Journal of Intelligent Systems and Internet of Things
Vol. 18, No. 02, PP. 273-289, 2026
DOI: https://doi.org/10.54216/JISIoT.180219

Advanced Deep Learning Model for Image Captioning Using Customized Vision Transformer with Global Optimization Algorithm

Suleman Alnatheer [1*], Mohammed Altaf Ahmed[1]
[1]Department of Computer Engineering, College of Computer Engineering & Sciences, Prince Sattam bin Abdulaziz University, Alkharj-11942, Saudi Arabia
Emails: s.alnatheer@psau.edu.sa; m.altaf@psau.edu.sa

Abstract

In the image-captioning field, the excellence of produced captions is vital for the effectual interaction of visual content. Image Captioning is the main task, which unites computer vision (CV) and natural language processing (NLP), where it goals to produce graphic legends for images. A dual-fold procedure depends on precise image perception and alters language understanding both semantically and syntactically. It is gradually challenging to stay up with the modern study and consequences in image captioning owing to the developing amount of knowledge accessible on the topic. This analysis examines into deep learning (DL) to tackle the tasks challenged by individuals with graphic impairments, targeting to improve their visual insight via advanced technologies. By tradition, the visually impaired have trusted physical support and adaptive helps for understanding and navigating visual content. With the beginning of DL, there is a unique chance to develop this scenery. In this paper, we offer an Advanced Deep Learning Method for Image Captioning Based Using Customized Transformer with a Global Optimization Algorithm (ADLIC-CTGOA). The foremost aim of ADLIC-CTGOA model is to focus on the initiation of the effectual textual image captioning of an input image. Initially, the ADLIC-CTGOA method employs preprocessing phase to enhances both image and text data: images undergo noise removal and contrast enhancement to improve quality, while text is processed by removing numbers, converting to lowercase, and text vectorization. Next, the customized swin transformer is employed for feature extraction to capture fine-grained visual features from images. In addition, the BERT Transformer model is deployed for image captioning process. To enhance the performance of proposed technique, the chaotic Aquila optimization (CAO) technique was applied for parameter tuning for enhancing the performance. A wide sort of simulation studies are executed to ensure the improved performance of ADLIC-CTGOA system. The comparative result exploration reported the betterment of the ADLIC-CTGOA model on recent approaches in terms of different evaluation measures.

Received: March 18, 2025 Revised: June 10, 2025 Accepted: August 03, 2025

Keywords: Image Captioning; Global Optimization Algorithm; Swin Transformer; BERT; Natural Language Processing; Deep Learning

1. Introduction

In quickly developing the digital environment, growing the connections between both Natural Language Processing (NLP) and Computer Vision (CV) has given growth to transformative inventions. The intersection of NLP and CV is a difficult challenge for producing descriptive text from picture captioning, or visual inputs [1]. Substantial developments in this domain have come from the natural language comprehensive models and grouping of image processing. An essential component of visual knowledge in image captioning involves comprehending images and offering a natural language report of them. Image captioning is an extensive challenge

in CV and NLP, which allows to accomplishment of the multimodal transformation from imagery to text [2]. A major resource of data, several images were transferred and stored numerically on the Internet. At the same time, social communication mainly based on natural language, which allows the system to describe the visual world for child education, visually impaired individuals, natural human-computer interaction, and retrieval of information [3]. Depending on an input image, this method repeatedly creates the textual descriptions. The domain of artificial intelligence (AI), automatically created image descriptions captivated substantial attention [4].

Recent investigation expressed that transformers might be effectively employed for vision challenges with competitive performance [5]. Diverse from CNNs that collect global information by loading several convolutions, vision transformers (ViTs) taking benefit of the self-attention mechanism to take non-local dependency, and spatial patterns [6]. This enables ViTs to collect valuable global information without handcrafting layer-wise local feature extraction of CNNs and consequently attains higher performance. ViT is employed to remove important image characteristics. The self-attention device permits the method to consider relations among diverse kinds of images, allowing it to represent and understand difficult visual information [7]. Alternatively, ViTs can accomplish much greater performance than convolution-based methods trained with huge volumes of data.

Several studies were dedicated to automated image captions; also it is characterized into numerous methods. The image captioning method is specified from the same images using captioning, and the retrieval-based approach recognizes visually comparable imageries utilizing the captioning from training datasets [8]. Many kinds of examination are depend on Deep Learning (DL) and Machine Learning (ML) approaches. A Deep Neural Networks (DNN) method has been employed for the image captioning method because of its effective approximation competencies [9]. The image captioning approach has immensely grown up owing to the significant growth of the DNN method. Over recent times, CNN has acquired substantial attention in CV challenges like object detection, and image classification. Moreover, Recurrent Neural Network (RNN) acts a vital role in NLP. Although numerous kinds of investigations were organized, yet needed to determine effective image captioning approaches for enhanced performance [10].

This study offers an Advanced Deep Learning Method for Image Captioning Based Using a Customized Transformer with Global Optimization Algorithm (ADLIC-CTGOA). Initially, the ADLIC-CTGOA method applies pre-processing phase to enhances both image and text data: images undergo noise removal and contrast enhancement to improve quality, while text is processed by removing numbers, converting to lowercase, and text vectorization. Next, the customized swin transformer is applied for extraction of feature to arrest fine-grained visual features from images. In addition, the BERT Transformer system has been deployed for image captioning process. To improve the model's performance, the chaotic Aquila optimization (CAO) system has been applied for parameter tuning for enhancing the performance. A wide sort of simulation analyses are applied to safeguard the heightened performance of ADLIC-CTGOA technique.

2. Related Works

In [11], the Multi-lingual Voice-Based Image Caption Generator (MVBICG) method was intended to offer real-world image descriptions for types of voice in several languages according to consumer needs. By the usage of MVBICG, the depictions might be attained as voice output in several dialects. It controls the recent improvements in DL, mainly CNN for image extraction of feature and RNN with attention systems for natural language generation. Safiya and Pandian [12] presents an innovative structure relating CV with voice-based image captioning utilizing DL methods. The presented method applies LSTM for NLP and the VGGNet-16 method for image processing. A two-fold method is utilized; the initial stage is involved to expose an input image to pre-processing utilizing a pre-established VGGNet16 method. Bayisa et al. [13] present a single structure, which can manage several challenges in CV with fine-tuning competencies for language tasks and other specific vision. The projected method applies an adjusted DenseNet201 as a feature extraction for the network backbone, task-specific head for implication, and an encoder-decoder structure. To enhance precision and tackle overfitting, improved normalization methods, and data augmentation are applied.

Solomon and Abebe [14] present a hybridized attention-based DNN method. This approach contains an Inception v3 CNN encoding to remove image characteristics, a visual attention device to take substantial characteristics, and a Bi-directional GRU (BiGRU) with attention for decoding to create the image caption. Wasi et al. [15] implements a hybrid suggested method over Sentiment Analysis (SA) for longer textual sentences. Social media is a huge resource of thoughts, which can be employed over SA utilizing DL methods, enhancing suggested methods, and overcoming language difficulties. This study addresses tasks in Bangla SA, like the linguistic nuances and scarcity of datasets, presenting a method that associates Bi-LSTM, CNN, and LSTM for heightened classification of text sequence. Safiya and Pandian [16] implements a new structure that incorporates CV with voice-based photo captioning through DL methods. In this study, 3 performance methods was organized, ResNet-152, MobileNet-V3, and VGGNet-16 to recognize the highest performing method for hardware utilization. These methods are

proficient for utilizing the Flickr8k, 2k, and Flickr30k custom databases. Afterward, the performed VGGNet16 method has been employed on a Raspberry Pi 4B with a GPU.

Cao et al. [17] aimed at image captioning depending on LSTM, and CNN utilizing ViT or DenseNet201 trained on LSTM as the decoder and the ImageNet1K image classification database as an image encoding. Pre-training has been achieved on the Flickr8k database. Afterward choosing the best method as pre-trained weight method for COCO database, tuning optimizer has achieved the COCO database, and an attention mechanism were utilized to intend the ablation test. Kim et al. [18] present a CT image captioning method, which employs a distilGPT-2, and 3D-CNN method. In this research, 4 groups of language and 3D-CNN methods were analysed and compared for their performance. Moreover, employing the loss function and altering penalty values through training has been investigated.

3. Proposed Methods

In this study, we offers an ADLIC-CTGOA system. The foremost purpose of the ADLIC-CTGOA system is to focus on the group of an effectual textual image captioning of an input imageries. It contains distinct kinds of processes involved as pre-processing of image and text, swin transformer based feature extraction, image captioning using the BERT transformer model, and global optimization algorithm using CAO. Fig. 1 represents the entire process of ADLIC-CTGOA technique.

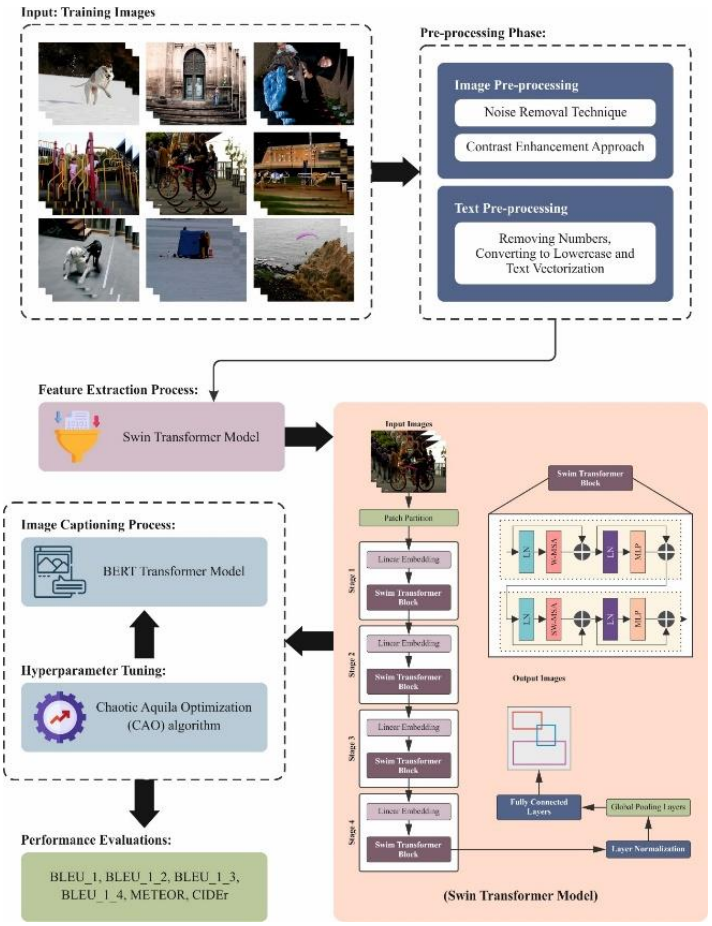

Figure 1. Overall Process of ADLIC-CTGOA model

A. Preprocessing of Image and Text

Initially, the ADLIC-CTGOA method applies the pre-processing phase enhance both image and text data.

i) Image Pre-processing

Noise Removal: Gaussian Filter

The process is completed by building the image with a Gaussian function, which delivers better weights to pixels nearer to the focal point of the convolution kernel and smaller weights to those furthest away [19]. These mains to

a phenomenon called as a blurring effect, which decreases the occurrence of higher-frequency noise and trivial fluctuations in pixel intensity. Gaussian filters are very effective when declining random noise and protecting edge integrity is essential. So, they are main in tasks like image improvement, edge recognition, and making for additional analytic phases such as feature extraction or segmentation. Eq. (1) displays the 2D Gaussian function.

$$Gauss(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

Contrast Enhancement: CLAHE

This approach is based on the concept that an image is divided into numerous non-overlapping regions around equivalent size. Tile calls are the element of operation for CLAHE rather than the whole image [20]. To change the contrast color of every tile, an adaptable method has been applied to define the task of the present state of the histogram. The histogram of the estimated value of the region develops using the contrasted tile counts. To prevent artificial barriers from making, nearby tiles were blended. The contrast should be decreased mainly in homogenous regions to avoid the amplification in an image. Gray scale histogram equalization on image x, whereas z refers to grayscale value among (0, 256). How to search probabilities at pixel image from level z in an image as:

$$P_x(z) = P_{(x-z)} = \frac{n_Z}{n_0} \rightarrow 0 \leq z < V \quad (2)$$

The value of V characterizes the complete intensity range of the image (256), n symbolizes the group's total number of image pixels, and $Px(z)$ denotes the index signifying the image histogram pixels. Tiles (of dimensions 8x8) have been applied to divide the image into smaller halves, and the histogram equalization procedure was used for all pairs of nearby tiles. The results that are collected in a focused region to make the histogram will used in that region. Limiting the contrast to reduce the noise was significant. Formerly applied histogram equalization, some bins whose contrast is greater than the threshold have been eliminated and reallocated uniformly to another bin; thereafter, bilinear interpolation was employed to remove artifacts in the border of tiles. After utilizing the optimizer method for every segmentation image, 3 images ($I_n, n = 1,2\ and\ 3$) have been formed and joined to make one image based on the relationship.

$$I_T(x,y) = CLAHE\big(I_1(x,y)\big) + CLAHE\big(I_2(x,y)\big) + CLAHE\big(I_3(x,y)\big) \quad (3)$$

The CLAHE method has been used again for the image. This method permits improving contrast by

$$I_{Te}(x,y) = CLAHE\big(I_T(x,y)\big) \quad (4)$$

ii) Text Pre-processing

Text data should be cleaned and altered into a format appropriate for ML tasks, like sentiment analysis or classification [21]. This is accomplished over the below mentioned steps:

- Removal of numbers: Any numeral values do not donate to the semantic meaning of the text are eliminated. This makes sure that the technique concentrates on the text's linguistic content.
- Lower Case Conversion: Transforming every text to a small letter regulates an input, removing differences caused by capitalization. For instance, "Machine" and "machine" would be preserved the similarly.
- Text Vectorization: This procedure converts text into a statistical format, which can be treated by ML techniques. General models contain:
- Word Embedding: Every word in the text is drawn to a vector of actual numbers, acquiring semantic relations among words. Pre-trained embedding's such as Word2Vec or GloVe are generally employed.
- TF-IDF: A numerical model that weighs the significance of words depend upon their frequency and reverse document frequency.
- Bag of Words: Signifies text as a matrix of word counts, where every document is denoted as vector.

This is classically followed by positioning parameters like a highest vocabulary dimension (MAX_VOCAB = 10,000) that limits the amount of single tokens measured, and a maximum sequence length (MAX_SEQ_LEN = 25), which safeguards even input dimensions across the database.

For images, an input size is usually standardized (IMG_SIZE = 224) to fit the technique's needs. Furthermore, the embedding size (EMBEDDING_DIM = 512) describes the dimension of the feature vector, acquiring rich semantic data. Appropriate pre-processing certifies the effective performance of model, decreasing the computational load and improving predictive accuracy.

B. Feature Extraction using Swin Transformer

Next, the customized swin transformer was employed for feature extraction to capture fine-grained visual features from images. The ST is an inspiring AI method tailored to computer vision (CV) [22]. It relies on the Transformer

method and presents dual main ideas shifted window attention mechanism and hierarchical feature maps. These progressions assistance effectively processing significant scale image data, forming it a promising device for difficult CV tasks. ST uses hierarchical feature maps to efficiently characterize various stages of characteristics in images, resulting in complete context understanding and an enhanced understanding of an input data.

The ST's four-step structure includes separating an input image into patch layers consisting of managed by transformer blocks completely. The resultant patches are transferred to the block of transition, keeping a similar amount of patches. During this next phase, layers of patch merging were applied to make a hierarchical method by sub-sampling and decreasing the token counts. Neighbourhood 2x2 features of the patches are joined to gain a 4C dimensional feature vector, which can be transformed with linear layers while maintaining a solution of $H/8$x$W/8$. This feature transformation and patch merging method is twice repeated in the following phases, leading to an output resolutions of $H/16$x$W/16$ and $H/32$x$W/32$, individually. In general, this framework allows the ST to efficiently handle image information and take contextual data at various scales, taking part in its greater performance in different tasks of vision.

The ST Blocks (STBs) contain dual successive multi-head self-attention (MSA) components: shifted window-based MSA ($SW-MSA$) and $W-MSA$. Formerly all this MSA unit, the Layer Norm (LN) layer has been applied. Then, there is a dual-layer multi-layer perceptron (MLP) using GELU nonlinearity in the middle. All modules contain a connection using the layer of LN. During Eqs. (5) & (6), MSA contains a quadratic computation efficiency concerning the token counts. This configuration considerably increases the ST performance and makes it more effective in comparison with the normal Transformer.

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{5}$$

$$\Omega(W-MSA) = 4hwC^2 + 2M^2hwC \tag{6}$$

The initial section shows a quadratic link concerning the patch quantity, represented as hw, while the next section establishes a linear dependence if the M value is continuous (normally set to 7 naturally). Calculating global self-attention turns out to be excessively costly for a higher value of hw, while WSA is flexible.

During these successive STBs, The SW separating method can be implemented to changeover among dual configurations. This method uses overlapping windows for presenting cross-window links effectively computing non-overlapping windows. During this initial module, a regular window dividing tactic has been utilized, and an 8x8 feature mapping is separated into 2x2 size windows 4x4($M=4$). Subsequently, the next component gives a window configuration by transferring the window by $\left(\left\lfloor\frac{M}{2}\right\rfloor, \left\lfloor\frac{M}{2}\right\rfloor\right)$. The blocks of the Transformer are calculated in Eq. (7)

$$\begin{aligned} \hat{z}^l &= W-MSA\left(LN(z^{l-1})\right) + z^{l-1}, \\ z^l &= MLP\left(LN(\hat{z}^l)\right) + \hat{z}^l, \\ \hat{z}^{l+1} &= SW-MSA\left(LN(z^l)\right) + z^l, \\ z^l &= MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \end{aligned} \tag{7}$$

Here z^l and $\hat{z}^l$ characterize output features for lth block from $(S)W-MSA$ and MLP components, correspondingly. $W-MSA$ and $SW-MSA$ signify $W-MSA$ with standard and shifted window partitioning patterns, correspondingly.

The ST accepts a specific structure to improve computational complexity in comparison with conventional Transformer methods. It attains these utilizing cyclic shifting operations among moved token blocks. Therefore, every block may work with masks utilized for feature maps. This method permits the ST to procedure reduced blocks of information rather than the complete feature maps directly, resulting in more effective extraction of features and avoiding computational complexity in sliding window.

The ST uses a self-attention mechanism, which integrates relative positioning bias to take relationships amid locations. The function of attention includes mapping values (V), queries (Q), and keys (K), to output vectors. The output matrix has been gained over this calculation procedure, which can be expressed in Eq. (8)

$$Attention\ (Q,K,V) = SoftMax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{8}$$

Here the value (V), query (Q), and key (K) matrices are of size R^{M^2xd}, while d characterizes the size, and M^2 denotes patches count in a window. During ST, related positions are described along with every axis inside the

range $[-M+l, M-1]$. The relative positioning can be parameterized as a balanced matrix $\hat{B}C\ R^{(2M-1)x(2M-1)}$, and the matrix elements B were gained from $\hat{B}$.

C. Image Captioning Using BERT Transformer Model

In addition, the BERT Transformer model is deployed for an image captioning procedure. This technique presented an AM in the study of NLP and carried revolutionary changes [23]. Moreover, the AM transformer contains the encoding and decoding. This encoding accompanies several auto-regressive stages. The constant sequence $z(z_1,..,z_n)$ has been gained afterward maps an input $(x_1,..,x_n)$ and helps to produce the output $(y_1,..,y_m)$ utilizing the decoding. Dual important models of transformers are shown below:

The Stacks of Encoder-Decoder: This encoding stack of a transformer has $N = 6$ layers, which include dual sub-layers. They have the responsibility to maintain the multiple head self-attention and a feedforward system namely point-by-point fully connected. The dimension of the sublayers is 512 and is Layer Norm $(x + \text{Sublayer}\ (x))$.

Multi-Head Attention: This method contains 3 various tasks wants to be implemented. When the decoder output layers are given to the following encoding layers. Formerly the layer of self-attention makes the values, queries, and memory key, utilizing an output of the preceding layer of an encoder result. The decoder utilizes the auto-regressive way and hides at the Softmax input. Using the values and keys in dimension d_k and the queries, this self-attention has been computed. Eventually, afterward-masking input of the Softmax, this decoder keeps the autoregressive properties of scaled dot product attention.

BERT is a transformer-based approach pre-trained on the biggest corps of English data utilizing a self-supervised method [24]. This pre-training comprises no human labeling, utilizing raw texts. In cyberbullying recognition context, the BERT model is pre-training is mathematically labelled through its dual major goal. In MLM, S is selected, and nearly 15% of words are arbitrarily masked. Assume $S = \{w_1, w_2, \dots, w_N\}$ be a sentence using N words.

$$L_{MLM}(S, S') = -\sum_{i=1}^{N} m_i \log P(w_i|S') \quad (9)$$

Now, m_j represents the indicator function, which can be 1 when the word w_j is masked and if not 0. This method varies from conventional recurrent neural networks (RNNs), which procedure words consecutively, and auto-regressive methods namely GPT, which mask upcoming tokens. During NSp, this model connects dual sentences, A and B, in pre-training. Assume Concat (A, B) signify the concatenation of Sentences A and B, and y remain a binary label. The NSP objective function is:

$$L_{NSP}(A, B) = -\log P(y|Concat(A, B)) \quad (10)$$

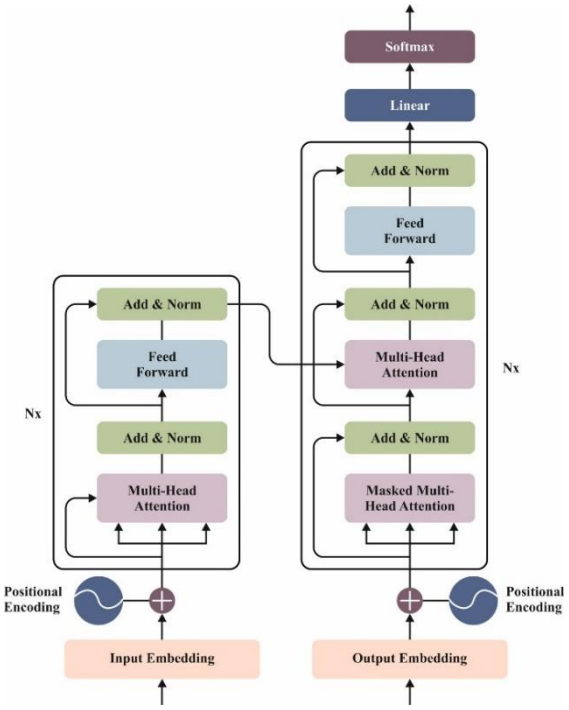

Figure 2. Structure of BERT Transformer Model

This capability is important in perception the conversation flow in online connections, which represents basic element in recognizing samples of cyberbullying. BERT successfully learns composite language patterns that is important for cyberbullying recognition. For the downstream task, the feature vector $F(S)$ removed from BERT for sentence S is applied.

$$L_{cyberbullying}(S, label) = Loss\big(C\big(F(S)\big), label\big) \quad (11)$$

Now, $label$ specifies if the sentence S has cyberbullying content. By adjusting BERT on cyberbullying data, this method studies to distinguish the subtle language signals of online harassment. The BERT approach experiences dual main stages: pre-training and fine- tuning. During two phases, this method uses related structures, apart from the output layers. The pre-training stage includes utilizing a set of primary model parameters. Once it arrives at adjustment, each of these parameters are accurately fine-tuned to enhance performance. Fig. 2 represents the structure of the BERT transformer model.

D. Global Optimization Algorithm using CAO

To optimize the model's performance, the CAO algorithm is applied for parameter fine-tuning to enhance performance. The current optimizer for the population-based swarm intelligent algorithm is the AO method [25]. The most famous predatory bird, which formerly existed in the northern hemisphere. The Aquila's back and body are golden in color. Aquila hooks various prey, mostly rabbits, hares, marmots, and squirrels, utilizing her strength and speediness, in addition to its stronger feet and wider claws. The AO mimics the 4 different searching methods and these methods are mathematically shown:

Stage 1: Extended exploration (X_1): The Aquila flies higher beyond the surface of the earth in this stage for properly scanning the region before jumping upright once it has identified the prey. These behaviors can be formulated in dual mathematical expressions as demonstrated:

$$X_1(l+1) = X_B(l) * \left(1 - \frac{l}{L}\right) + (X_{mean}(l) - X_B(l) * rand) \quad (12)$$

$$X_{mean}(l) = \frac{1}{M}\sum_{i=1}^{M} X_i\,(l), \forall M = 1,2,\dots,Dim \quad (13)$$

Whereas $X_{mean}(l)$ signifies the mean position of the current solution at $i\,th$ iteration by utilizing Eq. (13), X_B denotes the globally optimal solution in these iterations, $rand$ denotes randomly generated values, which exist in [0, 1], l represents the present iteration, L denotes an iteration count, M definite the size of the population, Dim state the size of the dimension.

Stage 2: Narrowed exploration (X_2): The common searching techniques of Aquilas include this specific stage. Contour flying has been attached with a short glide for attacking the prey. The succeeding Eq. (14) is upgraded to Aquila locations:

$$X_2(l+1) = X_B(l) * Levy(D) + X_R(l) + (y - x) * rand \quad (14)$$

Here X_R direct the arbitrary position of the Aquila, D signifies the dimensional area, Levy signifies the function of the levy probability distribution that was assessed by using Eqs. (15) to (18).

$$Levy(D) = s * \frac{ll * \sigma}{|v|^{\frac{1}{\beta}}} \quad (15)$$

$$\sigma = \frac{\Gamma(1+\beta) * sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) * \beta * 2^{\frac{\beta-1}{2}}} \quad (16)$$

Now, s and β represent the constant among 0.01 and 1.5, correspondingly. u and v means a random number exists in (0, 1). The succeeding dual equations are applied to compute the y and x values. Its mathematical formulation is expressed below:

$$y = \gamma * \cos(\theta) \quad (17)$$

$$x = \gamma * sin(\theta) \quad (18)$$

Here γ and θ are defined by utilizing Eqs. (19), (20), and (21).

$$\gamma = \gamma_1 + V * D_1 \quad (19)$$

$$\theta = -W * D_1 + \theta_1 \quad (20)$$

$$\theta_1 = \frac{3 * \pi}{2} \tag{21}$$

Whereas γ_1 captures a value among (1, 20); V and W are equivalent to 0.0265 and 0.005, respectively; and D_1 represents the randomly produced number in the interval of 1 to the dimensions.

Stage 3: Expanded exploitation (X_3): this stage includes finding the prey position such that agents can introduce low-flying preemptive strikes up and down. Next are some possibilities, of which agents can assault their prey:

$$X_3(l+1) = \left(X_B(l) - X_{mean}(l)\right) * \alpha - rand + \left((\mathcal{UB} - \mathcal{LB}) * rand + \mathcal{LB}\right) * \delta \tag{22}$$

Whereas α and δ are exploitation set parameters, and $\mathcal{LB}$ and $\mathcal{UB}$ signifies the lower and upper boundaries.

Stage 4: Narrowed exploitation (X_4): This stage includes the capability of the Aquila to rapidly track and attack its target utilizing escaping path lights that are computed by using Eqs. (23), (24), (25), and (26).

$$X_4(l+1) = QF * X_B(l) - (P_1 * X(l) * rand) - P_2 * Levy(D) + rand * P_1 \tag{23}$$

$$QF(l) = l^{\frac{2*rand-1}{(1-\gamma)^2}} \tag{24}$$

$$P_1 = 2 * rand - 1 \tag{25}$$

$$P_2 = 2 * \left(1 - \frac{l}{L}\right) \tag{26}$$

Whereas $QF(l)$ signifies the quality factor, P_1 represents the different signals of AO, and P_2 signifies the pursued target flighting slopes.

In recent decades, most metaheuristic methods have applied arbitrary parameters. This random parameter depends on likelihood distributions that are frequently Gaussian or uniform. In recent times, the chaotic concept has been employed to increase this parameter. Chaos is a famous phenomenon. Some changes in the initial chaos point might lead to nonlinear changes in upcoming behaviours. Chaos can be described by 3 main things: sensitivity, ergodicity, and quasi-stochastic to beginning conditions. Quasi-stochastic behaviour is the capacity for replacing variables, which are randomly using chaotic mapping value. The capability of chaotic variables for searching non-repeated for every state inside the particular limit is normally stated as the ergodicity feature. Lastly, regard to beginning conditions can be designated as some slight modification in the initial beginning points, which might lead to various behaviours. Joining both of these features contains the possibility to intensely enhance the efficacy of metaheuristic methods.

Here, we define chaotic maps, which have been applied to make chaotic groups. This collection of maps using a primary point of 0.7 is selected to establish a diverse behaviour. The initial point might be some number among (0, 1). While z_k indicates the $k\,th$ quantity. The additional parameters contain $e, f,$ and μ, are controller parameters. This parameter is applied to control the dynamical chaotic behavior systems.

This section presents the new CAO, which utilizes chaotic mapping for utilizing for chaotic random variables. This performance influences the novel AO key parameter. The presented CAO method pseudocode is exposed in Algorithm1. The mathematical expression of the presented CAO model can be defined as shown:

$$X_1(l+1) = X_B(l) * \left(1 - \frac{l}{L}\right) + (X_{mean}(l) * R_c - X_B(l) * rand) \tag{27}$$

$$X_3(l+1) = (X_B(l) - X_{mean}(l) * R_c) * \alpha - rand + ((\mathcal{UB} - \mathcal{LB}) * rand + \mathcal{LB}) * \delta \tag{28}$$

Whereas R_c denotes chaotic random variable. Eqs. (27) and (28) represent upgraded formulations of the presented CAO model.

Algorithm 1: Pseudo-code of the presented CAO
Start
Input: The Optimization problem information
Initializing the population X of the AO.
Initializing the chaotic maps.
Initializing AO parameters.
Initializing CAO parameters.

while $l < L$ do

Calculate the value of the fitness function.

Selecting the optimal candidate solution X_B.

for $(i = 1,2, \ldots, M)$ do

Updated the current mean position.

Calculate the parameters.

if $\left(l \leq \left(\frac{2}{3} * L\right)\right)$ then

if $rand \leq 0.5$ then

Updated current position using Eq. (27).

else

Updated current position utilizing Eq. (14).

if $rand \leq 0.5$ then

Updated current position utilizing Eq. (28).

else

Updated current position utilizing Eq. (23).

end if

end if

end if

end for

end while

Return optimal solution.

End CAO

The fitness selection is the considerable feature persuading the efficacy of the CAO model. The hyperparameter range method holds the solution-encoding model to value an effectiveness of the candidate solution. Here, the CAO system reflects accuracy as general principle to construct the function of fitness. Its mathematical expression is formulated below:

$$Fitness = \max(P) \tag{24}$$

$$P = \frac{TP}{TP + FP} \tag{25}$$

Here, TP denotes the true and FP represents the false positive values.

4. Performance Analysis

The performance evaluation of ADLIC-CTGOA algorithm is verified using dual datasets namely Flickr8k [26] and Flickr30k [27]. The Flickr8k dataset consists of 8000 size and the Flickr30k dataset has 31000 size, which is depicted in Table 1. Fig. 3 signifies the sample images. Fig. 4 depicts the sample of actual and predicted captions.

Table 1: Details of Dataset

Dataset	Size
"Flickr8k"	"8000"
"Flickr30k"	"31000"

Figure 3. Sample Images

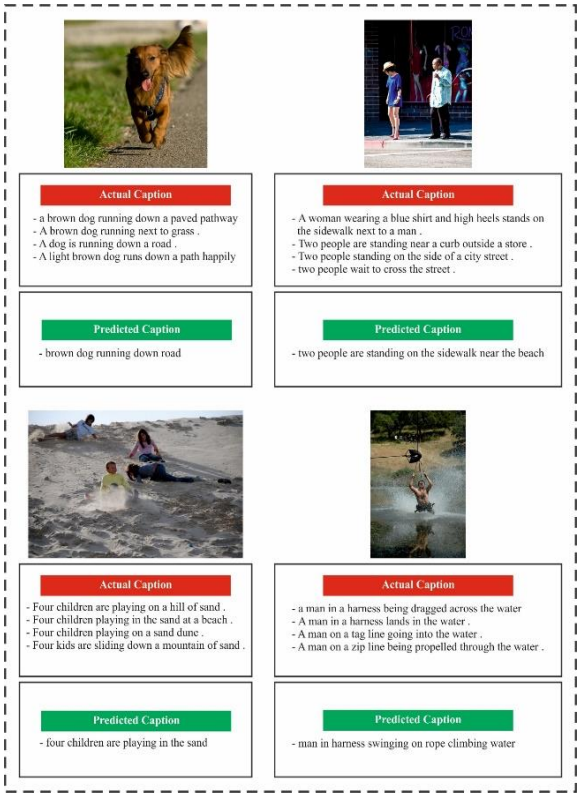

Figure 4. Sample of Actual and Predicted Caption

Table 2 represents the comparative analysis of the ADLIC-CTGOA model on Flickr8k dataset with existing models [28, 29].

Fig. 5 provides the comparison study of the ADLIC-CTGOA model on Flickr8k dataset with existing models in terms of BLEU_1, BLEU_1_2, BLEU_1_3, and BLEU_1_4. The table values indicate that ADLIC-CTGOA model has got effectual performance. According to BLEU_1, the ADLIC-CTGOA model has gained maximum BLEU_1 value of 74.25%, while the NIC model, Soft-Attention technique, Hard-Attention system, SCA-CNN-VGG, CNN, LSAHCNN-ICS, and AIC-SSAIDL models have obtained lesser BLEU_1 values of 58.58%, 60.88%, 63.00%, 65.55%, 67.83%, 70.48%, and 72.62%, respectively. Moreover, according to BLEU_1_2, the ADLIC-CTGOA system has attained enhanced BLEU_1_2 value of 74.25%, while the NIC system, Soft-Attention technique, Hard-Attention model, SCA-CNN-VGG, CNN, LSAHCNN-ICS, and AIC-SSAIDL models have obtained lesser BLEU_1_2 values of 43.61%, 45.52%, 46.95%, 50.41%, 52.31%, 53.97%, and 56.84%, respectively. Besides, according to BLEU_1_4, the ADLIC-CTGOA algorithm has reached maximal BLEU_1_4 value of 33.56%, while the NIC technique, Soft-Attention system, Hard-Attention method, SCA-CNN-VGG methodology, CNN, LSAHCNN-ICS, and AIC-SSAIDL models have obtained lesser BLEU_1_4 values of 18.68%, 20.57%, 23.56%, 24.76%, 26.93%, 29.83%, and 32.06%, respectively.

Table 2: Comparative study of ADLIC-CTGOA system on Flickr8k dataset with existing methods

Flickr8K Database						
Methods	BLEU_1	BLEU_1_2	BLEU_1_3	BLEU_1_4	METEOR	CIDEr
NIC	58.58	43.61	32.62	18.68	14.78	31.65
Soft-Attention	60.88	45.52	34.82	20.57	16.79	34.23
Hard-Attention	63.00	46.95	36.82	23.56	18.47	37.01
SCA-CNN-VGG	65.55	50.41	39.73	24.76	21.74	38.63
CNN Model	67.83	52.31	41.78	26.93	24.36	41.54
LSAHCNN-ICS	70.48	53.97	44.01	29.83	26.98	43.94
AIC-SSAIDL	72.62	56.84	46.40	32.06	30.16	46.43
ADLIC-CTGOA	74.25	58.59	48.08	33.56	31.96	48.00

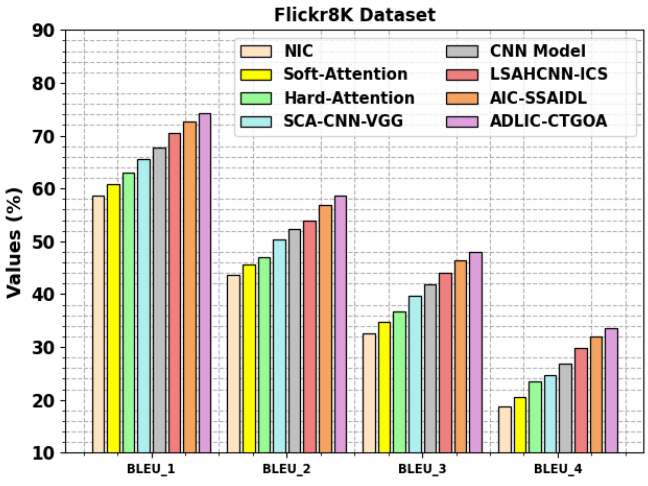

Figure 5. BLEU_1, BLEU_1_2, BLEU_1_3, and BLEU_1_ analysis of ADLIC-CTGOA model on Flickr8k dataset

Fig. 6 provides the comparison study of the ADLIC-CTGOA model on Flickr8k dataset using existing techniques in terms of METEOR, and CIDEr. The table values indicate that ADLIC-CTGOA model has effectual performance. According to METEOR, the ADLIC-CTGOA model attained enhanced METEOR value of 31.96%, while the the NIC system, Soft-Attention model, Hard-Attention method, SCA-CNN-VGG approach, CNN technique, LSAHCNN-ICS, and AIC-SSAIDL models have obtained lesser METEOR values of 14.78%, 16.79%, 18.47%, 21.74%, 24.36%, 26.98%, and 30.16%, respectively. Furthermore, based on CIDEr, the ADLIC-CTGOA model has obtained a higher CIDEr value of 31.96%, whereas the the NIC system, Soft-Attention model, Hard-Attention method, SCA-CNN-VGG approach, CNN technique, LSAHCNN-ICS, and AIC-SSAIDL models have obtained lesser CIDEr values of 31.65%, 34.23%, 37.01%, 38.63%, 41.54%, 43.94%, and 46.43%, respectively.

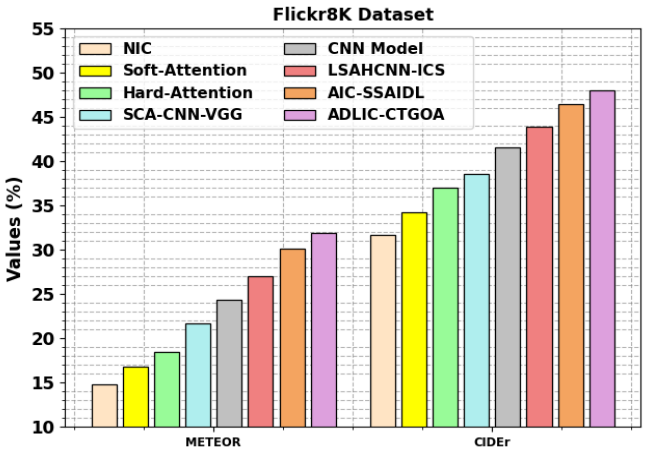

Figure 6. METEOR and CIDEr analysis of ADLIC-CTGOA model on Flickr8k dataset

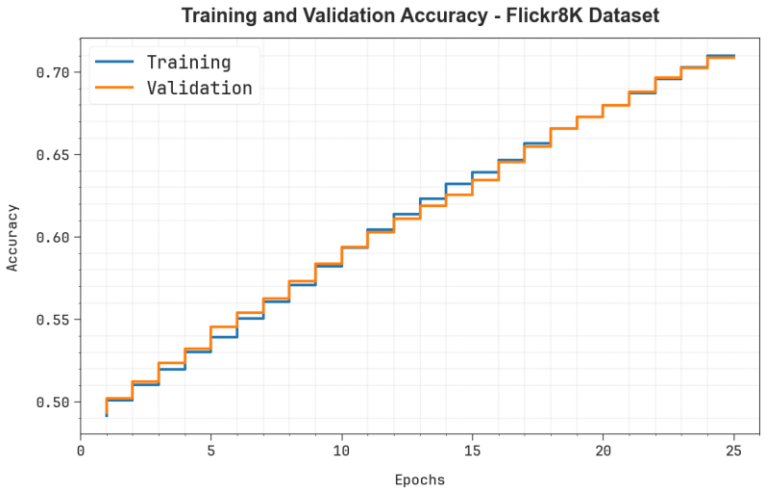

Figure 7. $Accu_y$ curve of ADLIC-CTGOA model on Flickr8k database

In Fig. 7, the training (TRAN) $accu_y$ and validation (VALD) $accu_y$ outcomes of the ADLIC-CTGOA method using the Flickr8k dataset is depicted. The values are computed through 0-25 epochs. The figure indicated that both values demonstrate an increasing trend through multiple iterations. In addition, the both values remain close through the epochs notifying lesser overfitting and showing higher performance of the ADLIC-CTGOA technique.

In Fig. 8, the TRA loss (TRALOS) and VAL loss (VALLOS) graph of ADLIC-CTGOA approach using Flickr8k database is shown. The values of loss are computed across 0-25 epochs. It is demonstrated that the values of both show a diminishing trend between data fitting and generalization. The continuous decrease securities the superior performance of the ADLIC-CTGOA system.

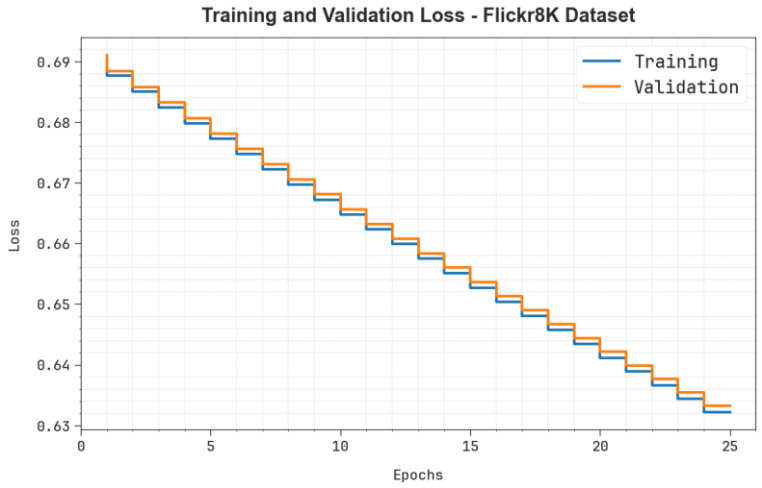

Figure 8. Loss curves of ADLIC-CTGOA model on Flickr8k database

Table 3 signifies the comparative study of the ADLIC-CTGOA model on the Flickr30k dataset with existing models in terms of BLEU_1, BLEU_1_2, BLEU_1_3, BLEU_1_4, CIDEr, and METEOR. The abbreviation of BLEU (Bilingual Evaluation Understudy), METEOR (metric for the evaluation of machine translation output), and CIDEr (Consensus-based Image Description Evaluation).

Fig. 9 provides the comparison study of ADLIC-CTGOA model on the Flickr30k dataset with existing techniques. The table values indicate that the ADLIC-CTGOA model has effectual performance. According to BLEU_1, the ADLIC-CTGOA approach has gained maximum BLEU_1 value of 73.30%, while the NIC system, Soft-Attention model, Hard-Attention method, SCA-CNN-VGG approach, CNN technique, LSAHCNN-ICS, and AIC-SSAIDL models have obtained lesser BLEU_1 values of 59.31%, 61.61%, 63.43%, 65.06%, 68.05%, 69.57%, and 71.74%, respectively. Additionally, according to BLEU_1_2, the ADLIC-CTGOA technique has attained enhanced BLEU_1_2 value of 64.45%, while the NIC system, Soft-Attention model, Hard-Attention method, SCA-CNN-VGG approach, CNN technique, LSAHCNN-ICS, and AIC-SSAIDL models have obtained lesser BLEU_1_2 values of 48.84%, 52.00%, 54.16%, 56.46%, 57.91%, 60.73%, and 62.90%, respectively. In addition, according to BLEU_1_3, the ADLIC-CTGOA method has reached improved BLEU_1_3 value of 55.75%, while the NIC system, Soft-Attention model, Hard-Attention method, SCA-CNN-VGG approach, CNN technique, LSAHCNN-ICS, and AIC-SSAIDL models have obtained lesser BLEU_1_3 values of 39.24%, 41.78%, 44.44%, 47.38%, 49.72%, 51.40%, and 54.13%, respectively.

Fig. 10 provides the comparison study of the ADLIC-CTGOA model on the Flickr30k dataset using existing techniques in terms of METEOR, and CIDEr. The table values indicate that the ADLIC-CTGOA model has effectual performance. According to METEOR, the ADLIC-CTGOA model has obtained maximum METEOR value of 36.93%, while the NIC, Soft-Attention technique, Hard-Attention system, SCA-CNN-VGG method, CNN technique, LSAHCNN-ICS approach, and AIC-SSAIDL models have obtained lesser METEOR values of 22.32%, 23.96%, 26.76%, 28.83%, 29.05%, 33.00%, and 35.42%, respectively. In addition, based on CIDEr, the ADLIC-CTGOA model has obtained higher CIDEr value of 63.88%, whereas the NIC model, Soft-Attention system, Hard-Attention technique, SCA-CNN-VGG approach, CNN methodology, LSAHCNN-ICS method, and AIC-SSAIDL models have obtained lesser CIDEr values of 38.03%, 39.96%, 42.89%, 45.84%, 56.83%, 59.86%, and 62.21%, respectively.

Table 3: Comparative study of ADLIC-CTGOA model on Flickr30k database with existing methods

Flickr30K Database						
Methods	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	CIDEr
NIC	59.31	48.84	39.24	28.12	22.32	38.03
Soft-Attention	61.61	52.00	41.78	29.65	23.96	39.96
Hard-Attention	63.43	54.16	44.44	31.57	26.76	42.89

SCA-CNN-VGG Model	65.06	56.46	47.38	34.36	28.83	45.84
CNN	68.05	57.91	49.72	36.62	29.05	56.83
LSAHCNN-ICS System	69.57	60.73	51.40	38.42	33.00	59.86
AIC-SSAIDL	71.74	62.90	54.13	41.50	35.42	62.21
ADLIC-CTGOA	73.30	64.45	55.75	43.12	36.93	63.88

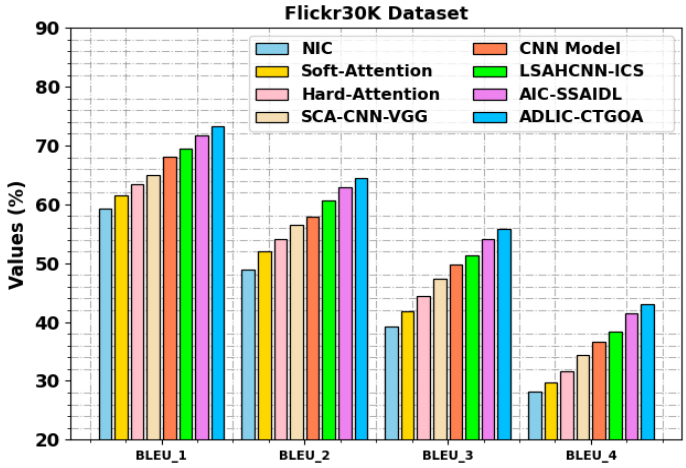

Figure 9. BLEU_1, BLEU_1_2, BLEU_1_3, and BLEU_1_ analysis of ADLIC-CTGOA model on Flickr30k dataset

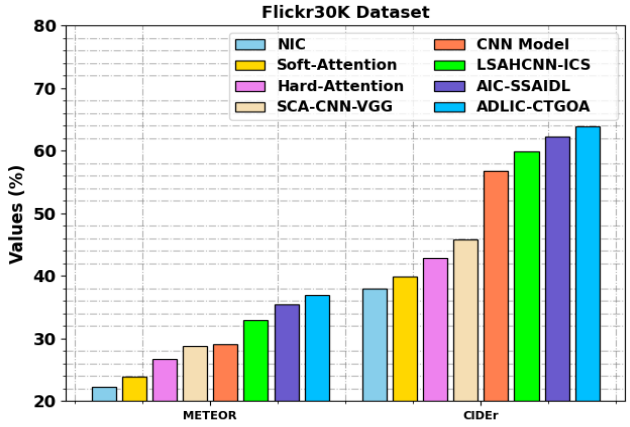

Figure 10. METEOR and CIDEr analysis of ADLIC-CTGOA model on Flickr30k dataset

In Fig. 11, the TRAN $accu_y$ and VALD $accu_y$ graph of ADLIC-CTGOA technique using the Flickr30k dataset is illustrated. The $accu_y$ values are computed across an interval of 0-25 epochs. The figure indicated that the values of both display an increasing movement indicating the capacity of the ADLIC-CTGOA technique. Likewise, the both values remain close over the epochs of the ADLIC-CTGOA system, assuring constant prediction on unseen samples.

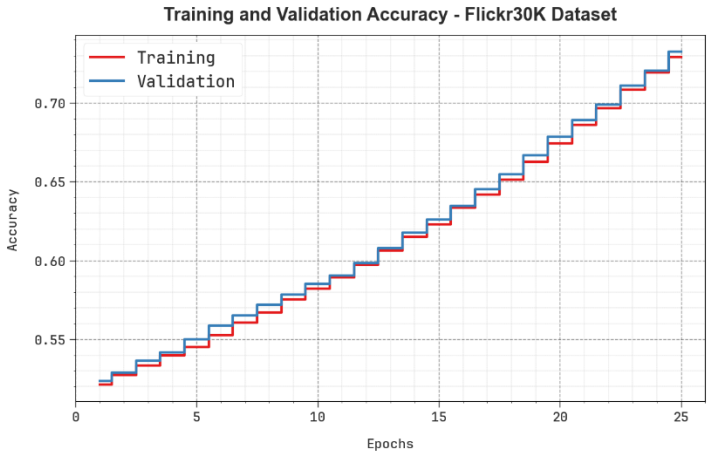

Figure 11. $Accu_y$ curve of ADLIC-CTGOA model on Flickr30k database

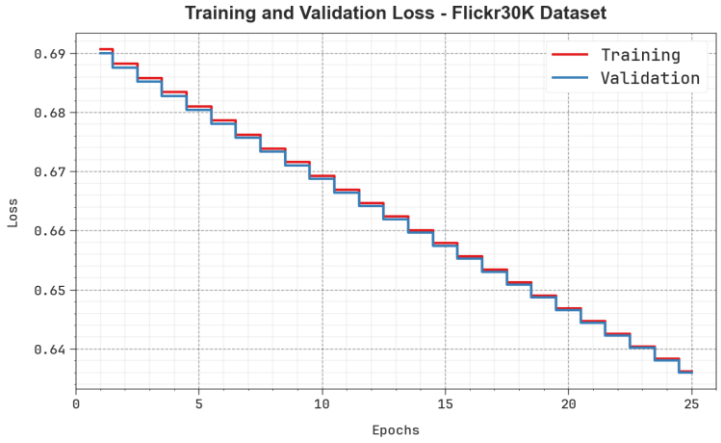

Figure 12. Loss curve of ADLIC-CTGOA model on Flickr30k database

In Fig. 12, the TRALOS and VALLOS graphs of ADLIC-CTGOA model using the Flickr30k database is demonstrated. The values of loss are computed through an interval of 0-25 epochs. It is exemplified that the values of both signify a reducing trend, which notified the competency of the ADLIC-CTGOA approach in harmonizing an equilibrium among data fitting and generalization. The persistent decrease assurances the superior performance of the ADLIC-CTGOA technique.

5. Conclusion

In this paper, we offer an ADLIC-CTGOA technique. The foremost aim of ADLIC-CTGOA system is to focus on the production of an effectual textual image captioning of the input images. It contains distinct kinds of processes involved as pre-processing of image and text, swin transformer-based feature extraction, image-captioning using the BERT transformer model, and global optimization algorithm using CAO. Initially, the ADLIC-CTGOA method applies a pre-processing phase that enhances both image and text data: images undergo noise removal and contrast enhancement to improve quality, while text is processed by removing numbers, converting to lowercase, and text vectorization. Next, the customized swin transformer is employed for feature extraction to capture fine-grained visual features from images. In addition, the BERT Transformer model is deployed for the image captioning process. To optimize the model's performance, the CAO algorithm is applied for hyperparameter tuning to enhance performance. A wide sort of simulation analyses are implemented to safeguard the improved performance of ADLIC-CTGOA system. The comparative outcome study reported the betterment of the ADLIC-CTGOA technique on recent approaches in terms of different evaluation measures.

Data Availability Statement: The data that support the findings of this study are openly available in the Kaggle repository at https://www.kaggle.com/datasets/adityajn105/flickr8k and https://www.kaggle.com/d atasets /hsankesara/ flickr-image-dataset, reference number [26, 27].

Acknowledgments: “The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2024/01/31833)”

Funding: “The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2024/01/31833)”

Conflicts of Interest: “The authors declare no conflict of interest.”

References

[1] Deorukhkar, K.P. and Ket, S., "Image Captioning using Hybrid LSTM-RNN with Deep Features," *Sensing and Imaging*, vol. 23, no. 1, p. 31, 2022.

[2] Liu, A.A., Zhai, Y., Xu, N., Nie, W., Li, W. and Zhang, Y., "Region-aware image captioning via interactive learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3685–3696, 2021.

[3] Tiwary, T. and Mahapatra, R.P., "An accurate generation of image captions for blind people using an extended convolutional atom neural network," *Multimedia Tools and Applications*, pp. 1–30, 2022.

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

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

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

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

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

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

[10] Mishal, S.M. and Hamad, M.M., "Text Classification Using Convolutional Neural Networks," 2022.

[11] Sangolgi, V.A., Patil, M.B., Vidap, S.S., Doijode, S.S., Mulmane, S.Y. and Vadaje, A.S., "Enhancing Cross-Linguistic Image Caption Generation with Indian Multilingual Voice Interfaces using Deep Learning Techniques," *Procedia Computer Science*, vol. 233, pp. 547–557, 2024.

[12] Safiya, K.M. and Pandian, R., "Computer Vision and Voice Assisted Image Captioning Framework for Visually Impaired Individuals Using Deep Learning Approach," in *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*, Oct. 2023, pp. 1–7.

[13] Bayisa, L.Y., Wang, W., Wang, Q., Ukwuoma, C.C., Gutema, H.K., Endris, A. and Abu, T., "Unified deep learning model for multitask representation and transfer learning: image classification, object detection, and image captioning," *International Journal of Machine Learning and Cybernetics*, pp. 1–21, 2024.

[14] Solomon, R. and Abebe, M., "Amharic Language Image Captions Generation Using Hybridized Attention-Based Deep Neural Networks," *Applied Computational Intelligence and Soft Computing*, vol. 2023, no. 1, p. 9397325, 2023.

[15] Wasi, A.A., Fahim, E.H., Inova, N.T., Fahim, A.A. and Preeti, T.T., "Hybrid recommendation system of intelligent captioning using deep learning networks," Doctoral dissertation, Brac University, 2024.

[16] Safiya, K.M. and Pandian, R., "Real-Time Photo Captioning for Assisting Blind and Visually Impaired People Using LSTM Framework," *IEEE Sensors Letters*, vol. 7, no. 11, pp. 1–4, 2023.

[17] Cao, X., Zhao, Y. and Li, X., "Optimizing image captioning algorithm to facilitate English writing," *Education and Information Technologies*, vol. 29, no. 1, pp. 1033–1055, 2024.

[18] Kim, G.Y., Oh, B.D., Kim, C. and Kim, Y.S., "Convolutional neural network and language model-based sequential CT Image captioning for intracerebral hemorrhage," *Applied Sciences*, vol. 13, no. 17, p. 9665, 2023.

[19] Nandan, D., Kanungo, J. and Mahajan, A., "An error-efficient Gaussian filter for image processing by using the expanded operand decomposition logarithm multiplication," *Journal of ambient intelligence and humanized computing*, pp. 1–8, 2024.

[20] Mahdi, T.F. and Daway, H.G., "MRI Image Enhancement Using Multilevel Image Thresholds Based on Contrast-limited Adaptive Histogram Equalization," *Iraqi Journal of Science*, 2024.

[21] Rasheed, F., Anwar, M. and Khan, I., "Detecting cyberbullying in Roman Urdu language using natural language processing techniques," *Pakistan Journal of Engineering and Technology*, vol. 5, no. 2, pp. 198–203, 2022.

[22] Pascal, I., "A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images," *International Journal of Machine Learning and Cybernetics*, pp. 1–19, 2024.

[23] Aurpa, T.T. and Ahmed, M.S., "An ensemble novel architecture for Bangla Mathematical Entity Recognition (MER) using transformer-based learning," *Heliyon*, vol. 10, no. 3, 2024.

[24] Mahdi, M.A., Fati, S.M., Hazber, M.A., Ahamad, S. and Saad, S.A., "Enhancing Arabic Cyberbullying Detection with End-to-End Transformer Model," *CMES-Computer Modeling in Engineering & Sciences*, vol. 141, no. 2, 2024.

[25] Gopi, S. and Mohapatra, P., "Chaotic Aquila Optimization algorithm for solving global optimization and engineering problems," *Alexandria Engineering Journal*, vol. 108, pp. 135–157, 2024.

[26] "Flickr8k Dataset". [Online]. Available: https://www.kaggle.com/datasets/adityajn105/flickr8k.

[27] "Flickr Image Dataset". [Online]. Available: https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset.

[28] Arasi, M.A., Alshahrani, H.M., Alruwais, N., Motwakel, A., Ahmed, N.A. and Mohamed, A., "Automated Image Captioning Using Sparrow Search Algorithm With Improved Deep Learning Model," *IEEE Access*, 2023.

[29] Alnashwan, R.O., Chelloug, S.A., Almalki, N., Issaoui, I., Motwakel, A. and Sayed, A., "Lighting Search Algorithm With Convolutional Neural Network-Based Image Captioning System for Natural Language Processing," *IEEE Access*, 2023.