Computers, Materials & Continua
DOI: 10.32604/cmc.2023.033091
Article

Tech Science Press

Natural Language Processing with Optimal Deep Learning-Enabled Intelligent Image Captioning System

Radwa Marzouk[1], Eatedal Alabdulkreem[2], Mohamed K. Nour[3], Mesfer Al Duhayyim[4,*], Mahmoud Othman[5], Abu Sarwar Zamani[6], Ishfaq Yaseen[6] and Abdelwahed Motwakel[6]

[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 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 Sciences, College of Computing and Information System, Umm Al-Qura University, Saudi Arabia
[4]Department of Computer Science, College of Sciences and Humanities-Aflaj, Prince Sattam bin Abdulaziz University, Saudi Arabia
[5]Department of Computer Science, Faculty of Computers and Information Technology, Future University in Egypt, New Cairo, 11835, Egypt
[6]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia
*Corresponding Author: Mesfer Al Duhayyim. Email: m.alduhayyim@psau.edu.sa

Received: 07 June 2022; Accepted: 12 July 2022

Abstract: The recent developments in Multimedia Internet of Things (MIoT) devices, empowered with Natural Language Processing (NLP) model, seem to be a promising future of smart devices. It plays an important role in industrial models such as speech understanding, emotion detection, home automation, and so on. If an image needs to be captioned, then the objects in that image, its actions and connections, and any silent feature that remains under-projected or missing from the images should be identified. The aim of the image captioning process is to generate a caption for image. In next step, the image should be provided with one of the most significant and detailed descriptions that is syntactically as well as semantically correct. In this scenario, computer vision model is used to identify the objects and NLP approaches are followed to describe the image. The current study develops a Natural Language Processing with Optimal Deep Learning Enabled Intelligent Image Captioning System (NLPODL-IICS). The aim of the presented NLPODL-IICS model is to produce a proper description for input image. To attain this, the proposed NLPODL-IICS follows two stages such as encoding and decoding processes. Initially, at the encoding side, the proposed NLPODL-IICS model makes use of Hunger Games Search (HGS) with Neural Search Architecture Network (NASNet) model. This model represents the input data appropriately by inserting it into a predefined length vector. Besides, during decoding phase, Chimp Optimization Algorithm (COA) with deeper Long Short Term Memory (LSTM) approach is followed to concatenate the description sentences produced by the method. The application of HGS and COA algorithms helps in accomplishing proper parameter tuning for NASNet and LSTM models respectively. The proposed NLPODL-IICS model was experimentally validated with the help of two benchmark datasets. A widespread comparative analysis confirmed the superior performance of NLPODL-IICS model over other models.

Keywords: Natural language processing; information retrieval; image captioning; deep learning; metaheuristics

1 Introduction

Internet users have encountered huge volumes of images from multiple sources like advertisements, internet, document diagrams, and news articles on a daily basis. These images, sourced from multiple domains, have to be interpreted by the viewers [1]. However, most of the images lack description while human-beings mostly realize it without having to provide detailed captions. But it is the responsibility of automation technologies to unravel few types of image captions, when individuals require automated generation of image captions for them [2]. Image captioning is a vital phenomenon for countless reasons. One of the common applications of image captioning is to train a visual model with the help of captioned images. This way, a language model can be trained with the help of captions offered through descriptions [3]. When learning a multimodal joint representation of captions and images, a semantic resemblance of captions and images is measured which suggests the most descriptive caption to the provided input image. The crucial part in visual language modeling is capturing the semantic associations between text and image modalities. For instance, image captioning was utilized for automated image indexing [4]. Image indexing is significant for Content-Based Image Retrieval (CBIR). Thus, it has been applied in different fields, i.e., education, biomedicine, military, library science, web searching, and e-commerce. Social media platforms like Twitter and Facebook straight away produce descriptions from images. This information involves the description of place, the thing worn by the user in image, and, significantly, the activities occurring in a specific locality [5,6].

If well-formed sentences are to be made, it needs syntactic as well as semantic understanding of the language [7]. Image understanding is heavily relied upon the acquisition of image features. The methods utilized for image understanding are classified into two categories one such as conventional Machine Learning (ML)-related methods and deep ML-related methods. In conventional ML-based methods, handcrafted features like Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), and Local Binary Patterns (LBPs) and the amalgamation of these features are broadly utilized. In such methods, the features can be derived from input data. Then, the features are transferred to the classifiers such as Support Vector Machine (SVMs) [8] for object classification. Since handcrafted features are task-specific, it is challenging to derive the features from huge and varied data sets. Additionally, real-world data like video and images are complicate in nature and involve distinct semantic explanations. When it comes to deep ML-based related methods, features are extracted from the training data automatically. These types of methods can manage huge volumes of multimedia data that includes videos and image sets. For instance, Convolutional Neural Networks (CNNs) [9,10] have been broadly utilized as a classifier whereas feature learning method, namely, Softmax is utilized for categorization. CNN is generally tracked by Recurrent Neural Networks (RNNs) to produce the captions [11,12].

The current study develops a Natural Language Processing with Optimal Deep Learning Enabled Intelligent Image Captioning System (NLPODL-IICS) model. The aim of the presented NLPODL-IICS model is to provide proper description for the input images. To attain this, the proposed NLPODL-IICS model makes use of Hunger Games Search (HGS) with Neural Search Architecture Network (NASNet) model to appropriately represent the input by inserting it into a predefined length vector. Besides, on decoding side, Chimp Optimization Algorithm (COA) with in-depth Long Short Term Memory (LSTM) approach is applied for the concatenation of the description provided for the input image. The proposed NLPODL-IICS model was experimentally validated using two benchmark datasets.

2 Literature Review

In the study conducted earlier [13], the authors suggested an end-to-end trainable deep Bi-LSTM method to address the issue of image captioning. By merging deep CNN with two separate LSTM networks, the method gains the ability to learn long-term visual-language connections with the help of future context information and history at high level semantic space. The researchers explored in-depth multi-modal bidirectional methods in which they raised the depth of nonlinearity transition in distinct means so as to learn hierarchical visual language embedding. Das et al. [14] provided a proof-of-concept demo for caption generation. This generative system depends on deep recurrent structure, blended with pre trained image-to-vector method Inception V3 through CNN and word-to-vectors system called word2vec, through skip-gram method.

In literature [15], the researchers suggested a hybrid system involving multilayer CNN to produce synonyms that can explain about the images and LSTM to structure the sentences with precise meaning using the keywords generated earlier. CNN compared the target image against a huge dataset of training images, after producing a precise explanation with the help of trained captions. The research scholars in the study conducted earlier [16] aimed at visual attention for which they proposed an advanced technique for image captioning in computer vision research zone. The researchers understood the influence exerted by distinct hyper-parameters over encoder-decoder visual attention structure with regards to efficiency. Nogueira et al. [17] suggested a technique in accordance with encoder–decoder structure using CNNs to extract the features from Gated Recurrent Units (GRUs) and reference images so as to describe the images. This method implies Part-of-Speech (PoS) examination and the possible function for weight generation in GRU. The above-mentioned methodology further executes the knowledge, transmitted at the time of validation stage using the k-Nearest Neighbour (kNN) algorithm.

In the study conducted earlier [18], a new Hyperparameter-Tuned DL for Automated Image Captioning (HPTDL-AIC) method was suggested. The proposed HPTDL-AIC method has two major parts such as decoder and encoder. Here, the encoder unit uses Faster SqueezNet by following RMSProp method to portray the input image by inserting it into a pre-defined length vector. Meanwhile, the decoder part uses Bird Swarm Algorithm (BSA) with LSTM technique so as to concentrate on the generation of descriptive sentences. In an earlier study [19], image captioning and semantic segmentation were widely inspected on the basis of advanced and traditional methods. In this study, the researchers detailed about the application of DL in segmentation examination of both 3D and 2D images utilizing FCN and other high-level hierarchical feature extraction methodologies. At first, every preliminary domain and the concept were explained after which semantic segmentation was deliberated together with appropriate features, evaluation criteria, and existing datasets.

3 The Proposed NLPODL-IICS Model

In this study, a novel NLPODL-IICS model has been developed to provide appropriate description for the input image. Initially, at encoding side, HGS is employed with NASNet model to create a proper representation for the input image through its insertion at a predefined length vector. Besides, on decoding side, COA with deeper LSTM model is used for the concatenation of the description sentences produced for the input image. Fig. 1 shows the overall processes involved in NLPODL-IICS method.

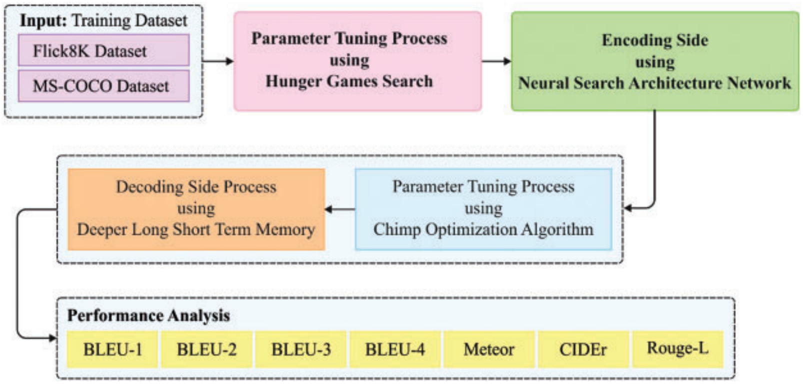
Input: Training Dataset
Flick8K Dataset
MS-COCO Dataset
Parameter Tuning Process using Hunger Games Search
Encoding Side using Neural Search Architecture Network
Decoding Side Process using Deeper Long Short Term Memory
Parameter Tuning Process using Chimp Optimization Algorithm
Performance Analysis
BLEU-1
BLEU-2
BLEU-3
BLEU-4
Meteor
CIDEr
Rouge-L

Figure 1: Overall process of NLPODL-IICS approach

3.1 Data Pre-Processing

At first, the data is pre-processed at different levels as listed below.

- The dataset text contains words with different letter cases. This creates a problem for the elements which sound similar to the words with various capitalizations. This scenario enhances the vocabulary problem and brings difficulty in achieving the outcomes. So, it is important to change the complete text to lower case so as to prevent these problems.
- Since punctuation marks increase the difficulty in achieving the desired outcome, it is removed from the dataset.
- The numerical data, present in the text, is a challenging issue for elements since it increases vocabulary issues. So, it is also removed.
- Refers to first and the last orders: The word tokens such as '< start >' and '< end >' denote the first and last words of all the sentences represent the primary and final ends of the predicted order to the element.
- Tokenization: Clean text can be divided into constituent words whereas the dictionary contains the whole vocabulary. So, word-to-index and index-to-word coherence is attained.
- Vectorization: In order to solve distinct and lengthy sentence issues, short sentence is padded according to the length of the long sentence order.

3.2 NASNet Based Feature Extraction

In this study, NASNet approach is utilized to create visual features of the input images. Google Team 2018 introduced NASNet architectur, a Convolutional Neural Network [20] and mentioned that 'the presented architecture is 1.2% better in top-1 accuracy when compared to the human-developed architecture that has nine billion less FLOPS—a 28% reduction in computation demand from the preceding model'. This model is employed as a feature extractor due to less computation demand and high accuracy. This model comprises of convolution cells. Reduction cell and normal cell are the two major functions of this model. In Reduction Cell, the feature maps are returned with width and height decreased by a factor of two. At the same time, Normal Cell returns a feature map with the help of normal cells for similar input dimension.

Then, the final Normal_A_cell, global_avg_pooling2d, flatten, and global_max_pooling2d layers are fed. The dimension of the tensor is reduced to Nx1056 in initial two layers, where the image count is trained to be N. These layers feed the dropout layers that turned off certain neurons to avoid over-fitting issues in the network. In order to optimally fine-tune the hyperparameters involved in NASNet model, HGS algorithm is applied. In nature, the survival of the animals is decided by their hunger-driven activity. In this regard, the decision on their motion plays a vital role in terms of survival. HGS approach considers behavioral choices as a set of game rules for the reality whereas this scenario is mathematically modelled on the basis of hunger-driven activity, so that it can be utilized as an effective meta-heuristic optimization method. Game rule mimics the co-operations among animals at the time of foraging [21]. But, it also considers the reluctance of an individual animal. The game rules are given below.

$$\overline{X(t+1)} = \begin{cases} Game_1 \ : \overline{X(t)} \cdot (1 + randn(1)), \ r_1 < C \\ Game_2 : \overline{W_1} \cdot \overline{X_b} + \overline{R} \cdot \overline{W_2} \cdot |\overline{X_b} - \overline{X(t)}|, r_1 > C, \ r_2 > E \\ Game_3 : \overline{W_1} \cdot \overline{X_b} - \overline{R} \cdot \overline{W_2} \cdot |\overline{X_b} - \overline{X(t)}|, r_1 > C, \ r_2 < E \end{cases} \tag{1}$$

In Eq. (1), random numbers within [0, 1] are represented by r_1 and r_2. Another random number that satisfies the normal distribution is *randn*. At the same time, t denotes the existing iteration, $\overline{W_1}$ and $\overline{W_2}$ represent hunger weight, $\overline{X_b}$ represents the position of optimal individual and $\overline{X(t)}$ indicates the position of every individual in the existing iteration. The choice between the designated rules ($Game_1$, $Game_2$ *and* $Game_3$) is defined, according to the variable, C. The variable of $\overline{R}$ is evaluated by the formula, $\overline{R} = 2 \times shr \times$ rand$-shr$ where $hr = 2 \times (1 - (t/T))$. In this description, rand signifies a random integer that lies in the interval of $[0, 1]$, and T characterizes the maximal number of iterations. The variation of each position is controlled by E variable and is determined by $E = sech(|F(i) - BFif|)$ in which $F(i)$ indicates the fitness value of every individual, and $i \in 1, 2, \ldots, n$. sech embodies a hyperbolic function and is evaluated by $sech(x) = (2/(e^x + e^{-x}))$. Considering the overall architecture described in Eq. (1), it is possible to mention two search classifications. The initial one signifies the self-dependent individual whereas the next one simulates the teamwork. Consequently, the individual reached an improved diversification result.

Also, the starvation feature of every individual is simulated in HGS approach. The hunger weight, shown in Eq. (1), is formulated in the following equations while N indicates the number of individuals and *Shng* denotes the sum of hungry feelings exhibited by every individual.

$$\overline{W_1(i)} = \begin{cases} hng(i) \cdot \dfrac{N}{Shng} \times r_4, r_3 < C \\ 1, \ r_3 > C \end{cases} \tag{2}$$

$$\overline{W_2(i)} = \left(1 - e^{-|hng(i) - Shng|}\right) \times r_5 \times 2 \tag{3}$$

In these equations, r_3, r_4 and r_5 represent random integers that lie in the range of [0, 1]. The value of $hng(t)$ can be determined as given below.

$$hng\left(i\right)=\begin{cases}0,\ Allfit\left(i\right)==BFit\\ hng\left(i\right)+H,\ Allfit\left(i\right)!=BFit\end{cases}\tag{4}$$

From Eq. (4), $Allfit\left(i\right)$ denotes the fitness of every individual in existing iteration. The subsequent term is utilized for defining the variable of H (hunger sensation).

$$H=\begin{cases}LH\times\left(1+r\right),\ TH<LH\\ TH,\ TH\geq LH\end{cases}\tag{5}$$

In Eq. (5), $H=((F(i)-BFit)/(WF-BFit))\times r_6\times 2\times(UB-LB)$. Now, WF denotes the Worst Fitness values and LH indicates lower limit. The feature space-based upper and lower limits are denoted as UB and LB correspondingly.

3.3 Image Caption Generation

Finally, in-depth LSTM model is utilized to generate proper captions for the test images. The new achievements of deep CNN in object detection and image classification determine that deep and hierarchical model are better at portraying representations than the shallower ones. This inspired the investigation of deep LSTM architecture in terms of learning bi-directional visual-language embedding [22]. LSTM is regarded as a configuration of more than one hidden layer that is extended in time. It is previously known as a deeper network. However, it can also achieve the 'horizontal depth' form, in which the network weight W is reutilized at every time step and is limited to learning the representation of features namely increasing 'vertical depth' of the network. In parallel, instead of stacking more than one LSTM layer, MLP is added as an intermediate transition among the LSTM layers. The direct stacking of more than one LSTM layer results in Bi-S-LSTM. Additionally, an FC layer is also utilized as an intermediate transition layer. The objective is to learn the additional hidden transition function, F_h.

$$h_t^{l+1}=F_h\left(h_t^{l-1},h_{t-1}^{l}\right)=Uh_t^{l-1}+Vh_{t-1}^{l},\tag{6}$$

In Eq. (6), h_t^l represents the hidden state of l^{th} layer at t time, U and V denote the matrices interconnected with transition layer. For readability, one direction is regarded for training with overwhelming bias terms. Likewise, in Bi-F-LSTM, a hidden transition function F_h is learnt with the help of the following equation.

$$h_t^{l+1}=F_h\left(h_t^{l-1}\right)=(\phi_r(Wh_t^{l-1}\oplus\left(V\left(Uh_t^{l-1}\right)\right),\tag{7}$$

Eq. (7), $\oplus$ indicates the operator that concatenates h_t^{l-1} and the abstraction to a long hidden state. ($\emptyset_r$ signifies the Relu function towards transition layer that implements ($\emptyset_r\left(x\right)=\max\ (0,\ x)$. Fig. 2 depicts the framework of LSTM.

In order to adjust the hyperparameters related to deeper LSTM model, COA algorithm is used. COA is a population-based algorithm that is stimulated by Chimps' exclusive sexual behavior in group hunting. Chimps are one of two completely indigenous African families of giant gorillas that are well-known for its high intelligence. This is because, their brain to body ratio is high compared to human beings. These characteristics are used to develop COA as a mathematic operation for chimpanzee hunting. Hunting method is classified into two different stages such as exploitation and exploration. Exploration step contains chase, Drive, and block while exploitation stage focuses on attacking the prey as mentioned below.

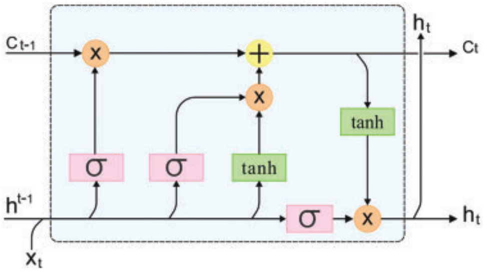

Figure 2: LSTM structure

Khishe et al. [23] exploited the subsequent formula to simulate the pursuit and driving of the prey.

$$D = \left| b \cdot U_{prey}\left(t\right) - e \cdot U_{chimp}\left(t\right) \right| \tag{8}$$

$$U_{chimp}\left(t+1\right) = U_{prey} - f \cdot D \tag{9}$$

In this equation, U_{chimp} and U_{prey} indicate the position vectors for chimps and the prey, correspondingly. The coefficient vector is represented as e whereas f is computed by the following equation

$$b = 2.r_2 \tag{10}$$

$$e = Chaotic_{value} \tag{11}$$

$$f = 2.a.r_1 - a \tag{12}$$

Eq. (7) is utilized in both global and local regimes by modifying the arbitrary vector and it gets nonlinearly decreased from 2.5 to 0 through iteration count $(b,\ e,\ and\, f)$. r_1 and r_2 represent random integer lies within the interval of [0–1].

The main focus of the hunting stage is to use a satisfactory location (prey position) that makes the hunting process, easier. Consequently, the position of the prey is recognized as the optimal solution vector. Since the colony is separated into four classes, the best chimpanzee from every group is carefully chosen. The finest chimpanzees established in this method, are the finest predictors of the prey's likely position. Another chimpanzee in neighborhood gets upgraded to the spot randomly [24]. The mathematical expression of the hunting method is formulated as given below.

$$\begin{gathered} D_{attacker} = \left| b \cdot U_{attacker} - e \cdot U_{chimp}\left(t\right) \right|, D_{Barrier} \\ = \left| b \cdot U_{Barrier} - e \cdot U_{chimp}\left(t\right) \right|, \\ D_{Chaser} = \left| b \cdot U_{Chaser} - e \cdot U_{chimp}\left(t\right) \right|, D_{Driver} \\ = \left| b \cdot U_{Driver} - e \cdot U_{chimp}\left(t\right) \right| \end{gathered} \tag{13}$$

$$\begin{gathered} U_1\left(t+1\right) = U_{attacker} - f \cdot D_{attacker}, U_2\left(t+1\right) \\ = U_{Barrier} - f \cdot D_{Barrier}, \\ U_3\left(t+1\right) = U_{Chaser} - f \cdot D_{Chaser},\ U_4\left(t+1\right) \\ = U_{Driver} - f \cdot D_{Driver} \end{gathered} \tag{14}$$

$$U\left(t+1\right)=\frac{U_1+U_2+U_3+U_4}{4} \tag{15}$$

Once the target stops moving, the final component of the hunting procedure occurs i.e., attack upon prey. Once the coefficient a_n successively decreases, the vector f is formulated as discussed herewith. Further, a decreases from 2.5 to 0 in CHOA, and f lies in the range of $[-1, 1]$. Therefore, the novel location of the chimpanzees can be anywhere between their existing location and the location nearby prey. In exploration phase, it is best to avoid striking the local solution as detailed herewith.

The aim of diversifying the chimpanzee is to quickly identify the position of the prey; thus, the large value of the vector f is more than 1 or less than -1 to attain the objective. In terms of chimp mobility, if $|f|. > 1$, the chimp gets disseminated all over the environment when searching for high quality solutions. Further, COA uses other significant coefficients, the b vector of Eq. (γ), in order to increase the exploration phase and prevent local minima issue in the last repetition. The b vector value lies within the interval of $[0, 2]$.

As mentioned earlier, the sexual behavior of the chimpanzees' is the major reason for their erratic action during final phase. The researchers applied six chaos maps such as Bernoulli, Quadratic, Gauss/Mouse, Logistic, Tent, and Singer maps for mathematical implementation. During optimization, the researchers applied 50% switch possibility to switch between the updating model and chaotic patterns in order to update the location of the chimpanzees.

$$U_{chimp}\left(t+1\right)=\begin{cases}U_{prey}\left(t\right)-f\cdot D\mu<0.5\\ Chaotic_{value}\ \mu>0.5\end{cases} \tag{16}$$

COA starts by creating a group of arbitrary solutions for a problem dimension with a size of $N \times Dim$, in which N denotes the population size and Dim refers to problem variable as given below.

$$U_{chimp_{i,d}}=randx\left(ub_d-lb_d\right)+lb_d i=1,2,\ldots,N,d=1,2,3,\ldots Dim \tag{17}$$

4 Experimental Validation

In this section, the performance of the proposed NLPODL-IICS approach was experimentally validated with the help of two benchmark datasets such as Flickr8K [25] and MSCOCO [26] caption dataset. Flickr dataset commonly includes Flickr8k and 30 K datasets. It contains 8,000 images with human actions. An image in the dataset comprises of five sentences of textual descriptions. MSCOCO dataset contains the data gathered from several objects with conditions. Some of the sample images are portrayed in Fig. 3.

The proposed NLPODL-IICS model was experimentally validated on Flick8k dataset under different measures and the results are shown in Table 1 and Fig. 4. The experimental outcomes imply that the proposed NLPODL-IICS approach demonstrated improved image captioning performance over other models. With respect to BLEU-1, the proposed NLPODL-IICS model attained an increased BLEU-1 of 0.733, whereas MRNN, Google-NICG, L-Bi-linear, DVSM, ResNet-50, and VGA16 models obtained less BLEU-1 values such as 0.570, 0.646, 0.629, 0.618, 0.688, and 0.634 respectively. Also, with respect to BLEU-2, NLPODL-IICS methodology attained a high BLEU-2 of 0.485, whereas MRNN, Google-NICG, L-Bi-linear, DVSM, ResNet-50, and VGA16 approaches obtained the least BLEU-2 values such as 0.345, 0.470, 0.399, 0.338, 0.440, and 0.421 correspondingly.

Figure 3: Sample Images

Table 1: Results of the analysis of NLPODL-IICS technique under distinct measures on Flick8k dataset

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MRNN	0.570	0.345	0.269	0.146
Google-NICG	0.646	0.470	0.325	0.180
L-Bi-linear	0.629	0.399	0.272	0.219
DVSM	0.618	0.338	0.281	0.179
ResNet-50	0.688	0.440	0.404	0.306
VGA16	0.634	0.421	0.399	0.241
NLPODL-IICS	0.733	0.485	0.425	0.342

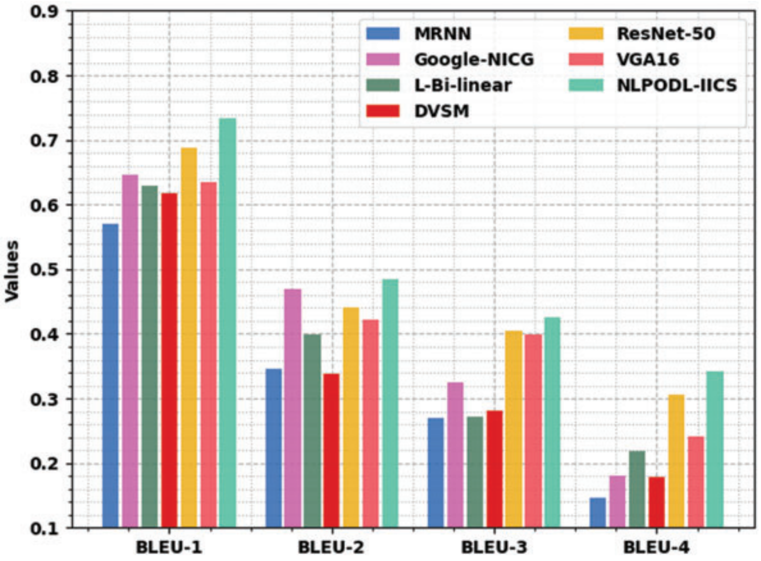

Figure 4: Results of the analysis of NLPODL-IICS technique on Flick8k dataset

In terms of BLEU-4, the proposed NLPODL-IICS approach attained the maximal BLEU-4 of 0.342, whereas MRNN, Google-NICG, L-Bi-linear, DVSM, ResNet-50, and VGA16 methodologies obtained less BLEU-4 values such as 0.146, 0.180, 0.219, 0.179, 0.306, and 0.241 correspondingly.

Table 2 provides the comparative analysis results accomplished by the proposed NLPODL-IICS model on Flick8k dataset [27]. Fig. 5 shows the comparative meteor examination results achieved by the proposed NLPODL-IICS approach and other existing approaches on Flick8k dataset. The figure indicates that ANIC model achieved the least meteor value of 16.00. At the same time, Google-NIC and DenseNet models reported improved meteor values such as 20.00 and 20.00 respectively. Followed by, SCSTIN and SCSTALL models produced reasonable meteor values such as 27.00 and 24.00 correspondingly. However, the proposed NLPODL-IICS model achieved superior performance with a meteor value of 28.

Table 2: Comparative analysis results of NLPODL-IICS approach and other recent methods on Flick8k dataset

Methods	Meteor	CIDEr	Rouge-L
SCSTIN	27.00	156.00	42.00
SCSTALL	24.00	155.00	48.00
Google-NIC	20.00	157.00	43.00
ANIC	16.00	160.00	49.00
DenseNet	20.00	165.00	44.00
NLPODL-IICS	28.00	177.00	53.00

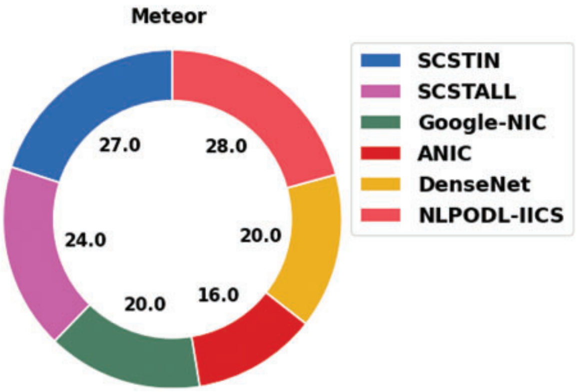

Figure 5: Meteor analysis results of NLPODL-IICS approach on Flick8k dataset

Fig. 6 demonstrates the relative CIDEr investigation outcomes accomplished by the proposed NLPODL-IICS approach and other recent techniques on Flick8k dataset. The figure implies that SCSTALL model achieved the least CIDEr value of 155.00. In addition, SCSTIN and Google-NIC models reported the maximal CIDEr values such as 156.00 and 157.00 correspondingly. Likewise, ANIC and DenseNet approaches produced reasonable CIDEr values such as 160.00 and 165.00 correspondingly. At last, the proposed NLPODL-IICS algorithm attained the maximum CIDEr value of 177.

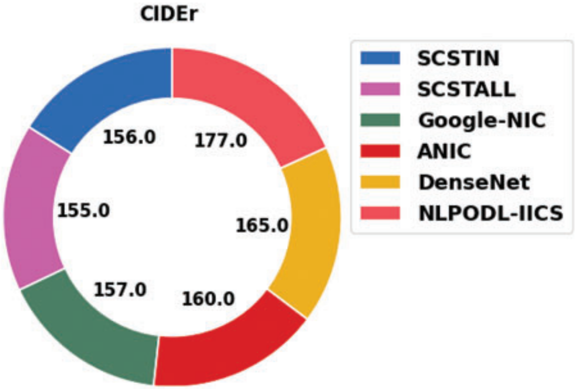

Figure 6: CIDEr analysis results of NLPODL-IICS approach on Flick8k dataset

Fig. 7 depicts the comparative rouge-L analysis results accomplished by the proposed NLPODL-IICS system and other existing methodologies on Flick8k dataset. The figure indicates that Google-NIC system accomplished a low rouge-L value of 43.00. Simultaneously, SCSTIN and DenseNet models reported enhanced rouge-L values such as 42.00 and 44.00 correspondingly. Moreover, SCSTALL and ANIC models produced reasonable rouge-L values such as 48.00 and 49.00 respectively. Finally, the proposed NLPODL-IICS methodology attained superior performance with a rouge-L value of 53.00.

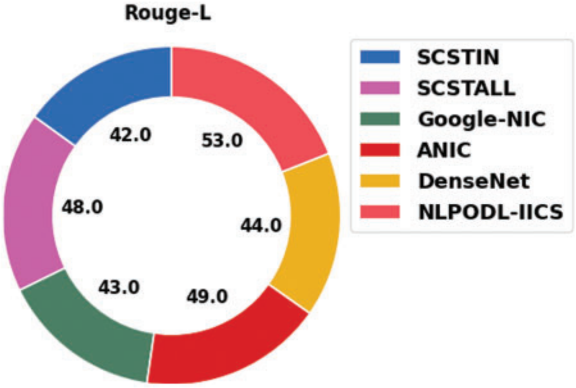

Figure 7: Rouge-L analysis results of NLPODL-IICS approach on Flick8k dataset

Both Training Accuracy (TA) and Validation Accuracy (VA) values, attained by the proposed NLPODL-IICS system on Flick8k dataset, are demonstrated in Fig. 8. The experimental outcomes imply that the proposed NLPODL-IICS method gained the maximum TA and VA values. To be specific, VA seemed to be higher than TA.

Both Training Loss (TL) and Validation Loss (VL) values, achieved by the proposed NLPODL-IICS methodology on Flick8k dataset, are portrayed in Fig. 9. The experimental outcomes infer that the proposed NLPODL-IICS approach achieved the least TL and VL values. To be specific, VL seemed to be lower than TL.

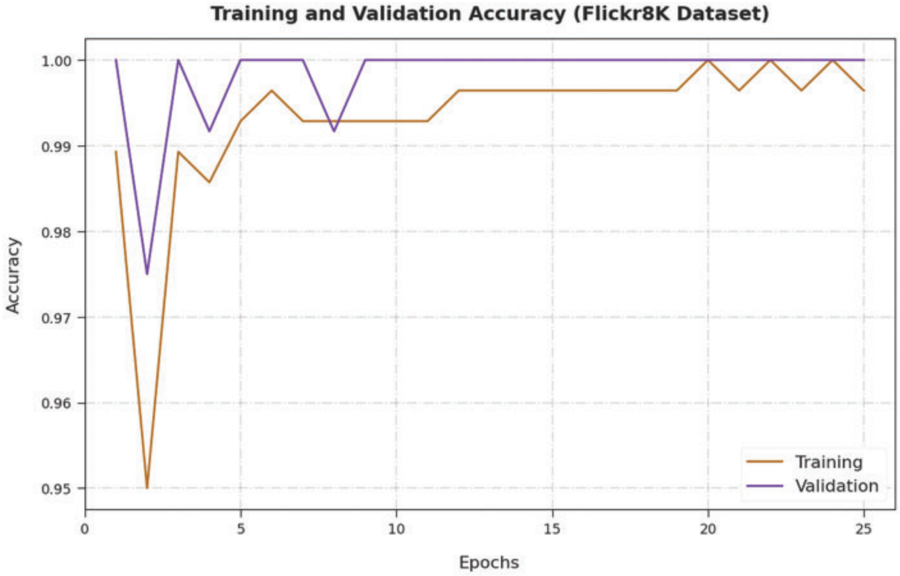

Figure 8: TA and VA analysis results of NLPODL-IICS system on Flick8k dataset

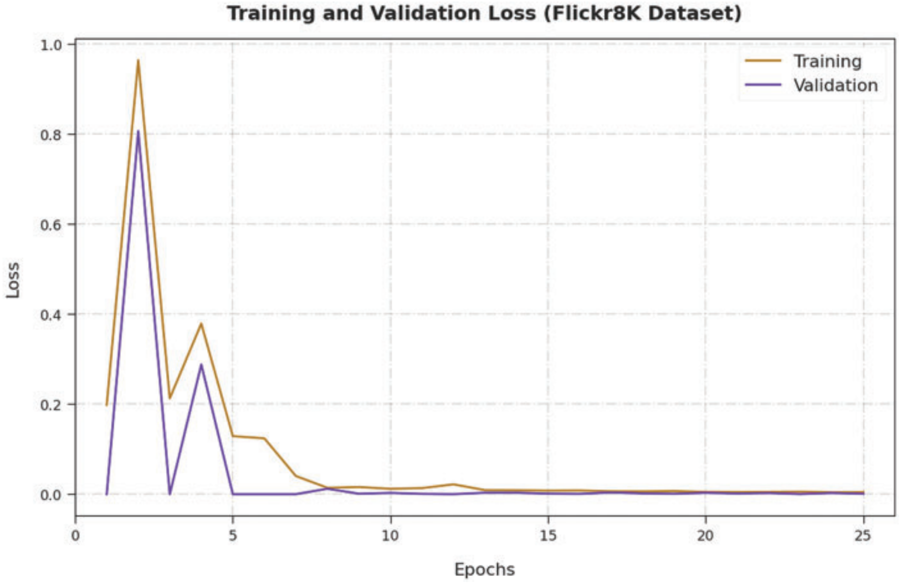

Figure 9: TL and VL analysis results of NLPODL-IICS system on Flick8k dataset

The experimental validation of the proposed NLPODL-IICS approach on MSCOCO 2014 dataset under different measures was conducted and the results are shown in Table 3 and Fig. 10. The experimental outcomes infer that the proposed NLPODL-IICS algorithm depicted the maximum image captioning performance over other models. With regard to BLEU-1, the proposed NLPODL-IICS model attained an increased BLEU-1 of 0.784, whereas MRNN, Google-NICG, L-Bi-linear, DVSM, ResNet-50, and VGA16 models obtained low BLEU-1 values such as 0.509, 0.707, 0.767, 0.684, 0.699, and 0.747 correspondingly. Moreover, in terms of BLEU-2, the proposed NLPODL-IICS model attained the maximum BLEU-2 of 0.605, whereas MRNN, Google-NICG, L-Bi-linear, DVSM, ResNet-50, and VGA16 approaches reached low BLEU-2 values such as 0.354, 0.442, 0.504, 0.442, 0.521, and 0.552 correspondingly. Additionally, with respect to BLEU-4, NLPODL-IICS algorithm attained the maximum BLEU-4 of 0.373, whereas MRNN, Google-NICG, L-Bi-linear, DVSM, ResNet-50, and VGA16 systems obtained low BLEU-4 values such as 0.114, 0.256, 0.206, 0.190, 0.336, and 0.310 correspondingly.

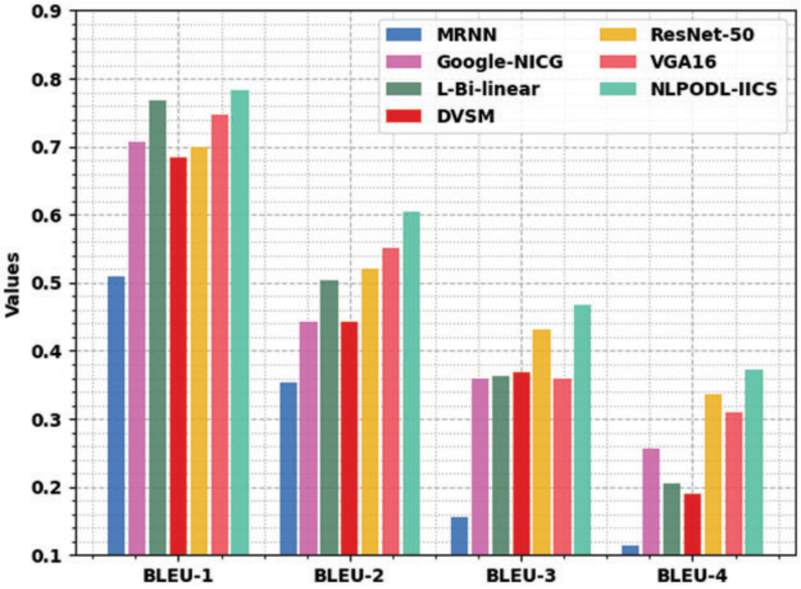

Figure 10: Results of the analysis of NLPODL-IICS technique on MSCOCO 2014 dataset

Table 3: Results of the analysis of NLPODL-IICS technique under distinct measures on MSCOCO 2014 dataset

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MRNN	0.509	0.354	0.155	0.114
Google-NICG	0.707	0.442	0.360	0.256
L-Bi-linear	0.767	0.504	0.363	0.206
DVSM	0.684	0.442	0.369	0.190
ResNet-50	0.699	0.521	0.431	0.336
VGA16	0.747	0.552	0.359	0.310
NLPODL-IICS	0.784	0.605	0.467	0.373

Table 4 demonstrates the detailed comparative analysis results attained by the proposed NLPODL-IICS approach upon MSCOCO 2014 dataset. Fig. 11 showcases the comparative meteor analysis results achieved by NLPODL-IICS algorithm and other recent techniques on MSCOCO 2014 dataset. The figure infers that ANIC approach accomplished the least meteor value of 21.00. Besides, Google-NIC and DenseNet models reported high meteor values such as 28.00 and 26.00 respectively. Followed by, SCSTIN and SCSTALL techniques produced reasonable meteor values such as 27.00

and 32.00 correspondingly. At last, the proposed NLPODL-IICS methodology attained a superior performance and yielded a meteor value of 34.00.

Table 4: Comparative analysis results of NLPODL-IICS approach and other recent methods on MSCOCO 2014 dataset

Methods	Meteor	CIDEr	Rouge-L
SCSTIN	27.00	113.00	52.00
SCSTALL	32.00	120.00	62.00
Google-NIC	28.00	108.00	58.00
ANIC	21.00	103.00	57.00
DenseNet	26.00	115.00	62.00
NLPODL-IICS	34.00	123.00	64.00

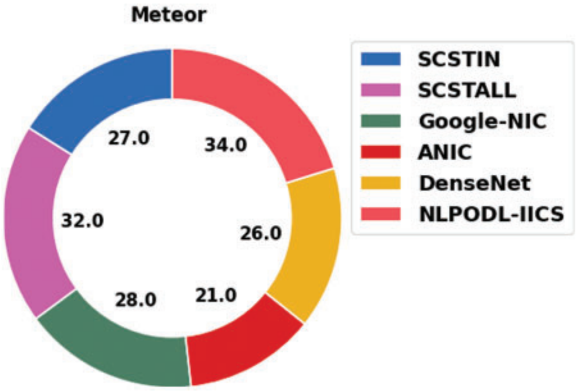

Figure 11: Meteor analysis results of NLPODL-IICS approach on MSCOCO 2014 dataset

Fig. 12 shows the comparative CIDEr investigation results achieved by the proposed NLPODL-IICS system and other existing techniques on MSCOCO 2014 dataset. The figure expose that SCSTALL technique accomplished the least CIDEr value of 120.00. At the same time, SCSTIN and Google-NIC systems reported enhanced CIDEr values such as 113.00 and 108.00 respectively. Followed by, ANIC and DenseNet systems yielded reasonable CIDEr values such as 103.00 and 115.00 correspondingly. But, the proposed NLPODL-IICS approach attained a supreme performance with a CIDEr value of 123.00.

Fig. 13 illustrates the comparative rouge-L examination results accomplished by the proposed NLPODL-IICS system and other existing methods on MSCOCO 2014 dataset. The figure shows that Google-NIC model accomplished a low rouge-L value of 58.00. Followed by, SCSTIN and DenseNet models yielded superior rouge-L values such as 52.00 and 62.00 correspondingly. Afterward, SCSTALL and ANIC techniques produced reasonable rouge-L values such as 62.00 and 57 respectively. Lastly, the proposed NLPODL-IICS algorithm attained a superior performance with a rouge-L value of 64.00. Based on the detailed results and discussion, it is found that NLPODL-IICS approach is an effective tool for image captioning process.

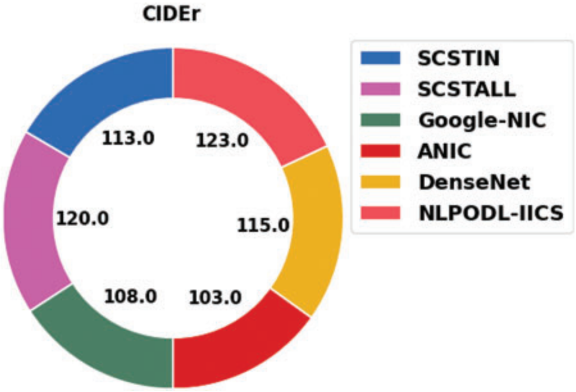

Figure 12: CIDEr analysis results of NLPODL-IICS approach on MSCOCO 2014 dataset

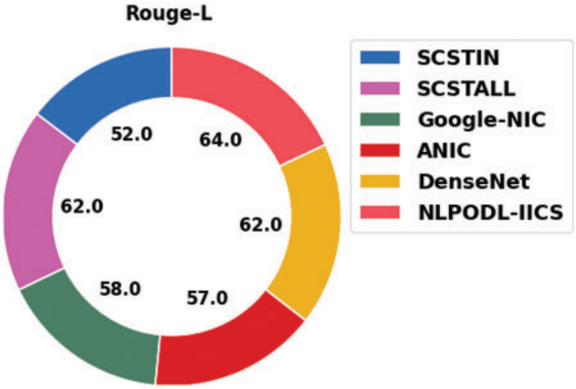

Figure 13: Rouge-L analysis results of NLPODL-IICS approach on MSCOCO 2014 dataset

5 Conclusion

In this study, a novel NLPODL-IICS model has been developed to create proper description for the input image. To attain this, the proposed NLPODL-IICS follows two stages such as encoding and decoding. Initially, at encoding side, HGS is employed with NASNet model to create a proper representation for the input image by inserting the image into a predefined length vector. Besides, on decoding side, COA with deeper LSTM model is applied upon the input image for the concatenation of the description of input images. The application of HGS and COA algorithms helps in accomplishing proper fine-tuning of the parameters involved in NASNet and LSTM models respectively. The proposed NLPODL-IICS model was experimentally validated using two benchmark datasets. A widespread comparative analysis was conducted and the results confirmed the superior performance of NLPODL-IICS model over other models. In future, the presented method can be designed with summarization techniques to improve the performance.

Funding Statement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R161), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would like to thank the | Deanship of Scientific Research at Umm Al-Qura University | for supporting this work by Grant Code: (22UQU4310373DSR33).

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

References

[1] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj and R. K. Mishra, "Image captioning: A comprehensive survey," in *2020 Int. Conf. on Power Electronics & IoT Applications in Renewable Energy and Its Control (PARC)*, Mathura, Uttar Pradesh, India, pp. 325–328, 2020.

[2] M. Chohan, A. Khan, M. Saleem, S. Hassan, A. Ghafoor *et al.,* "Image captioning using deep learning: A systematic literature review," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, pp. 278–286, 2020.

[3] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni *et al.,* "From show to tell: A survey on deep learning-based image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022. https://doi.org/10.1109/TPAMI.2022.3148210.

[4] Y. Cui, G. Yang, A. Veit, X. Huang and S. Belongie, "Learning to evaluate image captioning," in *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 5804–5812, 2018.

[5] T. Yao, Y. Pan, Y. Li and T. Mei, "Exploring visual relationship for image captioning," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Springer, Cham. pp. 684–699, 2018.

[6] Y. Li, T. Yao, Y. Pan, H. Chao and T. Mei, "Pointing novel objects in image captioning," in *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 12489–12498, 2019.

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

[8] J. Gu, J. Cai, G. Wang and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Louisiana, New Orleans, USA, vol. 32, no. 1, pp. 6837–6844, 2018.

[9] I. Abunadi, M. M. Althobaiti, F. N. Al-Wesabi, A. M. Hilal, M. Medani *et al.,* "Federated learning with blockchain assisted image classification for clustered UAV networks," *Computers, Materials & Continua*, vol. 72, no.1, pp. 1195–1212, 2022.

[10] G. Hoxha, F. Melgani and B. Demir, "Toward remote sensing image retrieval under a deep image captioning perspective," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4462–4475, 2020.

[11] A. M. Hilal, H. Alsolai, F. N. Al-Wesabi, M. K. Nour, A. Motwakel *et al.,* "Fuzzy cognitive maps with bird swarm intelligence optimization-based remote sensing image classification," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, 2022.

[12] K. Deorukhkar and S. Ket, "A detailed review of prevailing image captioning methods using deep learning techniques," *Multimedia Tools and Applications*, vol. 81, no. 1, pp. 1313–1336, 2022.

[13] C. Wang, H. Yang and C. Meinel, "Image captioning with deep bidirectional LSTMs and multi-task learning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 2 s, pp. 1–20, 2018.

[14] S. Das, L. Jain and A. Das, "Deep learning for military image captioning," in *2018 21st Int. Conf. on Information Fusion (FUSION)*, Cambridge, pp. 2165–2171, 2018.

[15] 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.

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

[17] T. D. C. Nogueira, C. D. N. Vinhal, G. D.C. Júnior, M. R. D. Ullmann and T. C. Marques, "A Reference-based model using deep learning for image captioning," *Multimedia Systems*, 2022. https://doi.org/10.1007/s00530-022-00937-3.

[18] M. Omri, S. A. 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, pp. 288, 2022.

[19] A. Oluwasammi, M. U. Aftab, Z. Qin, S. T. Ngo, T. V. Doan *et al.,* "Features to text: A comprehensive survey of deep learning on semantic segmentation and image captioning," *Complexity*, vol. 2021, pp. 1–19, 2021.

[20] K. Radhika, K. Devika, T. Aswathi, P. Sreevidya, V. Sowmya *et al.,* "Performance analysis of NASNet on unconstrained ear recognition," in *Nature Inspired Computing for Data Science, Studies in Computational Intelligence Book Series*, Cham: Springer, vol. 871, pp. 57–82, 2020.

[21] H. Nguyen and X. N. Bui, "A novel hunger games search optimization-based artificial neural network for predicting ground vibration intensity induced by mine blasting," *Natural Resources Research*, vol. 30, no. 5, pp. 3865–3880, 2021.

[22] A. Sagheer and M. Kotb, "Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems," *Scientific Reports*, vol. 9, no. 1, pp. 19038, 2019.

[23] M. Khishe and M. Mosavi, "Chimp optimization algorithm," *Expert Systems with Applications*, vol. 149, pp. 113338, 2020.

[24] M. Kaur, R. Kaur, N. Singh and G. Dhiman, "SChoA: A newly fusion of sine and cosine with chimp optimization algorithm for HLS of datapaths in digital filters and engineering applications," *Engineering with Computers*, 2021. https://doi.org/10.1007/s00366-020-01233-2.

[25] N. H. Phan, V. D. T. Hoang and H. Shin, "Adaptive combination of tag and link-based user similarity in flickr," in *Proc. of the Int. Conf. on Multimedia-MM '10*, Firenze, Italy, pp. 675, 2010.

[26] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 652–663, 2017.

[27] 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.