



Image caption generation using a dual attention mechanism

Roshni Padate^{a,*}, Amit Jain^a, Mukesh Kalla^a, Arvind Sharma^b

^a Department of Computer Science and Engineering, Sir padampat singhania university, Udaipur-Chittorgarh, Bhatewar, Rajasthan, India

^b Department of Mathematics, Sir padampat singhania university, Udaipur-Chittorgarh, Bhatewar, Rajasthan, India

ARTICLE INFO

Keywords:

Image captioning

Inception v3

CNN

BI-LSTM

SI-EFO optimization

Meteor score

ABSTRACT

In order to create a statement that accurately captures the main idea of an ambiguous visual, which is said to be a significant and demanding task? Conventional image captioning schemes are categorized into 2 classes: retrieval-oriented schemes and generation-oriented schemes. The image caption generating system should provide precise, fluid, natural, and informative phrases as well as accurately identify the content of the image, such as scene, object, relationship, and properties of the object in the image. However, it can be challenging to accurately express the image's content when creating image captions because not all visual information can be used. In this article, a new image captioning model is introduced that includes 3 main phases like (1) Extraction of Inception V3 features (2) Dual (Visual and Textual) attention generation and (3) generation of image caption. Convolutional Neural Network (CNN) is used to generate visual attention after first deriving initial V3 features. The input texts for the associated images, on the other hand, are analyzed and given to LSTM for the creation of textual attention. To create image captions, Bidirectional LSTM (BI-LSTM) is used to combine textual and visual attention. The Self Improved Electric Fish Optimization (SI-EFO) algorithm is used in particular to optimize the weights of the BI-LSTM. In the end, several measures confirm that the implemented system has improved. The adopted model is 35.21%, 33.76%, 39.52%, 29.69%, 30.12%, 21.49%, and 31.71% better than GAN-RL, LSTM, GRU, EC + GOA, EC + CMBO, EC + DA, EC + EFO models.

1. Introduction

The objective of picture captioning is to automatically generate human-like and content-rich summaries of input photos. Due to its vast application in video recovery and faster baby learning, it has spurred a significant increase in a study in both industry and academia (Huang and Hu, 2019; Wu et al., 2018). Compared to other Computer Vision (CV) tasks (object detection, picture classification), training an efficient image captioned technique is more challenging since it necessitates a thorough grasp of the necessary entities in an image and their associations (Yang et al., 2019; Tan and Chan, 2019). The level of uncertainty is reduced in words (Padate et al., 2022). In most cases, an encoder–decoder strategy is able as the core to offer summaries of images. For encoding the pixel-level data into thick dimensions, it uses a Convolutional NN (CNN) encoder. The decoder is deployed for translating this data into the ordinary language (Yuan et al., 2019; Fan et al., 2018; Chen et al., 2020; Wei et al., 2020).

In recent times, spatial attention methods on CNN feature maps were discovered, in which the attention method characteristically makes a spatial map that emphasizes the sections of images that are relevant to each formed word (Guan and Wang, 2018; Kinghorn et al., 2018; Liu et al., 2018). The dynamic Transformer module (DTRM) is modeled to acquire not only the local features but also the significant

context information (Tang et al., 2022). Using a context aggregation sub network that can collect global structural data, the DRCDN first predicts the coarse clear image (Zhang and He, 2020). As a result, we develop the dynamic contrast loss to detect identity information with dynamic factors for differentiating hard or easy samples (Si et al., 2022). On the other hand, most previous encoder–decoder tasks were trained via Cross-Entropy (XE) minimization which generally causes exposure bias problems. For addressing the exposure bias problem, the current analysis recommends integrating RL schemes into the conventional captioning scheme (Christie et al., 2017; Xiao et al., 2019). However, RL-oriented approaches are far from complete because they just take into account the results of the evaluation methods and do not make an effort to account for the diversity of the formed sentences (Deng et al., 2020; Su et al., 2019; Wei et al., 2020; George and Rajakumar, 2013; Rajakumar and George, 2012).

Currently, the analysts have discovered the usage of conditional GAN in image captioning for resolving the above problem which has attained success in producing variety and spontaneous sentences. Usually, both their discriminator and generator acquire the low-level convolution features as ocular inputs (Bang and Kim, 2020; Wang et al., 2020; Li et al., 2020). However, CNN layers have a low-level receptive field convolutional feature that only creates local (simple-structured) items,

* Corresponding author.

E-mail address: roshnilad.123@gmail.com (R. Padate).

Nomenclature

BI-LSTM	Bidirectional Long Short-Term Memory
CV	Computer Vision
CNN	Convolutional NN
CMBO	Cat Mouse-Based Optimization
DG-GAN	Dual Generator Generative Adversarial Network
DA	Dragonfly
DenseNet	Dense Convolutional Network
EFO	Electric Fish Optimization
GAN	Generative Adversarial Networks
GRU	Gated Recurrent Units
GOA	Grasshopper Optimization Algorithm
IGGAN	Interactions Guided Generative Adversarial Network
MAGAN	Multi-Attention GAN
NN	Neural Network
RMCNet	ResNet with Multi-scale module and adaptive Channel attention
RL	Reinforcement Learning
SI-EFO	Self Improved Electric Fish Optimization
SGC	Scene Graph Captioner
TA-LSTM	Triple Attention-LSTM
VD-SAN	Visual-Densely Semantic Attention Network
XE	Cross-Entropy

making it difficult to distinguish between non-local objects (Yilmaz and Sen, 2020; Rajakumar, 2013a,b; Swamy et al., 2013).

The research that has been offered has the following contribution:

1. Deploys a new image caption generation approach that deploys the inception v3 method for extracting the features.
2. Exploits dual attention mechanism (CNN + LSTM) for generating visual attention and textual attention.
3. Deploys an optimized BI-LSTM model for generating captions, where the weights are tuned via the SI-EFO model.

The article is set up as follows: Section 2 summarizes the project. Section 3 summarizes the suggested model, while Section 4 describes the features of Inception-v3. The CNN + LSTM dual attention technique for producing visual and textual attention is described in Section 5. The proposed SI-EFO model for BI-LSTM optimization is in Section 6. Further, Section 7 illustrates outcomes, and the paper is concluded by Section 8.

2. Literature review

This part holds the literature survey. The eight papers are selected and briefly explained their method, features, and challenges.

2.1. Related works

Cao et al. (2020) suggested using the Interactions Guided Generative Adversarial Network (IGGAN) to label photos unsupervised. This method brought object-to-object communication and multi-scale feature portrayals together. The image was encrypted using ResNet with a Multi-scale module and Adaptive Channel attention to produce a reliable description of features (RMCNet). In this case, IGGAN was used to rebuild the sentence and pictures that were created. Additionally, the created technique generated sentences devoid of any manually labeled image-caption combination.

Wei et al. (2020) have create a Multi-Attention approach for more efficient image captioning, both non-local and local evidence was used. A Multi-Attention GAN (MAGAN), which had a discriminator and generator, was also created based on this paradigm. The devised discriminator was used to tell if the produced sentences were produced by a computer or a human, while the adopted generator was designed to produce more precise sentences.

Yang et al. (2020) developed an efficient method based on the EnsCaption model that aimed to improve an “ensemble of generation-oriented and retrieval-oriented picture captioning systems” through the use of a new Dual Generator Generative Adversarial Network (DG-GAN). The discriminator was trained to give recovered and produced picture captions lower ranking scores as part of the adversarial training procedure, but the caption re-ranking and caption generation enhanced the retrieved labels by providing them a better ranking score. The outcomes of the research revealed that the new method’s effectiveness was significantly greater than the methods that were being evaluated.

Zhao et al. (2019) created a multimodal fusion method to produce descriptions that accurately depict the contents of images. The proposed method, therefore, uses four networks: “CNN for feature extraction, picture attribute extraction, language modeling, and recurrent network.” When compared to existing methods that predict words based on hidden states, the proposed model made use of image features and also created long-term relationships between recorded words. Finally, the simulated results served as proof that the suggested procedure had been developed.

He et al. (2019) aimed to merge the prediction and extraction phases firmly for image representation and a novel Visual-Densely Semantic Attention Network (VD-SAN) model was suggested for image captioning. Particularly, the whole captioning system included DenseNet and it moreover included a Long Short Term Memory (LSTM) for creating the captions. At last, analyses were done and the attained results demonstrated the substantial developments of the established scheme over the extant approaches.

Xu et al. (2019) deployed a Scene Graph Captioner (SGC) model for captioning images by captivating the visual scenes by designing the characteristics of objects, and relations among them. SGC was then used once a technique had first been exploited. Additionally, a cutting-edge method for embedding a scene graph was employed, captivating both graph topology and semantic ideas. Finally, LSTM was used to translate data into texts.

Two strategies were used by Zhu et al. (2018) to improve the performance of sequence learning models. The visual context data was initially controlled at all LSTM phases using a novel attention technique known as Triple Attention-LSTM (TA-LSTM). Then, LSTM was remodeled for attaining improved performance over benchmark Long Short Term Memory (LSTM). At last, arithmetical experimentations have established the enhancement of modeled approach.

Shetty et al. (2018) developed a video and image caption scheme depending upon Neural Network (NN) approach. The suggested approach was very much improved by deploying the distinct features from the scene context, objects, and locations. Consequently, a novel network was deployed for electing the optimum captions. At last, the enhancement of the developed scheme was proven from the simulated outcomes.

Katiyar and Borgohain (2021) have developed 17 distinct convolutional neural networks that have been tested on two well-liked frameworks for creating image captions: the first is based on the Neural Image Caption (NIC) generation model, while the second is based on the Soft-Attention framework. We observe that the Convolutional Neural Network model’s performance in feature extraction for the Image Caption Generation task does not necessarily correspond with its accuracy on the Object Recognition task. This is measured by the number of parameters.

Ye et al. (2021) have suggested a method for creating image captions based on the Optimized Bidirectional Long Short-Term Memory (B-LSTM) model. The PMFO (Proposed Moth Flame Optimization)

Table 1
Reviews on extant image captioning schemes.

Author	Deployed schemes	Features	Challenges	Results
Cao et al. (2020)	IGGAN	<ul style="list-style-type: none"> • Feature representation is fast • Reasonable sentences are generated. 	<ul style="list-style-type: none"> • Tolerates due to insufficient common sense reasoning. 	<ul style="list-style-type: none"> • The rouge value is 30.5
Wei et al. (2020)	MAGAN	<ul style="list-style-type: none"> • Cider and Meteor both have high scores. 	<ul style="list-style-type: none"> • The use of memory is increased by more discriminators. 	<ul style="list-style-type: none"> • The rouge score is 0.564
Yang et al. (2020)	EnsCaption model	<ul style="list-style-type: none"> • Cider scores highly and ROUGE scores highly. 	<ul style="list-style-type: none"> • More thought should be given to the ranking process. 	<ul style="list-style-type: none"> • The rouge score is 59.0
Zhao et al. (2019)	CNN	<ul style="list-style-type: none"> • The optimal caption is chosen • Accurateness is improved 	<ul style="list-style-type: none"> • No attention was given to the captions of the denser image. 	<ul style="list-style-type: none"> • The rouge score is 53.6
He et al. (2019)	VD-SAN	<ul style="list-style-type: none"> • Superior identification of attributes. • Enhanced captioning system. 	<ul style="list-style-type: none"> • It might confound the caption generator of LSTM. 	<ul style="list-style-type: none"> • The rouge score is 53.9.
Xu et al. (2019)	LSTM model	<ul style="list-style-type: none"> • The scene graph is accurate. • Enhanced visual attention. 	<ul style="list-style-type: none"> • Need concern on graph similarity. 	<ul style="list-style-type: none"> • The rouge score is 48.8
Zhu et al. (2018)	TA-LSTM	<ul style="list-style-type: none"> • Creates legible sentences. • Precise explanation of images. 	<ul style="list-style-type: none"> • Needs consideration on virtual asset safety. 	<ul style="list-style-type: none"> • The cider score is 54.8
Shetty et al. (2018)	CNN	<ul style="list-style-type: none"> • Elects optimal captions • Enhanced correctness. 	<ul style="list-style-type: none"> • It could not count the objects precisely. 	<ul style="list-style-type: none"> • The rouge score is 0.609
Katiyar and Borgohain (2021)	CNN	<ul style="list-style-type: none"> • The generated caption's diversity is improved. • The performance efficiency is improved. 	<ul style="list-style-type: none"> • The distortion has occurred 	<ul style="list-style-type: none"> • The rouge score is 46.8
Ye et al. (2021)	B-LSTM	<ul style="list-style-type: none"> • Enhanced performance efficiency • A higher cider score is achieved 	<ul style="list-style-type: none"> • Need concentration on the memory utilization of the system. 	<ul style="list-style-type: none"> • The cider score is 70.2

version using a correlation-based logarithmic spiral updates (PMFO). The results of the performance investigation show that the B-LSTM surpasses cutting-edge techniques for generating titles.

2.2. Review

The reviews of image captioning models are shown in Table 1. The first IGGAN method was created by Cao et al. in 2020, and while it initially produces logical sentences with strong feature representation, it suffers from a lack of commonsense thinking. The MAGAN technique utilized in Wei et al. (2020) produced high Cider and Meteor scores; nevertheless, the utilization of memory increases as the number of discriminators increases. Yang et al.'s EnsCaption algorithm (Yang et al., 2020) produced high ROUGE and Cider scores. The ranking procedure, however, calls for further consideration. The CNN method, which was created by Zhao et al. in 2019, selects the best caption with increased accuracy, however, it must take into account denser image captions. In addition, He et al. (2019) used VD-SAN, which offered superior attribute identification. Likewise, the LSTM scheme employed by Xu et al. (2019) ensured precise scene graphs with enhanced visual attention. Nevertheless, it has to focus on employing the resemblance amid graphs. TA-LSTM was deployed (Zhu et al., 2018) that created readable sentences with an accurate explanation of images; however, virtual asset security has to be concern more. CNN was presented in Shetty et al. (2018) that offered enhanced accuracy with the optimal selection of captions; but, it could not count the objects precisely. CNN was deployed in Katiyar and Borgohain (2021) which improves the performance efficiency but distortion is occurred. Bi-directional LSTM (B-LSTM) was presented in Ye et al. (2021) that provides the higher cider score but the memory utilization is high.

2.3. Research gap

The knowledge of image details is increased when visual attention is added to the encode-decode framework, although the expression of image content may not be comprehensive. The creation of the image caption must therefore be guided by some high-quality additional textual material. In actuality, attribute predictor typically acquires the additional textual information. It can be regarded as integrating these two separate processes – the gathering of more textual data and the creation of the image description – to create a whole trainable phase. This work suggests a dual attention approach to generate image captions in order to address the aforementioned difficulties. The textual attention mechanism is used to improve the integrity of the information, while the visual attention mechanism is used to improve understanding of image details. Extensive experimental findings are used to evaluate our proposed model, which incorporates the dual attention mechanism.

3. Proposed model for image captioning: A stepwise elucidation

The created picture captioning model consists of three crucial steps.

- The image data is first used to retrieve Inception V3 features throughout feature extraction.
- The Inception V3 features that were derived are then fed into CNN to produce visual attention.
- Further, the input texts for corresponding images are processed and subjected to LSTM for the generation of textual attention.
- Further, BI-LSTM is exploited to generate image captions by combining both visual and textual attention.
- To achieve more accurate findings, the BI-weights LSTM's are specifically adjusted using the SI-EFO framework.

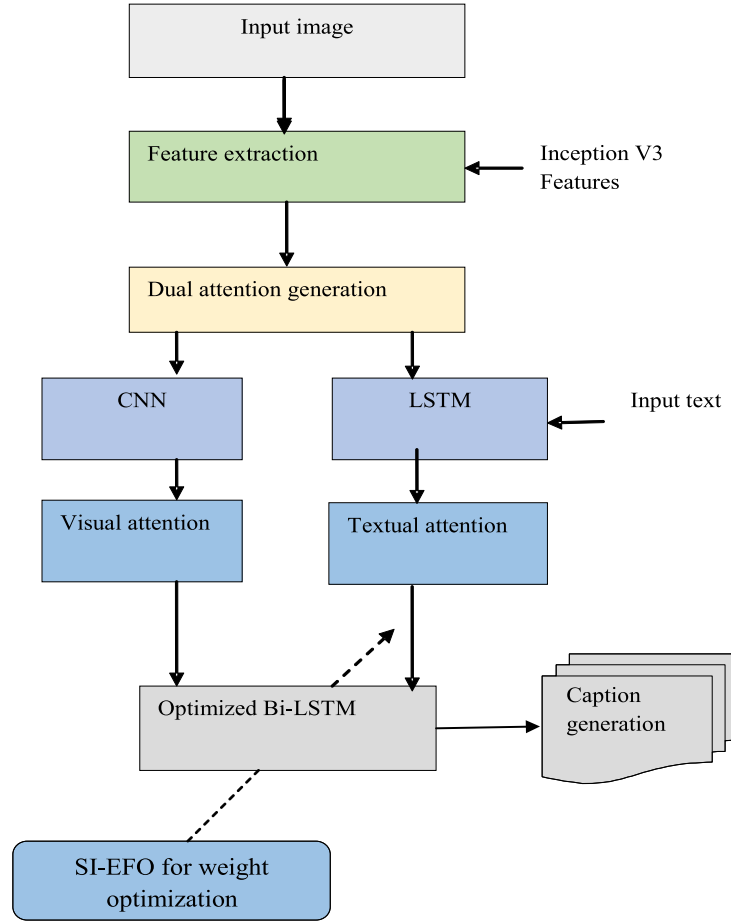


Fig. 1. Illustrative revelation for suggested image Caption generation approach.

The overall representation of the proposed SI-EFO-oriented structure is shown in Fig. 1.

4. Extracting inception-v3 features

In this Inception-v3 model, the elements were concatenated to enhance feature extraction (Ji et al., 2019).

4.1. Inception-v3

“Inception-v3 is a classical deep network and the network is composed of 11 Inception modules of five kinds in total. Each module is designed by experts with a convolutional layer, activation layer, pooling layer, and batch normalization layer”. Accordingly, the module takes up the multi-scale theory. Each module contains a number of branches with different kernel sizes, such as “(1 × 1, 3 × 3, 5 × 5, and 7 × 7)”. 4 parallel layers, including 1 × 1 convolution, 3 × 3 convolutions, 5 × 5 convolutions, and 7 × 7 max pooling, make up the inception model. These filters produce and combine several feature map scales and transfer the amalgamations to the subsequent phase; 1 × 1 convolutions were used in each initialization module to reduce the dimensions prior to the application of “computationally expensive 3 × 3 and 5 × 5 convolutions”.

Next to the exclusion of definite classifying layers, a set of classifying layers were adjoined to the new network graph. The data were overfitted when multiple deep layers of convolutions were used in a model. The conceptualization V1 model employs the idea of stacking several filters of various sizes on the same level to prevent this. Because parallel layers are utilized in place of deep layers in inception models, the model is larger rather than deeper.

These derived Inception-v3 features are then provided as input to CNN for generating visual attention. On the other hand, the input texts for corresponding images are processed and subjected to LSTM for the generation of textual attention. Fig. 2 shows the inception V3 model.

5. Dual attention mechanism via CNN + LSTM for generation of visual and textual attention

A detailed description of the CNN, LSTM, and Bi-LSTM are provided in Sections 5.1–5.3.

5.1. CNN classifier

The three layers are used in [CNN].

- Pooling layer
- Fully connected layer.
- Convolution layer

Fig. 3 shows the CNN architecture. All neurons are linked with nearby neurons in previous layer. At position (r, t) in l th layer of linked w th feature map, the features are evaluated as in Eq. (1).

$$B_{r,t,w}^l = D_w^l + W_w^{lT} P I_{r,t}^l \quad (1)$$

In Eq. (1), W_w^l implies weight, D_w^l refers to bias of w th filter linked to l th layer. In addition, at center location (r, t) of l th layer, the patch input is referred by $P I_{r,t}^l$. The activation value $(act_{r,t,w}^l)$ linked with convolutional features $B_{r,t,w}^l$ is assessed as specified in Eq. (2).

$$act_{r,t,w}^l = (B_{r,t,w}^l) act \quad (2)$$

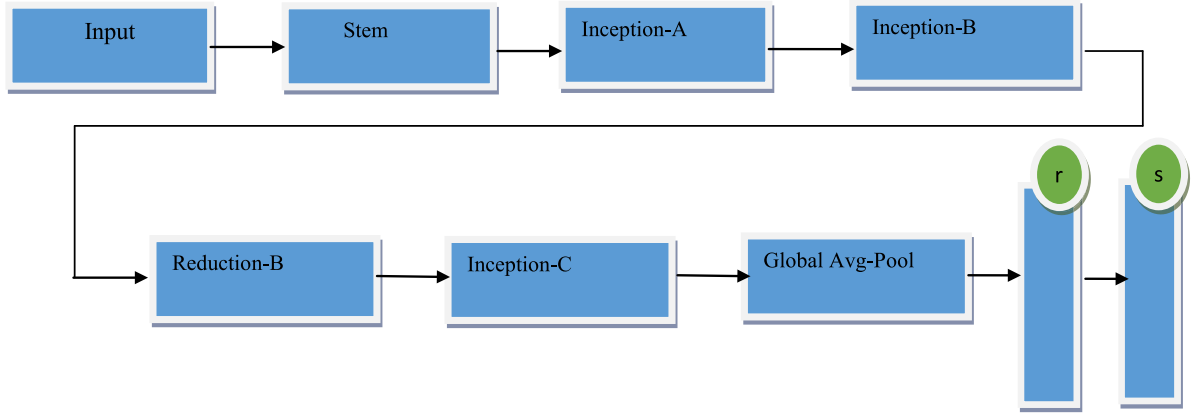


Fig. 2. Architecture of inception V3 model.

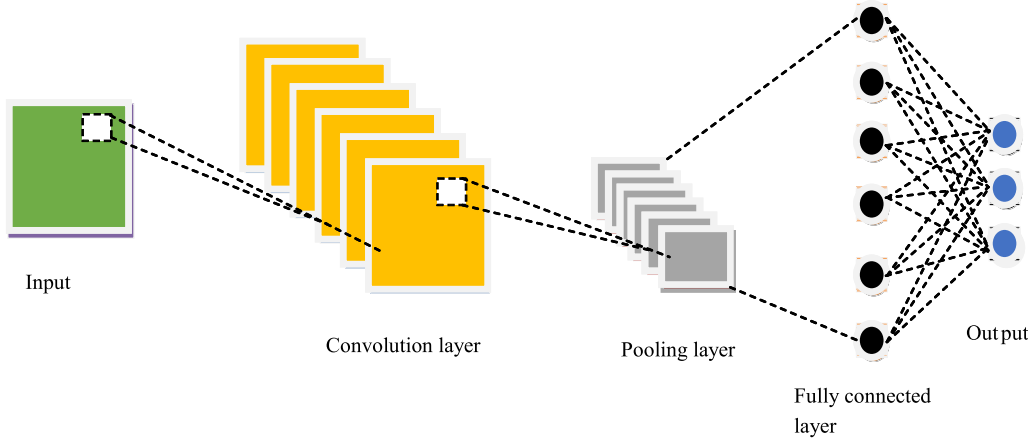


Fig. 3. Architecture of CNN model.

Pooling layer: It does the down-sampling function. For every pooling function $pool(\bullet)$ linked to $act_{m,h,w}^l$, the $C_{r,t,w}^l$ value is assessed as shown in Eq. (3), wherein, $NN_{r,t}$ implies neighbor's near location (r, t) .

$$C_{r,t,w}^l = \{pool(act_{m,h,w}^l), for all (m,h) \in NN_{r,t}\} \quad (3)$$

The prediction results take place at CNN's output layer. The CNN loss is indicated by $Loss$ and is revealed in Eq. (4).

$$Loss = \frac{1}{wn} \left\{ \sum_{h=1}^{wn} l(\theta; C^{(h)}, F^{(h)}) \right\} \quad (4)$$

The common constraint linked with W_w^l and D_w^l is implied by θ . Accordingly, wn counts of output-input relation exist $\{(PI^{(h)}, C^{(h)}) ; h \in [1, \dots, wn]\}$. The h th input feature, the labels and output is specified by $PI^{(h)}$, $C^{(h)}$ and $F^{(h)}$ in that order.

As a result, the output obtained from CNN offers the generated visual attention.

5.2. LSTM classifier

The input texts for corresponding images are processed and subjected to LSTM for generation of textual attention. In fact, LSTMs (Kırbaç et al., 2020) are focused in processing time sequence for memorizing and detecting, which are deployed for predicting the resultants at upcoming time phases (Katiyar and Borgohain, 2021; Bockrath et al., 2019). The benchmark LSTM approach mostly includes “an input gate, forget gate, output gate, and memory cell state and three gates”. The

3 gates contain neurons, which are trained for deciding which data have to be forgotten and permitted to output or input depending upon the present input and preceding output and the memory data. LSTMs encompass a separated cell state that learned the management of memory in a refined way. The memory lane is the cell state of the network, where, data that requires to be memorized could be deployed and updated by LSTMs at diverse times in future.

The attained outputs from LSTM resulted in the generation of textual attention. Further, the visual attention, as well as textual attention, are combined and subjected to BI-LSTM for image caption generation. Fig. 4 depicts the LSTM architecture model.

5.3. BI LSTM networks

The generated visual attention and textual attention are provided to the BI-LSTM approach. The BI-LSTM takes account of the “backward LSTM layer and a forward LSTM layer”. The Bidirectional method (Zhou et al., 2020) lessened data from both sides (backward as well as forward) is extensively exploited in dealing with longer sequences. The LSTM model consists of a series of recurred LSTM cells. All cells of LSTM include 3 multiplicative units, which represent the “forget gate, the input gate, and the output gate”. This unit allows the LSTM memory cells for stocking up and transmits data for an extensive time era. Assume the parameters H and C that point at the hidden and cell state in that order. (X_t, C_{t-1}, H_{t-1}) and (H_t, C_t) point at the input and output layers in that order. Fig. 5 shows the architecture of the Bi-LSTM model.

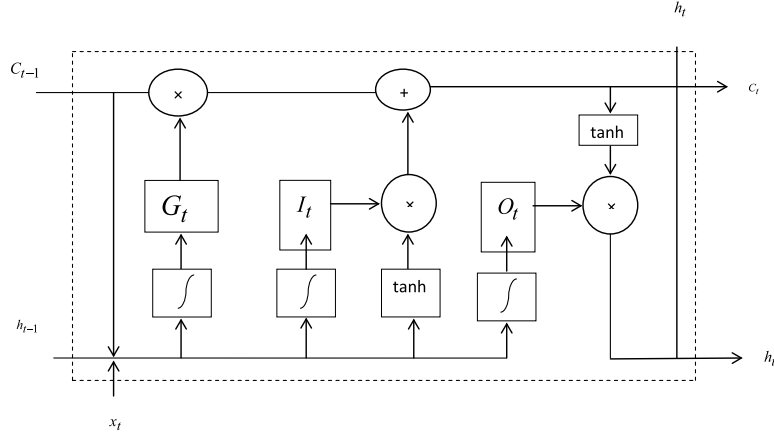


Fig. 4. LSTM model.

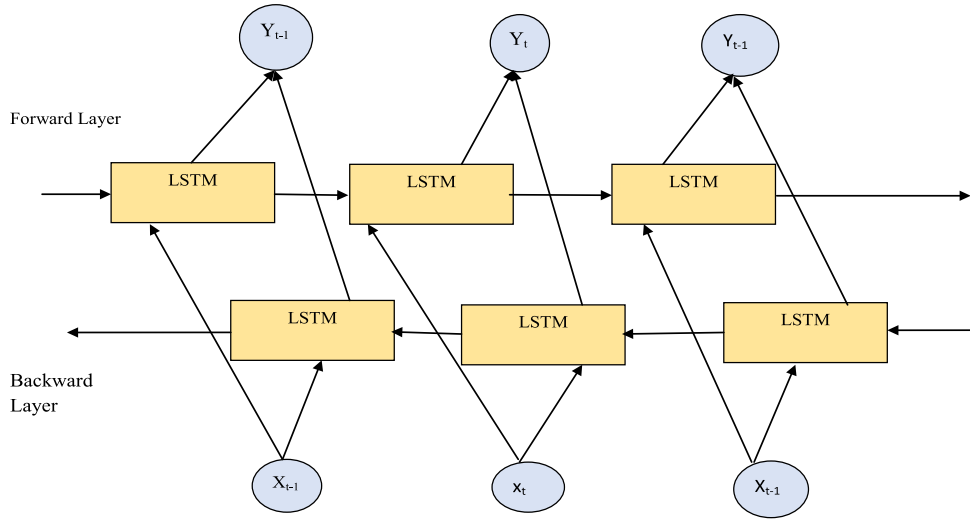


Fig. 5. Bi-LSTM model.

The output, input, and forget gate indicates O_t , I_t , F_t respectively, at time t . F_t is mostly used by LSTM to sort the data that should be avoided. The sorted data show distinct partial features connected to the directions of the prior glance; F_t is represented as in Eq. (5).

$$F_t = [X_t J_{FI} + B_{MF} + B_{IF} + J_{M_{t-1}MF}] \cdot \sigma \quad (5)$$

In Eq. (5), the (J_{MF}, B_{MF}) , and (J_{IF}, B_{IF}) denote weight and bias constraints to map hidden and input layers to forget gates, while σ denotes an inference of the activation function.

The LSTM makes use of the input gate as illustrated in Eq. (6) through (8), where the terms (J_{MG}, B_{MG}) and (J_{IG}, B_{IG}) denote weight and bias constraints to map the hidden and input layers to the cell gate, respectively. (J_{MI}, B_{MI}) and (J_{II}, B_{II}) denote weight and bias restrictions to map input and hidden layers to I_t .

$$G_t = [B_{MG} + J_{IG}X_t + J_{MG}M_{t-1} + B_{IG}] \times \tanh \quad (6)$$

$$I_t = [B_{MI} + J_{II}X_t + J_{MI}M_{t-1} + B_{II}] \times \sigma \quad (7)$$

$$D_t = D_{t-1}F_t + I_tG_t \quad (8)$$

$$O_t = \sigma \times [B_{OM} + B_{OI} + J_{MI}M_{t-1} + J_{IO}X_t] \quad (9)$$

$$M_t = [O_t \tanh(D_t)] \quad (10)$$

Accordingly, the LSTM cell obtains the output hidden layer from the output gate as exposed in Eqs. (9) and (10), in which, (J_{MO}, B_{MO}) and

Table 2
Hyper parameters of classifiers.

Methods	Parameter
LSTM	epoch
	batch_size
	verbose
	loss:categorical_crossentropy
CNN	epoch
	verbose
	batch_size
	activation: relu
	optimizer: adam

(J_{IO}, B_{IO}) denotes weight and bias to map the hidden and input layer to O_t . This work, it is planned to optimize the weights of Bi-LSTM by a new SI-EFO algorithm. Table 2 provides the hyper parameters of the classifier.

6. Proposed SI-EFO model for Bi-LSTM optimization

The weight of the Bi-LSTM is optimally tuned by using SI-EFO. The objective function is provided in Section 6.1.

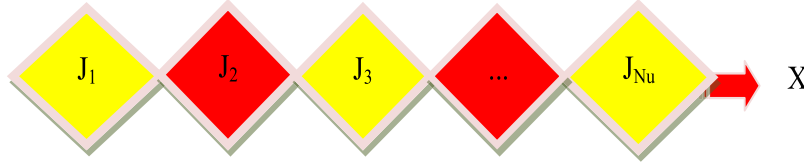


Fig. 6. Solution encoding.

6.1. Objective function

Solution Encoding: The weight of BI-LSTM is denoted by $j(J)$ are chosen optimally using the SI-EFO scheme. Fig. 6 shows the representation for solutions, wherein, nu Nu represent the entire count of BI-LSTM weights. The objective (Obj) of the concerned work is given in Eq. (11), wherein Er signifies the error. To minimize the loss is our objective function.

$$Obj = Min(Er) = Min(Loss) \quad (11)$$

6.2. Proposed SI-EFO algorithm

Although the conservative EFO (Yilmaz and Sen, 2020) plan has various additions, it has set limitations. Therefore, certain modifications are required, and SI-EFO is launched in order to overcome the drawbacks of standard EFO. Self-improvement is typically considered to be effective in conservative optimization methods (Rajakumar, 2013a,b; Swamy et al., 2013; George and Rajakumar, 2013; Rajakumar and George, 2012; Wagh and Gomathi, 2019; Halbhavi et al., 2019; Jadhav and Gomathi, 2019). The steps of the proposed SSI-EHO are as follows.

At first, the populace of electric fish is arbitrarily spread by the searching space, in which, X_{uv} implies u th individual's position in N at D -dimension space, $X_{\max v}$ and $X_{\min v}$ implies upper and lower limits for v dimension respectively, $\phi \in (0, 1)$ implies random values (Yilmaz and Sen, 2020).

$$X_{uv} = X_{\min v} + (X_{\max v} - X_{\min v}) \phi \quad (12)$$

The frequency plays the foremost task in the SI-EFO scheme that balanced exploration and exploitation and is employed to determine passive or active electrolocation.

Active electrolocation: The active ranges of u th individual (R_i) is described by amplitude (B_i). The active range calculation is shown in Eq. (13). Traditionally, distance is calculated depending upon X_{uv} and X_{kv} , but, in the presented SI-EFO scheme, the clark-oriented distance is calculated among two points (X_{uv}, X_{kv}) is done as shown in Eq. (14). In addition, arithmetic crossover takes place for identifying the best individual.

$$R_i = (X_{\max v} - X_{\min v}) B_i \quad (13)$$

$$Dis = \sqrt{\sum_{v=1}^{Dis} \left(\frac{X_{uv} - X_{kv}}{|X_{uv}| + |X_{kv}|} \right)^2} \quad (14)$$

If a single neighbor exists in the active area, EFO employs Eq. (15); otherwise, Eq. (16) is employed, wherein, k implies a randomly elected individual from u th individual neighbor.

$$X_{uv}^{can} = X_{uv} + \phi (X_{kv} - X_{uv}) \quad (15)$$

$$X_{uv}^{can} = X_{uv} + \phi R_i \quad (16)$$

Here, $\phi \in (-1, 1)$ in Eq. (16) implies random count and X_{uv}^{can} refers to candidate location of u th individual.

Passive electro-location: Traditionally, the novel location is updated with X_{rv} and X_{uv} . Herein proposed SI-EFO, the location update

occurs depending upon levy update as exposed in Eq. (17), where; $levy(\beta)$ symbolize levy update.

$$X_{uv}^{new} = X_{uv} + \phi (X_{Rv} - X_{uv}) + levy(\beta) \quad (17)$$

The last step in passively relocation is to change an individual's constraint u th to increase the likelihood that a trait will change, as in Eq. (18), where $ran(0, 1)$ denotes to arbitrary values taken from a homogenous distribution.

$$X_{uv}^{can} = X_{\min v} + \phi (X_{\max v} - X_{\min v}) \quad (18)$$

$$ran(0, 1) \leq ran(0, 1)$$

If the v th constraint value of the u th person exceeds the search space boundaries, it is shifted to the space boundaries that it exceeds.

$$X_{uv}^{can} = \begin{cases} X_{\min v} & X_{uv}^{can} < X_{\min v} \\ X_{uv}^{can} & X_{\max v} > X_{uv}^{can} > X_{\min v} \\ X_{\max v} & X_{uv}^{can} > X_{\max v} \end{cases} \quad (19)$$

The pseudo-code for SSI-EHO approach is given by Algorithm 1.

7. Results and discussions

This paper proposes an EC+SI-EFO scheme for image caption generation. The below sections hold the simulation procedure, analysis on bleu score, analysis of varied scores, convergence analysis, and conclusion part.

7.1. Simulation procedure

“Python” was used to carry out the analysis after the proposed EC + SI-EFO scheme for the picture caption generation approach was provided. As a result, the new approach's performance was evaluated in comparison to existing models, including GAN-RL (Yan et al., 2020), LSTM (Malhotra et al. 2015), GRU (Shi et al., 2020), EC + GOA (Saremi et al., 2017), EC + CMBO (Dehghani et al., 2021), EC + DA (Jafari and Chaleshtari Here, the performance analysis was carried out by changing the LP, which has four possible values: 60, 70, 80, and 90. In order to depict the effectiveness of the chosen approach, several ratings were also looked at, and convergence analysis and statistical analysis were performed. Additionally, a cost study was conducted to confirm the effectiveness of the suggested approach. Fig. 7 exposes a sample illustration of the image caption.

7.2. Dataset description

Here, the analysis was done using “Flickr Dataset” downloaded from <https://www.kaggle.com/hsankesara/flickr-image-dataset>. The Flickr 30k dataset is a well-known benchmark for sentences-based image descriptions. This study introduces Flickr30k Entities, a system for linking mentions of the same entities across several tags for the same image and connecting them to 276k manually annotated bounding boxes. Flickr30k Entities adds 244k co-reference chains to the 158k captions from Flickr30k. Such comments are required to progress automatic visual description and grounded speech interpretation. In this case, 60% of the data is used for training and 40% is used for testing.

Algorithm 1: SSI-EHO

Start
 Initializing the population
 Compute the quality of individuals
 Repeat
 Divide the population into passive and active electro locations based on the frequency values of every individual
 Compute clark based distance evaluation as shown in Eq. (14)
 Apply arithmetic crossover for identifying the best individual
 Perform active electro location as shown in Eq. (15) and Eq. (16)
 Perform passive electro location based on proposed formulation based upon levy as shown in Eq. (17)
 Update the amplitude and frequency values for every individual
 Till the termination condition is met



	“1. The dogs are in the snow in front of a fence	Two brown dogs are running in the snow	A dog running through the snow	A brown dog is running through the snow	A brown and brown dog is running in the snow	A dog is running in the snow
	2. The dogs play on the snow					
	3. Two brown dogs playfully fight in the snow					
	4. Two brown dogs playfully fight in the snow					
	5. Two dogs playing in the snow					
	1. A brown and white dog swimming toward some in the pool	A black and white dog	A dog in a black blue dog in a pool	A white dog is jumping into a pool	A black dog is playing in the water	A dog in a white dog in a water
	2. A dog in a swimming pool swims toward somebody we cannot see	dog swimmin g in a pool				

Fig. 7. Sample image representation showing (a) Input image (b) Ground truth (c) BLSTM+PMFO (d) BLSTM+MFO (e) BLSTM (f) LSTM and (g) C-RNN.

7.3. Analysis of bleu score

Table 3 shows the bleu score obtained by the developed EC + SI-EFO approach over the conservative schemes regarding the varied count of LPs that are altered from 60, 70, 80 and 90. Every obtained result for EC + SI-EFO scheme has shown high values than GAN-RL, LSTM, GRU, EC + GOA, EC + CMBO, EC + DA, EC + EFO, CNN and Bi-LSTM models. Chiefly, the presented model has accomplished better resultants when BLEU score is 1. Especially, when the BLEU score is 1, the adopted approach at 90th LP is 3.26%, 3.01%, 3.9%, 4.37%, 3.19%, 0.28%, 2.49%, 4.45%, and 3.23% better than GAN-RL, LSTM, GRU, EC + GOA, EC + CMBO, EC + DA, EC + EFO models. Similarly, on noticing the results, the EC + SI-EFO have accomplished a high bleu score 1 value of 0.767158, whereas, the traditional models such as GAN-RL, LSTM, GRU, EC + GOA, EC + CMBO, EC + DA, EC + EFO have attained a lower bleu score value of 0.765718, 0.760694, 0.741589, 0.74477, 0.749262, 0.766655 0.75341, 0.66578 and 0.755758 at 70th LP. Thus, the improvement of the adopted EC + SI-EFO approach is established from the investigation outcomes.

7.4. Analysis of varied scores

Fig. 8 displays the cider score obtained by the EC + SI-EFO technique (a). The provided system is compared to GAN-RL, LSTM, GRU, EC + GOA, EC + CMBO, EC + DA, and EC + EFO in this instance. Similarly, Fig. 8 shows the better score produced by the new EC + SI-EFO method over the previous approaches considering several LPs (b). According to the findings, the suggested framework has outperformed the traditional plan in terms of cider score. Specifically, the introduced EC + SI-EFO model has accomplished better values at 80th LP and 90th LP. For example, the presented model is 35.21%, 33.76%, 39.52%, 29.69%, 30.12%, 21.49% and 31.71% better than GAN-RL, LSTM, GRU, EC + GOA, EC + CMBO, EC + DA, EC + EFO, CNN and Bi-LSTM models when the LP is 90. Likewise, the executed method have attained a higher Meteor score (~0.37795). When the LP is 90, the traditional models, including GAN-RL, LSTM, GRU, EC + GOA, EC + CMBO, EC + DA, and EC + EFO, have attained minimum cider score values of 0.326037357, 0.333219223, 0.326926918, 0.340777353, 0.330250744, 0.348939688, 0.33181801, 0.377954448, 0.34256, and 0.340567. In the same way, from Fig. 8(c), the offered



	<p>3. A dog swims in a pool near a person</p> <p>4. Small dog is paddling through the water in a pool</p> <p>5. The small brown and white dog is in the pool</p>					
	<p>1. A man and a woman in festive costumes dancing</p> <p>2. A man and a woman with feathers on her head dance</p> <p>3. A man and a woman wearing decorative costumes and dancing in a crowd of onlookers</p> <p>4. One performer wearing a feathered headdress dancing with another performer in the streets</p> <p>5. Two people are dancing with drums on the right and a crowd behind them</p>	A group of people playing in a parade	A group of people in a red shirt in a red and red and red shirt	'A group of people are standing in a parade	A group of people are walking on a red shirt	A man in a red shirt is in a red shirt and a red shirt and a red shirt and a red shirt and a red shirt and a red shirt and a red shirt
	<p>1. A couple of people sit outdoors at a table with an umbrella and talk</p> <p>2. Three people are sitting at an outside picnic bench with an umbrella</p> <p>3. Three people sit at an outdoor cafe</p> <p>4. Three people sit at an outdoor table in front of a building painted like the Union Jack</p> <p>5. Three people sit at a picnic table outside of a building painted like a union jack</p>	A group of people sit on top of a bunch of people	A man in a red red shirt is are are is a red a red dog	A man in a blue jacket and a blue jacket and a pink jacket , and a pink jacket , and a pink jacket in front of a building	A man in a black shirt is walking on a building	A man in a red shirt and a red shirt is a red shirt and a red shirt and a red shirt and a red shirt and a red shirt and

Fig. 7. (continued).

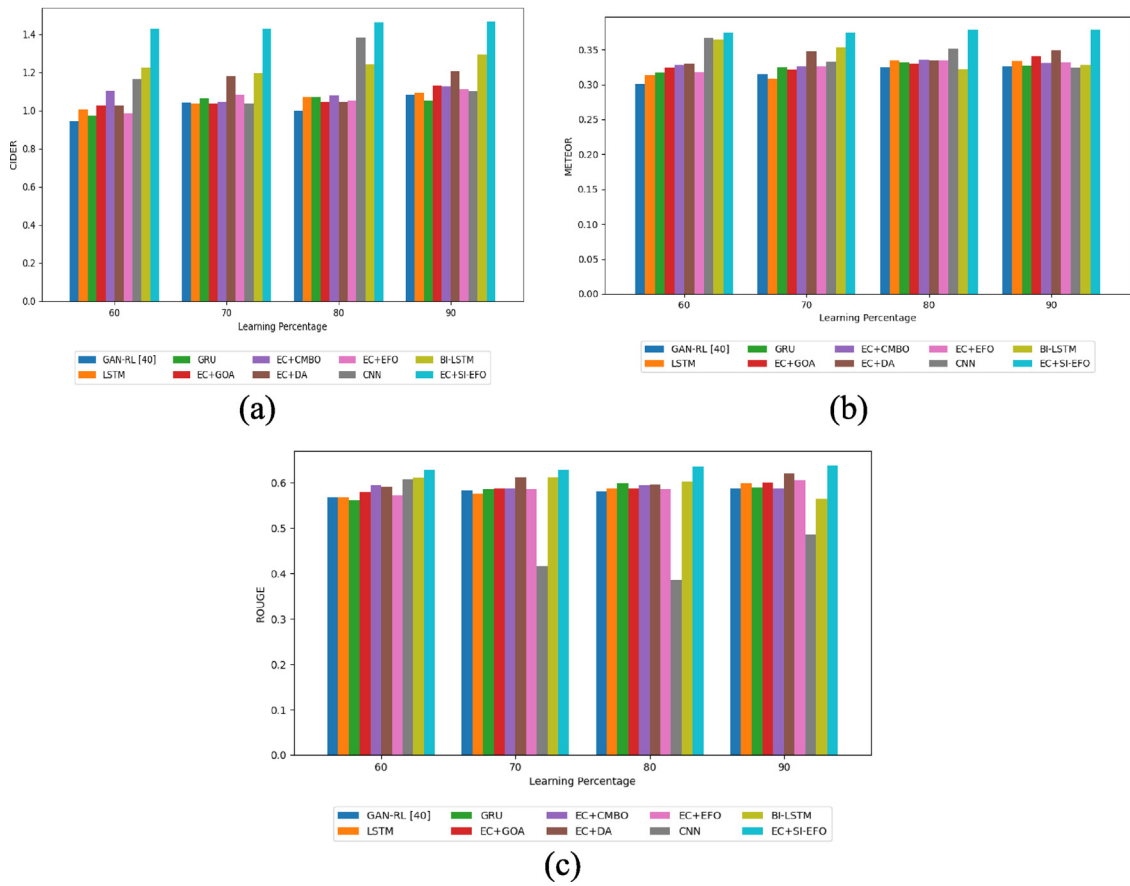


Fig. 8. Analysis using developed approach over extant schemes regarding (a) Cider (b) Meteor and (c) Rouge scores.

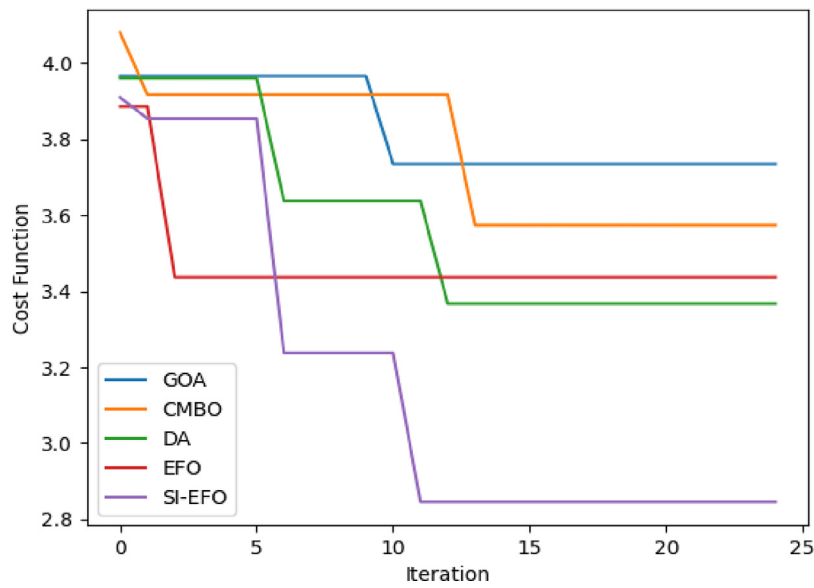


Fig. 9. Convergence analysis of SI-EFO scheme over evaluated models.

Table 4

Statistical analysis of EC + SI-EFO model over conventional models regarding BLEU 1 score.

	GAN-RL	LSTM	GRU	EC + GOA	EC + CMBO	EC + DA	EC + EFO	EC + SI-EFO	CNN	BI-LSTM
Mean	0.74859	0.73998	0.73756	0.73849	0.74894	0.75046	0.73982	0.76763	0.503443	0.572601
Median	0.74618	0.73906	0.74064	0.73993	0.74846	0.75236	0.74519	0.76772	0.458216	0.534537
Std-Dev	0.01078	0.01486	0.01210	0.00621	0.00343	0.01624	0.01481	0.00078	0.284397	0.325194
Min	0.73628	0.72109	0.71795	0.72931	0.74467	0.73046	0.71547	0.76660	0.121457	0.220426
Max	0.76572	0.76069	0.75099	0.74477	0.75415	0.76666	0.75341	0.76846	1.101246	1.293767

Table 5
Analysis of computational time.

Methods	Time
GOA	681.4753
CMBO	679.7402
DA	1434.129
EFO	686.7663
SI-EFO	702.9282

Table 6
Confidence time interval analysis.

Methods		
GOA	7.33857	7.3284
CMBO	7.35503	7.33768
DA	7.35696	7.33899
EFO	7.35267	7.32903
SI-EFO	7.35188	7.33051
GOA	7.33857	7.3284

7.8. Confidence time interval

The confidence interval is the range of values that, if you repeated your experiment or re-sampled the population in the same manner, you would anticipate your estimate to fall within a specific proportion of the time. Table 6 depicts the confidence time interval. A measure derived from observable data in statistics is called a confidence interval (CI). This provides a range of values for an unidentified parameter. The interval also includes a confidence level, which expresses the probability that an estimated interval.

8. Conclusion

In this research new image caption generation scheme was used, where; inception v3 scheme was employed for deriving the features from the input image. Further, the derived features were classified via Convolutional Neural Network (CNN) for the visual generation. In the meantime, the input texts of corresponding images were classified via Long Short Term Memory (LSTM) for a textual generation. The generated visual attention and textual attention are provided to the Bidirectional Long Short Term Memory (BI-LSTM) approach that generated the image captions. Furthermore, for attaining exact captions, an optimized BLSTM scheme was introduced, where the weights were tuned optimally using the SI-EFO scheme. Finally, the suggested scheme's superiority to the traditional plans according to several measures was proved. In particular, the introduced Ensemble Classifier-Self Improved Electric Fish Optimization (EC + SI-EFO) model has accomplished better values at 80th LP and 90th LP. For example, the presented model was 35.21%, 33.76%, 39.52%, 29.69%, 30.12%, 21.49% and 31.71% better than Generative Adversarial Network-Reinforcement Learning (GAN-RL), Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), Ensemble Classifier Grasshopper Optimization Algorithm (EC + GOA), Ensemble Classifier Cat Mouse-Based Optimization (EC + CMBO), Ensemble classifier Dragon fly Algorithm (EC + DA), Electric Fish Optimization (EC + EFO) approaches when the LP is 90. Therefore, the professionalism of the proposed approach's perfection was established. The effectiveness, generality, and resilience of presented models were assessed. On both generation and retrieval tasks, our models produce outcomes that are extremely competitive or cutting edge. Additionally, we illustrate how the learned alignments closely match human intuition and how the learnt attention can be used to increase the interpretability of the model generation process. The textual focus strengthens the information's integrity and the visual attention improves understanding of image features. The creation of the image caption needs to be guided by some excellent additional textual material. In actuality, attribute predictor typically acquires the additional textual information. The proposed model's drawback is

that the fusion process is not taken into account. Future efforts will concentrate mostly on how to enhance the quality of the image labels and combine visual and textual attention to enhance the development of image captions. Additionally, we intend to investigate the role that attention positions visual and textual play in the creation of image captions.

CRedit authorship contribution statement

Roshni Padate: Conceptualization, Methodology. **Amit Jain:** Resources, Data curation. **Mukesh Kalla:** Formal analysis. **Arvind Sharma:** Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

- Bang, S., Kim, H., 2020. Context-based information generation for managing UAV-acquired data using image captioning. *Autom. Constr.* 112, 103116.
- Bockrath, S., Rosskopf, A., Koffel, S., Waldhör, S., Srivastava, K., Lorentz, V.R., 2019. State of charge estimation using recurrent neural networks with long short-term memory for lithium-ion batteries. In: *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society*, Vol. 1. IEEE, pp. 2507–2511.
- Cao, S., An, G., Zheng, Z., Ruan, Q., 2020. Interactions guided generative adversarial network for unsupervised image captioning. *Neurocomputing* 417, 419–431.
- Chen, X., Zhang, M., Wang, Z., Zuo, L., Li, B., Yang, Y., 2020. Leveraging unpaired out-of-domain data for image captioning. *Pattern Recognit. Lett.* 132, 132–140.
- Christie, G., Laddha, A., Agrawal, A., Antol, S., Goyal, Y., Kochersberger, K., Batra, D., 2017. Resolving vision and language ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. *Comput. Vis. Image Understand.* 163, 101–112.
- Dehghani, M., Hubálovský, Š., Trojovský, P., 2021. Cat and mouse based optimizer: a new nature-inspired optimization algorithm. *Sensors* 21 (15), 5214.
- Deng, Z., Jiang, Z., Lan, R., Huang, W., Luo, X., 2020. Image captioning using DenseNet network and adaptive attention. *Signal Process., Image Commun.* 85, 115836.
- Fan, C., Zhang, Z., Crandall, D.J., 2018. Deepdiary: Lifelogging image captioning and summarization. *J. Vis. Commun. Image Represent.* 55, 40–55.
- George, A., Rajakumar, B.R., 2013. APOGA: An adaptive population pool size based genetic algorithm. *AASRI Proc.* 4, 288–296.
- Guan, J., Wang, E., 2018. Repeated review based image captioning for image evidence review. *Signal Process., Image Commun.* 63, 141–148.
- Halbhavi, S., Kodad, S.F., Ambekar, S.K., Manjunath, D., 2019. Enhanced invasive weed optimization algorithm with chaos theory for weightage based combined economic emission dispatch. *J. Comput. Mech. Power Syst. Control* 2, 19–27.
- He, X., Yang, Y., Shi, B., Bai, X., 2019. VD-SAN: Visual-densely semantic attention network for image caption generation. *Neurocomputing* 328, 48–55.
- Huang, G., Hu, H., 2019. C-rnn: a fine-grained language model for image captioning. *Neural Process. Lett.* 49 (2), 683–691.
- Jadhav, A.N., Gomathi, N., 2019. DIGWO: Hybridization of dragonfly algorithm with improved grey wolf optimization algorithm for data clustering. *Multimedia Res.* 2 (3), 1–11.
- Ji, Q., Huang, J., He, W., Sun, Y., 2019. Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images. *Algorithms* 12 (3), 51.
- Katiyar, S., Borgohain, S.K., 2021. Comparative evaluation of CNN architectures for image caption generation. *arXiv preprint arXiv:2102.11506*.
- Kinghorn, P., Zhang, L., Shao, L., 2018. A region-based image caption generator with refined descriptions. *Neurocomputing* 272, 416–424.
- Kırbaş, I., Sözen, A., Tuncer, A.D., Kazancıoğlu, F.S., 2020. Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. *Chaos Solitons Fractals* 138, 110015.
- Li, R., Liang, H., Shi, Y., Feng, F., Wang, X., 2020. Dual-CNN: A convolutional language decoder for paragraph image captioning. *Neurocomputing* 396, 92–101.
- Liu, Q., Chen, Y., Wang, J., Zhang, S., 2018. Multi-view pedestrian captioning with an attention topic CNN model. *Comput. Ind. Eng.* 97, 47–53.
- Padate, Roshni, et al., 2022. High-level and low-level feature set for image caption generation with optimized convolutional neural network. *Technology* 67.

- Rajakumar, B.R., 2013a. Impact of static and adaptive mutation techniques on the performance of genetic algorithm. *Int. J. Hybrid Intell. Syst.* 10 (1), 11–22.
- Rajakumar, B.R., 2013b. Static and adaptive mutation techniques for genetic algorithm: a systematic comparative analysis. *Int. J. Comput. Sci. Eng.* 8 (2), 180–193.
- Rajakumar, B.R., George, A., 2012. A new adaptive mutation technique for genetic algorithm. In: *2012 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, pp. 1–7.
- Saremi, S., Mirjalili, S., Lewis, A., 2017. Grasshopper optimisation algorithm: theory and application. *Adv. Eng. Softw.* 105, 30–47.
- Shetty, R., Tavakoli, H.R., Laaksonen, J., 2018. Image and video captioning with augmented neural architectures. *IEEE MultiMedia* 25 (2), 34–46.
- Shi, T., Huang, S., Chen, L., Heng, Y., Kuang, Z., Xu, L., Mei, H., 2020. A molecular generative model of ADAM10 inhibitors by using GRU-based deep neural network and transfer learning. *Chemometr. Intell. Lab. Syst.* 205, 104122.
- Si, T., He, F., Zhang, Z., Duan, Y., 2022. Hybrid contrastive learning for unsupervised person re-identification. *IEEE Trans. Multimed.*
- Su, J., Tang, J., Lu, Z., Han, X., Zhang, H., 2019. A neural image captioning model with caption-to-images semantic constructor. *Neurocomputing* 367, 144–151.
- Swamy, S.M., Rajakumar, B.R., Valarmathi, I.R., 2013. Design of hybrid wind and photovoltaic power system using opposition-based genetic algorithm with cauchy mutation.
- Tan, Y.H., Chan, C.S., 2019. Phrase-based image caption generator with hierarchical LSTM network. *Neurocomputing* 333, 86–100.
- Tang, W., He, F., Liu, Y., 2022. YDTR: infrared and visible image fusion via y-shape dynamic transformer. *IEEE Trans. Multimed.*
- Wagh, M.B., Gomathi, N., 2019. Improved GWO-CS algorithm-based optimal routing strategy in VANET. *J. Netw. Commun. Syst.* 2 (1), 34–42.
- Wang, H., Wang, H., Xu, K., 2020. Evolutionary recurrent neural network for image captioning. *Neurocomputing* 401, 249–256.
- Wei, Y., Wang, L., Cao, H., Shao, M., Wu, C., 2020. Multi-attention generative adversarial network for image captioning. *Neurocomputing* 387, 91–99.
- Wu, C., Wei, Y., Chu, X., Su, F., Wang, L., 2018. Modeling visual and word-conditional semantic attention for image captioning. *Signal Process., Image Commun.* 67, 100–107.
- Xiao, F., Gong, X., Zhang, Y., Shen, Y., Li, J., Gao, X., 2019. DAA: Dual LSTMs with adaptive attention for image captioning. *Neurocomputing* 364, 322–329.
- Xu, N., Liu, A.A., Liu, J., Nie, W., Su, Y., 2019. Scene graph captioner: Image captioning based on structural visual representation. *J. Vis. Commun. Image Represent.* 58, 477–485.
- Yan, S., Xie, Y., Wu, F., Smith, J.S., Lu, W., Zhang, B., 2020. Image captioning via hierarchical attention mechanism and policy gradient optimization. *Signal Process.* 167, 107329.
- Yang, M., Liu, J., Shen, Y., Zhao, Z., Chen, X., Wu, Q., Li, C., 2020. An ensemble of generation-and retrieval-based image captioning with dual generator generative adversarial network. *IEEE Trans. Image Process.* 29, 9627–9640.
- Yang, J., Sun, Y., Liang, J., Ren, B., Lai, S.H., 2019. Image captioning by incorporating affective concepts learned from both visual and textual components. *Neurocomputing* 328, 56–68.
- Ye, Z., Khan, R., Naqvi, N., Islam, M.S., 2021. A novel automatic image caption generation using bidirectional long-short term memory framework. *Multimedia Tools Appl.* 80 (17), 25557–25582.
- Yilmaz, S., Sen, S., 2020. Electric fish optimization: a new heuristic algorithm inspired by electrolocation. *Neural Comput. Appl.* 32 (15), 11543–11578.
- Yuan, A., Li, X., Lu, X., 2019. 3G structure for image caption generation. *Neurocomputing* 330, 17–28.
- Zhang, S., He, F., 2020. DRCDN: learning deep residual convolutional dehazing networks. *Vis. Comput.* 36 (9), 1797–1808.
- Zhao, D., Chang, Z., Guo, S., 2019. A multimodal fusion approach for image captioning. *Neurocomputing* 329, 476–485.
- Zhou, X., Lin, J., Zhang, Z., Shao, Z., Chen, S., Liu, H., 2020. Improved itracker combined with bidirectional long short-term memory for 3D gaze estimation using appearance cues. *Neurocomputing* 390, 217–225.
- Zhu, X., Li, L., Liu, J., Li, Z., Peng, H., Niu, X., 2018. Image captioning with triple-attention and stack parallel LSTM. *Neurocomputing* 319, 55–65.