Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.035529
Article

Tech Science Press

Red Deer Optimization with Artificial Intelligence Enabled Image Captioning System for Visually Impaired People

Anwer Mustafa Hilal[1,*], Fadwa Alrowais[2], Fahd N. Al-Wesabi[3] and Radwa Marzouk[4,5]

[1]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia
[2]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O.Box 84428, Riyadh, 11671, Saudi Arabia
[3]Department of Computer Science, College of Science & Art at Mahayil, King Khalid University, Mahayil, Saudi Arabia
[4]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O.Box 84428, Riyadh, 11671, Saudi Arabia
[5]Department of Mathematics, Faculty of Science, Cairo University, Giza, 12613, Egypt
*Corresponding Author: Anwer Mustafa Hilal. Email: a.hilal@psau.edu.sa

Received: 24 August 2022; Accepted: 14 December 2022

Abstract: The problem of producing a natural language description of an image for describing the visual content has gained more attention in natural language processing (NLP) and computer vision (CV). It can be driven by applications like image retrieval or indexing, virtual assistants, image understanding, and support of visually impaired people (VIP). Though the VIP uses other senses, touch and hearing, for recognizing objects and events, the quality of life of those persons is lower than the standard level. Automatic Image captioning generates captions that will be read loudly to the VIP, thereby realizing matters happening around them. This article introduces a Red Deer Optimization with Artificial Intelligence Enabled Image Captioning System (RDOAI-ICS) for Visually Impaired People. The presented RDOAI-ICS technique aids in generating image captions for VIPs. The presented RDOAI-ICS technique utilizes a neural architectural search network (NASNet) model to produce image representations. Besides, the RDOAI-ICS technique uses the radial basis function neural network (RBFNN) method to generate a textual description. To enhance the performance of the RDOAI-ICS method, the parameter optimization process takes place using the RDO algorithm for NasNet and the butterfly optimization algorithm (BOA) for the RBFNN model, showing the novelty of the work. The experimental evaluation of the RDOAI-ICS method can be tested using a benchmark dataset. The outcomes show the enhancements of the RDOAI-ICS method over other recent Image captioning approaches.

Keywords: Machine learning; image captioning; visually impaired people; parameter tuning; artificial intelligence; metaheuristics

1 Introduction

Image Captioning as a service has assisted persons with visual disabilities in studying the images they take and making sense of images they encounter in digital atmospheres [1]. Applications enable visually impaired people (VIPs) to take images of their environments and upload them to descriptions of that images [2]. These applications use a human-in-the-loop technique for generating descriptions. To avoid dependence on a human, there comes a necessity for automating the Image captioning procedure. Inappropriately, the present Image captioning methods were built utilizing crowdsourced, large, publicly available datasets that were accumulated and created in a contrived setting [3]. Therefore, such methods execute poorly on images snapped by VIPs largely due to the images snapped by visually impaired persons varying dramatically from the images presented in the data [4].

Image captioning mostly includes computer vision (CV) and natural language processing (NLP). CV is helpful in understanding and recognizing the condition in an image; NLP will convert semantic knowledge into a descriptive line [5]. Retrieving the semantic substance of a photo and interacting with it in a structure humans can recognize is extremely complicated. The Image captioning technique not just grants data but also reveals the connection between the substances. Image captioning consists of several applications—for instance, as an aid advanced to guide an individual with visual disabilities when travelling alone [6]. This can be made possible by varying the scenario into text and converting text into voice messages. Image captioning was even using mass interaction for automated generation of the image caption, which can be posted or to describe a video [7]. Furthermore, automatic Image captioning may foster the Google image search approach by varying the Image to a caption and, after, by leveraging the keywords for further relevant searches.

In the conventional machine learning (ML) technique, Input data can be employed for feature extraction. ML technique was utilized for Image captioning; however, it is not highly effective [8]. The reason can be deriving handcrafted features, namely SIFT (Scale-invariant feature transform), LBPs (Local Binary Pattern), and HOG (Histogram of oriented gradients), from huge data was not so easy and feasible [9]. Since realistic images were highly complicated and data is much diversified. So, the ML algorithm for Image captioning was not a highly efficient technique. But in the past 6–7 years, several deep learning (DL) papers were published for Image captioning. Several techniques were modelled utilizing the DL structure [10].

This article introduces a Red Deer Optimization with Artificial Intelligence Enabled Image Captioning System (RDOAI-ICS) for VIPs. The presented RDOAI-ICS technique aids in generating image captions for VIPs. The presented RDOAI-ICS technique utilizes a neural architectural search network (NASNet) model to produce image representations. Besides, the RDOAI-ICS technique uses the radial basis function neural network (RBFNN) method to generate a textual description. To enhance the performance of the RDOAI-ICS method, the parameter optimization procedure takes place using the RDO algorithm for NasNet and the butterfly optimization algorithm (BOA) for the RBFNN method. The experimental evaluation of the RDOAI-ICS method is tested using a benchmark dataset.

2 Related Works

In [11], an improvised image captioning method, which includes Image captioning, object detection, and color analysis, was modelled to produce textual descriptions of imageries mechanically. In an encoder-decoder method for Image captioning, the visual geometry group (VGG16) was utilized as a long short-term memory (LSTM) network with attention. An encoder can be employed as a decoder. Moreover, Mask region-based convolution neural network (CNN), including OpenCV, was employed for colour analysis and object detection. The incorporation of color recognition and Image captioning can be executed to offer superior descriptive image details. Besides, the generated textual sentence can be transformed into speech. Bhalekar et al. [12] devise an image captioning mechanism that produces comprehensive captions, derives text from imagery, if any, and utilizes it as a part of the caption to offer a highly accurate description of the imagery. The devised method will use LSTM and CNNs to extract the image features to produce respective sentences related to the learned image features. Additionally, utilizing text-extracting components, the derived text was encompassed in the image description, and the captions can be offered in audio form.

In [13], the authors use a fusion of visual data and high-level semantic data for Image captioning. The authors devise a hierarchical deep neural network (DNN) with top and bottom layers. The former will extract the Image's visual and high-level semantic data and detect areas correspondingly. In contrast, the latter compiles both having adaptive attention systems for caption generation. Rane et al. [14] devised a new implementation of smart spectacles related to Optical Character Recognition (OCR) and Image Captioning to ease navigation. The mechanism has a camera that can be entrenched in spectacles, an Image Captioning component, Text-To-Speech (TTS) module, and an OCR module. The Image captioning module assists in identifying signboards or notices in the immediate vicinity, whereas the OCR element will help read text on the sign boards. Text-To-Speech was employed chiefly for translating this data into a format that can help to visually impaired, like voice notifications.

Kim et al. [15] devised a technique for generating multiple captions utilizing a variational autoencoder (VAE), the single generative method. Due to an image feature serving a significant role while producing captions, a technique for extracting an image's Caption Attention Maps (CAMs) was modelled. CAM can be anticipated to be a latent distribution. In [16], the authors implemented and designed a cost-effective academic assistance cane, predominantly for blind individuals, related to an edge-cloud collaboration, scheme CV, and sensors. The authors have also devised an object detection and image captioning function with high-speed processing ability related to an edge-cloud collaboration technique to enhance the user experience. In [17], the authors offer an end-wise method that considers RNN as the decoder and deep CNN as the encoder. For superior image captioning extraction, the authors devise a highly modularized multi-branch CNN that rises when preserving the number of hyper-parameters unchanged. This technique offers a devised network with parallel sub-modules of the same structures.

Though several models are existed in the literature to carry out the Image captioning process, it is still desirable to improve the classifier outcomes. Owing to the continual deepening of the model, the number of parameters of DL models also surges quickly, which results in model overfitting. At the same time, different hyperparameters significantly impact the efficiency of the CNN model. Since the trial and error method for hyperparameter tuning is tedious and erroneous, metaheuristic algorithms can be applied. Therefore, in this work, the RDO algorithm is employed for the parameter selection of the NASNet model.

3 The Proposed Model

In this study, a new RDOAI-ICS technique has been developed for VIP image caption generation. The presented RDOAI-ICS technique aids in generating image captions for VIPs. Fig. 1 demonstrates the overall working process of the RDOAI-ICS system. As shown in the figure, the input data is initially preprocessed in various forms to make it compatible. In addition, the features are extracted by the optimal NASNet model in which the hyperparameter tuning process takes place by the RDO algorithm. Finally, the RBF model performs the classification process.

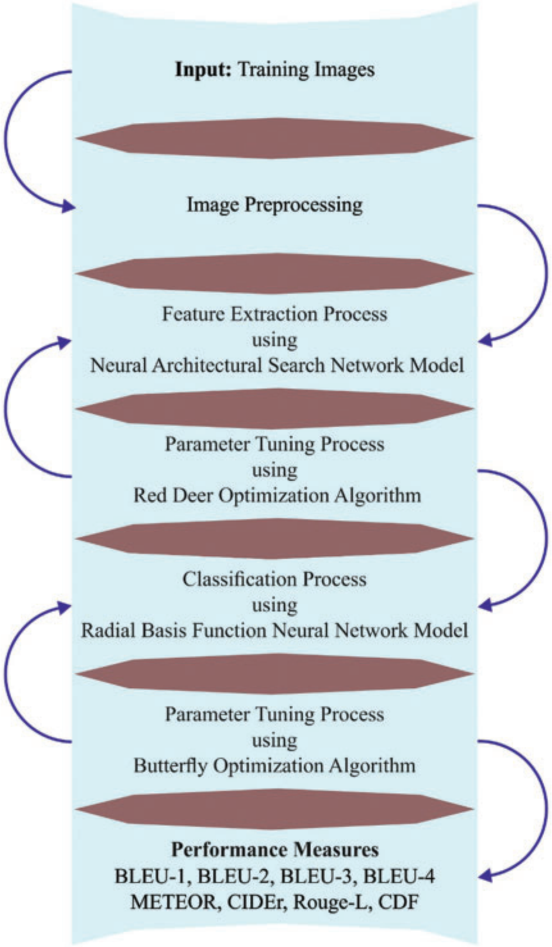
Input: Training Images
Image Preprocessing
Feature Extraction Process
using
Neural Architectural Search Network Model
Parameter Tuning Process
using
Red Deer Optimization Algorithm
Classification Process
using
Radial Basis Function Neural Network Model
Parameter Tuning Process
using
Butterfly Optimization Algorithm
Performance Measures
BLEU-1, BLEU-2, BLEU-3, BLEU-4
METEOR, CIDEr, Rouge-L, CDF

Figure 1: Overall process of RDOAI-ICS system

3.1 Pre-Processing

In the initial stage, data preprocessing can be executed in various stages.

- Numerical values are removed from the input text.
- Every character is transformed into lower case character
- Perform tokenization
- Removal of punctuation marks to decrease complexity;
- Vectorization (for turning the original strings into integer series in which every integer denotes the index of a word in a glossary).

3.2 Feature Extraction Using Optimal NASNet

The presented RDOAI-ICS technique used the NASNet model to produce image representations. Researchers analyzed a technique for directly learning structured models on the data. They applied the search technique for finding the cell or better convolution layer on CIFAR-10 and later used that cell to ImageNet by stacking several copies [18].

The cell employed in this architecture is called normal and reduction cells. Normal ones are convolution cells that return feature maps of a similar dimension. In contrast, reduction cells were convolution cells that return feature maps with a two-fold reduction in feature map height and width. The controller RNN is applied to search the architecture of normal and reduction cells.

The reduction and normal cells exist on the three presented NASNet architectures, such as NASNet-A, NASNet-B, and NASNet-C. For the ImageNet data, the test results were from 2.4% to 3.59% errors in prediction. The NASNet Mobile is the version of NASNet that is trained with ImageNet data. The study used NASNet Mobile, where the final global average pooling has shape $(1, 1056)$.

The hyperparameter tuning of the NASNet method takes place by the RDA. The stages in discovering the better solution in RDA were discussed in the following [19]: A random population of size N_{p0p} can be firstly produced for representing the *RDs*. A 1 xN array of RD is given below:

$$RD = [R_1,\ R_2,\ R_3 \ldots,\ R_N] \tag{1}$$

where R represent a solution to the problem.

The better RD from the population is grouped as male D, N_{male} and the residual population was determined as hinds (female RD), N_{hind}:

$$N_{hind} = N_{pop} - N_{male} \tag{2}$$

At the beginning of roaring amongst male RDs stag, male RD can rise their attractiveness by roaring. Then, compared fitness values among male RDs and adjacent male RDs, and if the latter has the better fitness values, its location in the solution is upgraded.

Male RD exudes dissimilar features reliant on their capability to mate, roar, and attract hinds. For that, only r percent of the best male is carefully chosen as male commander, N_{Com}. The remaining male RDs are represented as stags, N_{stag} as follows:

$$N_{Com} = round\{\gamma^{*}N_{male}\} \tag{3}$$

$$N_{stag} = N_{male} - N_{Com} \tag{4}$$

Every commander randomly approaches each stag for a one-to-one battle. Afterwards, all fight, two novel solutions (routes) were produced alongside 2 primary ones that pertained to the stag and commander. The best fitness values replace the commander in the solution. In other words, the fighting procedure allows better male RDs to be selected as commanders. The multiple hinds allotted to the harem were proportionate to the commander's power, which can be described using the objective function. In hindsight, the best fitness values of the commander, more hinds would be under their control as follows:

$$V_n = v_n - \max_i^{N_{Com}}\{v_i\} \tag{5}$$

In Eq. (5), v_n indicates the objective function of *the* n^{th} commander, and V_n illustrates the normalized value, which is given below:

$$P_n = \left| \frac{V_n}{\sum_{i=1}^{Ncom} V_{\iota}} \right| \tag{6}$$

The number of hinds for each harem possessed by all the commanders can be evaluated using the following expression:

$$N \cdot harerm_n = round\left\{P_n^* N_{hind}\right\} \tag{7}$$

In Eq. (10), $N \cdot harerm_n$ characterizes the number of hinds, N_{hind}, in the n^{th} harem.

The mating procedure was significant in differentiating an RD population. Later, the commander and stags mate with hinds.

- Mating of commander having α percentage of hinds in his harem

Every commander arbitrarily mates with α percentage of hinds in its harem as follows:

$$N \cdot harerm_n^{mate} = round\ \{\alpha^* N \cdot harern_n\} \tag{8}$$

Whereas $N \cdot harern_n^{mate}$ indicates the number of hinds chosen for mating in all the harems, n.

- Mating of commander with β percentage of hinds in other harems

Next, every commander of the harem mates with β percentage of hinds from arbitrarily selected harem k, except its harem.

$$N \cdot harem_k^{mate} = round\{\beta^*\ N \cdot harem_n\} \tag{9}$$

In Eq. (9), $N \cdot harem_k^{mate}$ signifies the number of hinds in the k^{th} harem.

The significance of this mating stage was for the commander to raise the size of its territory.

- Mating of a stag with the nearby hind

Finally, every stag mate with the hind found in the neighbourhood. The distance among all the stags and *i-th* hinds is evaluated as follows, and hinds nearby the stag can be selected for the mating procedure.

$$d_i = \sqrt{\sum\nolimits_{J \in J} \left(stag_j - hind_J^i\right)^2} \tag{10}$$

The distance between every stag and *i-th* hinds is represented by d_i.

The upcoming generation of RD is later chosen in two ways. Initially, a percentage of the better solution is chosen as the male RD that involves the commander and stag. Next, the remaining population, viz., hinds, are selected from the population of hinds and offspring produced formerly through the roulette wheel selection approach. Fig. 2 illustrates the steps involved in the RDO technique.

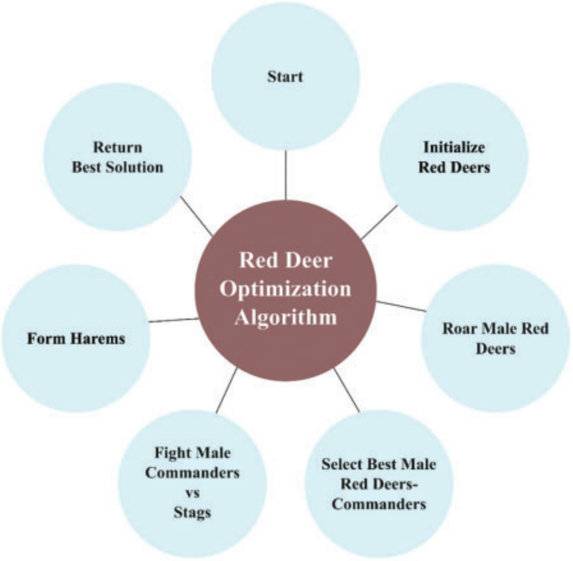
Start
Initialize Red Deers
Roar Male Red Deers
Select Best Male Red Deers-Commanders
Fight Male Commanders vs Stags
Form Harems
Return Best Solution
Red Deer Optimization Algorithm

Figure 2: Steps involved in the RDO technique

3.3 Image Captioning Using RBFNN Model

At this stage, the RDOAI-ICS technique exploited the RBFNN model to generate textual descriptions. RBFNN is a specific kind of feedforward network that exactly uses a single hidden layer (HL) [20]. The presented method is used increasingly in regression and classification problems. The presented method aims to convert information from a non-linear to a linear format before implementing the classification. The radial basis function (RBF) raises the dimension of the feature vector and implements classification via converting *the* d-dimension feature vector to the f-dimension feature vector if $f > d$. The RBFNN is an adapted artificial neural network (ANN) that applies RBF in HL. A neuron that implies a cluster in the HL is created in the learning process and is represented by the RBF as an activation function. There are three types of RBFNN, namely Quadratic, inverse, and Gaussian. The basic concerns over using RBFNN are the width of clusters and the determination of the centroids in the HL. The structure of RBF is comprised of output, input, and HLs. RBFNN has a single HL, and it is represented as a feature vector. The RBF is used for the formation of HL. The major feature of RBF is that the membership values of the pattern subsequently decrease or increase with decreasing or increasing in the distance from a centroid. The RBF offers improved accuracy with fast convergence for dense data. RBF is used in linear and non-linear modules. A cluster can be formed in the HL, also known as neurons or nodes. The number of clusters formed is characterized as $H_1, H_2, \ldots, H_j$. Each cluster signifies a set of corresponding class datasets. The k-th cluster H_k is demonstrated as $H_k = \left[c_{k_1},\ c_{k_2},\ \ldots,\ c_{k_n}\right]$.

The subsequent step shows the formation of a cluster in RBFNN:

- Step 1: The input layer receives n-dimension input $X = [x_1,\ x_2,\ \ldots,\ x_n]$ and is forwarded to the hidden layer.
- Step 2: Output of the hidden layer uses the Gaussian function. Where σ refers to the width of a cluster as follows.

$$\psi = exp\left(\frac{\Sigma_{k=0}^{n}(X_i - C_{ij})^2}{2 * \sigma^2}\right) \tag{11}$$

- Step 3: The gradient descent algorithm is utilized in the determination of weights between output and hidden layers. W_{ij} characterizes the weight between *i-th* HLs and *j-th* output class layers.
- Step 4: The output layer allocates input to a specific cluster, and the output of *i-th* nodes for m classes can be determined as follows.

$$y_i = f\left(\sum_{j=0}^{J} W_{ij} * \psi_j\right) \tag{12}$$

Whereas $= 1, 2, \ldots, m$.

3.4 Parameter Tuning Using BOA

To enhance the performance of the RDOAI-ICS model, the parameter optimization process takes place using BOA for the RBFNN model. In the BOA approach, every butterfly in the population is regarded as an independent search individual that discharges a specific concentration of fragrance [21]. Consequently, motion fitness changes when a single butterfly moves from one position to a novel position during the search process and spreads the fragrance during the movement. Every butterfly perceived the fragrance of another butterfly; however, the scent gradually decayed with distance, and the butterfly moved to the position with the stronger fragrance. This is the major difference between the BOA and other metaheuristics.

In the BOA, the value of fragrance concentration f is defined regarding three variables: the power index (a), the sensory modality (c), and the stimulus intensity (I). The perceptual morphology is the butterfly perception of the fragrance, i.e., initialization constant, and is generally utilized as an optimization variable. The stimulus intensity (I) is derived from the fitness function. Furthermore, the power index (a) is a constant that differs within the interval of zero and one, and it is formulated in the following equation:

$$f = cI^a \tag{13}$$

The key phases of BOA are briefly discussed in the following:

(a) Initialization phase, the objective function can be determined based on the time jerk requirement of the trajectory planning. Next, the trajectory is enhanced by modifying the control point of the NURBS. Therefore, it is needed to predetermine the initial population, sensory modality, power index, and switching probability and to compute the respective fitness value.

(b) Iteration stage, the location of the butterfly in the solution space, is redistributed, and thus the fragrance and fitness value of every butterfly must be re-estimated. In this phase, performing a global or local search is essential. In the global search, the BF moves to the butterfly g^* with the maximum fragrance value and the mathematical expression are given below:

$$x_i^{t+1} = x_i^t + \left(r^2 \times g^* - x_i^t\right) \times f_i \tag{14}$$

In Eq. (14), x_i^{t+1} and x_i^t indicate the solution for the *i-th* butterflies in iterations $t + 1$ and t, and r shows a random value that lies in $[0, 1]$. In this work, the global search of the JADE-GL tuned bbf technique is employed to implement a large global search, and it is formulated in the following equation:

$$x_i^{t+1} = x_i^t + \left(r\left(g_i^* - x_i^t\right) + (1 - r)\left(x_{r1}^t - x_{r2}^t\right)\right) \times f_i \tag{15}$$

In Eq. (15), $r1, r2$ indicate arbitrary numbers within 1. The population size n, x_i^t represents the solution respective to the *i-th* butterflies in *the t-th* iteration, g^* represents the optimum solution for the current iteration, f_i designates the fragrance produced by the *i-th* butterflies. R indicates a random value within [0, 1].

Once a butterfly could not sense the fragrance discharged by another butterfly, it takes a random walk and it is evaluated as follows:

$$x_i^{t+1} = x_i^t + \left(r^2 \times x_j^t - x_k^t\right) \times f_i \tag{16}$$

In Eq. (16), correspondingly, x_i^t and x_k^t denote the jth and kth solutions respective to the *t-th* iterations. Furthermore, r indicates a random value within [0, 1]. A switching probability p is used to switch between local and global searches.

(c) Termination phase, where the optimum solution is attained. This condition is accomplished once the number of iterations reaches a certain value or the attained solution could fulfil the requirement.

4 Results and Discussion

The proposed model is simulated using Python 3.6.5 tool on PC i5-8600k, GeForce 1050Ti 4 GB, 16 GB RAM, 250 GB SSD, and 1TB HDD. The parameter settings are learning rate: 0.01, dropout: 0.5, batch size: 5, epoch count: 50, and activation: ReLU.

4.1 Dataset Details

This section inspects the experimental validation of the RDOAI-ICS model on two datasets, namely Flickr8K and MS COCO 2014 datasets. Fig. 3 depicts some sample images with image captions. Flickr8K dataset is a benchmark collection for sentence-based image description and search, consisting of 8,000 images paired with five different captions that provide clear descriptions of the salient entities and events. COCO is a large-scale object detection, segmentation, and captioning dataset.

Crowd watching air ballons at night

A man climbs a rocky wall

Figure 3: Sample images with captions

4.2 Result Analysis

The Image captioning outcomes of the RDOAI-ICS model on the Flickr8K dataset is given in Table 1 and Fig. 4 [22]. The experimental outcomes revealed that the RDOAI-ICS model had outperformed other Image captioning approaches under all measures. For instance, on bilingual evaluation understudy (BLEU)-1 metric, the RDOAI-ICS model has exhibited a higher BLEU-1 of 69.86%. In contrast, the modified recurrent neural network (MRNN), GNICG, ResNet-50, k-nearest neighbor (KNN), VGG-16, hyperparameter tuned DL (HPTDL), and MODLE-AICT models have attained lower BLEU-1 of 58.66%, 65.84%, 66.69%, 59.08%, 63.43%, 67.66%, and 67.78% respectively. Similarly, on the BLUE-2 metric, the RDOAI-ICS method has manifested a higher BLEU-2 of 48% whereas the MRNN, GNICG, ResNet-50, KNN, VGG-16, HPTDL, and MODLE-AICT methodologies have acquired lower BLEU-2 of 31.49%, 43.99%, 44.96%, 36.85%, 44.81%, 43.23%, and 45.83% correspondingly. Concurrently, on the BLUE-3 metric, the RDOAI-ICS method has shown a higher BLEU-3 of 40.78% whereas the MRNN, GNICG, ResNet-50, KNN, VGG-16, HPTDL, and MODLE-AICT techniques have achieved lower BLEU-3 of 24.01%, 26.61%, 28.99%, 24.19%, 36.67%, 35.57%, and 39.15% correspondingly.

Table 1: Result analysis of RDOAI-ICS approach with existing algorithm under Flickr8K dataset [22]

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MRNN algorithm	58.66	31.49	24.01	15.68
GNICG algorithm	65.84	43.99	26.61	16.06
ResNet-50 algorithm	66.69	44.96	28.99	17.08
KNN algorithm	59.08	36.85	24.19	16.07
VGG-16 algorithm	63.43	44.81	36.67	25.27
HPTDL algorithm	67.66	43.23	35.57	22.13
MODLE-AICT algorithm	67.48	45.83	39.15	26.05
RDOAI-ICS	69.86	48.00	40.78	28.50

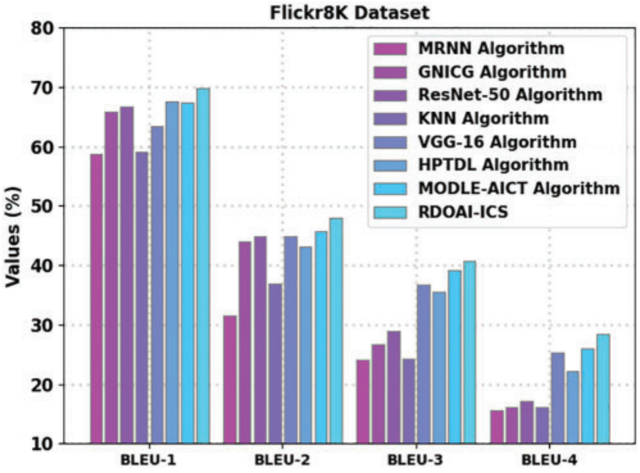

Figure 4: Result analysis of RDOAI-ICS approach under Flickr8K dataset

A comparative analysis of the RDOAI-ICS model on the Flickr8K dataset is given in Table 2 and Fig. 5 [22]. The experimental outcomes displayed by the RDOAI-ICS algorithm have exhibited other Image captioning techniques in all measures. For example, the RDOAI-ICS approach has manifested a higher METEOR of 30.03% on METEOR. In contrast, the SCST-IN, SCST-ALL, GNIC, Dense CNN, HPTDL, and MODLE-AICT techniques have gained lower METEOR of 18.94%, 24.12%, 20.52%, 19.65%, 27.29%, 28.46% correspondingly. Also, on CIDEr, the RDOAI-ICS method has illustrated higher CIDEr of 178.07%. In contrast, the SCST-IN, SCST-ALL, GNIC, Dense CNN, HPTDL, MODLE-AICT algorithms have reached lower CIDEr of 161.66%, 154.92%, 152.65%, 161.86%, 172.34%, 175.93% correspondingly. At the same time, on Rouge-L, the RDOAI-ICS approach has revealed a higher Rouge-L of 53%, whereas the SCST-IN, SCST-ALL, GNIC, Dense CNN, HPTDL, MODLE-AICT techniques have reached lower Rouge-L of 51.38%, 41.38%, 45.37%, 50.44%, 43.54%, 47.46% correspondingly.

Table 2: Comparative analysis of RDOAI-ICS approach with existing algorithm under Flickr8K dataset

Methods	METEOR	CIDEr	Rouge-L
SCST-IN algorithm	18.94	161.66	51.38
SCST-ALL algorithm	24.12	154.92	41.38
GNIC algorithm	20.52	152.65	45.37
Dense CNN algorithm	19.65	161.86	50.44
HPTDL algorithm	27.29	172.34	43.54
MODLE-AICT algorithm	28.46	175.93	47.46
RDOAI-ICS	30.03	178.07	53.00

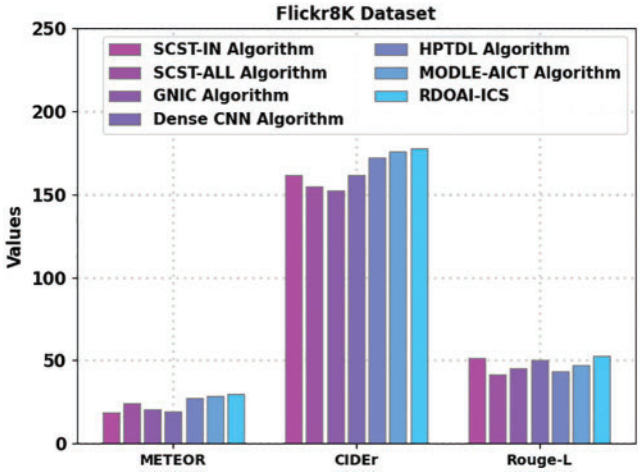

Figure 5: Comparative analysis of RDOAI-ICS approach under Flickr8K dataset

The training accuracy (TRA) and validation accuracy (VLA) acquired by the RDOAI-ICS algorithm under Flickr8K Dataset is displayed in Fig. 6. The experimental result denoted the RDOAI-ICS approach has gained maximal values of TRA and VLA. The VLA is greater than TRA.

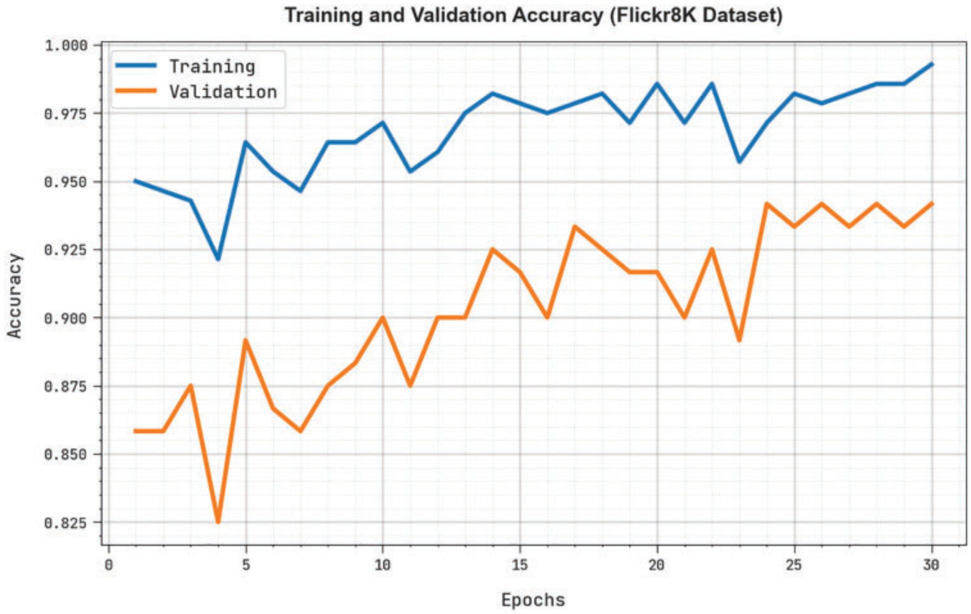

Figure 6: TRA and VLA analysis of RDOAI-ICS algorithm under Flickr8K dataset

The training loss (TRL) and validation loss (VLL) obtained by the RDOAI-ICS algorithm under Flickr8K Dataset are exhibited in Fig. 7. The experimental result represents the RDOAI-ICS method has manifested minimal values of TRL and VLL. Particularly, the VLL is lesser than TRL.

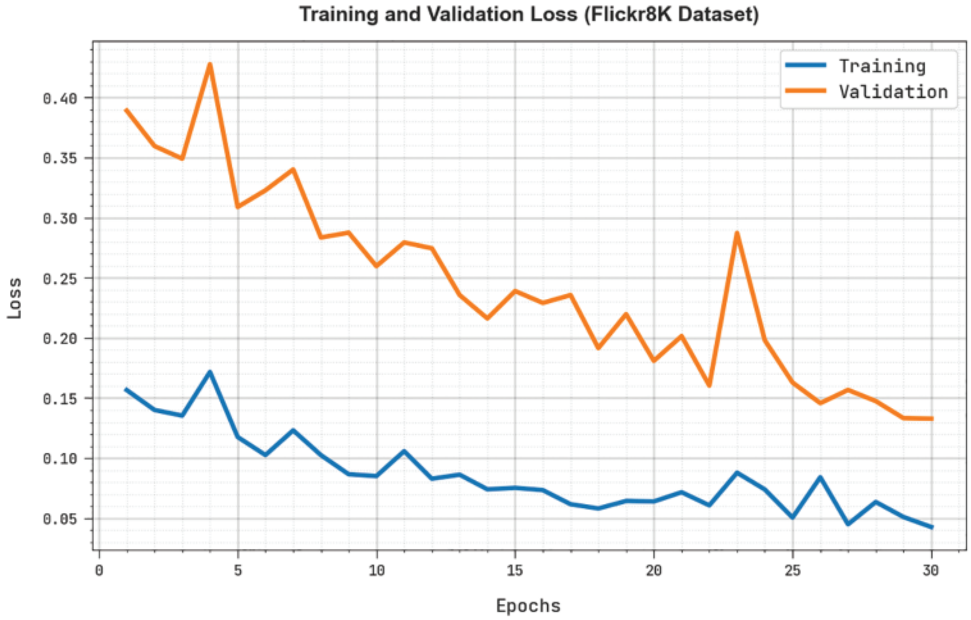

Figure 7: TRL and VLL analysis of RDOAI-ICS algorithm under Flickr8K dataset

The Image captioning outcomes of the RDOAI-ICS algorithm on the MS COCO 2014 dataset are given in Table 3 and Fig. 8. The experimental results show that the RDOAI-ICS approach has displayed other Image captioning techniques in all measures. For example, on the BLUE-1 metric, the RDOAI-ICS methodology has revealed a higher BLEU-1 of 77.56%, whereas the MRNN, GNICG, ResNet-50, KNN, VGG-16, HPTDL, and MODLE-AICT techniques have gained lower BLEU-1 of 49.31%, 67.05%, 72.85%, 64.55%, 72.60%, 69.01%, and 76.06% correspondingly. Also, on the BLUE-2 metric, the RDOAI-ICS technique has manifested a higher BLEU-2 of 61.76%, whereas the MRNN, GNICG, ResNet-50, KNN, VGG-16, HPTDL, and MODLE-AICT algorithms have reached lower BLEU-2 of 28.86%, 48.55%, 51.32%, 43.86%, 59.79%, 55.08%, and 58.62% correspondingly. Concurrently, on the BLUE-3 metric, the RDOAI-ICS method has shown a higher BLEU-3 of 47.94% whereas the MRNN, GNICG, ResNet-50, KNN, VGG-16, HPTDL, and MODLE-AICT approaches have achieved lower BLEU-3 of 16.67%, 35.93%, 33.72%, 31.64%, 43.70%, 42.43%, and 45.90% correspondingly.

Table 3: Result analysis of RDOAI-ICS approach with existing algorithm under MS COCO 2014 dataset

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
KNN algorithm [22]	49.31	28.86	16.67	11.11
G-NICG algorithm [23]	67.05	48.55	35.93	25.65
L-Bilinear algorithm [22]	72.85	51.32	33.72	26.21
ResNet-50 algorithm [24]	64.55	43.86	31.64	23.77
VGA16 algorithm [22]	72.60	59.79	43.70	34.14
VGG-16 algorithm [22]	69.01	55.08	42.43	30.36
HPTDL algorithm [24]	76.06	58.62	45.90	35.56
RDOAI-ICS	77.56	61.76	47.94	37.51

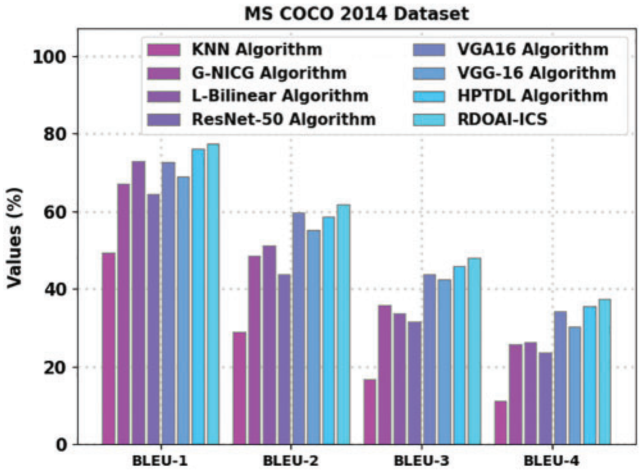

Figure 8: Result analysis of RDOAI-ICS approach under MS COCO 2014 dataset

A comparative study of the RDOAI-ICS method on the Flickr8K dataset is given in Table 4 and Fig. 9. The experimental outcomes show the RDOAI-ICS algorithm has revealed other image captioning approaches under all measures. For example, the RDOAI-ICS approach has shown a higher METEOR of 37.46% on METEOR. In contrast, the SCST-IN, SCST-ALL, GNIC, Dense CNN, HPTDL, and MODLE-AICT approaches have reached lower METEOR of 20.66%, 23.78%, 21.38%, 23.02%, 26.55%, 35.21% correspondingly. Also, on CIDEr, the RDOAI-ICS algorithm has displayed higher CIDEr of 126.21%. In contrast, the SCST-IN, SCST-ALL, GNIC, Dense CNN, HPTDL, and MODLE-AICT approaches have reached lower CIDEr of 109.42%, 113.07%, 110.37%, 110.46%, 121.83%, 124.14% correspondingly. At the same time, on Rouge-L, the RDOAI-ICS algorithm has manifested a higher Rouge-L of 62.15%, whereas the SCST-IN, SCST-ALL, GNIC, Dense CNN, HPTDL, MODLE-AICT techniques have attained lower Rouge-L of 51.49%, 59.55%, 51.76%, 59.38%, 57.18%, 59.90% correspondingly.

Table 4: Comparative analysis of RDOAI-ICS approach with existing algorithm under MS COCO 2014 dataset

Methods	METEOR	CIDEr	Rouge-L
SCST-IN algorithm [22]	20.66	109.42	51.49
SCST-ALL algorithm [22]	23.78	113.07	59.55
GNIC algorithm [23]	21.38	110.37	51.76
Dense CNN algorithm [23]	23.02	110.46	59.38
HPTDL algorithm [24]	26.55	121.83	57.18
MODLE-AICT algorithm [22]	35.21	124.14	59.90
RDOAI-ICS	37.46	126.21	62.15

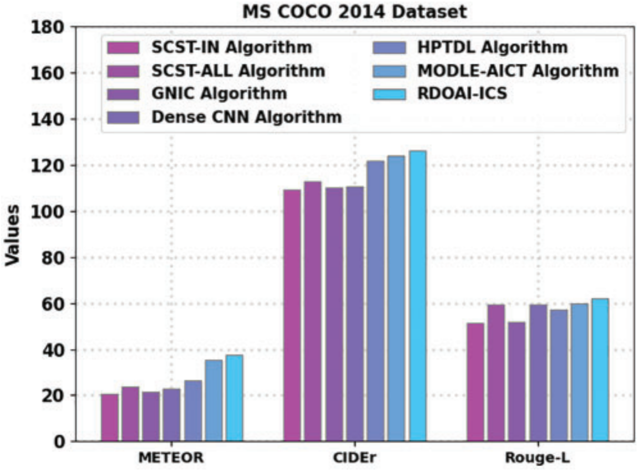

Figure 9: Comparative analysis of RDOAI-ICS approach under MS COCO 2014 dataset

The TRA and VLA attained by the RDOAI-ICS approach under MS COCO 2014 Dataset is shown in Fig. 10. The experimental result indicated the RDOAI-ICS method has gained maximal values of TRA and VLA. The VLA is greater than TRA.

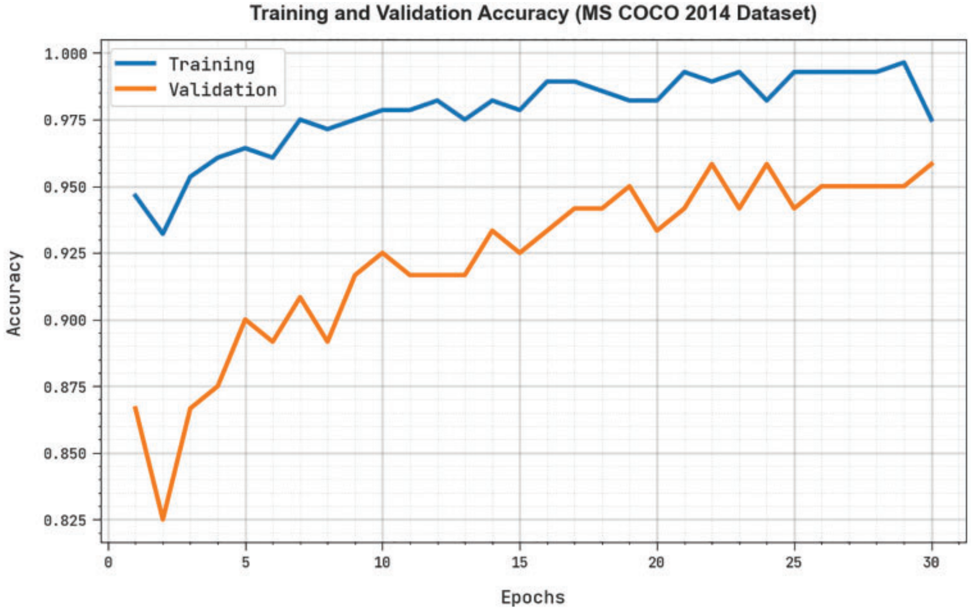

Figure 10: TRA and VLA analysis of RDOAI-ICS algorithm under MS COCO 2014 dataset

The TRL and VLL obtained by the RDOAI-ICS algorithm in MS COCO 2014 Dataset are exhibited in Fig. 11. The experimental outcome indicates the RDOAI-ICS methodology has displayed the least values of TRL and VLL. Particularly, the VLL is lesser than TRL. These results affirmed the enhanced image captioning outcomes of the RDOAI-ICS model on two datasets.

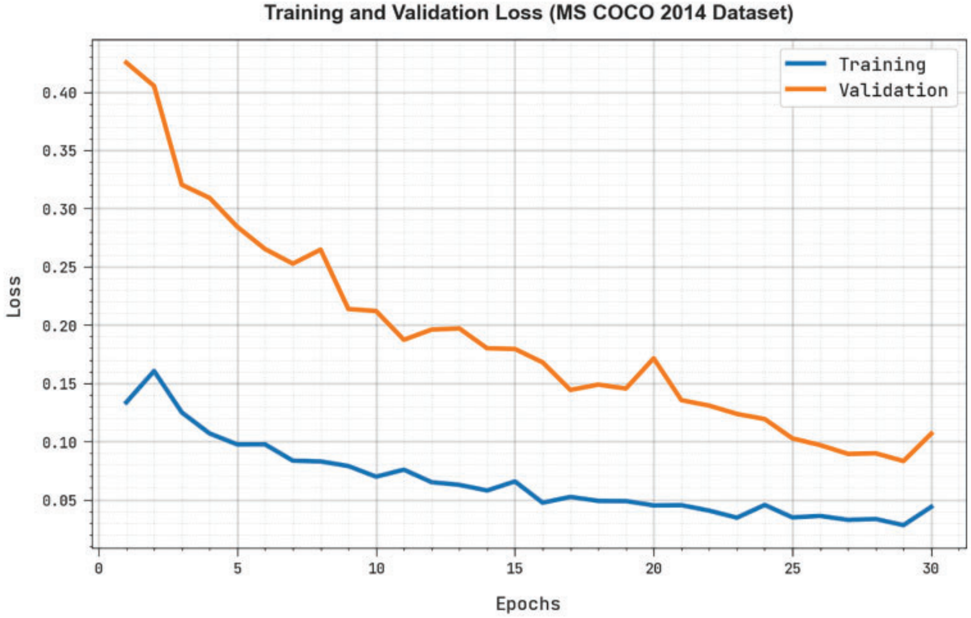

Figure 11: TRL and VLL analysis of RDOAI-ICS algorithm under MS COCO 2014 dataset

5 Conclusion

This study devised a new RDOAI-ICS algorithm for image caption generation for VIPs. The presented RDOAI-ICS technique aids in generating image captions for VIPs. The presented RDOAI-ICS technique used the NASNet model to produce image representations to accomplish this. Besides, the RDOAI-ICS technique exploited the RBFNN model to generate textual descriptions. To enhance the performance of the RDOAI-ICS method, the parameter optimization procedure takes place using the RDO algorithm for NasNet and BOA for the RBFNN model. The experimental evaluation of the RDOAI-ICS method can be tested using a benchmark dataset and the outcomes show the enhancements of the RDOAI-ICS method over other recent Image captioning approaches with CIDEr of 161.66 and 126.21 on Flickr8k and MS COCO 2014 datasets respectively. Thus, the RDOAI-ICS model can be applied for real-time caption generation, which aids the navigation of the VIPs. In future, hybrid DL models can be employed to generate captions automatically.

Funding Statement: The authors extend their appreciation to the King Salman center for Disability Research for funding this work through Research Group no KSRG-2022-017.

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

References

[1] D. Virmani, C. Gupta, P. Bamdev and P. Jain, "Iseeplus: A cost effective smart assistance archetype based on deep learning model for visually impaired," *Journal of Information and Optimization Sciences*, vol. 41, no. 7, pp. 1741–1756, 2020.

[2] P. Chun, T. Yamane, and Y. Maemura, "A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage," *Computer-aided Civil Engineering*, vol. 37, no. 11, pp. 1387–1401, 2022.

[3] H. Lu, R. Yang, Z. Deng, Y. Zhang, G. Gao *et al.,* "Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, no. 1s, pp. 1–18, 2021.

[4] K. Iwamura, J. Y. L. Kasahara, A. Moro, A. Yamashita and H. Asama, "Image captioning using motion-cnn with object detection," *Sensors*, vol. 21, no. 4, pp. 1270, 2021.

[5] S. Kalra and A. Leekha, "Survey of convolutional neural networks for image captioning," *Journal of Information and Optimization Sciences*, vol. 41, no. 1, pp. 239–260, 2020.

[6] J. A. Alzubi, R. Jain, P. Nagrath, S. Satapathy, S. Taneja *et al.,* "Deep image captioning using an ensemble of CNN and LSTM based deep neural networks," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 4, pp. 5761–5769, 2021.

[7] T. Jaiswal, "Image captioning through cognitive IOT and machine-learning approaches," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 9, pp. 333–351, 2021.

[8] S. K. Mishra, R. Dhir, S. Saha and P. Bhattacharyya, "A hindi image caption generation framework using deep learning," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 2, pp. 1–19, 2021.

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

[10] D. S. L. Srinivasan and A. L. Amutha, "Image captioning-a deep learning approach," *International Journal of Applied Engineering Research*, vol. 13, no. 9, pp. 7239–7242, 2018.

[11] Y. H. Chang, Y. J. Chen, R. H. Huang and Y. T. Yu, "Enhanced image captioning with color recognition using deep learning methods," *Applied Sciences*, vol. 12, no. 1, pp. 209, 2021.

[12] M. Bhalekar and M. Bedekar, “D-CNN: A new model for generating image captions with text extraction using deep learning for visually challenged individuals,” *Engineering, Technology & Applied Science Research*, vol. 12, no. 2, pp. 8366–8373, 2022.

[13] Y. Su, Y. Li, N. Xu and A. A. Liu, “Hierarchical deep neural network for image captioning,” *Neural Processing Letters*, vol. 52, no. 2, pp. 1057–1067, 2020.

[14] C. Rane, A. Lashkare, A. Karande and Y. S. Rao, “Image captioning based smart navigation system for visually impaired,” in *Int. Conf. on Communication Information and Computing Technology*, Mumbai, India, pp. 1–5, 2021.

[15] B. Kim, S. Shin and H. Jung, “Variational autoencoder-based multiple image captioning using a caption attention map,” *Applied Sciences*, vol. 9, no. 13, pp. 2699, 2019.

[16] Y. Ma, Y. Shi, M. Zhang, W. Li, C. Ma *et al.,* “Design and implementation of an intelligent assistive cane for visually impaired people based on an edge-cloud collaboration scheme,” *Electronics*, vol. 11, no. 14, pp. 2266, 2022.

[17] S. He and Y. Lu, “A modularized architecture of multi-branch convolutional neural network for image captioning,” *Electronics*, vol. 8, no. 12, pp. 1417, 2019.

[18] F. Martínez, F. Martínez and E. Jacinto, “Performance evaluation of the NASNet convolutional network in the automatic identification of COVID-19,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 10, no. 2, pp. 1–12, 2020.

[19] N. Unnisa and M. Tatineni, “Adaptive deep learning strategy with red deer algorithm for sparse channel estimation and hybrid precoding in millimeter wave massive MIMO-OFDM systems,” *Wireless Personal Communications*, vol. 122, no. 4, pp. 3019–3051, 2022.

[20] H. Zhang, X. Zhang and R. Bu, “Radial basis function neural network sliding mode control for ship path following based on position prediction,” *Journal of Marine Science and Engineering*, vol. 9, no. 10, pp. 1055, 2021.

[21] S. Arora and S. Singh, “Butterfly optimization algorithm: A novel approach for global optimization,” *Soft Computing*, vol. 23, no. 3, pp. 715–734, 2019.

[22] M. Al Duhayyim, S. Alazwari, H. A. Mengash, R. Marzouk, J. S. Alzahrani *et al.,* “Metaheuristics optimization with deep learning enabled automated image captioning system,” *Applied Sciences*, vol. 12, no. 15, pp. 7724, 2022.

[23] 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.

[24] Y. Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang, “Automatic image captioning based on ResNet50 and LSTM with soft attention,” *Wireless Communications and Mobile Computing*, vol. 2020, pp. 1–7, 2020.