applied sciences

MDPI

*Article*

# Metaheuristics Optimization with Deep Learning Enabled Automated Image Captioning System

**Mesfer Al Duhayyim [1,*], Sana Alazwari [2], Hanan Abdullah Mengash [3], Radwa Marzouk [3], Jaber S. Alzahrani [4], Hany Mahgoub [5,6], Fahd Althukair [7] and Ahmed S. Salama [8]**

[1] Department of Computer Science, College of Sciences and Humanities-Aflaj, Prince Sattam Bin Abdulaziz University, Al-Kharj 16278, Saudi Arabia
[2] Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; alazwari.s@tu.edu.sa
[3] Department of Informaion Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia; hamengash@pnu.edu.sa (H.A.M.); rmMarzouk@pnu.edu.sa (R.M.)
[4] Department of Industrial Engineering, College of Engineering at Alqunfudah, Umm Al-Qura University, Mecca 24382, Saudi Arabia; jszahrani@uqu.edu.sa
[5] Department of Computer Science, College of Science & Art at Mahayil, King Khalid University, Abha 62529, Saudi Arabia; hmahgoub@kku.edu.sa
[6] Department of Computer Science, Faculty of Computers and Information, Menoufia University, Shibin El Kom 32511, Egypt
[7] Department of Electrical Engineering and Computer Sciences, College of Engineering, University of California, Berkeley, CA 94720, USA; falthukair@berkeley.edu
[8] Department of Electrical Engineering, Faculty of Engineering & Technology, Future University in Egypt, New Cairo 11845, Egypt; asalama@fue.edu.eg
* Correspondence: m.alduhayyim@psau.edu.sa

check for updates

**Citation:** Al Duhayyim, M.; Alazwari, S.; Mengash, H.A.; Marzouk, R.; Alzahrani, J.S.; Mahgoub, H.; Althukair, F.; Salama, A.S. Metaheuristics Optimization with Deep Learning Enabled Automated Image Captioning System. *Appl. Sci.* **2022**, *12*, 7724. https://doi.org/10.3390/app12157724

Academic Editor: Andrea Prati

Received: 21 May 2022
Accepted: 29 July 2022
Published: 31 July 2022



**Abstract:** Image captioning is a popular topic in the domains of computer vision and natural language processing (NLP). Recent advancements in deep learning (DL) models have enabled the improvement of the overall performance of the image captioning approach. This study develops a metaheuristic optimization with a deep learning-enabled automated image captioning technique (MODLE-AICT). The proposed MODLE-AICT model focuses on the generation of effective captions to the input images by using two processes involving encoding unit and decoding unit. Initially, at the encoding part, the salp swarm algorithm (SSA), with a HybridNet model, is utilized to generate effectual input image representation using fixed-length vectors, showing the novelty of the work. Moreover, the decoding part includes a bidirectional gated recurrent unit (BiGRU) model used to generate descriptive sentences. The inclusion of an SSA-based hyperparameter optimizer helps in attaining effectual performance. For inspecting the enhanced performance of the MODLE-AICT model, a series of simulations were carried out, and the results are examined under several aspects. The experimental values suggested the betterment of the MODLE-AICT model over recent approaches.

**Keywords:** image captioning; natural language processing; deep learning; machine learning; metaheuristics

## 1. Introduction

Presently, a significant number of images have been produced from many origins such as advertisements, the internet, document diagrams, and news articles. Such origins have images which viewers should analyze themselves [1]. Many images do not contain descriptions; however, human beings can mostly understand them without having any detailed captions. However, machinery should make an interpretation in certain forms of image captions whenever human beings require automated image captions from it. Image captioning is considered significant on numerous grounds [2]. For instance, it is utilized for automated image indexing. Image indexing plays a vital role in content-based image

retrieval (CBIR), and thus is implemented in numerous areas involving digital libraries, biomedicine, the military, education, web searching, and commerce. Mass media platforms such as Twitter and Facebook could straight away produce descriptions from images [3]. The descriptions might involve places (e.g., beach, cafe), things that are worn, and, most significantly, the activities that are taking place.

Image captioning basically involves natural language processing (NLP) and computer vision. Computer vision is helpful in recognizing and understanding the situation in an image [4]; NLP transforms semantic knowledge into a descriptive line. Retrieval of the semantic matter of an image and communicating it in a structure which human beings can understand becomes extremely complex. The complete image captioning method not only gives information, but it also further reveals the connection among the substances [5]. Image captioning consists of numerous applications—for example, as an aid advanced for guiding the person having visual disabilities while traveling alone [6]. This is made possible by changing the scenario into text and transforming the text into voice messages. Image captioning is also utilizes mass communication for the automatic generation of the caption for an image which is posted or to explain a video [7,8]. Moreover, automated image captioning might enhance the Google image search method by changing the image to a caption and, after, by utilizing the keywords for additional related searches [9].

Image realization mostly relies on acquiring image features. The methods utilized for understanding the image of two categories are deep machine learning (ML)-related methods and old machine learning-related methods (2). In conventional ML, handcrafted features, namely the histogram of oriented gradients (HOG), local binary patterns (LBPs), and scale-invariant feature transform (SIFT), and a grouping of these features, were broadly utilized. In such methods, features have been derived from the input unit [10]. They were passed afterward to a classifier such as support vector machines (SVMs) for article classification. Furthermore, real-world data such as video and images become complicated and contain diverse semantic interpretations [11]. Conversely, in ML-related methods, features were studied automatically from training data, and they could manage a big and varied set of videos and images. For instance, convolutional neural networks (CNNs) were broadly employed for feature learning, and a classifier such as Softmax can be utilized for categorization. CNN can be usually tracked by recurrent neural networks (RNNs) for generating captions [12].

This study develops a metaheuristic optimization with a deep learning-enabled automated image captioning technique (MODLE-AICT). The proposed MODLE-AICT model aims for the generation of effective captions to the input images by using two processes involving an encoding unit and a decoding unit. At the encoding part, the salp swarm algorithm (SSA) with a HybridNet model is utilized to generate effectual input image representation using fixed-length vectors. Then, the decoding part includes a bidirectional gated recurrent unit (BiGRU) model to generate descriptive sentences. For examining the enhanced performance of the MODLE-AICT model, a series of simulations were carried out, and the results are examined under several aspects.

## 2. Prior Image Captioning Techniques

In Zhao et al. [13], a fine-grained, structured attention-related technique was suggested when using the structural features of semantic matters in high-resolution distant sensing images. The segmentation is mutually trained with captioning in a unified outline with no need for pixel-wise annotations. Hoxha et al. [14] provide an RSIR technique which mainly focuses on exploiting and producing written descriptions to precisely define the relations among the matters and their features in RS images including captions (e.g., sentences). The initial level focuses to encrypt the image's visual characteristics and later convert the encrypted features to a textual description which sums up the image content-containing captions. The next level focuses on converting the produced textual descriptions as to semantically useful feature vectors. Lastly, estimating the likeness among the textual

descriptions' vectors of query images with that of archive images restores images of high likeness to the query image.

Wang et al. [15] suggested an end wise trainable deep bidirectional LSTM (Bi-LSTM) method for addressing the issue. With the combination of two separate LSTM networks and a deep CNN (DCNN), this methodology can learn long-term visual–language interactions with the help of future and historical context data at a high level semantic area. In Chang et al. [16], an advanced image captioning method—with image captioning, object detection, and color analysis—was suggested for the automated generation of the textual descriptions of images. In an encrypted–decrypted method for image captioning, VGG16 can be utilized as an encoder and an LSTM network and can be employed as a decoder.

Xiong et al. [17] recommend a hierarchical transformer-related medical imaging report generation technique. This presented technique has two parts: one is ann image encoder that extracts heuristic visual features through a bottom-up attention algorithm; the other is a non-recurrent captioning decoding technique that enhances computational efficacy through parallel computation. Wang et al. [18] suggested an original methodology to indirectly design the association between areas of interest in an image by a graph NN along with the original context-aware attention system for guiding attention selection by completely memorizing formerly attended visual contents.

In Al-Malla et al. [19], the authors introduced a new method to apply the Generative Adversarial Network into sequence generation. The greedy decoding method is utilized for generating an effective baseline reward for self-critical training. The visual and semantic relationship of diverse objects are combined into local-relation attention. The authors in [20] developed an attention-based encoder–decoder deep model which utilizes convolutional features derived from a CNN model that is pre-trained on ImageNet (Xception) along with the object features derived by the YOLOv4 model, pre-trained on MS COCO. The authors also introduced a novel positional encoding scheme for object features, termed the "importance factor".

## 3. The Proposed Model

In this study, a new MODLE-AICT technique has been developed for the generation of effective captions to the input images by using two processes involving an encoding unit and a decoding unit. Primarily, at the encoding part, the SSA with a HybridNet model is utilized to generate effectual input image representation using fixed-length vectors. In addition, the decoding part includes a BiGRU model that is used to generate descriptive sentences. Figure 1 show cases of the block diagram of the MODLE-AICT algorithm.

### 3.1. Data Pre-Processing

At the preliminary level, data pre-processing is performed in different stages as given below.

- Lower case conversion;
- Removal of punctuation marks to decrease complexity;
- Removal of numeric values;
- Tokenization;
- Vectorization (to turn the original strings into integer sequences where each integer represents the index of a word in a vocabulary).

### 3.2. Feature Extraction: HybridNet Model

In this work, the HybridNet model is utilized for generating visual features of the input images. Generally, classification requires intra-class, in-variant features, while reconstruction requires the preservation of each dataset. In order to overcome these shortcomings, HybridNet includes the unsupervised path ($E_u$ and $D_u$) and the discriminative path ($E_c$ and $D_c$). These two $E_u$ and $E_c$ encoders take an $x$ input image and generate $h_c$ and $h_u$ representations, whereas decoders $D_c$ and $D_u$ take $h_c$ and $h_u$, respectively, as an input to generate $\hat{x}$ and $\hat{x}$ partial reconstructions. At last, the $C$ classifier produces a class prediction

by means of a discriminative feature: $\hat{y} = C(h_c)$. Although both paths may have analogous architecture, they have to perform complementary and different roles. The discriminative path needs to extract $h_c$ discriminative feature that must be ultimately well crafted to effectively execute a classifier task and produce a $\hat{x}_c$ partial reconstruction that should not be accurate; meanwhile, retaining each dataset is not a behavior that we want to inspire [21]. As a result, the role of unsupervised paths is complementary to the discriminative path through p in $h_u$ the data lost in $h_c$. Consequently, it produces $\hat{x}$ complementary reconstruction, while, integrating $\hat{x}$ and $x$, the last reconstruction $\hat{x}$ is closer to $x$. The architecture of HybridNet is formulated by using the below expression:

$$h_c = E_c(x)\ \hat{x}_c = D_c(h_c)\ \hat{y} = C(h_c)$$

$$h_u = E_u(x)\ \hat{x}_u = D_u(h_u)\ \hat{x} = \hat{x}_c + \hat{x}_u$$

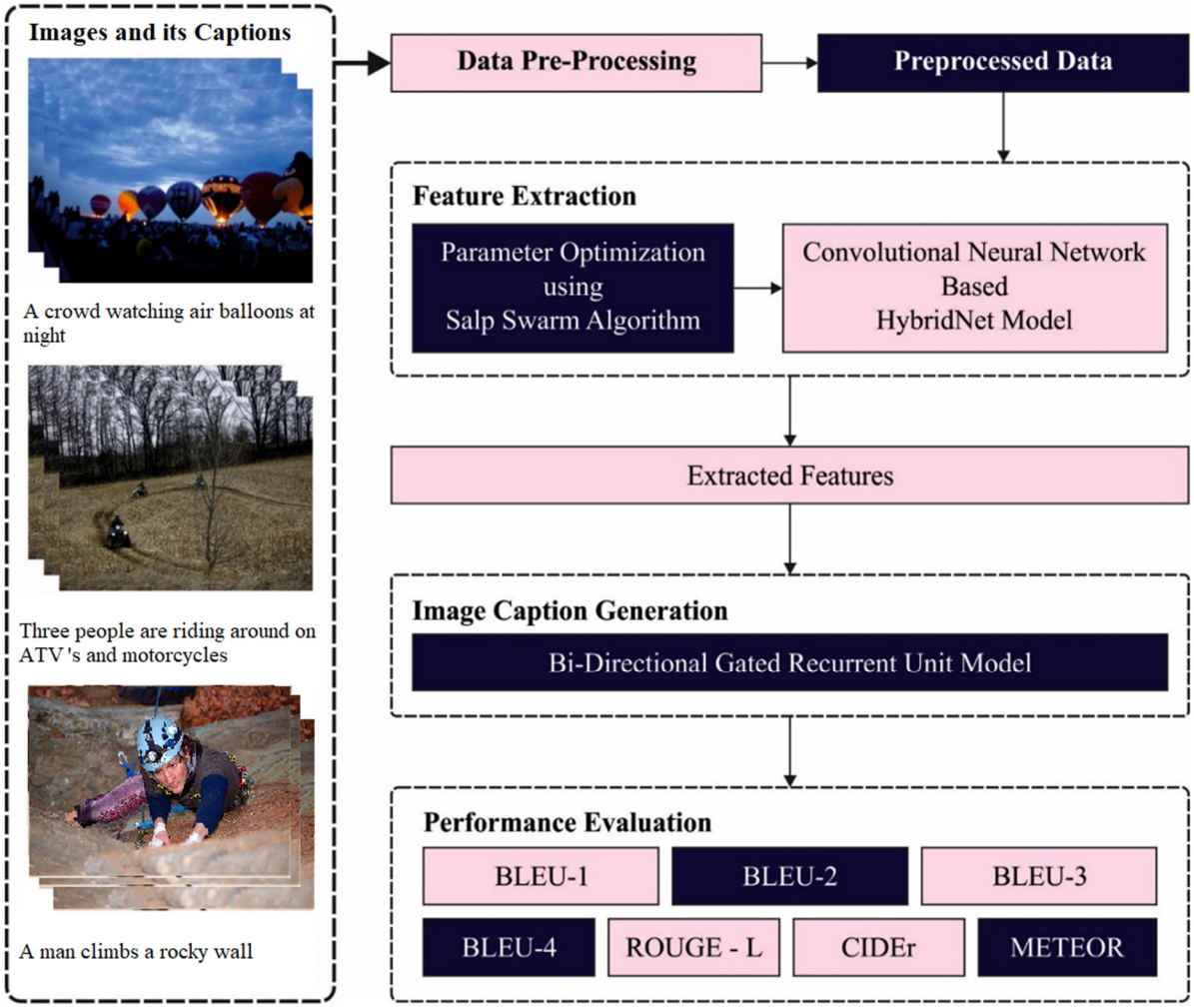


**Figure 1.** Block diagram of MODLE-AICT approach.

It should be noted that the ultimate role of reconstruction is to regularize for the discriminative encoding. The major contribution and challenge of the study is to establish a method to guarantee that both paths would actually perform in such way.

The two major problems that we address are the discriminative path to emphasize discriminative features and the fact that we need these two paths to contribute and co-operate to the reconstruction. In fact, with this framework, we might create two paths that work individually: a reconstruction path $\hat{x} = \hat{x}_u = D(E(x))$ and a classification path $\hat{y} = C(E(x))$ and $\hat{x}_c = 0$. We resolve the issue by using the encoder and decoder architecture along with a proper training and loss function. The HybridNet model has two data paths,

with one generating a class prediction and both generating partial reconstruction that needs to be integrated. In these subsections, we resolve the problem of training this structural design proficiently. It encompasses terms for stability with $\Omega_{stability}$; classification with $\mathcal{L}_{cls}$; last reconstruction with $\mathcal{L}_{rec}$; and intermediate reconstruction with $\mathcal{L}_{rec-interb,l}$ (for layer $l$ and branch $b$).

Moreover, it is followed by a branch complementarity training model. All the terms are weighted through a $\lambda$ variable, respectively:

$$\mathcal{L} = \lambda_c \mathcal{L}_{cls} + \lambda_r \mathcal{L}_{rec} + \sum_{b\epsilon\{c,u\},l} \lambda_{rb,l} \mathcal{L}_{rec-interb,l} + \lambda_s \Omega_{stability} \tag{1}$$

HybridNet architecture is trained on partially labelled data comprised of unlabeled images $\mathcal{D}_{unsup} = \left\{x^{(k)}\right\}_{k=1..N_u}$ and labelled pairs $\mathcal{D}_{sup} = \left\{\left(x^{(k)}, y^{(k)}\right)\right\}_{k=1..N_s}$. All the batches are comprised of $n$ instances, separated into unlabeled images $n_u$ from $\mathcal{D}_{unsup}$ and labelled images $n_s$ from $\mathcal{D}_{sup}$. The classification term is a regular cross-entropy term employed only on the $n_s$-labelled instance, as follows:

$$\ell_{cls} = \ell_{CE}(\hat{y}, y) = -\sum_i y_i \log \hat{y}, \mathcal{L}_{cls} = \frac{1}{n_s} \sum_k \ell_{cls}\left(\hat{y}^{(k)}, y^{(k)}\right) \tag{2}$$

### *3.3. Hyperparameter Optimization*

In order to effectually tune the hyperparameters related to the HybirdNet model, the SSA is exploited. To resolve optimization problems, motion behavior of SSA can be mathematically modelled [22]. Salps are sea creatures that have barrel-shaped, jelly-like bodies and move from place to place by driving water through their bodies from one side to the other sides. They exist as colonies and travel together like chains. Leader and follower are the two most important classes of salps. Leaders lead the chain in a forwarding direction, while followers follow the leader synchronously and in harmony. Similar to a swarm intelligent model, SSA begins with an arbitrary initialization of the swarm of N salps. Variable n is considered to be measured, x symbolizes the position of salp, and y defines the food source-specifying objective of swarm in searching region. Leader salp describes their position by the subsequent formula:

$$x_{i^1} = \begin{cases} y_i + r_1((ub_i - lb_i)r_2 + lb_i), \ r_3 \geq 0, \\ y_i - r_1((ub_i - lb_i)r_2 + lb_i), \ r_3 < 0, \end{cases} \tag{3}$$

In Equation (3), in $i$-th parameter, $x_{i^1}$—position of initial salp; $y_i$—position of food. $ub_i$ and $lb_i$—upper and lower bounds, and $r_1$,$r_2$,$r_3$—arbitrary number.

Among three arbitrary numbers, $r_1$ inhabits the lead position because it balances exploitation and exploration at the time of searching process. It can be formulated as follows:

$$r_1 = 2e(\frac{4I}{L})^2 \tag{4}$$

In Equation (4), $l$ shows existing iteration; $L$—the formerly determined amount of iterations; $r_2$,$r_3$—arbitrary integer lies within [0, 1]. To update the location according to Newton's law of motion, the following mathematical expressions are utilized for followers:

$$X_l^j = 0.5\lambda t^2 + \delta_0 t \tag{5}$$

where $\geq 2$,$x_l^j$—position of $j$-th salp in $i$-th parameter, $t$—time, $\delta_0$—initial speed.

$$\lambda = \frac{\delta_{final}}{\delta_0}, \text{where } \delta = \frac{x - x_0}{r} \tag{6}$$

Assume that $\delta_0 = 0$, $t$—iteration in an optimization issue; the abovementioned formula is transformed into succeeding expression:

$$x_l^j = 0.5\left(x_l^j + x_l^{j-1}\right) \tag{7}$$

In Equation (7), $j \geq 2$. This equation demonstrates that follower salps describe the position according to the preceding salps and their own position. When some salps escape from the restricted searching space, they are carried back within the limitation as follows:

$$X_l^j = \begin{cases} l^j & if x_t^j \leq l^j \\ u^j & if_{X_l^j} \geq u^j \\ \chi_l^j & \text{otherwise} \end{cases} \tag{8}$$

The abovementioned expression is repeatedly executed until the ending condition is met. Note that the food source is sometimes upgraded by exploring and exploiting space around an existing solution, which might determine the best solution. Salp chains, during optimization, have the capacity to move toward global optimum solutions as illustrated in Algorithm 1.

**Algorithm 1** Pseudocode of SSA

1: Input: maximum iterations $L$, population size $m, ub, lb, l = I$
2: Initialization of salp position $\{u_1, u_2, u_3, \ldots\ldots\ldots, u_m\}$
3: While (stopping criteria is not fulfilled)
4: Determine fitness of all salps
5: Arrange salp position based on fitness value
6: Define $F$ as optimal place for present population
7: Upgrade $C_l$
8: For every salp position $(u_i)$
9: If $(i \leq m/2)$ upgrades the position of leading salps
10: Else upgrade the position of follower salp
11: end
12: end
13: Change the salp which crosses higher and lower limits
14: end
15: Display optimum output

### 3.4. Image Captioning

In this study, the decoding part includes the BiGRU model to generate descriptive sentences. A recurrent neural network (RNN) has been successfully used to handle data sequences in different areas [23]. In RNN, the input sequence $= (x_1, \ldots, x_T)$, hidden vector sequence $h = (h_1, \ldots, h_T)$, and output vector sequence $y = (y_1, \ldots, y_T)$ are derived by the given equations:

$$h_\tau = \Phi(Ux_t + WWh_{\tau-1} + b) \tag{9}$$

$$y_t = Vh_t + c \tag{10}$$

Let $\Phi$ be the activation function, and the popular activation function is generally an element-wise application of the sigmoid function. $U$ refers to the input-hidden weight matrixes, $W$ stands for the hidden-hidden weight matrixes, and, in Equation (10), $b$ denotes the hidden bias vector, $V$ signifies the hidden-output weight matrixes, and $c$ denotes the output bias vector. It is nearly impossible to capture the long-term dependency of RNN, as the gradient tends to explode or vanish. Therefore, some research workers have made every effort to develop a more complex activation function to resolve the shortcomings. For instance, the LSTM unit is initially proposed for capturing the long-term dependency. In recent years, other variants of the recurrent unit, such as GRU, is also devised, which are

easier to calculate and have good performance of generalization compared to that of the LSTM unit. Figure 2 depicts the framework of GRU.

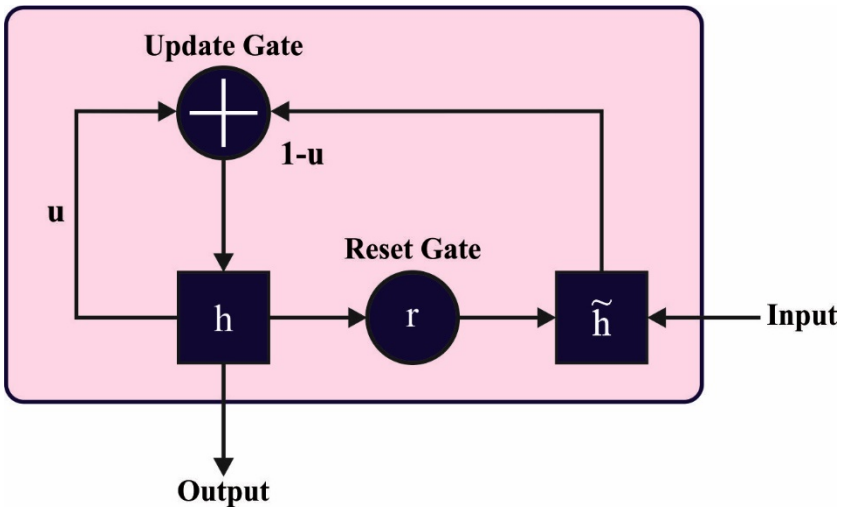


**Figure 2.** Architecture of GRU.

LSTM makes use of an output gate for controlling the exposure of the quantity of memory content.

$$h_t = o_t tanh(c_t) \tag{11}$$

In Equation (11), the output gate is represented as $o_t$ and is calculated as follows:

$$0_t = \sigma(W_0 \cdot [h_{t-1},\ x_t,\ c_l] + b_0) \tag{12}$$

In Equation (12), the logistic function is indicated as $\sigma$. The memory cell $c_t$ is preserved by adding some new memories and eliminating (forgetting) current memories:

$$c_t = f_t c_{t-1} + i_\tau \widetilde{c} + b_c \tag{13}$$

The $\widetilde{c}_t$ new memories are given by:

$$\widetilde{c} = tanh(W_c \cdot [h_{t-1},\ x_t]) \tag{14}$$

The extent to add and remove memories can be controlled by the input gate $i_t$ and the forget gate $f_t$. The forget gate can be calculated by the following equation:

$$f_t = \sigma\Big(W_f \cdot [h_{t-1},\ x_t,\ c_{t-1}] + b_f\Big) \tag{15}$$

and $i_t$ is calculated as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1},\ x_t,\ c_{t-1}] + b_i) \tag{16}$$

From the equation, the corresponding bias vector is indicated as $b$. As with the LSTM unit, GRU uses the gate for controlling the data stream inside a unit; however, there is no memory cell. The $h_r$ hidden state is a linear integration of new hidden states $\widetilde{h}_t$ and the preceding hidden state $h_{t-1}$:

$$h_t = (1 - z_t)h_{t-1} + z_t \widetilde{h}_t \tag{17}$$

In Equation (17), the update gate $z_t$ controls how much its new activation is upgraded. It is calculated as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1},\ x_t]) \tag{18}$$

The $\widetilde{h}_t$ new activation is calculated as follows:

$$\widetilde{h} = tanh(W_h \cdot [r_t \odot h_{t-1},\ x_t]) \tag{19}$$

In Equation (19), the forget gate $r_t$ is the same as the update unit in LSTM:

$$r_l = \sigma(W_r \cdot [h_{t-1},\ x_t]) \tag{20}$$

While typical RNN exploits the preceding data, the bi-directional RNN (BRNN) processes information in two directions. The $y$ output of BRNN is attained by measuring the $\overrightarrow{h_r}$ forward hidden sequence and $\overleftarrow{h_t}$ backward sequence as follows:

$$\overrightarrow{h}_t = \Phi\left(WW_{\underset{\chi h}{\rightarrow}} x_t + W_{\underset{h h}{\rightarrow\rightarrow}} \overrightarrow{h}_{t-1} + b_{\underset{h}{\rightarrow}}\right) \tag{21}$$

$$\overleftarrow{h}_t = \Phi\left(W_{\underset{x h}{\leftarrow}} x_t xh + W_{\underset{h h}{\leftarrow\leftarrow}} \overleftarrow{h}_{t-1} + b_{\underset{h}{\leftarrow}}\right) \tag{22}$$

$$y_t = W_{\underset{h y}{\rightarrow}} \overrightarrow{h}_t + W_{\underset{h y}{\leftarrow}} \overleftarrow{h}_{t-1} + b_y \tag{23}$$

Integrating BRNN with GRU provides BIGRU that is utilized for accessing the long-term data sequence in two directions. As a fault, a diagnoses issue can be generally regarded as a classification issue, and cross entropy is adapted as the loss function. The weighted cross entropy is represented as:

$$f(\theta) = -\sum_{n=1}^{N} w_n \sum_{i=1}^{M} y_i log(\hat{y}_i) \tag{24}$$

In Equation (24), $\theta$ indicates the neural network parameter, $N$ represents the sample count, the number of faults is represented as $M$, and the true label can be indicated as $y_i$ and the predicted probability is represented as $\hat{y}_i$.

## 4. Performance Validation

The experimental validation of the MODLE-AICT model is tested using the Flick8K dataset (https://www.kaggle.com/adityajn105/flickr8k/activity, accessed on 13 March 2022) and MS-COCO 2014 dataset [24]. A comparison study is also made with recent models [25–31]. A few sample images are depicted in Table 1. It contains 8000 images that are each paired with five different captions which provide clear descriptions of the salient entities and events.

### 4.1. Performance Measures

To validate the performance of the presented model, a set of four metrices are utilized, such as BLEU, Meter, CIDEr, and Rouge-L. BLEU [25], a commonly utilized metric for estimating the quality of the produced text. For an effectual image captioning outcome, BLEU values are required to be high, and it is defined using Equation (25):

$$BP = \min\left(1, e^{1-\frac{r}{c}}\right)$$

$$BLEU_N = BP \times e^{\frac{1}{N}\sum_{n=1}^{N} logp_n} \tag{25}$$

where $BP$ denotes penalty factor, $r$ and $c$ represent length of the reference and generated sentences, respectively. METEOR metric relies on word recall rate and single precision weighted harmonic mean. It computes the reconciliation mean of accuracy and recalls amongst the optimum candidate and reference translations. It is defined as follows:

$$METEOR = (1 - Pen)F_{mean} \tag{26}$$

where $\alpha$, $\gamma$, and $\theta$ denotes default parameters.

**Table 1.** Sample Images and its captions.

| Sample Image | Different Captions |
| --- | --- |
|  | A crowd watching air balloons at night<br>A group of hot air balloons lit up at night<br>People are watching hot air balloons in the park<br>People watching hot air balloons<br>Seven large balloons are lined up at night-time near a crowd |
|  | A man climbs a rocky wall<br>A climber wearing a blue helmet and headlamp is attached to a rope on the rock face<br>A rock climber climbs a large rock<br>A woman in purple snakeskin pants climbs a rock<br>Person with blue helmet and purple pants is rock climbing |
|  | People on ATVs and dirt bikes are traveling along a worn path in a field surrounded by trees<br>Three people are riding around on ATVs and motorcycles<br>Three people on motorbikes follow a trail through dry grass<br>Three people on two dirt bikes and one four-wheeler are riding through brown grass<br>Three people ride off-road bikes through a field surrounded by trees |

The CIDEr index assumes each sentence as a "document" and represents in the form of a TF-IDF vector. It computes the cosine similarity amongst the created caption $(s_{ij})$ and original caption by the use of a score value.

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i)^T g^n(s_{ij})}{\|g^n(c_i)\| \times \|g^n(s_{ij})\|} \quad (27)$$

ROUGE is another similarity measurement model that is mainly based on the recall rate. It determines the co-occurrence probability of N-gram in the reference translation and the translation to be investigated. It is defined using Equation (28).

$$ROUGE - N = \frac{\sum_{S \in \{ReferencesSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferencesSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (28)$$

### *4.2. Result Analysis*

Table 2 and Figure 3 inspect a detailed result analysis of the MODLE-AICT model on the test Flickr8K dataset [23–28]. The results implied that the MODLE-AICT model has gained effectual outcomes over other models. For instance, based on BLEU-1, the MODLE-AICT model obtained a higher BLEU-1 of 69.06, whereas the M-RNN, G-NICG,

L-Bilinear, DVS, ResNet50, VGA-16, and HPTDL models attained a lower BLEU-1 of 59.18, 64.13, 65.96, 58.35, 62.65, 67.69, and 68.26, respectively. At the same time, based on BLEU-4, the MODLE-AICT technique has attained a higher BLEU-4 of 27.80, whereas the M-RNN, G-NICG, L-Bilinear, DVS, ResNet50, VGA-16, and HPTDL methods have acquired a lower BLEU-4 of 14.19, 16.12, 18.49, 17.08, 26.16, 23.25, and 26.71, correspondingly.

**Table 2.** Result analysis of MODLE-AICT algorithm with approaches on Flickr8K dataset.

| Methods | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| M-RNN Model [24] | 59.18 | 29.09 | 24.17 | 14.19 |
| G-NICG Model [26] | 64.13 | 42.64 | 27.11 | 16.12 |
| L-Bilinear Model [27] | 65.96 | 43.29 | 28.63 | 18.49 |
| DVS Model [28] | 58.35 | 37.98 | 25.54 | 17.08 |
| ResNet50 Model [23] | 62.65 | 46.28 | 37.26 | 26.16 |
| VGA-16 Model [23] | 67.69 | 44.34 | 33.99 | 23.25 |
| HPTDL Model [25] | 68.26 | 46.16 | 37.81 | 26.71 |
| MODLE-AICT | 69.06 | 47.26 | 38.78 | 27.80 |

Table 3 and Figure 4 examine a detailed classification analysis of the MODLE-AICT system on the test Flickr8K dataset. The results implied that the MODLE-AICT technique has attained effectual outcomes over other models. For example, based on METEOR, the MODLE-AICT approach has gained higher METEOR of 30, whereas the SCST-IN, SCST-ALL, G-NIC, A-NIC, DenseNet, and HPTDL methodologies have obtained a lower METEOR of 20, 23, 19, 21, 25, and 28, correspondingly. Meanwhile, based on Rouge-L, the MODLE-AICT approach has received a higher Rouge-L of 53, whereas the SCST-IN, SCST-ALL, G-NIC, A-NIC, DenseNet, and HPTDL algorithms have attained a lower Rouge-L of 49, 42, 43, 48, 43, and 46, correspondingly.

**Table 3.** Classification analysis of MODLE-AICT algorithm with approaches on Flickr8K dataset.

| Methods | METEOR | CIDEr | Rouge-L |
|---|---|---|---|
| SCST-IN Model [29] | 20.00 | 161.00 | 49.00 |
| SCST-ALL Model [29] | 23.00 | 154.00 | 42.00 |
| G-NIC Model [26] | 19.00 | 153.00 | 43.00 |
| A-NIC Model [26] | 21.00 | 160.00 | 48.00 |
| DenseNet Model [24] | 25.00 | 173.00 | 43.00 |
| HPTDL Model [25] | 28.00 | 175.00 | 46.00 |
| MODLE-AICT | 30.00 | 179.00 | 53.00 |

A comparison study of the MODLE-AICT model with recent models on the Flickr8K dataset is shown in Figure 5. The figure implied that the SCST-IN and SCST-ALL models have obtained lower performance than other models. This was followed by the G-NIC, A-NIC, and DenseNet models, which attained moderately closer results. Along with that, the HPTDL model accomplished a reasonable performance. However, the MODLE-AICT model has shown enhanced performance over other models on the test Flickr8K dataset.

The training accuracy (TA) and validation accuracy (VA) attained by the MODLE-AICT approach on the Flickr8K dataset is demonstrated in Figure 6. The experimental outcome implied that the MODLE-AICT technique has gained maximum values of TA and VA. Specifically, the VA seemed to be higher than TA.

The training loss (TL) and validation loss (VL) achieved by the MODLE-AICT methodology on the Flickr8K dataset are established in Figure 7. The experimental outcome inferred that the MODLE-AICT system accomplished the lowest values of TL and VL. Specifically, the VL seemed to be lower than TL.

Table 4 and Figure 8 depict the detailed results of the analysis of the MODLE-AICT system on the test MS-COCO 2014 dataset. The results implied that the MODLE-AICT approach obtained effectual outcomes over other models. For instance, based on BLEU-1,

the MODLE-AICT method gained a higher BLEU-1 of 75.12, whereas the M-RNN, G-NICG, L-Bilinear, DVS, ResNet50, VGA-16, and HPTDL methodologies received a lower BLEU-1 of 49.60, 67.92, 71.75, 63.86, 73.57, 70.30, and 74.28, correspondingly. Nonetheless, based on BLEU-4, the MODLE-AICT technique has gained a higher BLEU-4 of 34.75, whereas the M-RNN, G-NICG, L-Bilinear, DVS, ResNet50, VGA-16, and HPTDL methodologies acquired a lower BLEU-4 of 10.95, 24.94, 24.35, 23.29, 32.52, 30.06, and 33.96, correspondingly.

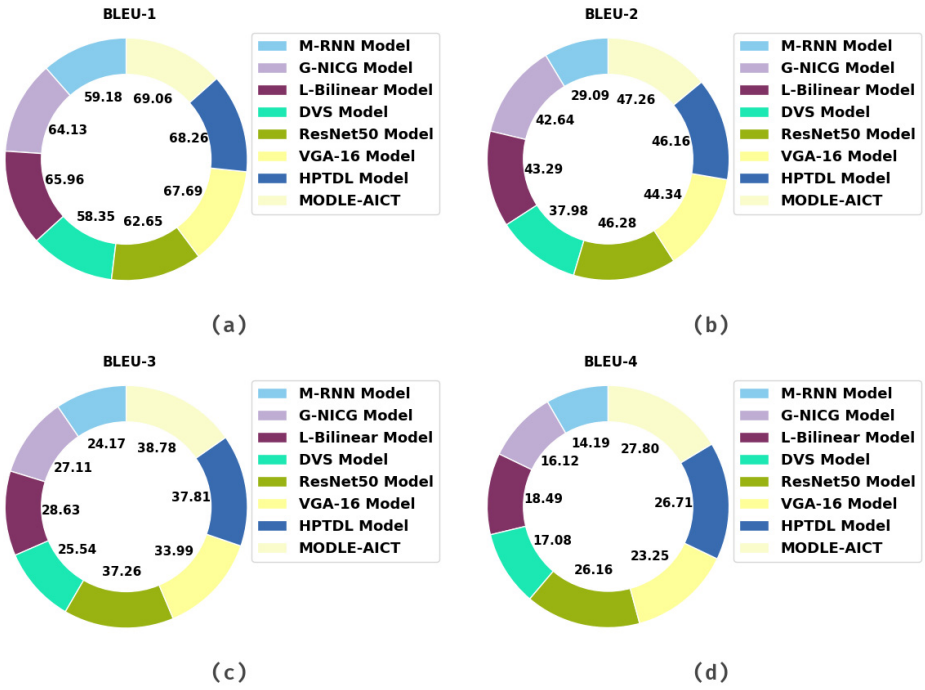


**Figure 3.** Result analysis of MODLE-AICT algorithm on Flickr8K dataset (**a**) BLEU-1, (**b**) BLEU-2, (**c**) BLEU-3, and (**d**) BLEU-4.

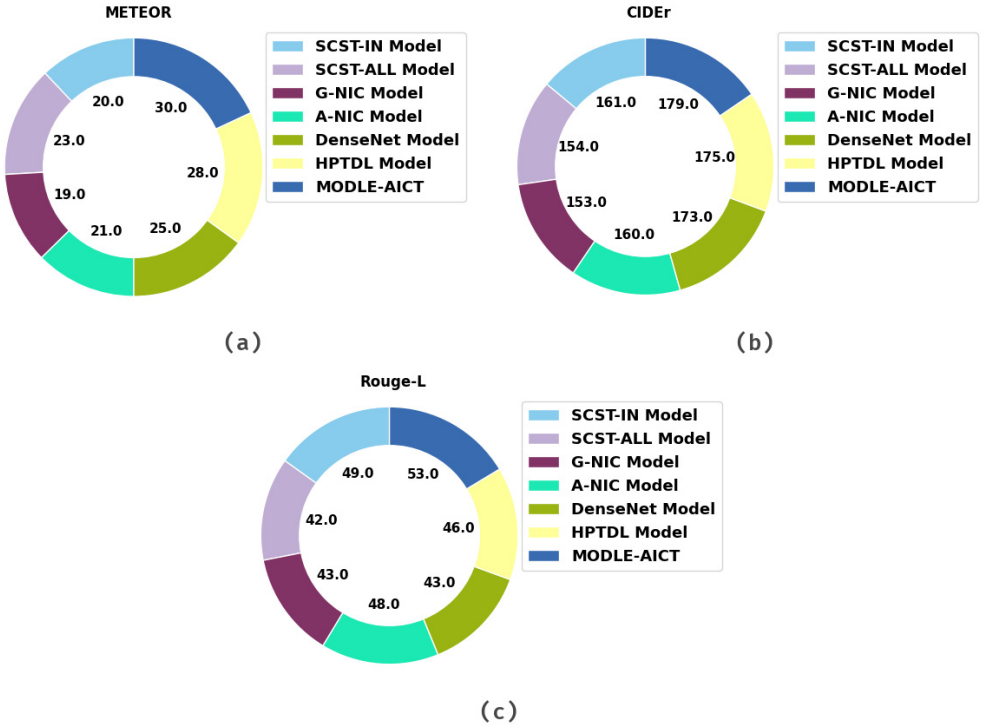


**Figure 4.** Result analysis of MODLE-AICT algorithm on Flickr8K dataset (**a**) METEOR, (**b**) CIDEr, and (**c**) Rouge-L.

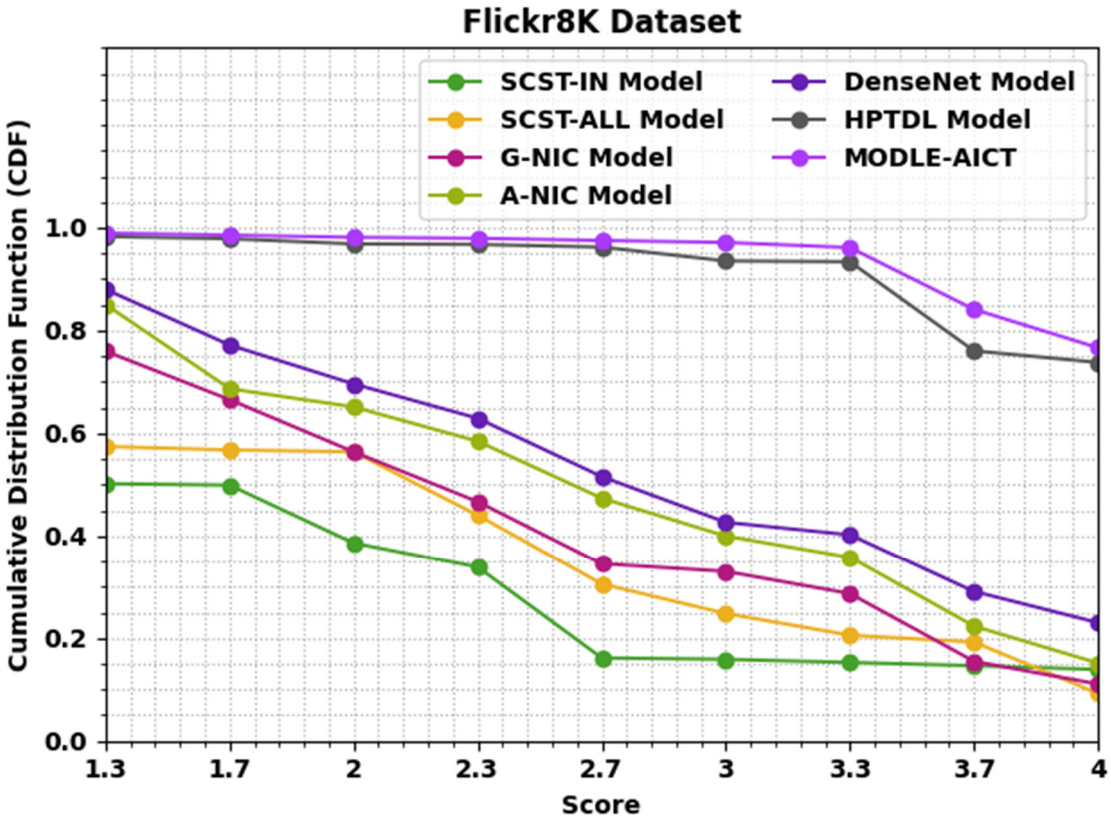


**Figure 5.** CDF analysis of MODLE-AICT technique with existing approaches on Flickr8K dataset.

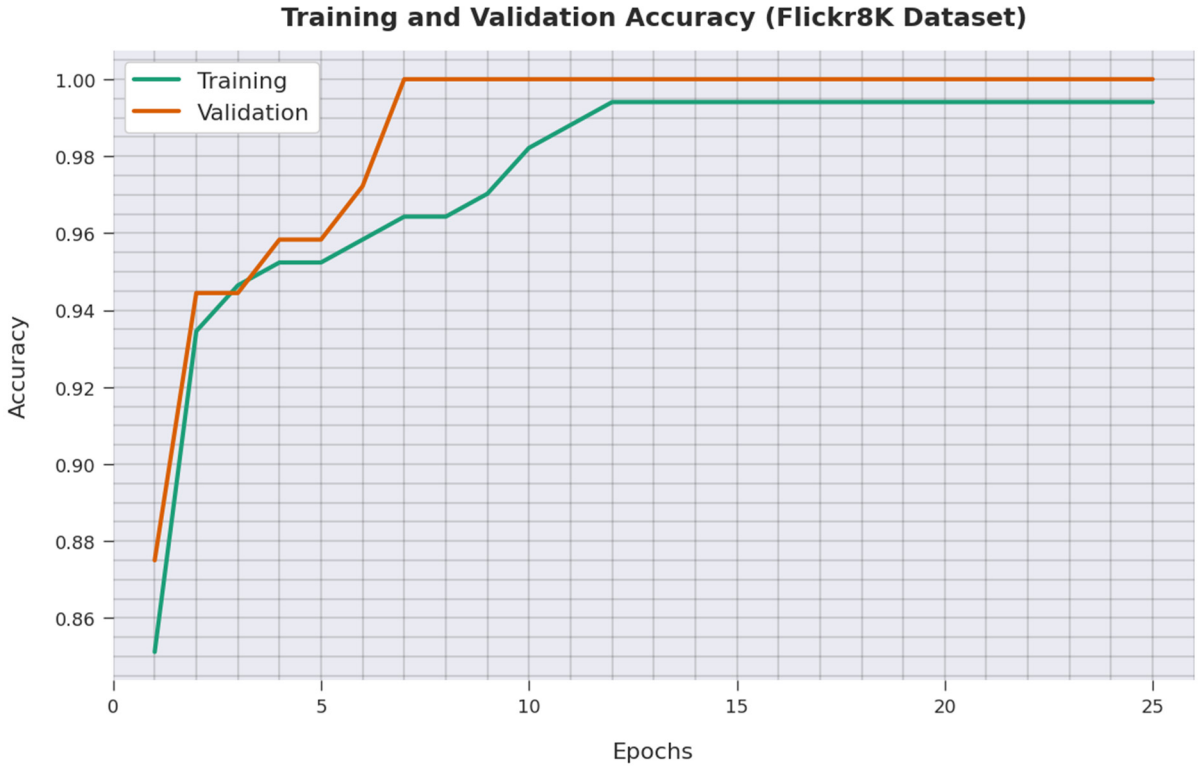


**Figure 6.** TA and VA analysis of MODLE-AICT technique on Flickr8K dataset.


**Figure 7.** TL and VL analysis of MODLE-AICT technique on Flickr8K dataset.

**Table 4.** Result analysis of MODLE-AICT algorithm with approaches on MS-COCO 2014 dataset.

| Methods | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| KNN Model [25] | 49.60 | 28.60 | 17.00 | 10.95 |
| G-NICG Model [26] | 67.92 | 46.23 | 34.03 | 24.94 |
| L-Bilinear Model [27] | 71.75 | 49.22 | 34.65 | 24.35 |
| DVS Model [28] | 63.86 | 44.98 | 32.58 | 23.29 |
| ResNet50 Model [23] | 73.57 | 57.21 | 42.05 | 32.52 |
| VGA16 Model | 70.30 | 53.72 | 40.45 | 30.06 |
| VGA-16 Model [23] | 74.28 | 59.39 | 43.33 | 33.96 |
| HPTDL Model [25] | 75.12 | 60.21 | 44.22 | 34.75 |

Table 5 and Figure 9 review a detailed classification analysis of the MODLE-AICT system on the test MS-COCO 2014 dataset. The results implied that the MODLE-AICT methodology acquired effectual outcomes over other models. For instance, based on METEOR, the MODLE-AICT methods have obtained a higher METEOR score of 37, whereas the SCST-IN, SCST-ALL, G-NIC, A-NIC, DenseNet, and HPTDL approaches have attained a lower METEOR score of 22, 25, 21, 24, 24, and 34, correspondingly. Meanwhile, based on Rouge-L, the MODLE-AICT technique acquired a higher Rouge-L of 63, whereas the SCST-IN, SCST-ALL, G-NIC, A-NIC, DenseNet, and HPTDL algorithms acquired a lower Rouge-L of 51, 59, 51, 58, 57, and 60, correspondingly.

A comparison study of the MODLE-AICT technique with recent models on the MS-COCO 2014 dataset is shown in Figure 10. The figure implied that the SCST-IN and SCST-ALL methodologies acquired a lower performance than the other models. Then, the G-NIC, A-NIC, and DenseNet approaches have gained moderately closer results. Moreover, the HPTDL approach has tried to accomplish reasonable performance. However, the MODLE-AICT system has shown an enhanced performance over other models on the test MS-COCO 2014 dataset.

The TA and VA attained by the MODLE-AICT technique on the MS-COCO 2014 dataset are demonstrated in Figure 11. The experimental outcome implied that the MODLE-AICT method has gained maximum values of TA and VA. Specifically, the VA seemed to be higher than TA.

**Table 5.** Classification analysis of MODLE-AICT algorithm with approaches to MS-COCO 2014 dataset.

| Methods | METEOR | CIDEr | Rouge-L |
|---|---|---|---|
| SCST-IN Model [29] | 22.00 | 109.00 | 51.00 |
| SCST-ALL Model [29] | 25.00 | 114.00 | 59.00 |
| G-NIC Model [26] | 21.00 | 111.00 | 51.00 |
| A-NIC Model [26] | 24.00 | 110.00 | 58.00 |
| DenseNet Model [24] | 24.00 | 122.00 | 57.00 |
| HPTDL Model [25] | 34.00 | 125.00 | 60.00 |
| MODLE-AICT | 37.00 | 129.00 | 63.00 |

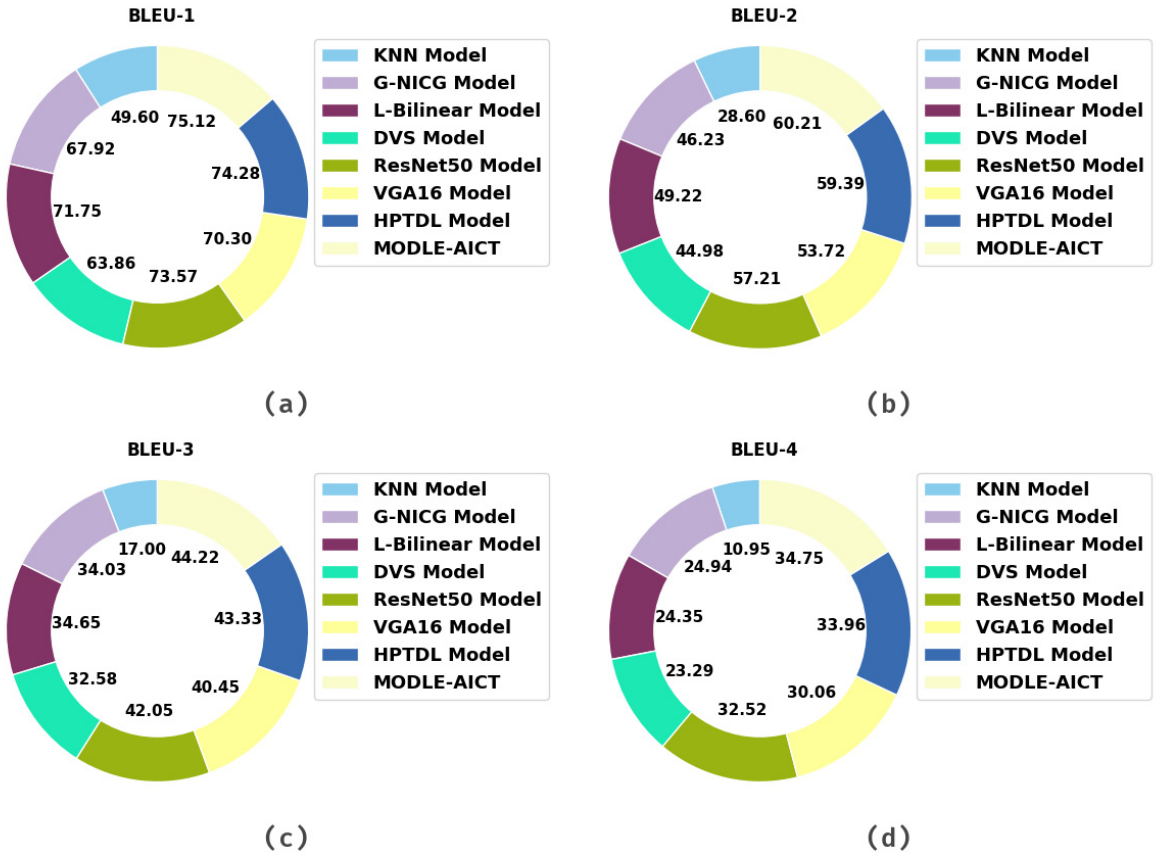


**Figure 8.** Result analysis of the MODLE-AICT algorithm on the MS-COCO 2014 dataset: (**a**) BLEU-1, (**b**) BLEU-2, (**c**) BLEU-3, and (**d**) BLEU-4.

The TL and VL achieved by the MODLE-AICT approach on the MS-COCO 2014 dataset are established in Figure 12. The experimental outcome inferred that the MODLE-AICT methodology has accomplished least values of TL and VL. Specifically, the VL seemed to be lower than TL. From the detailed results and discussion, it is assured that the proposed model has shown effective outcomes on the image captioning process.

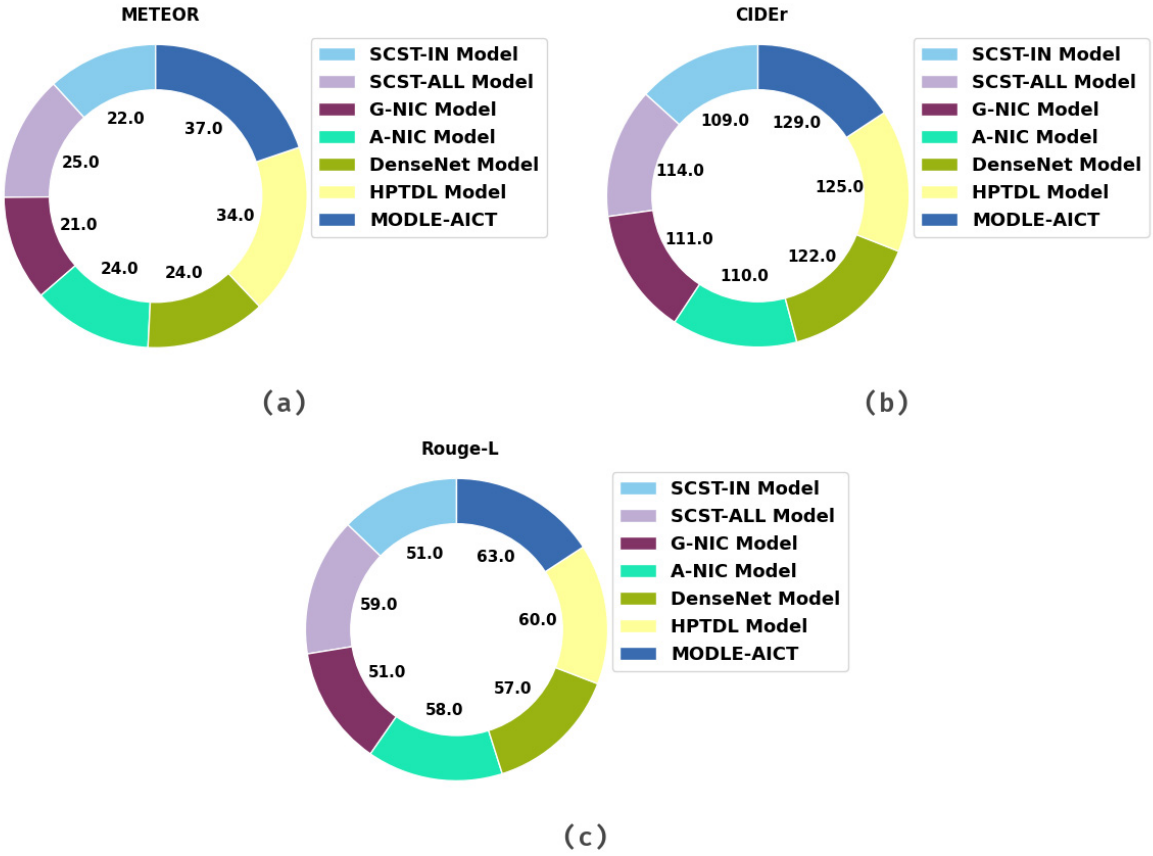


**Figure 9.** Result analysis of MODLE-AICT algorithm on MS-COCO 2014 dataset (**a**) METEOR, (**b**) CIDEr, and (**c**) Rouge-L.

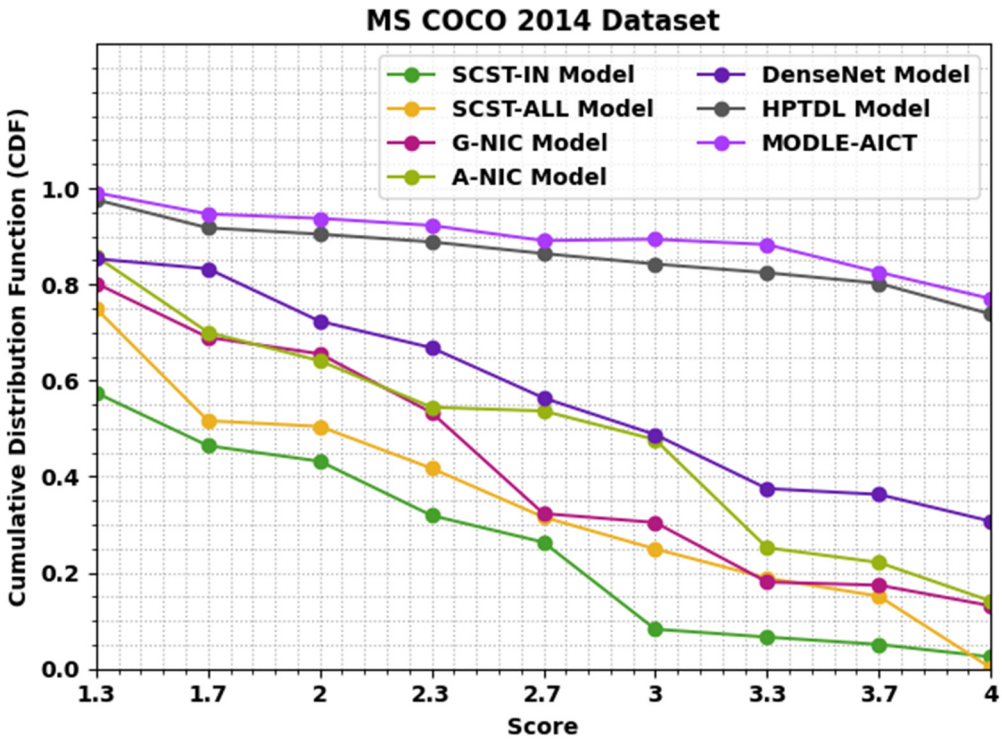


**Figure 10.** CDF analysis of MODLE-AICT technique with existing approaches on MS-COCO 2014 dataset.

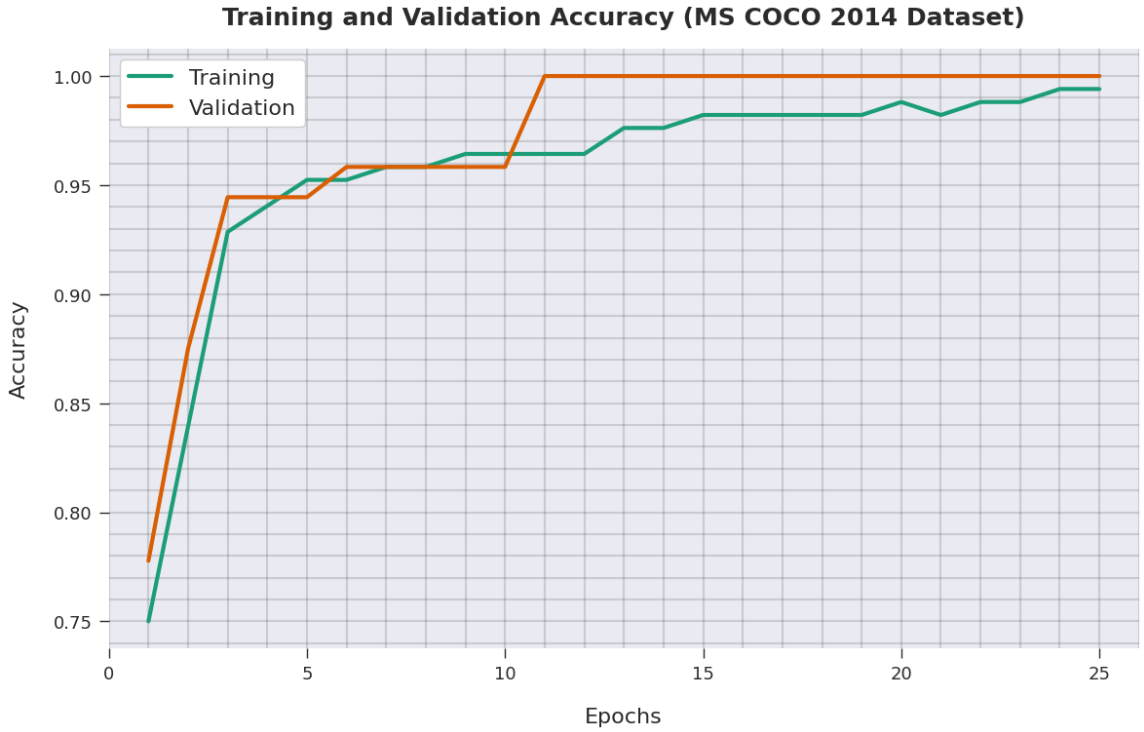


**Figure 11.** TA and VA analysis of MODLE-AICT technique on MS-COCO 2014 dataset.

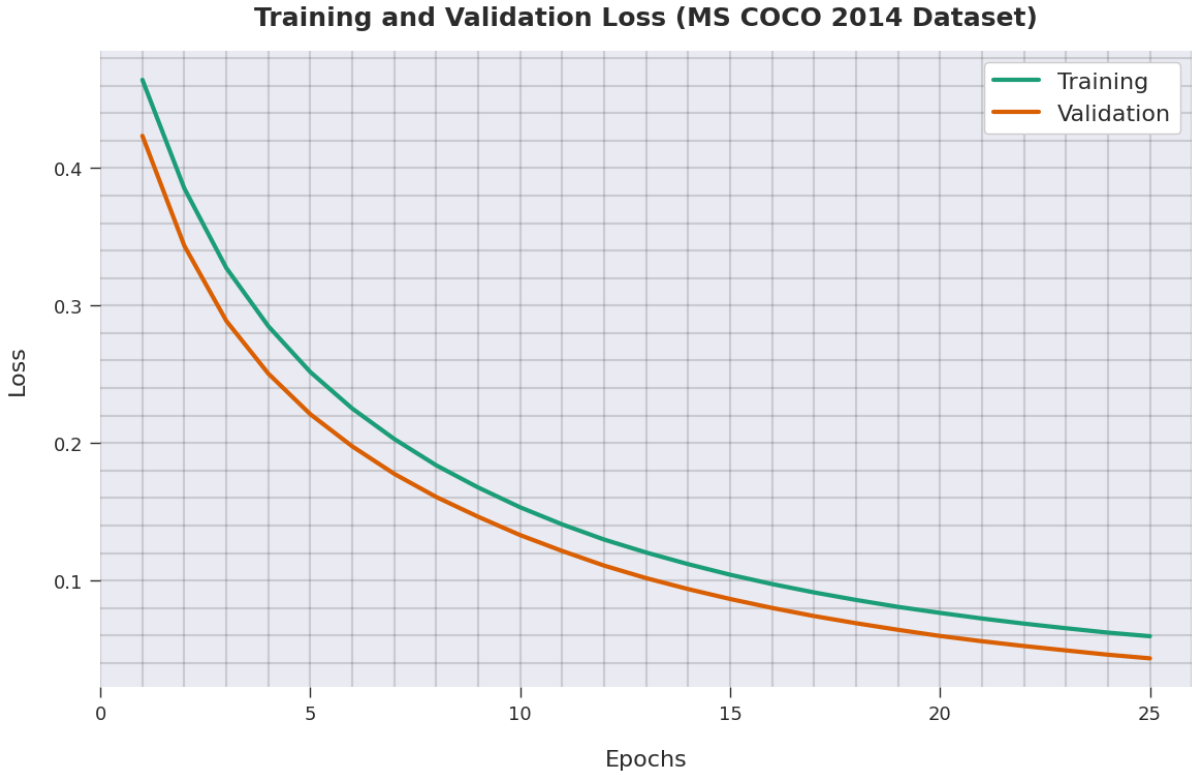


**Figure 12.** TL and VL analysis of the MODLE-AICT technique on MS-COCO 2014 dataset.

## 5. Conclusions

In this study, a novel MODLE-AICT technique was developed for the generation of effective captions to inputted images using two processes involving an encoding unit and a decoding unit. Primarily, at the encoding part, the SSA with a HybridNet model

is utilized to generate effectual input image representation using fixed-length vectors. In addition, the decoding part includes a BiGRU model used to generate descriptive sentences. The inclusion of an SSA-based hyperparameter optimizer helps in attaining effectual performance. For inspecting the enhanced performance of the MODLE-AICT model, a series of simulations are carried out, and the results are examined under several aspects. The experimental values implied the betterment of the MODLE-AICT model over recent approaches. Thus, the presented MODLE-AICT technique can be exploited as an effectual approach for image captioning. In future, ensemble DL-based fusion models can be designed to enhance the performance.

**Author Contributions:** Conceptualization, M.A.D. and H.A.M.; methodology, S.A.; software, R.M.; validation, J.S.A., H.M. and F.A.; formal analysis, A.S.S.; investigation, R.M.; resources, M.A.D.; data curation, R.M.; writing—original draft preparation, H.A.M., S.A. and J.S.A.; writing—review and editing, M.A.D. and R.M.; visualization, F.A.; supervision, M.A.D.; project administration, S.A.; funding acquisition, H.A.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Large Groups Project under grant number (46/43). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R114), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant Code: (22UQU4340237DSR33).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing not applicable to this article as no datasets were generated during the current study.

**Conflicts of Interest:** The authors declare that they have no conflict of interest. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

## References

1. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv. (CsUR)* **2019**, *51*, 1–36. [CrossRef]
2. Sharma, H.; Agrahari, M.; Singh, S.K.; Firoj, M.; Mishra, R.K. Image captioning: A comprehensive survey. In Proceedings of the 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), Mathura, India, 28–29 February 2020; pp. 325–328.
3. Stefanini, M.; Cornia, M.; Baraldi, L.; Cascianelli, S.; Fiameni, G.; Cucchiara, R. From show to tell: A survey on deep learning-based image captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**. [CrossRef] [PubMed]
4. Oluwasammi, A.; Aftab, M.U.; Qin, Z.; Ngo, S.T.; Doan, T.V.; Nguyen, S.B.; Nguyen, S.H.; Nguyen, G.H. Features to text: A comprehensive survey of deep learning on semantic segmentation and image captioning. *Complexity* **2021**, *2021*, 5538927. [CrossRef]
5. Wan, B.; Jiang, W.; Fang, Y.M.; Zhu, M.; Li, Q.; Liu, Y. Revisiting image captioning via maximum discrepancy competition. *Pattern Recognit.* **2022**, *122*, 108358. [CrossRef]
6. Anwer, H.; Hadeel, A.; Fahd, N.; Mohamed, K.; Abdelwahed, M.; Ani, K.; Ishfaq, Y.; Abu Sarwar, Z. Fuzzy cognitive maps with bird swarm intelligence optimization-based remote sensing image classification. *Comput. Intell. Neurosci.* **2022**, *2022*, 4063354.
7. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring visual relationship for image captioning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 684–699.
8. Abunadi, I.; Althobaiti, M.M.; Al-Wesabi, F.N.; Hilal, A.M.; Medani, M.; Hamza, M.A.; Rizwanullah, M.; Zamani, A.S. Federated learning with blockchain assisted image classification for clustered UAV networks. *Comput. Mater. Contin.* **2022**, *72*, 1195–1212. [CrossRef]
9. Huang, W.; Wang, Q.; Li, X. Denoising-based multiscale feature fusion for remote sensing image captioning. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 436–440. [CrossRef]
10. Chohan, M.; Khan, A.; Mahar, M.S.; Hassan, S.; Ghafoor, A.; Khan, M. Image captioning using deep learning: A systematic literature review. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*. [CrossRef]

11. Xu, N.; Zhang, H.; Liu, A.A.; Nie, W.; Su, Y.; Nie, J.; Zhang, Y. Multi-level policy and reward-based deep reinforcement learning framework for image captioning. *IEEE Trans. Multimed.* **2019**, *22*, 1372–1383. [CrossRef]
12. Lakshminarasimhan Srinivasan, D.S.; Amutha, A.L. Image captioning—A deep learning approach. *Int. J. Appl. Eng. Res.* **2018**, *13*, 7239–7242.
13. Zhao, R.; Shi, Z.; Zou, Z. High-resolution remote sensing image captioning based on structured attention. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
14. Hoxha, G.; Melgani, F.; Demir, B. Toward remote sensing image retrieval under a deep image captioning perspective. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4462–4475. [CrossRef]
15. Wang, C.; Yang, H.; Meinel, C. Image captioning with deep bidirectional LSTMs and multi-task learning. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2018**, *14*, 1–20. [CrossRef]
16. Chang, Y.H.; Chen, Y.J.; Huang, R.H.; Yu, Y.T. Enhanced Image Captioning with Color Recognition Using Deep Learning Methods. *Appl. Sci.* **2021**, *12*, 209. [CrossRef]
17. Xiong, Y.; Du, B.; Yan, P. Reinforced transformer for medical image captioning. In *International Workshop on Machine Learning in Medical Imaging*; Springer: Cham, Switzerland, 2019; pp. 673–680.
18. Chen, T.; Li, Z.; Wu, J.; Ma, H.; Su, B. Improving image captioning with Pyramid Attention and SC-GAN. *Image Vis. Comput.* **2022**, *117*, 104340. [CrossRef]
19. Al-Malla, M.A.; Jafar, A.; Ghneim, N. Image captioning model using attention and object features to mimic human image understanding. *J. Big Data* **2022**, *9*, 1–16. [CrossRef]
20. Wang, S.; Ye, X.; Gu, Y.; Wang, J.; Meng, Y.; Tian, J.; Hou, B.; Jiao, L. Multi-label semantic feature fusion for remote sensing image captioning. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 1–18. [CrossRef]
21. Robert, T.; Thome, N.; Cord, M. Hybridnet: Classification and reconstruction cooperation for semi-supervised learning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 153–169.
22. Mirjalili, S.; Gandomi, A.H.; Mirjalili, S.Z.; Saremi, S.; Faris, H.; Mirjalili, S.M. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. *Adv. Eng. Softw.* **2017**, *114*, 163–191. [CrossRef]
23. Liu, J.; Yang, Y.; Lv, S.; Wang, J.; Chen, H. Attention-based BiGRU-CNN for Chinese question classification. *J. Ambient. Intell. Humaniz. Comput.* **2019**, 1–12. [CrossRef]
24. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Int.* **2017**, *39*, 652–663. [CrossRef]
25. Chu, Y.; Yue, X.; Yu, L.; Sergei, M.; Wang, Z. Automatic image captioning based on ResNet50 and LSTM with soft attention. Wireless Communications and Mobile Computing. *Wirel. Commun. Mob. Comput.* **2020**, *2020*, 8909458. [CrossRef]
26. Wang, E.K.; Zhang, X.; Wang, F.; Wu, T.Y.; Chen, C.M. Multilayer dense attention model for image caption. *IEEE Access* **2019**, *7*, 66358–66368. [CrossRef]
27. Omri, M.; Abdel-Khalek, S.; Khalil, E.M.; Bouslimi, J.; Joshi, G.P. Modeling of Hyperparameter Tuned Deep Learning Model for Automated Image Captioning. *Mathematics* **2022**, *10*, 288. [CrossRef]
28. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2–7.
29. Mnih, A.; Hinton, G. Three new graphical models for statistical language modelling. In Proceedings of the ICML '07.: 24th International Conference on Machine Learning, Corvalis, OR, USA, 20–24 June 2007; pp. 641–648.
30. Karpathy, A.; Li, F. *Deep Visual-Semantic Alignments for Generating Image Descriptions*; Stanford University: Palo Alto, CA, USA, 2015.
31. Bujimalla, S.; Subedar, M.; Tickoo, O. B-SCST: Bayesian self-critical sequence training for image captioning. *arXiv* **2020**, arXiv:2004.02435.