captioning process, the AIC-SSAIDL technique utilizes an attention mechanism with long short-term memory (AM-LSTM) network. Finally, the hyperparameter selection of the AM-LSTM network is performed by the fruit fly optimization (FFO) algorithm. A wide range of experiments has been conducted on benchmark data to demonstrate the better performance of the AIC-SSAIDL method.

II. RELATED WORKS

Duhayyim et al. [11] established a meta-heuristic optimizer with a DL-empowered automated image caption technique (MODLE-AICT). The presented technique concentrates on generating the effectual caption to input images by utilizing two procedures containing encoder and decoding units. Firstly, at the encoder part, the SSA with HybridNet method was employed for generating effective input image descriptions utilizing fixed-length vectors demonstrating the novelty of works. Furthermore, the decoder part contains a bidirectional gated recurrent unit (BiGRU) technique utilized for generating descriptive sentences. Chaudhari and Devane [12] present an intelligent-based image captioning method. An implemented method contains some steps such as caption generation, word generation, and sentence formation. Primarily, the input image was exposed to the DL technique named convolutional neural network (CNN). Additionally, a group of sentences are designed with created words utilizing the long short-term memory (LSTM) technique. This work establishes a novel improved optimizer technique Rider with Randomized Bypass and Over-taker update (RR-BOU) for better selection.

Chu et al. [13] developed one joint AICRL technique which conducts the automatic image captions dependent upon LSTM and ResNet50 with soft attention. AICRL includes one encoding and one decoding. The encoding implements ResNet-50 dependent upon CNN that makes a varied representation of the provided image with embedded it as fixed length vector. The decoding can be planned with LSTM, recurrent neural network (RNN) and soft attention mechanism for selecting the concentration of attention on particular parts of images for predicting the next sentence. In [14], the authors presented a method that integrates a CNN and LSTM for boosting the accuracy of image captions by fusing text features accessible from an image with visual extraction features from recent approaches. The authors [15] introduce a Variational Autoencoder and Reinforcement Learning based two-stage Multi-task Learning Model (VRTMM) for remote sensing image captioning tasks. In the primary stage, the authors fine-tuned the CNN along with the variational autoencoder (VAE).

Al-Malla et al. [16] introduced an attention-based Encoder-Decoder deep structure which utilizes CNN based convolution features extraction model named Xception. Bai et al. [17] present a structure utilizing a CNN-based generation method for generating image captioning by utilizing conditional GAN (CGAN). Additionally, a multi-modal graph convolutional network (MGCN) was utilized for exploiting visual connections among objects to create the caption with semantic meaning, whereas the scene graph has been utilized as a bridge for connecting objects, attributes and visual relationship data combined for generating optimum captions.

Many automated devices are accessible in the related works for effectual image captioning. Even though the ML and DL approaches occurred in the previous studies, it is still required for enhancing the image captioning efficiency. Due to the continual deepening of the method, the count of parameters of DL approaches also enhances rapidly the outcome in model overfitting. Simultaneously, distinct hyperparameters are an important effect on the performance of the CNN approach. In particular, hyperparameters like batch size, epoch count, and learning rate selection are crucial to obtain efficient results. As the trial and error process for hyperparameter tuning is a tedious and erroneous method, meta-heuristic approaches are executed. Therefore, in this work, we employ SSA and FFO algorithms for the parameter selection of the MobileNetv2 and AM-LSTM models respectively.

III. THE PROPOSED MODEL

In this study, we have developed a new AIC-SSAIDL algorithm for the automated generation of textual captioning for the input images. It encompasses MobileNetv2-based feature vector generation, SSA-based hyperparameter tuning, AM-LSTM-based textual description generation, and FFO-based parameter optimization. Fig. 1 demonstrates the overall process of the AIC-SSAIDL approach.

A. FEATURE VECTOR PROCESS

The MobileNetv2 model is exploited at this stage to generate feature descriptors. MobileNetv2 is a mobile-enriched FC network that depends mainly on the inverted residual structure that has a bottleneck level connected with the residual connection. The MobileNet model is exploited in this stage for extracting features. Mostly, the CNN is gathered from input, fully connected (FC), convolution, pooling, and output layers [18]. In contrast with the standard NN, it features weighted sharing, local connection, and downsampling. It might effectively improve the efficiency of eliminating local features, avoid over-fitting, and reduce the network parameter. The convolutional layer is a building block of CNN, and local feature extraction was identified by interconnecting the input of every neuron to the local sensing area of the prior layer. The convolution function is categorized into activation and convolution, and the computation process is represented in Eq. (1):

$$T = f_k\left(\sum_{x,y,z=1}^{r} C_{x,y,z} w_{x,y,z}^{s} + b^{s}\right) \tag{1}$$

In Eq. (1), T and C represent the input and resultant of the convolution layer; f_k indicates the activation function of k^{th} layers; r and s characterize the sequence number of convolutions and the channel count; w and b denotes the weight and bias of convolution; x, y, and z characterize the dimensional of the input dataset.

In the activation function, Tanh, rectified linear unit (ReLU), Sigmoid, and Leaky ReLU are non-linear functions used to map the input and linear conversion for improving the non-linear expression ability of the model. Particularly, the gradient computation speed was very quick and, ReLU eliminates the vanishing gradient outcome of sigmoid purpose, hence it is most commonly used. Therefore, the ReLU was implemented in the convolution layer. The pooling layer has a feature mapping layer which reduces the resulting dimensional of the convolution layer to realize the downsampling of the local dataset and efficiently avoid overfitting. Average pooling, overlapping pooling, and max pooling are typical pooling algorithms. In such cases, max pooling was implemented to express the local feature, and various pooling and convolutional layers are exploited to realize the feature extraction.

In the FC layer, every neuron is FC to all the neurons from the front layer, and the prediction values are evaluated by the weighted sum of the inter-layer weighted coefficient. For the regression procedure, the abovementioned nonlinear activation functions are not fitting to the last FC layer. While it correspondingly maps the outcome within $(0, \infty)$, $(-1, 1)$, and $(0, 1)$ intervals. Hence, the linear activation and ReLU function are correspondingly used for resultant and FC layers to improvise the expression capability of the model.

In the MobileNetv2, followed by 19 residual bottleneck layers, the first FC layer with 32 filters is used. Training the model, process adding model parameters, building up the model, and basic model with MobileNetv2, amplification image generator, and storing for forthcoming approximation are six phases in the model progression. During training, the loss of 0.25 assured a random omission of 25% of the weight. This model drastically decreases the overfitting problem. The principal objective is to retain from gaining a wide knowledge of the input from using several weight models. A batch size of 32 images was used for this dataset. Consequently, 32 images have been learned in one cycle. Once the batch size is improvised, the model grows larger. But this minimizes module can categorize the uncommon classes. MobileNetv2 enhances efficacy over a broad size of the model. The MobileNetv2 is made up of n times as various recurrent layers. In the presented method, depthwise separable convolution is utilized that comprises pointwise and depthwise convolutional layers successively.

B. HYPERPARAMETER TUNING PROCESS

The SSA is used in this work for optimal hyperparameter selection of the MobileNetv2 model. The SSA is a metaheuristic technique which motivates the predation and anti-predatory behaviours of sparrows [19]. Especially, during foraging, joiner and discoverer are major two roles acted by individuals. The discoverer is accountable for guiding and searching for food, and the joiner forage by following the discoverer. A certain proportion of sparrows can be selected as the guarder that transfers the alarm signal and performs anti-predation action once they found out the danger. The location of the discoverer can be regenerated by using Eq. (2):

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^{t} \cdot \exp(-\dfrac{i}{\alpha \cdot T}) & R_2 < ST \\ X_{i,j}^{t} + O \cdot G & R_2 \geq ST \end{cases} \tag{2}$$

where $\alpha \in (0, 1]$ shows the random integer. O signifies an arbitrary parameter. $ST \in (0.5, 1]$ suggests a safety value. $R_2 \in (0, 1]$ determines a warning value. G represents the $1 \times d$ matrix whose value is 1. t denotes the present value of an update. T denotes the highest value of an update. X_{ij}^{t} shows the existing location of $i-th$ agents. X_{ij}^{t+1} represents the upgraded location of the i^{th} sparrow at the j^{th} dimension.

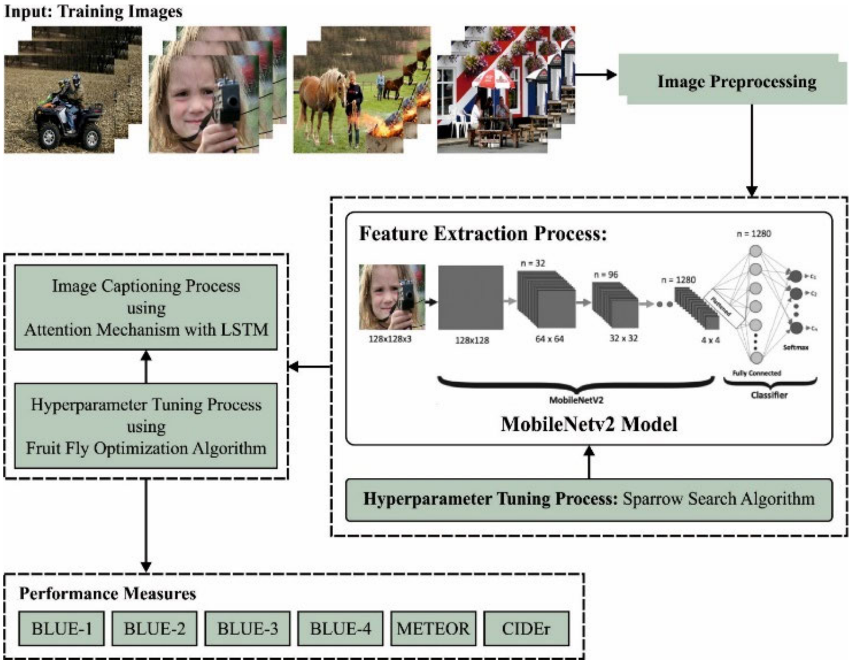

FIGURE 1. The overall process of the AIC-SSAIDL method.

The joiner location is redeveloped in Eq. (3):

$$X_{i,j}^{t+1} = \begin{cases} O \cdot \exp(\dfrac{x_w - X_{i,j}^t}{i^2}) & i > n/2 \\ X_b + |X_{i,j}^t - X_b| \cdot B \cdot G & otherwise \end{cases} \tag{3}$$

where X_b indicates the present optimal location of the discoverer. X_w defines the worst location of the sparrow, $andB$ represents the $1 \times d$ matrix whose values are corresponding to 1 or -1 and $A^+ = A^T \left(AA^T\right)^{-1}$. The location regeneration for the guarder is formulated by using Eq. (4):

$$X_{i,j}^{t+1} = \begin{cases} X_{best}^t + \beta \cdot \left|X_{i,j}^t - X_{best}^t\right|, & ft_j > ft_g \\ X_{i,j}^t + K \cdot \left(\dfrac{X_{i,i}^t - X_{worst}^t}{(f^{t_i} - f^{t_w}) + \varepsilon}\right), & ft_i = ft_g \end{cases} \tag{4}$$

where X_{best} represents the global optimum position. β and $K \in [-1, 1]$ characterize two arbitrary integers; ft_i shows the fitness value. ft_w and ft_g correspondingly denote the existing worst and optimum fitness value in the population, and ε shows the smallest number that is nearer to zero.

The SSA technique not only defines a fitness function to reach the highest performance of the classifier and also determines a positive value to indicate the better outcome of the solution candidate. The reduction of classification error rate is regarded as the fitness function, as given in Eq. (5).

$$\begin{aligned} \boldsymbol{fitness\ (x_i)} &= \boldsymbol{Classifier\ Error\ Rate\ (x_i)} \\ &= \frac{\boldsymbol{No.\ misclassified\ samples}}{\boldsymbol{Total\ no.\ samples}} * \mathbf{100} \end{aligned} \tag{5}$$

C. IMAGE CAPTIONING PROCESS

For the image captioning process, the AIC-SSAIDL technique makes use of the $AM - LSTM$ model. LSTM is a specific RNN architecture which is stable and effective for modelling long-term dependency in several prior researchers [20]. The LSTM is the cell state c_r that mechanism with three gate designs (forget, input, and output gates) for completing the accumulation of state data.

However, h_t implies the hidden layer (HL) at time t i.e., the resultant value of cell units A. χ_t denotes the input traffic flow order at t time, cell state was defined as c_t, $\oplus$ and $\otimes$ denotes the addition and multiplication of the matrix correspondingly, and arrow represents the conversion of matrices. For instance, arrows on h_r and χ_t imply that h_r and x_r are transformed once correspondingly, in other words, $WV_h \otimes h_r$ and $WV_\chi \otimes x_r$ are achieved, whereas WV_h and WV_χ denote the weighted matrixes of neural network (NN). σ wrapped by box denotes the sigmoid function, and two arrows derive collected to signify the vector addition. An offset vector b was executed whereas the two vectors are added. The input, output, and forget gates at t time are correspondingly formulated as i_r, 0_r and f_t, then the connection among input, output and forget gates, the input χ_t, HL h_r and cell state c_r is illustrated in Eqs. (6)-(8).

$$\begin{pmatrix} i_t \\ f_t \\ O_t \\ \tilde{C} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{pmatrix} WV \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} + \begin{pmatrix} b_i \\ b_f \\ b_0 \\ b_C \end{pmatrix} \tag{6}$$

$$c_r = f_t \otimes c_{r-1} + i_r \otimes \tilde{C} \tag{7}$$

$$h_r = 0_r \otimes \tanh c_r \tag{8}$$

whereas W denotes the weighted matrix from the HL and b refers to the offset vector.

Based on Eq. (6), i_t, 0_t, f_t, and $\tilde{C}$ are all created by simple NN functions. Because of the distinct activation functions Tanh and σ functions, the value range of parameters $\tilde{C}$ is $(-1, 1)$ and the value range of i_r, 0_r, f_t is [0, 1]. Based on Eq. (7), c_t was attained with the addition of cell states at the

Algorithm 1 Pseudocode of SSA

```
Begin
Determine $Iter_{\max}, NP, n, P_{dp}, sf, Gc, FS_U$ and $FS_L$
Initialize the flying squirrel location randomly
        $FS_{i,j} = FS_L + rand() * (FS_U - FS_L), i = 1, 2, \ldots, NP, j = 1, 2, \ldots, n$
Calculate fitness value
            $f_i = f_i(FS_{i,1}, FS_{i,2}, \ldots, \mathrm{FS}_{i,n}), i = 1, 2, \ldots, NP$
  While $Iter < Iter_{\max}$
    $[sorted - f, sorte - index] = sort(f)$
    $FS_{ht} = FS(sorte\_index(1))$
    $FS_{at}(1:3) = FS(sorte\_index(2:4))$
    $FS_{nt}(1:NP-4) = FS(sorte\_index(5:NP))$
    Generate novel location
    For $t = 1 : n1( n1 =$ total amount of squirrels on acorn tree)
      If $R_1 \geq P_{dp}$
        $FS_{at}^{new} = FS_{at}^{old} + d_g G_c(FS_{ht}^{old} - FS_{at}^{old})$
      Else
      $FS_{at}^{new} = random\ location$
      End
    End for
    For $t = 1 : n2( n2 =$ total amount of squirrels on a normal tree moving towards acorn tree)
      If $R_2 \geq P_{dp}$
        $FS_{nt}^{new} = FS_{nt}^{old} + d_g G_c(FS_{at}^{old} - FS_{nt}^{old})$
      Else
        $FS_{nt}^{new} = randomlocation$
      End
    end
    For $t = 1 : n3( n3 =$ total amount of squirrels on a normal tree moving towards hickory trees)
      If $R_3 \geq P_{dp}$
        $FS_{nt}^{new} = FS_{nt}^{old} + d_g G_c(FS_{ht}^{old} - FS_{nt}^{old})$
      Else
        $FS_{nt}^{new} = random\ location$
      End
    End
$S_c^t = \sqrt{\sum_{k-1}^{n}(FS_{atk}^t - FS_{htk})^2}, S_{c\min} = \frac{10B-6}{365^{Iter/(Iter_{\max})/2.5}}$
    If $s_c^t < s_{c\min}$
      $FS_{nt}^{new} = FS_L + Lévy(n) \times (FS_U - FS_L)$
    End
Calculate the fitness value of the novel location
$f_i = f_i(FS_{i,1}^{new}, FS_{i,2}^{new}, \ldots, FS_{i,n}^{new}), i = 1, 2, \ldots, NP$
$Iter = Iter + 1$
$End$
```

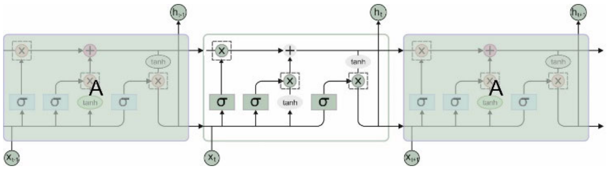

FIGURE 2. Framework of LSTM.

final moment c_{t-1} and C parameters. With the forgetting gate f_t and input gate i_r, the degree of accumulation is managed correspondingly, representative of the degree of cell states written and cleared. Based on Eq. (8), the output gate 0_t control the cell state c_r and defines the degree to that c_t is transferred to the last state h_t. One benefit of utilizing the memory cell and gates for controlling data flow and avoiding vanishing suddenly is a crucial issue for the RNN method. Fig. 2 represents the framework of LSTM.

The attention mechanism (AM) is originally utilized in machine translation that supports the network method allocating various weights for all the parts of the input, extracting very important and vital data, and creating the model with a more correct judgment, without taking the additional cost to computation and storing of the models. The data are predictive of the AM − LSTM method and are separated into subsequent stages.

(1) Compute the HL $h_i, i = 1, 2, \ldots, t$. Data $x_i, i = 1, 2, \ldots, t$ attain the HL output h_i with a typical LSTM network:

(2) Compute the distribution of attention $\alpha_i, i = 1, 2, \ldots, t$ as depicted in Eq. (9):

$$\alpha_i = soft\max\left(s\left(h_i, y\right)\right) = \frac{\exp\left(s\left(h_i, y\right)\right)}{\sum_{j=1}^{t}\exp\left(s\left(h_j, y\right)\right)} \quad (9)$$

whereas similarity was implemented as a scoring function of attention, i.e., $s\left(h_i, y\right) = h_i^T y$ and $s(h_i, y)$ implies the scoring function of attention ;

(3) Estimate the weighted average of characteristic value $h = \sum_{i=1}^{t} \alpha_i h_i$

(4) Determine the predictive value. h is input as FC and the output X_{t+1} implies the predictive value at $t + 1$ time.

It could be realized in the procedure that theAM − LSTM method chooses the feature data, rather than choosing only one of the t HLs, computing the weighted average of every t HL. Afterwards, the weighted data h is input as NN for the next computation. But the network parameter of

TABLE 1. Details of the dataset.

Dataset	Size	Reference
Flickr8k	8000	[22]
Flickr30k	31000	[23]
MSCOCO	164062	[24]

$AM - LSTM$ networks can be set after training, it is weighted $\alpha_1, \alpha_2, \ldots, \alpha_t$ and is modified with the change of input $x_1, x_2, \ldots, x_r$ under the testing. This allows the $AM - LSTM$ model for selecting suitable parameters to predict based on the alteration of input data and resolves the defect of the typical LSTM model.

At the final stage, the hyperparameter selection of the $AM - LSTM$ network is performed by the FFO algorithm. The FFO algorithm is a SI optimization technique stimulated by the foraging behaviours of the fruit fly (FF) [21]. A group of FFs firstly rely on the concentration of odour while finding food, to define the estimated distance between the target food and the FF itself. The general concept of the FFO in resolving the problems of searching for the optimum solution is the same as the FF search for food. The position data of the better FF can be defined using the smell-searching technique. But, the visual search technique can be distinct in that every FF in the search space is arbitrarily outwards from the position data of the optimum FF and later the smell and visual search phases are repeated until the optimum solution is attained. To further define the FFO technique, consider the optimization problems of binary function $g(x, y)$. The procedure of resolving the optimum solution of the binary operation by FFO is split into subsequent stages.

Step 1: Initialize the parameter of the FF swarm such as the initial location (x_0, y_0) of the FF swarm, N number of FFs in the swarm, the maximal amount $T_{\max}$ of iterations, and the searching step size L.

Step 2: Smell search procedure. The FF swarm begin from the first location (x_0, y_0) and random searches in each direction with step L for getting the updated location (x_i, y_i), $i = 1, 2, \ldots, N$. This process is equated in Eqs. (10)-(11).

$$x_i = x_0 + L \times Rand \tag{10}$$

$$y_i = y_0 + L \times Rand \tag{11}$$

From the expression, $Rand()$ denotes the randomly generated value of zero and one. Afterwards finishing the random search, the odour smells judgment value $Smell_i = g(x_i, y_i)$ of the existing location of the FF is evaluated, and g denotes the odour intensity function.

Step 3: Visual search process. The FF with better odour smell judgment value in the swarm is chosen as an optimum FF and the global optimum. Then, upgrade the value and the global optimum position data as soon as the FF is better than that of the global optimum value Smellbest. It is given in Eqs. (12)-(15).

$$[bestsmell, bestIndex] = \max\left(Smell_1, Smell_2, \cdots, Smell_N\right) \tag{12}$$

FIGURE 3. Sample images.

$$Smellbest = bestSmell \tag{13}$$

$$x_0 = bestIndex_x \tag{14}$$

$$y_0 = bestIndex_y \tag{15}$$

Step 4: Define the terminating criteria. If the maximal amount of iterations is attained, the process ends and outputs the optimum solution $(bestIndex_x, bestIndex_y)$ that results in the optimum value of a function (x, y), or else, return to step 2.

IV. RESULTS AND DISCUSSION

The proposed model is simulated using Python 3.6.5 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.

The image captioning results of the AIC-SSAIDL method can be tested on three datasets: Flickr8k, Flickr30k, and MSCOCO dataset as defined in Table 1. Fig. 3 represents the sample images. Flickr8k database is a novel benchmark gathered for sentence-based image description and search, containing 8,000 images which are all paired with 5 distinct captions that offer clear descriptions of salient entities and events. The Flickr30k database has developed a typical benchmark for sentence-based image description. This study proposes Flickr30k Entities that augment the 158k captions in Flickr30k with 244k co-reference chains, connecting mentions of similar entities across distinct captions for a similar image, and connecting them with 276k manually annotated bounding boxes. COCO is a large-scale object recognition, segmentation, and captioning database. COCO contains many features: Object segmentation, 330K images (>200K labelled), 1.5 million object samples, 91 stuff categories, 80 object categories, and 5 captions per image.

In Table 2, the overall image captioning results of the AIC-SSAIDL technique with recent models are made on the Flickr8k dataset [25], [26]. The experimental values portray the improvement of the AIC-SSAIDL technique. Fig. 4 provides a comparative investigation of the AIC-SSAIDL method in terms of BLUE on the Flickr8k dataset. The results

TABLE 2. Image captioning outcome of AIC-SSAIDL approach with other systems under the Flickr8k dataset.

Flickr8K Dataset						
Methods	BLUE-1	BLUE-2	BLUE-3	BLUE-4	METEOR	CIDEr
NIC [28]	58.25	43.26	32.28	18.36	14.44	31.32
Soft-Attention [25]	60.55	45.20	34.49	20.25	16.48	33.91
Hard-Attention [25]	62.69	46.61	36.48	23.21	18.13	36.67
SCA- CNN-VGG [27]	65.23	50.08	39.39	24.43	21.42	38.31
CNN Model [26]	67.50	51.97	41.43	26.61	24.03	41.24
AIC-SSAIDL	72.31	56.52	46.06	31.71	29.83	46.12

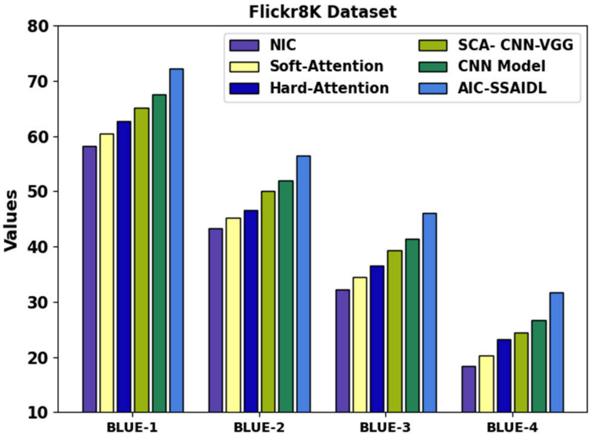

FIGURE 4. The BLUE outcome of AIC-SSAIDL approach under Flickr8k dataset.

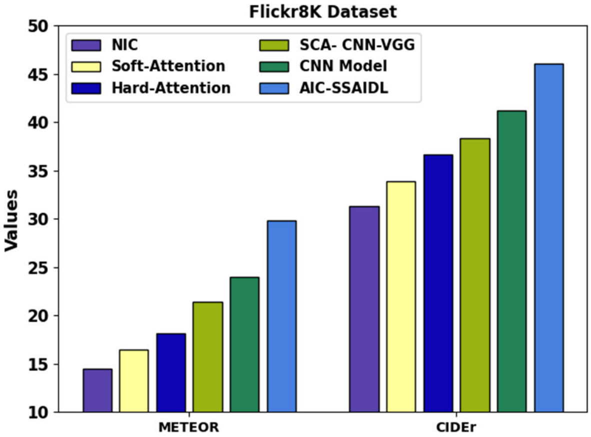

FIGURE 5. METEOR and CIDEr outcome of AIC-SSAIDL approach under Flickr8k dataset.

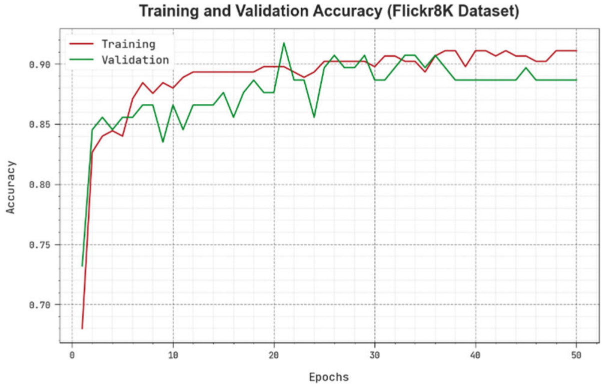

FIGURE 6. TACY and VACY outcome of AIC-SSAIDL approach under Flickr8k dataset.

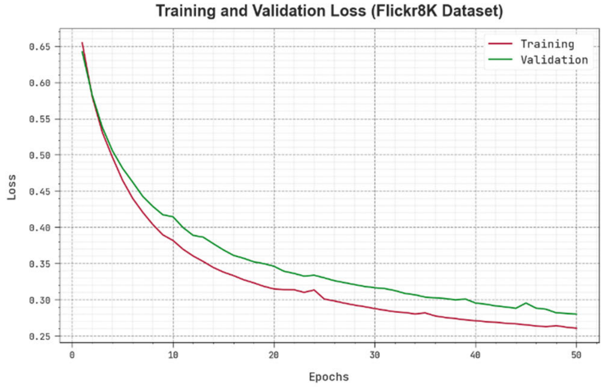

FIGURE 7. TLOS and VLOS outcome of AIC-SSAIDL approach under Flickr8k dataset.

indicate that NIC and soft-attention models reached lower BLUE values whereas certainly improvised BLUE values can be accomplished by the hard-attention and SCA-CNN-VGG models. Moreover, the CNN model reaches considerable outcomes with BLUE-1 of 67.50, BLUE-2 of 51.97, BLUE-3 of 41.13, and BLUE-4 of 26.61. Nevertheless, the AIC-SSAIDL technique reaches higher performance with BLUE-1 of 72.31, BLUE-2 of 56.52, BLUE-3 of 46.06, and BLUE-4 of 31.71.

A comparative METEOR and CIDEr examination of the AIC-SSAIDL technique on the Flickr8k dataset is given in Fig. 5. The experimental outcomes signify the enhanced performance of the AIC-SSAIDL technique. Based on METEOR, the AIC-SSAIDL technique reaches a higher METEOR of 29.83. Contrastingly, the NIC, soft-attention, hard-attention, SCA-CNN-VGG, and CNN models accomplish reducing METEOR of 14.44, 16.48, 18.13, 21.42, and 24.03, respectively. Likewise, based on CIDEr, the AIC-SSAIDL method obtains maximum CIDEr 46.12. Contrastingly, the NIC, soft-attention, hard-attention, SCA-CNN-VGG, and CNN techniques obtain minimum CIDEr of 31.32, 33.91, 36.67, 38.31, and 41.24, respectively.

The TACY and VACY of the AIC-SSAIDL method under the Flickr8k dataset are represented in Fig. 6. The figure inferred that the AIC-SSAIDL method has given superior performance with maximum values of TACY and VACY. Note that the AIC-SSAIDL model has obtained increased TACY outcomes

The TLOS and VLOS of the AIC-SSAIDL method under the Flickr8k dataset are represented in Fig. 7. The figure implied that the AIC-SSAIDL method has provided improved performance with minimum values of TLOS and VLOS. Note that the AIC-SSAIDL model has resulted in reduced VLOS outcomes.

In Table 3, the overall image captioning outcomes of the AIC-SSAIDL method with the recent method are

TABLE 3. Image captioning outcome of AIC-SSAIDL approach with other systems under the Flickr30k dataset.

Flickr30k Dataset						
Methods	BLUE-1	BLUE-2	BLUE-3	BLUE-4	METEOR	CIDEr
NIC [28]	58.98	48.53	38.92	27.79	21.96	37.69
Soft-Attention [25]	61.27	51.70	41.43	29.30	23.63	39.62
Hard-Attention [25]	63.10	53.82	44.13	31.25	26.46	42.54
SCA- CNN-VGG [27]	64.71	56.12	47.04	34.03	28.48	45.50
CNN Model [26]	67.70	57.56	49.40	36.29	28.72	56.51
AIC-SSAIDL	71.42	62.56	53.81	41.18	35.09	61.89

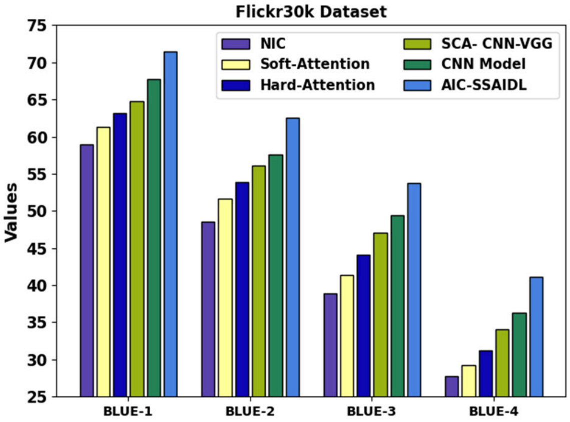

FIGURE 8. The BLUE outcome of AIC-SSAIDL approach under Flickr30k dataset.

made on the Flickr30k dataset. The experimental value portrays the improvement of the AIC-SSAIDL technique. Fig. 8 provides a comparative analysis of the AIC-SSAIDL technique in terms of BLUE on the Flickr30k dataset. The results indicate that NIC and soft-attention models reached lower BLUE values while certainly improvised BLUE values can be obtained by the hard-attention and SCA-CNN-VGG methods. Furthermore, the CNN model attains considerable outcomes with BLUE-1 of 67.70, BLUE-2 of 57.56, BLUE-3 of 49.40, and BLUE-4 of 36.29. Nevertheless, the AIC-SSAIDL technique reaches higher performance with BLUE-1 of 71.42, BLUE-2 of 62.56 BLUE-3 of 53.81, and BLUE-4 of 41.18.

A comparative METEOR and CIDEr examination of the AIC-SSAIDL method on the Flickr30k dataset is given in Fig. 9. The experimental outcomes signify the superior performance of the AIC-SSAIDL method. Based on METEOR, the AIC-SSAIDL technique reaches a higher METEOR of 35.09. Contrastingly, the NIC, soft-attention, hard-attention, SCA-CNN-VGG, and CNN models accomplish reducing METEOR of 21.96, 23.63, 26.46, 28.48, and 28.72, respectively. Similarly, based on CIDEr, the AIC-SSAIDL method attains higher CIDEr 61.89. Contrastingly, the NIC, soft-attention, hard-attention, SCA-CNN-VGG, and CNN models accomplish reducing CIDEr of 37.69, 39.62, 42.54, 45.50, and 56.51, respectively.

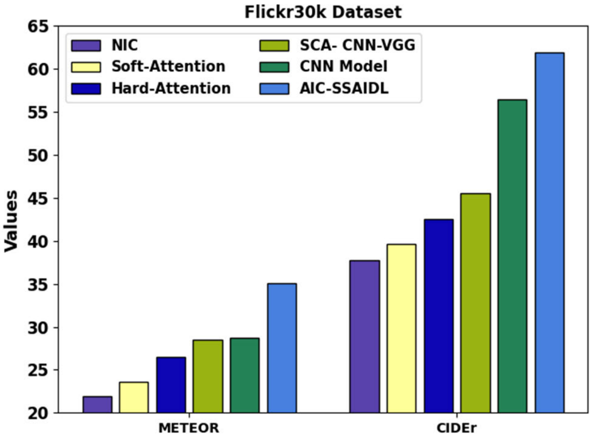

FIGURE 9. METEOR and CIDEr outcome of AIC-SSAIDL approach under Flickr30k dataset.

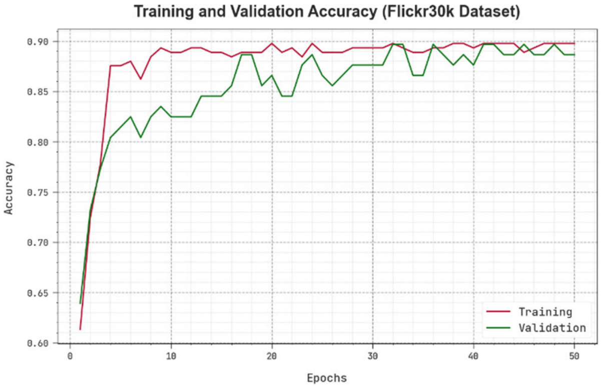

FIGURE 10. TACY and VACY outcome of AIC-SSAIDL approach under Flickr30k dataset.

The TACY and VACY of the AIC-SSAIDL method under the Flickr30k dataset are represented in Fig. 10. The figure inferred that the AIC-SSAIDL method has given superior performance with maximum values of TACY and VACY. Note that the AIC-SSAIDL model has obtained increased TACY outcomes.

The TLOS and VLOS of the AIC-SSAIDL method under the Flickr30k dataset are represented in Fig. 11. The figure implied that the AIC-SSAIDL method has proved superior performance with reduced values of TLOS and VLOS. Note that the AIC-SSAIDL model has resulted in the least VLOS outcomes.

In Table 4, the overall image captioning results of the AIC-SSAIDL method with recent models are made on the MSCOCO dataset. The experimental value portrays the improvement of the AIC-SSAIDL method. Fig. 12 provides a comparative analysis of the AIC-SSAIDL method in terms of BLUE on the MSCOCO dataset. The result indicates that NIC and soft-attention models attained minimum BLUE values whereas certainly improvised BLUE values can be obtained by the hard-attention and SCA-CNN-VGG methods. Moreover, the CNN method attains considerable outcomes with BLUE-1 of 76.01, BLUE-2 of 58.27, BLUE-3 of 42.64, and BLUE-4 of 32.68. Nevertheless, the AIC-SSAIDL method

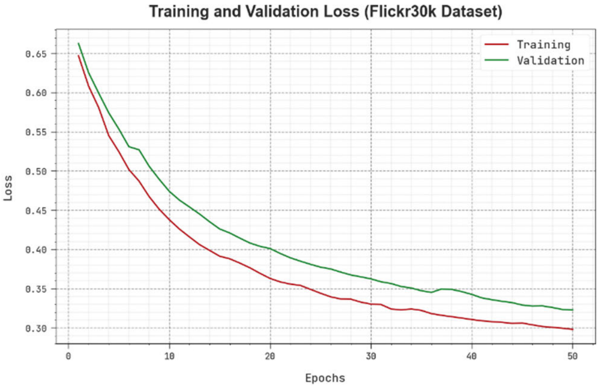

FIGURE 11. TLOS and VLOS outcome of AIC-SSAIDL approach under Flickr30k dataset.

TABLE 4. Image captioning outcome of the AIC-SSAIDL approach with other systems under the MSCOCO dataset.

MSCOCO Dataset						
Methods	**BLUE-1**	**BLUE-2**	**BLUE-3**	**BLUE-4**	**METEOR**	**CIDEr**
NIC [28]	62.35	48.37	33.56	23.04	19.35	68.93
Soft-Attention [25]	64.48	50.75	35.42	25.33	21.44	71.36
Hard-Attention [25]	66.95	52.15	37.43	28.23	24.23	88.85
SCA- CNN-VGG [27]	69.20	55.55	40.26	31.40	26.10	106.07
CNN Model [26]	76.01	58.27	42.64	32.68	29.44	117.47
AIC-SSAIDL	80.40	62.94	47.81	38.04	33.58	137.45

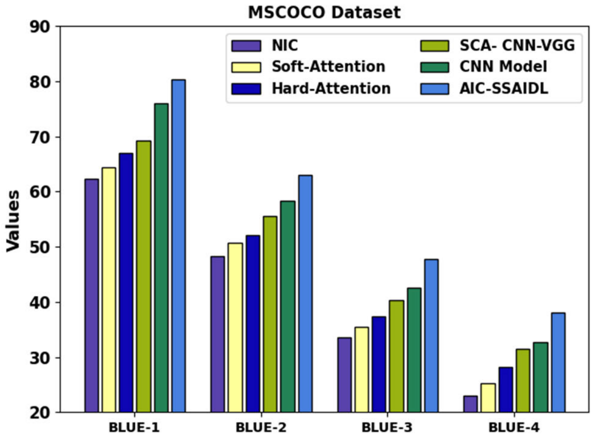

FIGURE 12. The BLUE outcome of AIC-SSAIDL approach under MSCOCO dataset.

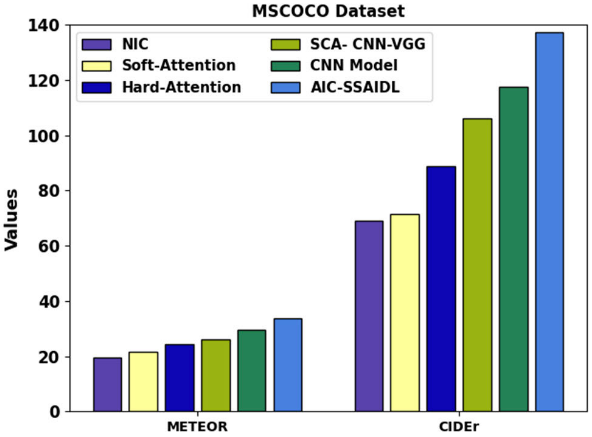

FIGURE 13. METEOR and CIDEr outcome of AIC-SSAIDL approach under MSCOCO dataset.

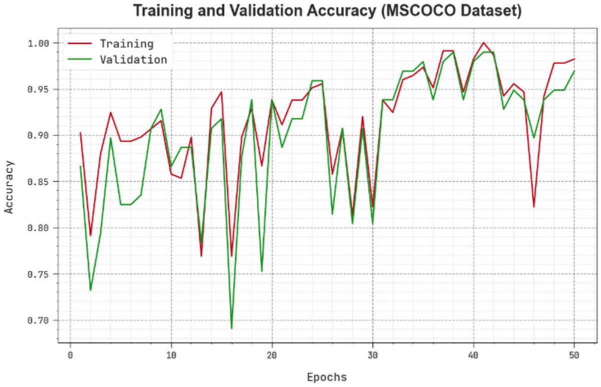

FIGURE 14. TACY and VACY outcome of AIC-SSAIDL approach under the MSCOCO dataset.

attains maximum performance with BLUE-1 of 80.40, BLUE-2 of 62.94 BLUE-3 of 47.81, and BLUE-4 of 38.04.

A comparative METEOR and CIDEr examination of the AIC-SSAIDL method on the MSCOCO dataset is illustrated in Fig. 13. The experimental outcomes signify the superior performance of the AIC-SSAIDL method. Based on METEOR, the AIC-SSAIDL technique reaches a higher METEOR of 33.58. Contrastingly, the NIC, soft-attention, hard-attention, SCA-CNN-VGG, and CNN methods attain reducing METEOR of 19.35, 21.44, 24.23, 26.10, and 29.44, respectively. Likewise, based on CIDEr, the AIC-SSAIDL method attains maximum CIDEr of 137.45. Contrastingly, the NIC, soft-attention, hard-attention, SCA-CNN-VGG, and CNN models accomplish reducing CIDEr of 68.93, 71.36, 88.85, 106.07, and 117.47, respectively.

The TACY and VACY of the AIC-SSAIDL method under the MSCOCO dataset are represented in Fig. 14. The figure inferred that the AIC-SSAIDL method has given enhanced performance with improved values of TACY and VACY. Note that the AIC-SSAIDL method has obtained maximum TACY outcomes.

The TLOS and VLOS of the AIC-SSAIDL system under the MSCOCO dataset are represented in Fig. 15. The figure implied that the AIC-SSAIDL method has proved improved performance with the lowest values of TLOS and VLOS. Note that the AIC-SSAIDL method has resulted in minimum VLOS outcomes. The above-mentioned results highlighted the improved image captioning performance of the AIC-SSAIDL technique.

V. CONCLUSION

In this study, a new AIC-SSAIDL method was introduced for the automated generation of textual captions for the input images. At the initial stage, the AIC-SSAIDL technique utilizes MobileNetv2 model-generated feature descriptors of the input images and its hyperparameter tuning process gets

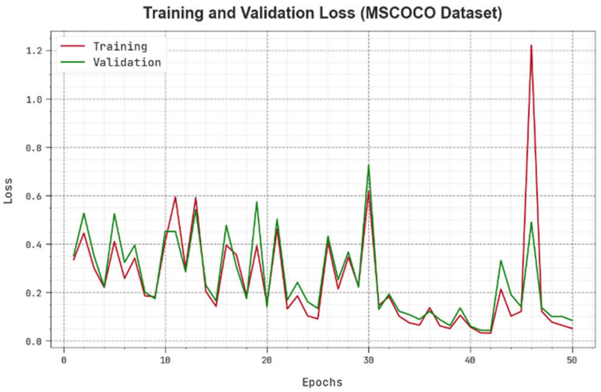

FIGURE 15. TLOS and VLOS outcome of AIC-SSAIDL approach under the MSCOCO dataset.

executed by the SSA. For the image captioning process, the AIC-SSAIDL technique exploits the AM-LSTM network. Finally, the hyperparameter selection of the AM-LSTM network is performed by the FFO algorithm. A series of simulations have been conducted on benchmark datasets to demonstrate the better performance of the AIC-SSAIDL technique. The extensive result analysis highlighted the enhanced captioning results of the AIC-SSAIDL method in terms of different evaluation measures. In the future, the performance of the AIC-SSAIDL algorithm can be boosted by the ensemble fusion process. In addition, the computation complexity of the proposed model can be examined in future. Moreover, the image captioning performance of the proposed model can be tested on large-scale datasets.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups Project under grant number (RGP2/61/44). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R237), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Research Supporting Project number (RSPD2023R608), 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] A. Singh, J. K. Raguru, G. Prasad, S. Chauhan, P. K. Tiwari, A. Zaguia, and M. A. Ullah, "Medical image captioning using optimized deep learning model," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–9, Mar. 2022.

[2] H. Kwon and S. Lee, "Toward backdoor attacks for image captioning model in deep neural networks," *Secur. Commun. Netw.*, vol. 2022, pp. 1–10, Aug. 2022.

[3] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 539–559, Jan. 2023.

[4] K. P. Deorukhkar and S. Ket, "Image captioning using hybrid LSTM-RNN with deep features," *Sens. Imag.*, vol. 23, no. 1, p. 31, Dec. 2022.

[5] A. Elhagry and K. Kadaoui, "A thorough review on recent deep learning methodologies for image captioning," 2021, *arXiv:2107.13114*.

[6] T. Tiwary and R. P. Mahapatra, "An accurate generation of image captions for blind people using an extended convolutional atom neural network," *Multimedia Tools Appl.*, vol. 82, pp. 3801–3830, Jul. 2022.

[7] A.-A. Liu, Y. Zhai, N. Xu, W. Nie, W. Li, and Y. Zhang, "Region-aware image captioning via interaction learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3685–3696, Jun. 2022.

[8] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, Nov. 2019.

[9] R. Castro, I. Pineda, W. Lim, and M. E. Morocho-Cayamcela, "Deep learning approaches based on transformer architectures for image captioning tasks," *IEEE Access*, vol. 10, pp. 33679–33694, 2022.

[10] Y. Zhang, X. Shi, S. Mi, and X. Yang, "Image captioning with transformer and knowledge graph," *Pattern Recognit. Lett.*, vol. 143, pp. 43–49, Mar. 2021.

[11] 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.

[12] C. P. Chaudhari and S. Devane, "Improved framework using rider optimization algorithm for precise image caption generation," *Int. J. Image Graph.*, vol. 22, no. 2, Apr. 2022, Art. no. 2250021.

[13] Y. Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang, "Automatic image captioning based on ResNet50 and LSTM with soft attention," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–7, Oct. 2020.

[14] 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.

[15] X. Shen, B. Liu, Y. Zhou, J. Zhao, and M. Liu, "Remote sensing image captioning via variational autoencoder and reinforcement learning," *Knowl.-Based Syst.*, vol. 203, Sep. 2020, Art. no. 105920.

[16] 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.

[17] 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.

[18] R. Indraswari, R. Rokhana, and W. Herulambang, "Melanoma image classification based on MobileNetV2 network," *Proc. Comput. Sci.*, vol. 197, pp. 198–207, Jan. 2022.

[19] J. Yuan, Z. Zhao, Y. Liu, B. He, L. Wang, B. Xie, and Y. Gao, "DMPPT control of photovoltaic microgrid based on improved sparrow search algorithm," *IEEE Access*, vol. 9, pp. 16623–16629, 2021.

[20] W. Fang, W. Zhuo, J. Yan, Y. Song, D. Jiang, and T. Zhou, "Attention meets long short-term memory: A deep learning network for traffic flow forecasting," *Phys. A, Stat. Mech. Appl.*, vol. 587, Feb. 2022, Art. no. 126485.

[21] Q. Zhang, C. Li, C. Yin, H. Zhang, and F. Su, "A hybrid framework model based on wavelet neural network with improved fruit fly optimization algorithm for traffic flow prediction," *Symmetry*, vol. 14, no. 7, p. 1333, Jun. 2022.

[22] Accessed: Apr. 14, 2023. [Online]. Available: https://www.kaggle.com/datasets/adityajn105/flickr8k

[23] Accessed: Apr. 2, 2023. [Online]. Available: https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset

[24] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, 2014, September, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2014, pp. 740–755.

[25] M. Omri, S. Abdel-Khalek, E. M. Khalil, J. Bouslimi, and G. P. Joshi, "Modeling of hyperparameter tuned deep learning model for automated image captioning," *Mathematics*, vol. 10, no. 3, p. 288, Jan. 2022, doi: 10.3390/math10030288.

[26] 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.

[27] 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. 5659–5667.

[28] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.