

Enhanced Caption Generation Model Using Hawk Swarm Optimization Based Bilstm Model

Sumedh Pundlikrao Ingale^{1*} and Gajendra Rambhau Bamnote²

Submitted: 27/01/2024 Revised: 05/03/2024 Accepted: 13/03/2024

Abstract: The accurate representation and contextual understanding are difficult problems to solve when integrating computer vision and language processing in the field of creating captions for visual data. Intricate subtitles and maintaining visual details are challenges for existing models also achieving high accuracy with less over fitting issues is difficult. Hence to overcome these difficulties develop a hawk swarm optimization based BiLSTM model (HSO-BiLSTM) to enhance the process of generating captions for visual data. To achieve this, here utilize two datasets: Flickr30k (d1) and COCO (d2). Initially, segregate the images and their corresponding transcripts. Subsequently, perform separate preprocessing for images and transcripts. For images, conduct distinct preprocessing steps, and then employ VGG-16 for feature extraction. In the case of transcripts, construct a vocabulary, tokenize the text, assign indices, and pad sequences. Afterward, integrate both sets of features to optimize a Bidirectional Long Short-Term Memory (BiLSTM) model. To enhance the effectiveness of the BiLSTM, utilize the Harris Hawk optimization (HHO) and Harmony Search optimization techniques for fine-tuning. The optimized BiLSTM is then employed to generate captions for the transcripts. The metrics for dataset 1 acquired values are 0.50, 0.32, 0.55 and 26.66, and similarly for dataset 2 acquired values are 0.49, 0.32, 0.57 and 26.58.

Keywords: Hawk swarm optimization, BiLSTM, tokenization, images, transcripts and VGG-16

1. Introduction

An increasing demand for systems that can both comprehend images and videos and succinctly explain them in natural language has emerged in the digital age due to the expansion of visual content. The blending of visual and textual information has prepared the way for the fascinating discipline of visual data caption creation, where computers work to connect the linguistic and visual worlds. By improving the precision, originality, and contextual relevance of generated captions for visual data, this mechanism aims to extend the frontiers of this subject by leveraging the power of machine learning [1]-[4]. This study aims to explore the complex intricacies of image recognition and language production by utilizing the capabilities of machine learning techniques. The objective is to create a model that not only recognizes the objects, sceneries, and activities in an image or video with accuracy, but also creates descriptions that are linguistically fluid, educational, and aware of the context [5]-[7]. This combination of computer vision and natural language processing has a lot of potential for a range of uses, from helping the blind to automating content production and improving data indexing.

The mechanism attempts to address the difficulties of

ambiguity, diversity, and semantic coherence in visual data captioning through an investigation of several machine learning architectures, unique training methodologies, and substantial datasets. In addition, the mechanism will examine the moral issues raised by AI-generated content in an effort to weigh such issues against the technology's possible societal effects [8][9]. In the end, it hopes to increase both theoretical knowledge and real-world applications of improving visual data caption production [10][11]. By utilizing machine learning's capabilities, it ushers in a time where machines can transmit the essence of the visual world in elegant and profound ways in addition to seeing and interpreting it.

Enhancing the production of visual data captions has a wide range of benefits that help visual content to be communicated in a way that is more precise, informative, and interesting.. Here are some of the main advantages: improved captions help viewers understand visual information more thoroughly by capturing minute details, context, feelings, and connections between objects and actions. People with visual impairments can more easily access visual content due to detailed captions, which enable them to "see" through text descriptions and interact with images and videos [12]. High-quality captions improve the user experience by adding more levels of detail and improving the appeal and engagement of the material. The discoverability and indexing of material are enhanced by accurate captions, making it simpler for users to locate certain visual assets using keywords and descriptions [13].

¹Prof. Ram Meghe Square, Anjangaon Bari Rd, Badnera, Amravati, Maharashtra 444701

²Prof. Ram Meghe Square, Anjangaon Bari Rd, Badnera, Amravati, Maharashtra 444701*

Corresponding Author Email: sumedh3003@gmail.com

Multilingual enhanced captions can be produced, enabling efficient communication between various linguistic populations. User preferences can be utilized to modify captions, resulting in a more customized experience and stronger interactions with the audience. By doing away with the necessity for manual annotation, automated caption generation speeds up and conserves resources throughout the creation of content [14]-[16]. Making captions can help with data annotation and labeling for machine learning activities, allowing for more precise model training for a range of applications. By adding verbal descriptions to accompany visual data, captions improve the quality and depth of the information offered and contributes to the enrichment of datasets. By adding more context and explanations, enhanced captions in tutorials, films, and instructional resources aid learners in understanding difficult subjects [17].

By highlighting essential points and facilitating comprehension, captions in video presentations and lectures improve the effectiveness of information transfer. Engaging captions on social media sites draw attention and promote conversation, which increases user engagement and sharing. By incorporating narrative elements, feelings, and theme descriptions, captions can enhance storytelling and increase the impact of visual content. Captions that are imaginative and captivating can improve brand identity, communicate ideas, and arouse feelings, so increasing marketing initiatives. Research in computer vision, natural language processing, and artificial intelligence is being fueled by the creation of caption generating models. By describing historical items, works of art and customs for future generations, captions can aid in the preservation of cultural heritage. Live events, newscasts, and video conferences can all be captioned in real-time to increase accessibility and audience reach [18]-[20]. Captions help create immersive experiences in entertainment and gaming by narrating stories, describing situations, and directing players. The architecture of the model, the training procedure, and methods to address issues like over fitting and language variety will all affect how well machine learning improves the creation of visual data captions [21].

The main aim of the research is to develop a HSO-BiLSTM model enhance the process of generating captions for visual data. To achieve this, here utilize two datasets: Flickr30k and COCO. Initially, segregate the images and their corresponding transcripts. Subsequently, perform separate preprocessing for images and transcripts. For images, conduct distinct preprocessing steps, and then employ VGG-16 for feature extraction. In the case of transcripts, construct a vocabulary, tokenize the text, assign indices, and pad sequences. Afterward, integrate both sets of features to optimize a BiLSTM model. To enhance the effectiveness of the BiLSTM, utilize the

Harris Hawk optimization and Harmony Search optimization techniques for fine-tuning. The optimized BiLSTM is then employed to generate captions for the transcripts.

➤ Hawk swarm optimization: The HSO model may be effectively used to integrate Harmony Memory, which enables hawks in HHO to better utilize the knowledge they have learned from previous successful harmony search (HS) solutions to improve their exploration strategy. The memory-based optimization of HS and the cooperative exploration behavior of HHO can both be utilized to the best of their abilities in the optimization process due to this integration. A group of excellent solutions that have been discovered throughout the optimization process are kept in Harmony Memory in HS.

➤ HSO based BiLSTM: The BiLSTM model undergoes tuning through hawk swarm optimization, enhancing the caption generation model with a focus on minimal complexity. This optimization effectively mitigates the problem of over fitting.

➤ The paper follows a structured framework. Section 2 reviews prior caption generation research, outlining its pros and cons. In the 3rd section, the novel caption generation mechanism is detailed. Sections 4 and 5 delve into hawk swarm optimization and experimental results, respectively, with Section 6 serving as the platform for concluding remarks.

2. Motivation

The goal of this research is to improve optimization techniques by combining harmony search with hawks' hunting behaviors. Complex spaces present a difficulty for traditional optimization. The study aims for a more robust method by combining the musical inspiration of harmony search with the flexible investigation of hawks. By balancing exploration and efficiency, this interdisciplinary fusion seeks to improve optimization for real-world problems.

2.1 Literature review

By incorporating a coverage mechanism into the attention-based framework to overcome concerns with over- and under-recognition, Teng Jiang et al.'s [22] goal was to improve the production of image captions. The coverage method keeps track of previous attention data, resulting in more accurate captions and more balanced visual content recognition. Additionally, the complexity that the coverage technique has imposed may make training more difficult and resource-intensive. Unlike conventional sequential methods, the hierarchical Long Short-Term Memory (phi-LSTM) structure introduced by Ying Hua Tan and Chee Seng [23] for image captioning adopts a hierarchical approach, producing captions by decoding from phrases to

sentences. This method better captures the fundamental structure of language. However, the increasing complexity of this model may necessitate the use of more computer resources for training and inference, which could potentially reduce its effectiveness in real-time applications. Xinwei He and colleagues [24] aimed to improve image caption generation by employing Part of Speech (PoS) tags to guide a Long Short-Term Memory (LSTM) based word generator. This model improves the model's capacity to match linguistic descriptions with image information, leading to improved image captioning performance across benchmark datasets. However, if the tagging accuracy is compromised or the approach is unable to handle complicated phrase structures correctly, this model may result in inferior performance. Guiguang Ding's [25] goal was to improve image captioning by adding reference data using a modified LSTM framework. This algorithm achieves significant performance gains by using reference data to weight words and enhance caption quality. However, if reference data is not representative or is labeled incorrectly, relying on it may limit its usefulness. In their efforts to enhance image captioning, Aihong Yuan and the team [26] developed a 3-gated model. The three gated structures in this model made it easier to integrate adaptively both global and local image characteristics, which enhanced image comprehension and led to more accurate caption production. However, the three gated components of the model's increased complexity may result in higher processing demands during both training and inference, thus reducing efficiency. Through a multi-stage architecture, Ling Cheng et al. [27] sought to improve automatic image caption generation. They introduced an innovative stack decoder model that collaborates with LSTM layers to optimize attention weights for both visual-level features and semantic-level attributes, resulting in an enhanced fine-grained image caption.. A more careful tweaking of the hyper parameters and a higher computational overhead during training and inference may be necessary due to the multi-stage architecture's greater complexity. To enhance image caption generation, Huawei Zhang et al. [28] incorporated a Bi-LSTM structure that leverages both preceding and succeeding information to make contextually precise predictions. This led to a notable improvement in image captioning performance over the original LSTM model. However, this model's increased complexity as a result of using both forward and backward decoders could lead to higher processing demands during training and inference, thereby compromising effectiveness and scalability. The goal of Xing Liu, Weibin Liu, and Weiwei Xing [2] was to improve the effectiveness of image description by merging local information to capture both particular and general image properties. This effectively combines local and global elements. But this approach might make

computations more difficult during both inference and training, which might have an impact on performance, especially for larger datasets.

2.2 Challenges

- The model architecture adds more computational complexity and parameters, which could need the use of more resources for training and inference.
- Due to the requirement to analyze sequences both forward and backward, training a BiLSTM model can take longer than training a unidirectional model.
- Particularly when using limited training data, the additional model complexity may make it more susceptible to over fitting.
- It can be difficult to capture long-range interdependence since context awareness may not always be achieved by harmonizing information from both directions.
- Finding the ideal sequence length for efficient bidirectional processing can be difficult and have an impact on the level of generated captions.

3. Enhanced caption generation model using proposed hawk swarm optimization based BiLSTM model

The primary objective of this research is to enhance the process of generating captions for visual data. To achieve this, here utilize two datasets: d1 and d2. Initially, segregate the images and their corresponding transcripts. Subsequently, perform separate preprocessing for images and transcripts. For images, distinct preprocessing steps are conducted, and then employ VGG-16 for feature extraction. In the case of transcripts, construct a vocabulary, tokenize the text, assign indices, and pad sequences. Afterward, integrate both sets of features to optimize a BiLSTM model. The optimized BiLSTM is then employed to generate captions for the transcripts. To amplify the prowess of the BiLSTM model, harness the potential of Harris Hawk optimization and Harmony Search techniques. These advanced strategies play a pivotal role in refining the model's performance. The culmination of efforts is the empowered and optimized BiLSTM, which serves as the engine to generate vivid and contextually accurate captions for the associated transcripts.

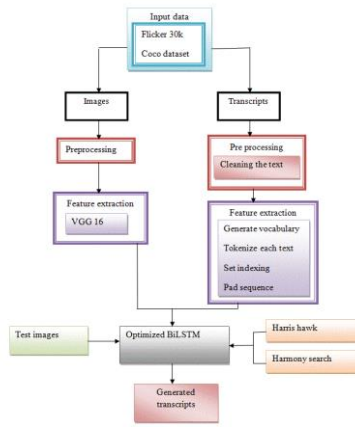


Fig. 1. Architecture of the proposed caption generation model

3.1 Input

The statistical analysis of the caption generating model, which takes the $d1$ and $d2$ as inputs, is given below.

$$M = \sum_{t=1}^h M_c + \sum_{t=1}^b M_d \quad (1)$$

Here, M_c denotes the first database, M_y for the dataset's data, with values ranging from 1 to h , M_z for the second database, and M_d for the dataset's data, with values ranging from 1 to b .

3.2 Initial segregation

Initial Segregation entails separating the selected datasets, like Flickr30k and COCO, into two independent categories: one for images and one for text transcripts. All future stages of data processing and analysis develop around this separation. Here establish a clear separation between visual content and the written descriptions that go with it by constructing these discrete sets, which makes it easier to focus on and effectively process data to extract features, improve models, and finally provide superior image captions. This separation guarantees that images and the text data they are associated with can be treated independently while maintaining their essential relationship for future study advances.

Lets define the following sets

M_{images} : Set of all image data from selected datasets (e.g., Flickr30k, COCO).

$M_{transcripts}$: Set of all text transcripts associated with the images in M_{images} .

The initial segregation can be conceptually represented as:

$$M_{images} \cap M_{transcripts} = \phi \quad (2)$$

$M_{images} \cap M_{transcripts}$, represents the intersection of the two sets, which is an empty set (ϕ), indicating that there are no common elements between the sets. This illustrates the clear separation between image data and text transcripts.

3.3 Image preprocessing

In the context of generating captions for visual data, image preprocessing describes the sequence of preliminary actions and changes performed on raw images prior to their use in the caption production process. These steps aim to improve the accuracy and relevance of the image data, making it better suited for feature extraction and captioning tasks that are crucial for producing accurate and descriptive captions. They accomplish this by providing clear, standardized, and informative image representations that can be effectively used by the following caption generation model. The preprocessed image is represented as MCI^* , MdI^*

3.4 Feature extraction

The process of converting raw image into compact and accurate numerical representations that capture the most important visual information is known as image feature extraction. In order to assess and decipher the visual information of image, this extraction makes use of pre-trained deep learning models like VGG-16. By utilizing image feature extraction, the caption creation process obtains a deeper understanding of the image content, resulting in more precise and contextually significant captions that precisely represent the visual situations. Extracting useful characteristics from the preprocessed images using the VGG-16 model, it is possible to create compact and accurate image representations by using the VGG-16 model, which is a potent tool for capturing pertinent visual data.

3.4.1 VGG 16

The VGG-16 models feature extraction procedure uses its deep Convolutional neural network (CNN) architecture to extract hierarchical and abstract visual characteristics from preprocessed images in the context of creating captions for visual data. VGG-16 can learn a wide variety of visual patterns because it has already been pre-trained on a sizable image dataset. An image travels through a number of Convolutional and pooling layers when it passes through the VGG-16 model. These layers successively analyze the image at various degrees of abstraction, extracting basic properties like edges and textures before moving on to more complex features like forms and individual object parts. The activations of the Convolutional layers include these properties. The fully connected (dense) layers in the VGG-16 model typically follow the Convolutional layers and further compress the

collected features into a compact representation. These dense layers aid in the feature vector encoding of the most important visual information, effectively summarizing the contents of the image. These VGG-16 traits that were extracted serve as a link between the textual and visual modes for creating captions. The caption generation model, such as a BiLSTM model, uses the compact and representative feature vectors as input, and as a result, they assist in the process of connecting visual information with linguistic descriptions. This feature extraction stage makes sure that the generated captions are based on the visual data of the image, improving their accuracy and coherence. The feature extracted output is denoted as $MCI^\#, Mdl^\#$

3.5 Text preprocessing

Text preprocessing involves preparing raw text descriptions for analysis by carrying out activities such as tokenization, removing special characters, assigning indices, and padding sequences in order to provide captions for visual data. Thus, reliable and coherent captions that effectively explain visual content are generated by integrating the text with visual elements and ensuring consistent input for caption generating models. The preprocessed text is represented as MCT^*, MdT^*

3.5.1 Cleaning the text

Cleaning the text entails enhancing and upgrading raw textual descriptions to create an organized, uniform format that facilitates the best natural language processing in the context of visual data. This entails activities like reducing unnecessary characters, uniformly changing text to lowercase, and removing excess spaces. In order to facilitate accurate and meaningful caption production that appropriately reflects the associated visual content, the objective is to develop a coherent and consistent textual input that can interact with visual characteristics in a seamless manner.

3.6 Feature extraction

Transcript feature extraction includes compressing raw textual descriptions into condensed numerical representations in order to create captions for visual data. Tokenization techniques capture word meaning, facilitating the integration of visual data for accurate caption generation. As a result, it is easier to provide exact and pertinent captions for the visual content, which improves contextual understanding.

3.6.1 Generate vocabulary

Generating vocabulary in transcripts is the process of compiling an extensive and well-organized list of the special words used in the textual descriptions. As a starting point for further NLP activities, this lexicon includes all unique words that may be located in the transcripts. The

vocabulary list assists in tagging words with indices, allowing for their numerical representation for additional processing and analysis.

3.6.2 Tokenize each text

Tokenization is the process of breaking each transcript's unique texts into tokens from the raw textual descriptions. Depending on the chosen tokenization approach, these tokens could be words, sub words, or even characters. Tokenization assists in breaking the text down into manageable components, allowing for later processing and analysis.

3.6.3 Set indexing

Setting indexing is the procedure of giving distinct numerical identifiers (indices) to tokens inside a dataset, frequently as a component of text preprocessing. Each token, which normally represents a word, has a unique index that acts as its numerical representation. This indexing makes it possible to store, retrieve, and manipulate textual material effectively when performing various operations. Setting indexing facilitates the conversion of words in transcripts into corresponding numerical values in the context of creating captions for visual data, making it easier for them to be integrated with visual elements for caption generation models.

3.6.4 Pad sequence

Padding sequences is the process of uniformly lengthening sequences in a dataset that have different lengths by inserting placeholder items (like zeros). In this situation, padding sequences are used to ensure that all captions are the same length by adding padding tokens to text transcripts. This makes it easier for the caption creation process to handle textual data consistently together with visual elements. The final feature extracted outcome is denoted as $MCT^\#, MdT^\#$

3.7 Working of BiLSTM model in generating captions for visual data

The vector $MCI^\#, Mdl^\#$ and $MCT^\#, MdT^\#$ serve as the input for the caption generation model. The BiLSTM model then uses these extracted features as input, processing them to create captions or transcripts that provide further explanation for the associated visual data. The BiLSTM, a sort of recurrent neural network, makes use of the contextual data included in the features to produce coherent and contextually relevant textual outputs, yielding accurate and understandable descriptions of the related visual content. The model may provide evocative and contextually appropriate captions for visual content by using the Bidirectional LSTM, which efficiently captures complex connections in textual data. An evolving RNN-based architecture is called LSTM. In addition to providing

a memory unit and gate mechanism to enable the use of longer distance information in sentences, it addresses the problem of gradient disappearance in RNNs. The door structure's design allows for selective saving. These are the LSTM models:

$$j_v = \sigma(M_j[e_{v-1}, a_u] + c_j) \quad (3)$$

$$g_v = \sigma(M_g[e_{v-1}, a_u] + c_g) \quad (4)$$

$$p_v = \sigma(M_p[e_{v-1}, a_u] + c_p) \quad (5)$$

$$\tilde{h}_u = \tanh(M_h[e_{v-1}, a_u] + c_h) \quad (6)$$

$$h_u = j_u * \tilde{h}_v + g_1 * h_{u-1} \quad (7)$$

$$e_v = p_u * \tanh(h_u) \quad (8)$$

In this case, the sigmoid activation function is denoted as σ . The hyperbolic tangent function is shown by the symbol \tanh . The unit input is symbolized by a_v . The output gate, the forget gate, and the input gate are each represented by j_u, g_u, p_u at time v . M and c stand for the input gate, forget gate, and output gate weights and bias, respectively. The symbol \tilde{h}_v indicates the input's current status. h_u represents the updated status at time v and e_u stands for the output at time u .

The input sequence is processed by the BiLSTM concurrently in two directions, forward and backward. This enables a more thorough understanding of the textual input by allowing the model to capture both prior and following context. LSTM units serve as the foundational building blocks of the BiLSTM. Due to their ability to maintain memory across lengthy sequences, LSTMs are well suited for identifying dependencies and recurring patterns in language. The BiLSTM receives its input sequence from the features that have been retrieved from the visual and textual data. The LSTM units process the sequence in both ways, taking into account both past and future characteristics, and each feature vector is guaranteed to match to a particular time step by the bidirectional processing.

By changing their internal states and gates, the LSTM units learn to capture relationships between the features. This helps the model to comprehend how words interact with one another, ensuring that the captions that are generated are accurate in terms of context. As the BiLSTM progresses through the input sequence, it acquires the capability to predict the subsequent word in the caption. This prediction is influenced by LSTM's memory retention

and bidirectional processing, which are informed by the accumulated comprehension of preceding words. During training, the BiLSTM is fine-tuned to minimize the disparity between predicted words and the actual words found in the training data's captions. This fine-tuning involves adjusting the model's internal parameters through techniques such as gradient descent and back propagation. The BiLSTM constructs words one at a time during inference (caption generation for new data) utilizing its acquired knowledge of context and word associations.

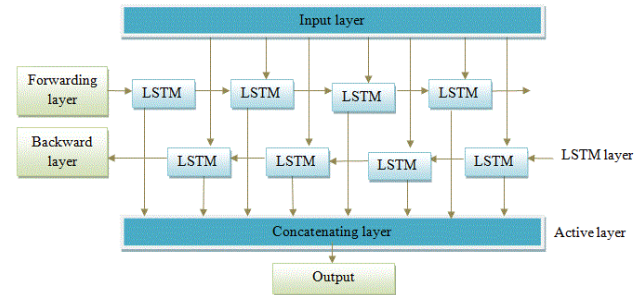


Fig. 2. Working of BiLSTM model in generating captions

4. Proposed hawk search optimization

Hawk search optimization fine-tunes the parameters, including weights and biases, of the classifier. An efficient method of utilizing the knowledge learned from previously successful HS solutions to improve the hawks' exploration approach in HHO [29] is to include Harmony Memory [30]. The memory-based optimization of HS and the cooperative exploration behavior of HHO can both be utilized to the best of their abilities in the optimization process due to this integration. A collection of high solutions that have been discovered during the optimization process is kept in Harmony Memory in HS. These solutions show areas of the search space where successful results were obtained. By adding HM into HHO, enables the hawks' movement to be guided by the combined knowledge of these solutions.

4.1 Motivation

The cooperative hunting behavior of Harris's hawks in the wild serves as the inspiration for the exploration phase of the HHO algorithm. The hawks in the flock spread out during this phase to search a large region for probable prey. The exploration phase of optimization aims to efficiently cover the solution space to find various and potentially useful solutions. The optimization memory is an essential component of the HS algorithm for directing the search process toward better solutions over time. The desire to improve the algorithm's performance and convergence by drawing on previously effective solutions is the motivation behind the inclusion of an optimization memory in Harmony Search.

The improvisational process of musicians is mimicked by harmony search in an intelligent way, and improvisation and optimization are probably comparable in the following ways:

1. Each decision variable has a corresponding musician.;
2. The value range of the decision variable is equivalent to the pitch range of a musical instrument.
3. The solution vector at a particular iteration corresponds to musical harmony at a particular moment in time.
4. The objective function aligns with the aesthetics of the audience.

The solution vector undergoes iterative improvement, much like the gradual enhancement of musical harmony over time. In general, HS is composed of the following five steps:

The following definition of the optimization problem:

$$\begin{aligned} & \text{minimize} \setminus \text{maximize } t(c), \\ & \text{subject to } c_f \in B_f, \dots, f = 1, 2, \dots, K \end{aligned} \quad (9)$$

In this scenario, the objective function is denoted as $t(c)$, The set of all (c_f) decision variables is C , B_f is the set of all potential values for $Q^{c_f} \leq B_f \leq T^{c_f}$, and the overall sum of the decision variables is K .

The HS's parameters are then initialized. These criteria are as follows:

1. Harmony Memory Size (HMS) refers to the quantity of solution vectors retained in the memory for harmonious combinations.
2. Where, $HMCR \in [0,1]$, Harmony Memory Considering Rate (HMCR);
3. Pitch-adjustment rate, or PAR, where $PAR \in [0,1]$;

In the following sections, these parameters will be further explained.

4.2 Initialize harmony memory

As depicted in Equation 10, the harmony memory (HM) is a matrix composed of solutions, with each harmony memory vector representing an individual solution. In this phase, solutions are randomly generated and rearranged within HM based on their objective function values, such as $t(c^1) \leq t(c^2) \dots \leq t(c^{HMS})$.

$$HM = \begin{bmatrix} c_1^1 & c_2^1 & \dots & c_K^1 & | & t(c^1) \\ c_1^2 & c_2^2 & \dots & c_K^2 & | & t(c^2) \\ \dots & \dots & \dots & \dots & | & \dots \\ c_1^{HMS} & c_2^{HMS} & \dots & c_K^{HMS} & | & t(c^{HMS}) \end{bmatrix} \quad (10)$$

4.3 Improvise New Harmony

In this stage, the HS algorithm creates a New Harmony vector (improvises) $c' = (c'_1, c'_2, c'_3, \dots, c'_K)$. This process relies on three fundamental operators: random consideration, pitch adjustment, and memory consideration. The values of the New Harmony vector are randomly inherited from the previous values stored in HM with a probability governed by HMCR in the memory consideration step. Consequently, the variable stored in HM plays a role in determining the value of decision variable (c'_1) . The subsequent decision variable, (c'_2) , is chosen from $c_2^1, c_2^2, c_2^3, \dots, c_2^{HMS}$. Similarly, the remaining decision variables $c_2^1, c_2^2, c_2^3, \dots, c_2^{HMS}$ are sequentially selected using the same method, also with the probability dictated by $HMCR \in [0,1]$

The utilization of HM in this process can be likened to a musician relying on their memory to compose a captivating melody. This cumulative phase ensures that successful harmonies are considered as integral elements of the New Harmony vectors.

According to the HMCR probability test, the remaining choice variable values are then randomly selected from among those $c'_f \in B_f$, where they are not chosen from HM. This scenario is termed random consideration (with a probability of $(1-HMCR)$), which enhances the range of solutions and encourages the system to persist in exploring a wide array of diverse solutions in pursuit of global optimality.

These two phases were summarized by the equation shown below. i.e., consideration of randomness and memory

$$c'_f = \begin{cases} c'_f \in \{c_f^1, c_f^2, c_f^3, \dots, c_f^{HMS}\} & \text{w.p. } HMCR \\ c'_f \in B_f & \text{w.p. } (1-HMCR) \end{cases} \quad (11)$$

Moreover, the exploration of additional promising solutions within the search space is achieved by employing the $PAR \in [0,1]$ operator to fine-tune each decision variable in the new harmony vector, $c' = (c'_1, c'_2, c'_3, \dots, c'_K)$, which was inherited from HM. These decision variables (c'_f) are examined and adjusted with a probability of PAR as specified in Equation 12.

$$c'_f \leftarrow \begin{cases} \text{Adjusting pitch w.p. } PAR \\ \text{Doing nothing w.p. } (1-PAR) \end{cases} \quad (12)$$

If a randomly generated number, $rnd \in [0,1]$, falls within the specified probabilities of PAR, the new decision variable, (c'_f) will be adjusted according to the following equation:

$$(c'_f) = (c'_f) \pm rand() * dg \quad (13)$$

In this context, dg represents a flexible distance bandwidth employed to enhance the effectiveness of HS, while $(rand)$ is a function generating a random value within the range $\in [0,1]$. The extent of adjustments or modifications to the constituent parts of the new vector is determined by dg . The value of dg varies based on whether the optimization problem is discrete or continuous. In essence, the way the parameter (PAR) influences the constituent parts of the New Harmony vector is akin to how musicians might slightly adjust their tone frequencies to achieve significantly improved harmonies. Consequently, this broadens the potential solutions within the search space and enhances the search capabilities.

4.4 Enhanced convergence

The HSO model may be effectively used to integrate Harmony Memory, which enables hawks in HHO to better utilize the knowledge they have learned from previous successful HS solutions to improve their exploration strategy. The memory-based optimization of HS and the cooperative exploration behavior of HHO can both be utilized to the best of their abilities in the optimization process due to this integration. A group of excellent solutions that have been discovered throughout the optimization process are kept in Harmony Memory in HS.

4.4.1 Update the harmony memory

For each New Harmony vector $t(c')$, the objective function is evaluated to update HM with the newly generated vector $c' = (c'_1, c'_2, c'_3, \dots, c'_K)$. If the objective function value of the new vector surpasses that of the worst harmony vector stored in HM, the new vector replaces the worst harmony vector. Otherwise, the new vector is discarded. Nevertheless, it's possible to explore alternative harmonies with the least similarity among the various harmonies in HM. To prevent premature convergence in HM, a limit on the maximum number of identical harmonies in HM could be imposed.

4.4.2 Exploration Phase of HHO

The exploratory stage of HHO is covered in this subsection. The Harris Hawks have strong sight that can track and recognize prey, but occasionally the prey is not readily apparent. The hawks have been watching and waiting for the prey for a long time in this situation. Hawks are considered as potential solutions in HHO, with the prey considered the best option after each iteration. Hawks use two tactics, which are depicted in equation (14), to locate prey by perching at certain spots and continuously

observing the surroundings, if $(g < 0.5)$ the hawks perch according to the position of the family. The hawks perch in a random location within the population area if $g \geq 0.5$.

The combined equation for HS's memory-based optimization and HHO's cooperative exploring behavior.

$$P = 0.5 \begin{cases} (B_h(i) - b_1 | B_h(i) - 2b_2 B(i)|), & g \geq 0.5 \\ (B_{rabbit}(i) - B_g(i)) - b_3, & + c' \in HM \wedge c^{worst} \notin HM \\ (M_T + b_4(R_T - M_T)), & g < 0.5 \end{cases} \quad (14)$$

Where, $B(i+1)$ indicates Hawks position in the following iteration in equation (1). $B_{rabbit}(i)$, describes the rabbit position. The location of hawks is indicated by the symbol $B(i)$. Random variables with values between 0 and 1 are b_1, b_2, b_3, b_4 and g . The lower bound and upper bound of random variables are denoted by the letters LB and UB. $B_v(i)$ stands for the location of each hawk at the v^{th} iteration, and G stands for the number of hawks in the search space. $B_g(i)$ or the average hawk position is represented by the following equation.

$$B_g(i) = \frac{1}{G} \sum_{v=1}^G B_v(i) \quad (15)$$

S.NO	Pseudo code of hawk search optimization
1	Initialization:
2	Initialize parameters: HMS, HMCR, PAR, max_ iterations, exploration_prob
3	Initialize HM with random solutions
4	Initialize best_ solution with a high value
5	for iteration in range(max_ iterations):
6	Create New Harmony vector (new_ solution) using HMCR and PAR
7	Tune new_ solution using PAR operator
8	if random() < exploration_prob:
9	Apply exploration phase to new_ solution
10	Evaluate the fitness of new_ solution
11	if new_ solution improves over the worst solution in HM:
12	Replace the worst solution in HM with new_ solution
13	if new_ solution is better than best_

	solution:
14	Update best _solution with new_ solution
15	return best _solution
16	# Exploration Phase
17	function Exploration Phase(solution):
18	if random() < 0.5:
19	Apply random consideration (modify a subset of variables)
20	else:
21	Apply memory consideration (modify variables based on HM)
22	return modified _solution
23	Terminate

5. Result and discussion




The HSO-BiLSTM model is employed to elevate the performance of the caption generation model, and a comparative analysis is conducted to assess its efficacy when juxtaposed with alternative methodologies.

5.1 Experimental setup

The experiment is conducted using a robust setup comprising an 8GB internal memory and the Windows 10 operating system as the foundation for the caption generation model.

5.2 Experimental results

The outcomes and conclusions attained through the testing and evaluation of the model's performance are referred to as experimental results in the context of a caption generating model. These findings provide the model's efficacy in generating captions for images or other types of data. Based on its training data and observed patterns, the model has produced various captions for the input image. It makes an attempt to analyze the scene by considering all of the potential outcomes. The generated captions offer many methods to describe the same visual material, demonstrating the variety that can occur in caption generation. The given input images are displayed in figures a) to i), and the various caption's results are displayed in figures j) to r).

	Caption generation model		
	Sample 1	Sample 2	Sample 3
			
















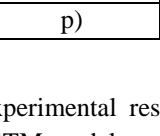
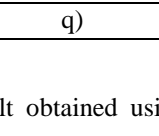
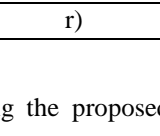
Input images	a)	b)	c)
			
	d)	e)	f)
			
	g)	h)	i)
			
			
m)	n)	o)	
			
p)	q)	r)	
			

Fig. 3. Experimental result obtained using the proposed HSO-BiLSTM model

5.3 Dataset description

5.3.1 D1

A popular benchmark dataset for image captioning jobs is the "Flickr30k" dataset. It comprises of 31,783 images gathered from the Flickr website, each with numerous human annotators' detailed annotations. The goal of this dataset is to make it easier to create and assess image captioning models.

5.3.2 D

A vast collection of images called the COCO dataset was created to further the study of how to comprehend objects and scenes in natural environments. The dataset, which includes a wide variety of visual information and settings, focuses on object detection and understanding within complicated scenes.

5.4 Performance analysis concerning TP for d1

Figure 4 shows the HSO-BiLSTM models bleu score, meteor, rouge score and spice in enhancing the caption generation model.

Figure 4a shows the HSO-BiLSTM approach's achievement of values for the bleu score of 0.36, 0.37, 0.44, 0.47, and 0.49 while maintaining a TP of 90 with epoch values of 100, 200, 300, 400, and 500.

As depicted in Figure 4b, the HSO-BiLSTM strategy demonstrates meteor scores of 0.20, 0.23, 0.29, 0.30, and 0.32, all while during TP 90.

As depicted in Figure 4c, the HSO-BiLSTM strategy demonstrates rouge scores of 0.39, 0.42, 0.45, 0.54 and 0.58 while during TP 90.

As depicted in Figure 4d, the HSO-BiLSTM approach obtains values of 19.00, 21.43, 23.17, 25.12, and 26.65 for the spice metric, while maintaining a TP of 90.

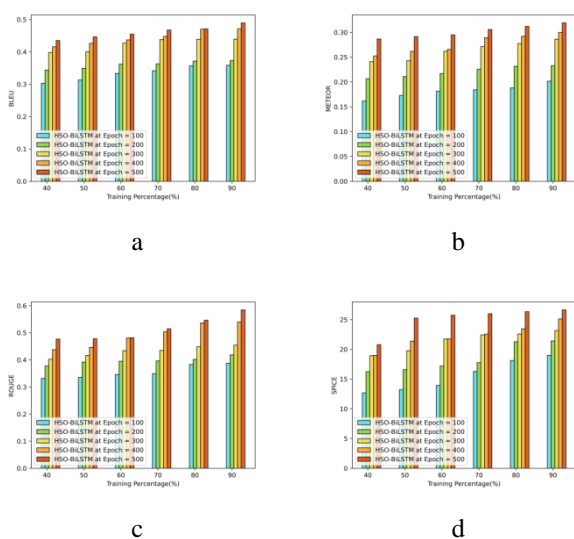


Fig.4. Performance analysis concerning d1 a) bleu score, b) meteor, c) rouge score, d) spice

5.5. Performance analysis concerning TP for d2

Figure 5 shows the HSO-BiLSTM models bleu score, meteor, rouge score and spice in enhancing the caption generation model.

Figure 5a illustrates the HSO-BiLSTM approach's achievement of values for the bleu score of 0.35, 0.38, 0.41, 0.47 and 0.51 while maintaining a TP of 90.

As depicted in Figure 5b, the HSO-BiLSTM strategy demonstrates meteor scores of 0.20, 0.24, 0.28, 0.31, and 0.32, all during a TP 90.

As depicted in Figure 5c, the HSO-BiLSTM strategy demonstrates rouge scores of 0.39, 0.41, 0.45, 0.54 and 0.58, all while during a TP 90.

As depicted in Figure 5d, the HSO-BiLSTM approach obtains values of 18.19, 18.66, 22.83, 24.83 and 26.05 for the spice metric, while maintaining a TP of 90.

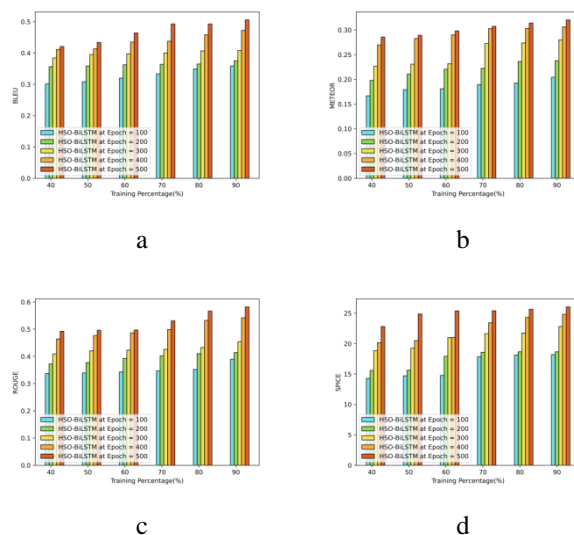


Fig.5. Performance analysis concerning d2 a) bleu score, b) meteor, c) rouge score, d) spice

5.6 Comparative methods

In a comparative analysis, various models including the gated recurrent unit model [31], LSTM model, BiLSTM model [32], as well as Harris hawks optimization based on BiLSTM [33], and harmony search optimization based on BiLSTM [34] are employed to showcase the effectiveness of the HSO-BiLSTM approach.

5.6.1 Comparative analysis concerning TP for d1

In Figure 6a), the bleu score for caption generation by the HSO-BiLSTM model is shown, maintaining a TP of 90. Surpassing the harmony search optimization–BiLSTM model by 2.53%, the HSO-BiLSTM model attains a bleu score of 0.50.

In Figure 6b), the meteor score for caption generation by the HSO-BiLSTM model is showcased, maintaining a TP of 90. Surpassing the harmony search optimization–BiLSTM model by 14.28%, the HSO-BiLSTM model achieves a meteor score of 0.32.

In Figure 6c), the rouge score for caption generation by the HSO-BiLSTM model is presented, maintaining a TP of 90. Surpassing the harmony search optimization–BiLSTM model by 2.73%, the HSO-BiLSTM model attains a rouge score of 0.55.

In Figure 6d), the spice score for caption generation by the HSO-BiLSTM model is showcased, maintaining a TP of 90. Outperforming the harmony search optimization–BiLSTM model by 13.87%, the HSO-BiLSTM model reaches a spice score of 25.66.

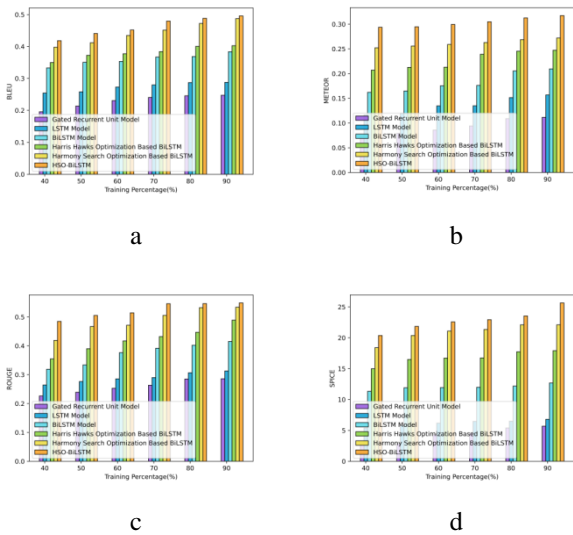


Fig. 6. Comparative analysis concerning d1 a) bleu score, b) meteor, c) rouge score, d) spice

5.6.2 Comparative analysis concerning TP for d1

In Figure 7a), the bleu score for caption generation by the HSO-BiLSTM model is shown, maintaining a TP of 90. Surpassing the harmony search optimization–BiLSTM model by 5.55%, the HSO-BiLSTM model attains a bleu score of 0.49.

In Figure 7b), the meteor score for caption generation by the HSO-BiLSTM model is showcased, maintaining a TP of 90. Surpassing the harmony search optimization–BiLSTM model by 10.21%, the HSO-BiLSTM model achieves a meteor score of 0.32.

In Figure 7c), the rouge score for caption generation by the HSO-BiLSTM model is presented, maintaining a TP of 90. Surpassing the harmony search optimization–BiLSTM model by 7.95%, the HSO-BiLSTM model attains a rouge score of 0.57.

In Figure 7d), the spice score for caption generation by the HSO-BiLSTM model is showcased, maintaining a TP of 90. Outperforming the harmony search optimization–BiLSTM model by 20.55%, the HSO-BiLSTM model reaches a spice score of 26.58.

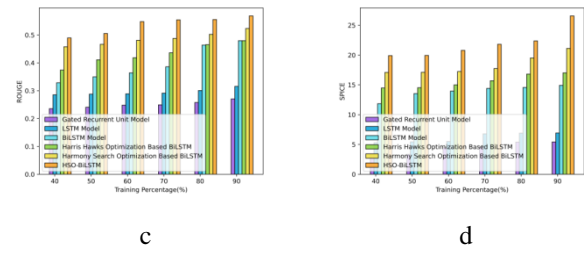
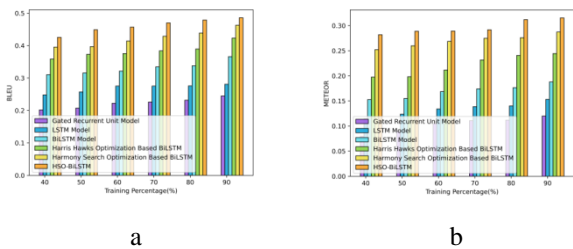


Fig. 7.Comparative analysis concerning a) bleu score, b) meteor, c) rouge score, d) spice

5.7 Comparative discussion

While databases 1 and 2 were utilized to assess TP 90, a comparative analysis is conducted to illustrate the superior performance of the HSO-BiLSTM models compared to existing models. The metrics for dataset 1 acquired values are 0.50, 0.32, 0.55 and 26.66, and similarly for dataset 2 acquired values are 0.49, 0.32, 0.57 and 26.58.

Table1: Comparative discussion table for D1 and D2

Models	D1				D2			
	Bleu score	meteor	rouge	Spice	Bleu score	meteor	rouge	Spice
Gated Recurrent Unit Model	0.25	0.19	0.26	5.8	0.24	0.17	0.27	5.4
LSTM Model	0.29	0.26	0.37	9.9	0.28	0.25	0.32	9.2
BiLSTM Model	0.38	0.32	0.46	8.7	0.37	0.31	0.49	8.2
Harris Hawks Optimization Based BiLSTM	0.42	0.45	0.59	7.1	0.41	0.44	0.58	7.0
Harmony Search Optimization Based BiLSTM	0.49	0.57	0.65	2.0	0.48	0.56	0.64	2.2
HSO-BiLSTM	0.50	0.32	0.55	26.66	0.49	0.32	0.57	26.58

6. Summary

In this research, the aim is to enhance the process of generating captions for visual data. Here use two datasets: Flickr30k and COCO which start by organizing images and their transcripts. Then, process images and transcripts separately. For images, enhance their features using VGG-

16. For transcripts, build a vocabulary, tokenize text, and prepare sequences. These enhanced features are combined to optimize a BiLSTM model. This optimized BiLSTM generates accurate captions for transcripts. To boost the model's power, advanced techniques like Harris Hawk optimization and Harmony Search are used. These techniques fine-tune the model for better performance. The result is an improved BiLSTM that crafts rich captions, making the transcripts come alive. This research transforms caption generation for visual data by combining cutting-edge methods into a powerful engine of accurate and meaningful captions. The metrics for dataset 1 acquired values are 0.50, 0.32, 0.55 and 26.66, and similarly for dataset 2 acquired values are 0.49, 0.32, 0.57 and 26.58.

References

- [1] X. Liu, W. Liu, and W. Xing, "Image Caption Generation with Local Semantic Information and Global Information." In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/S CI)*, pp. 680-685. IEEE, 2019.
- [2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A neural image caption generator," *Computer Vision and Pattern Recognition*, pp. 3156-3164, 2015.
- [3] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural Networks (m-RNN)," *International Conference On Learning Representations*, 2015.
- [4] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 28, no.10, pp. 2222-2232, 2017.
- [5] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," *International Conference on Computer Vision*, 2017, pp. 4904-4912.
- [6] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image Captioning with Semantic Attention," *Computer Vision and Pattern Recognition*, pp. 4651-4659, 2016.
- [7] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," *Multimodal Neural Language Models. International Conference on Machine Learning*, pp. 595-603, 2014.
- [8] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi, "Collective generation of natural image descriptions." In: *Annual meeting of the association for computational linguistics*, pp 359-368. 2012.
- [9] P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi, "Treetalk: composition and compression of trees for image descriptions." *Trans Assoc Comput Linguist.* 2(10):351-62, 2014.
- [10] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention." in: *Proceedings of the AAAI*, pp. 4133-4139, 2017.
- [11] Z. Yang, Y. Yuan, Y. Wu, W.W. Cohen, and R.R. Salakhutdinov, "Review networks for caption generation," in: *Proceedings of the NIPS*, pp. 2361-2369, 2016.
- [12] K. Fu, J. Jin, R. Cui, F. and Sha, C. Zhang, "Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts," *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (12), 2321-2334, 2017.
- [13] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (6), 1367-1381, 2018.
- [14] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in: *Proceedings of the IEEE ICCV*, pp. 22-29, 2017.
- [15] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in: *Proceedings of the CVPR*, pp. 4651-4659, 2016.
- [16] M. Hermans, and B. Schrauwen, "Training and analysing deep recurrent neural networks," in: *Proceedings of the NIPS*, pp. 190-198, 2013.
- [17] V. Yngve, "A model and an hypothesis for language structure," *Proc. Am. Philos. Soc.* 104 (5), 444-466, 1960.
- [18] J. Johnson, Karpathy A. Fei-Fei, and L. Denscap, "Fully convolutional localization networks for dense captioning." In: *IEEE Conference on computer vision and pattern recognition*, pp 4565-4574. 2016.
- [19] A. Karpathy, Li FF. "Deep visual-semantic alignments for generating image descriptions." In: *IEEE conference on computer vision and pattern recognition*, pp 3128-3137. 2015.
- [20] A. Krizhevsky, I. Sutskever, and GE. Hinton, "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems*, pp 1097-1105. 2012.
- [21] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, T. Berg, "Baby talk: understanding and generating simple image descriptions." In: *IEEE conference on computer vision and pattern recognition*, pp 1601-1608. 2011.
- [22] S. Ding, S. Qu, Y. Xi, A.K. Sangaiah, and S. Wan, "Image caption generation with high-level image

- features." *Pattern Recognition Letters*, 123: 89-95, 2019.
- [23] Y.H. Tan, and C.S. Chan, "Phrase-based image caption generator with hierarchical LSTM network." *Neurocomputing* 333: 86-100, 2019.
- [24] X. He, B. Shi, X. Bai, G.S. Xia, Z. Zhang, and W. Dong, "Image caption generation with part of speech guidance." *Pattern Recognition Letters*, 119: 229-237, 2019.
- [25] G. Ding, M. Chen, S. Zhao, H. Chen, J. Han, and Q. Liu, "Neural image caption generation with weighted training and reference." *Cognitive Computation* 11, no. 6: 763-777, 2019.
- [26] A.Yuan, X. Li, and X. Lu, "3G structure for image caption generation." *Neurocomputing* 330: 17-28, 2019.
- [27] L. Cheng, W. Wei, X. Mao, Y. Liu, and C. Miao, "Stack-VS: Stacked visual-semantic attention for image caption generation." *IEEE Access* 8: 154953-154965, 2020.
- [28] H. Zhang, C. Ma, Z. Jiang, and J. Lian, "Image Caption Generation Using Contextual Information Fusion With Bi-LSTM-s." *IEEE Access* 11: 134-143, 2022.
- [29] A.A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications." *Future generation computer systems* 97: 849-872, 2019.
- [30] O.M.D. Alia, and R. Mandava, "The variants of the harmony search algorithm: an overview." *Artificial Intelligence Review* 36: 49-68, 2011.
- [31] R. Dey, and F.M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks." *In 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pp. 1597-1600, 2017.
- [32] F. Shahid, A. Zameer, and M. Muneeb, "A novel genetic LSTM model for wind power forecast." *Energy* 223: 120069, 2021.
- [33] A.A. Sharfuddin, M.N. Tihami, and M.S. Islam, "A deep recurrent neural network with bilstm model for sentiment classification." *In 2018 International conference on Bangla speech and language processing (ICBSLP)*, pp. 1-4. IEEE, 2018.
- [34] R. J. Kavitha, C. Thiagarajan, P. Indira Priya, A. Vivek Anand, Essam A. Al-Ammar, Madhappan Santhamoorthy, and P. Chandramohan. "Improved Harris Hawks Optimization with Hybrid Deep Learning Based Heating and Cooling Load Prediction on residential buildings." *Chemosphere* 309: 136525, 2022.